Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 *Güz* 2019 *Cilt: 10- Sayı: 3*
Autumn 2019 *Volume: 10- Issue: 3*

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayınlanan hakemli ulusal bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

Iletişim
e-posta: epod@epod-online.org
Web: http://epod-online.org

Journal of Measurement and Evaluation in Education and Psychology (EPOD) is a national refereed journal that is published four times a year. The responsibility lies with the authors of papers.

Contact
e-mail: epod@epod-online.org
Web:http://epod-online.org

Dizinleme / Abstracting & Indexing
Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), TÜBİTAK TR DIZIN

Hakem Kurulu / Referee Board

Ahmet Salih ŞİMŞEK (Kırşehir Ahi Evran Üni.)
Ahmet TURHAN (American Institute Research)
Akif AVCU (Marmara Üni.)
Alperen YANDI (Abant İzzet Baysal Üni.)
Asiye ŞENGÜL AVŞAR (Recep Tayyip Erdoğan Üni.)
Ayfer SAYIN (Gazi Üni.)
Ayşegül ALTUN (Ondokuz Mayıs Üni.)
Arif ÖZER (Hacettepe Üni.)
Aylin ALBAYRAK SARI (Hacettepe Üni.)
Bahar ŞAHİN SARKIN (İstanbul Okan Üni.)
Belgin DEMİRUS (MEB)
Bengu BORKAN (Boğaziçi Üni.)
Betül ALATLI (Gaziosmanpaşa Üni.)
Beyza AKSU DÜNYA (Bartın Üni.)
Bilge GÖK (Hacettepe Üni.)
Bilge BAŞUSTA UZUN (Mersin Üni.)
Burak AYDIN (Recep Tayyip Erdoğan Üni.)
Burcu ATAR (Hacettepe Üni.)
Burhanettin ÖZDEMİR (Siirt Üni.)
Celal Deha DOĞAN (Ankara Üni.)
Cem Oktay GÜZELLER (Akdeniz Üni.)
Cenk AKAY (Mersin Üni.)
Ceylan GÜNDEĞER (Aksaray Üni.)
Çiğdem REYHANLIOĞLU KEÇEOĞLU (MEB)
Cindy M. WALKER (Duquesne University)
Çiğdem AKIN ARIKAN (Ordu Üni.)
David KAPLAN (University of Wisconsin)
Deniz GÜLLEROĞLU (Ankara Üni.)
Derya ÇAKICI ESER (Kırıkkale Üni)
Derya ÇOBANOĞLU AKTAN (Hacettepe Üni.)
Didem KEPIR SAVOLY
Didem ÖZDOĞAN (İstanbul Kültür Üni.)
Dilara BAKAN KALAYCIOĞLU (Gazi Üni.)
Dilek GENÇTANRIM (Kırşehir Ahi Evran Üni.)
Durmuş ÖZBAŞI (Çanakkele Onsekiz Mart Üni)
Duygu Gizem ERTOPRAK (Amasya Üni)
Duygu KOÇAK (Alanya Alaaddin Keykubat Üni.)
Ebru DOĞRUÖZ (Çankırı Karatekin Üni.)
Elif Bengi ÜNSAL ÖZBERK (Trakya Üni.)
Emine ÖNEN (Gazi Üni.)
Emrah GÜL (Hakkari Üni.)
Emre ÇETİN (Doğu Akdeniz Üni.)
Emre TOPRAK (Erciyes Üni.)
Eren Can AYBERK (Pamukkale Üni.)
Eren Halil Özberk (Trakya Üni.)
Ergül DEMİR (Ankara Üni.)
Erkan ATALMIS (Kahramanmaras Sütçü İmam Üni.)
Ersoy KARABAY (Kirşehir Ahi Evran Üni.)
Esin TEZBAŞARAN (İstanbul Üni)
Esin YILMAZ KOĞAR (Niğde Ömer Halisdemir Üni.)
Esra Eminoğlu ÖZMERCAN (MEB)
Fatih KEZER (Kocaeli Üni.)
Fatih ORCAN (Karadeniz Teknik Üni)
Fatma BAYRAK (Hacettepe Üni.)
Fazilet TAŞDEMİR (Recep Tayyip Erdoğan Üni.)
Funda NALBANTOĞLU YILMAZ (Nevşehir Üni.)
Gizem UYUMAZ (Giresun Üni.)
Gonca USTA (Cumhuriyet Üni.)
Gökhan AKSU (Adnan Menderes Üni.)
Gül GÜLER (İstanbul Aydın Üni.)
Gülden KAYA UYANIK (Sakarya Üni.)
Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan Sarıçam (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mehmet KAPLAN (MEB)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Metin BULUŞ (Adıyaman Üni)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Nail YILDIRIM (Kahramanmaraş Sütçü İmam Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Recep GÜR (Erzincan Üni.)

Hakem Kurulu / Referee Board

Ragıp TERZİ (Harran Üni.)
Recep Serkan ARIK (Dumlupınar Üni.)
Sakine GÖÇER ŞAHİN (University of Wisconsin Madison)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sedat ŞEN (Harran Üni.)
Seher YALÇIN (Ankara Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Selen DEMİRTAŞ ZORBAZ (Ordu Üni)
Selma ŞENEL (Balıkesir Üni.)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Serkan ARIKAN (Muğla Sıtkı Koçman Üni.)
Seval KIZILDAĞ (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sungur GÜREL (Siirt Üni.)
Sümeyra SOYSAL
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT AVCI (Kilis 7 Aralık Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Ufuk AKBAŞ (Hasan Kalyoncu Üni.)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Zekeriya NARTGÜN (Abant İzzet Baysal Üni.)
Zeynep ŞEN AKÇAY (Hacettepe Üni)

*Ada göre alfabetik sıralanmıştır. / Names listed in alphabetical order.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Eylül 2019, Sayı: 10-3)

Journal of Measurement and Evaluation in Education and Psychology (September 2019, Issue: 10-3)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 202-218

Examination of Educational Films Suggested by MEB to Teachers from the Perspective of Measurement and Evaluation *

Fatih KEZER ** Kübra ÇETİNER ***

Abstract
Seeking new approaches on in-service trainings, The Ministry of National Education has recently developed a number of vocational training programs in order to increase teachers' pedagogical formation skills. The films with educational content covered by this study have been suggested to teachers in those programs. Through these films, teachers are expected to gain new perspectives in a pedagogical sense and to recall existing ones. The relationship between the film outputs and the social behavior is a matter of curiosity. This study intended to examine the films proposed by the Ministry of National Education for teachers within the scope of the 'vocational study program' in terms of the elements of measurement and evaluation and the sub-texts they contain. Research was conducted in qualitative research design. Document analysis used in the research, The study evaluates the educational content of the films that were suggested to the teachers within the scope of September 2017 Professional Study Program which was created to increase the knowledge and skills of teachers and administrators working in pre-school, primary and secondary education institutions by the General Directorate of Teacher Training and Development of Ministry of National Education. In collecting data, a Film Evaluation Form which was developed by the researchers under 26 themes was used. Given that the films recommended by MoNE to teachers have already been watched by the majority of teachers, it is inevitable for teachers to be affected from the scenes, content and sub-texts in the educational content consciously/unconsciously and to acquire new patterns of behavior. Another important point, however, is that these well-known and popularized film are also watched by parents. They affect not only the teachers, but also behaviors of the students and parents. The result of the research, points to the fact that these films, which are proposed to teachers as part of in-service training, can lead to the acquisition of negative behaviors in view of measurement and evaluation and that they have implicit messages that can create negative perception.

Key Words: Measurement and evaluation, films with educational content, document analysis, content analysis.

INTRODUCTION

In its general sense, education is defined as the process of creating desired behavioral change in an individual in a planned and programmed manner (Ertürk, 1972; Sönmez, 2004) and as a system it consists of four elements (Demirel, 2005; Fitz-Gibbon & Morris, 1989). These four elements can be defined as inputs, process, outputs / products and control / evaluation. Whether the individual achieves the desired behavior or not is rendered possible by evaluation process. Measurement and evaluation are of great importance in terms of monitoring, controlling and improving the process (Demirel, 2005). It also provides more systematic and objective evidence for educational decisions (Linn & Gronlund, 1995). The determination of the behaviors to be acquired by individuals via education process can be insufficient from time to time with the conventional measurement and evaluation approaches. After the information acquired by individuals, different measurement techniques may be required to measure the skills that they gain through practice (Turgut & Baykul, 2010; Yıldırım, 1999). Considering the traditional and complementary measurement and evaluation approaches, it is possible to measure / evaluate the knowledge, skills and abilities of individuals more effectively and reliably by means of the diversity and quality of the tools. However, it is observed that teachers are insufficient from time to time in view of their measurement and evaluation skill which is one of the essential teacher

* This study was presented at the 6th International Congress on Measurement and Evaluation in Education and Psychology.
** Assist. Prof. Dr., Kocaeli University, Educational Faculty, Kocaeli-Turkey, fatih.kezer@kocaeli.edu.tr, ORCID ID: 0000-0001-9640-3004
*** Teacher, Ministry of National Education, Kocaeli-Türkiye, kubracetiner@gmail.com ORCID ID: 0000-0001-9374-7354

To cite this article:
Kezer, F. & Çetiner, K. (2019). Examination of educational films suggested by MEB to teachers from the perspective of measurement and evaluation. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 202-218. doi: 10.21031/epod.547240

Received: 31.03.2019
Accepted: 27.05.2019

competencies. The studies conducted in this field indicated that teachers consider measurement and evaluation activities as important (Anılan, Anagün, Atalay & Kılıç, 2016; Duban & Küçükyılmaz, 2008) but they also indicated that they have problems related to the measurement and evaluation process and that they lack the knowledge and skills in this area (Adıyaman, 2005; Anıl & Acar, 2008; Anılan et al., 2016; Bal, 2009; Çakan, 2004; Çoruhlu, Nas & Çepni, 2009; Duban & Küçükyılmaz, 2008; Evin-Gencel & Özbaşı, 2013; Gelbal & Kelecioğlu, 2007; Gömleksiz & Bulut, 2007; Güven, 2008; Kilmen, Akın Kösterelioğlu & Kösterelioğlu, 2007; Özenç, 2013; Güneş, Dilek, Hoplan, Celikoglu & Demir, 2010; Yanpar, 1992; Yapıcı & Demirdelen, 2007). In other words, most of the teachers state that they can use different measurement and evaluation approaches when they have sufficient knowledge and yet they cannot find an exemplary role model for these practices and that there is no one to guide them (Güneş et al., 2010). Also, teachers think that in-service training is not enough about the use of measurement and evaluation approaches and effective in-service training should be given to them (Anıl & Acar, 2008; Çakan, 2004; Gelbal & Kelecioğlu, 2007; Temel, 1991; Yanpar, 1992). They also emphasize the need to focus on practical examples in these in-service trainings (Anıl & Acar, 2008; Anılan et al., 2016; Güneş et al., 2010). Stiggns (2001) also emphasizes the same, positing that the lack of teacher skills in using different measurement and evaluation approaches is due to their low competencies resulting from the insufficiency of pre-service training.

The Ministry of National Education, which has been looking for new approaches for in-service trainings, has established vocational training programs (and updated the existing ones as well) in recent years in order to increase teachers' pedagogical formation skills. Films with educational content have been proposed to teachers in one of these programs. Through these films, teachers are expected to gain new perspectives in the pedagogical sense and to recall the existing ones.

The relationship between the outputs of films and the behavior of society is a matter of curiosity. One of the first studies on this subject was the study named "From Caligari to Hitler: A Psychological History of the German Film" by Siegfried Kracauer in 1947 (Güçhan, 1993). It is possible to deal with the relationship between cinema and society in two ways. On one hand, cinema is a mirror of the psychological, sociological, cultural, historical, political, social and economic structure of the society (Tolon, 1978). In other words, cinema is a product of social structure (Armağan, 1992). On the other hand, social behavior is also influenced by cinema. Films are one of the important sources for the dissemination of social information. However, it can be said that societies can reproduce themselves through films (Diken & Laustsen, trans. 2011; Özer, 2004; Yakar, 2013). Considering all media tools including films, visual and audio materials can have an impact on society's behaviors and thoughts. Audiences can develop different behaviors due to the explicit and implicit messages contained in the media (Şahin, 2011). According to Metz (1985), each image is a sentence in itself (as cited in Sivas, 2012).

Like all mass media, films are a powerful and non-formal educational resource (Güçhan, 1993). In his book "We're in the Money: Depression America and Its Films" (1971), Andrew Bergman (American writer, screenwriter and director) posited that one of the most important contributions of American cinema to American education was to teach that its own institutions have enough power to correct a mistake and to do the right thing, and that this contribution was made by reflecting success and hope on the screen (as cited in Güçhan, 1993).

When the relationship between the mass media and the society is examined, Raymond Williams' *Flow Theory* for television takes part in the related literature as an important point of view to take into account. Williams in "Television, Technology and Cultural Form" introduced a critical perspective with his flow theory to the relationship between television and society. According to Williams' theory, (Şentürk, 2009), all programs of a television channel have a conscious planning and the program contents can have effects on social perception and can transform the values of a society. Though TV is considered as a large narration tool, this large tool actually consists of small messages / narratives (Serttaş, 2014). *Cultivation Theory* which was introduced by George Gerbner in the 1970s, is based on the fact that viewers realize learning without being aware of the stimuli that are presented to them through the media. Gerbner described this 2 directional / dimensional learning as "Society is the

message" in 1974 and discussed it in that work with the same title (Gerbner, 1974). Gerbner states that the educative function of the stories in television is hidden within the thought *what most people do, what they think*. However, according to this theory, the degree of education varies based on frequency of exposure to a given message (Ercan & Demir, 2015). The symbolic environment exposed by mass media is interpreted as collective consciousness.

Films, both with their themes and scenes, leave sometimes indelible and permanent traces in the human memory over the years (Budak, 1986). Birkök (2008), emphasing the idea *seeing is believing*, states that the films are a better tools in understanding complex information than texts. In addition to this, knowledge transfer can be realized through behavioral models as well as formal education. With the development of technology, individuals are constantly exposed to visual media. With the widespread use of social media tools, film scenes can be shared for many purposes. Visual media addresses multiple sensory organs. Therefore, it makes learning something easier and it is more attractive for individuals.

The fact that individuals develop behavior in the light of the visual elements that they encounter brings with it another question. What messages do these visual elements contain? Considering the dimension of measurement and evaluation, the activities used in films, the effects of these activities on students, teacher behaviors and the similar elements will inevitably affect the audiences (especially the teachers in the in-service training as target audience) consciously / unconsciously. Having both positive and negative impact on audiences, what kind of content do these elements have or what messages do they contain? In the related literature, it is seen that film studies are carried out from a variety of perspectives focusing on education (Akcan & Polat, 2016; Akıncı-Yüksel, 2015; Beldağ & Kaptan, 2017; Hamarat, Işıtan, Özcan & Karaşahin, 2015; Kalaycı, 2015; Kaşkaya, Ünlü, Akar & Özturan-Sağırlı, 2011; Polat, 2011; Polat & Akcan, 2017; Yakar, 2013; Yıldırım, Tüzel & Yıldırım, 2016). However, there are not many studies examining the films from the perspective of measurement and evaluation (ME) activities. The remarkable study was done by Doğan in 2017 on the Hababam Sınıfı series. Even though the aim of the films is not to give educational messages to the audience, teachers, students and families encounter elements that are related to learning in these educational films. In this study, the films proposed by the Ministry of National Education for teachers within the scope of the vocational study program have been examined in view of the elements of measurement and evaluation, and of their sub-texts.

METHOD

The present research is a qualitative research which aims to examine the elements of educational content which are proposed by MoNE by teachers. Document analysis has been used in the research. Document analysis can be defined as the collection of visual and written materials (Sönmez & Alacapınar, 2013). It is the analysis of materials containing information about the facts and events intended to be investigated. These materials alone can be the data collection tool of a research. The biggest advantages of visual materials such as film, video and photography are that they can be monitored repeatedly by the researchers and that non-verbal expressions like (gestures and mimics, body language, facial movements, etc.) can be retained to be searched by other researchers (which might increase reliability and validity) (Yıldırım & Şimşek, 2016).

Study Material

In the study, educational materials which have been proposed to increase the pedagogical formation skills of teachers used as study materials. This study evaluated the films proposed by the General Directorate of Teacher Education and Development of MoNE to increase the knowledge and skills of teachers and administrators who are working in preschool, primary and secondary education institutions. These films with educational content have been proposed for teachers within the scope of Vocational Study Program which was formed in September 2017 (MEB, 2017). Since a film was re-listed in the original list of 30 films, and since three films were excluded from the study because they

were documentaries, the remaining 26 films were included in this study. The list of films is given in Table 1.

Table 1. List of Films with Educational Content Recommended by MoNE

Original Name	Turkish Name	Release Date	Country	Runtime (min.)
1.3 Idiots	3 Aptal	2009	Hindistan	170
2. AmericanTeacher	-	2011	ABD	81
3. Billy Elliot	-	2001	İngiltere, Fransa	110
4. The First Grader	Birinci Sınıf	2010	İngiltere, ABD, Kenya	103
5. Good Will Hunting	Can Dostum	1997	ABD	126
6. Monsieur Lazhar	Canım Öğretmenim	2011	Kanada	94
7. Hababam Sınıfı		1974	Türkiye	90
8. Hababam Sınıfı Dokuz Doğuruyor		1979	Türkiye	88
9. Hababam Sınıfı Güle Güle		1981	Türkiye	78
10. Hababam Sınıfı Sınıfta Kaldı		1975	Türkiye	91
11. Hababam Sınıfı Tatilde		1977	Türkiye	93
12. Hababam Sınıfı Uyanıyor		1976	Türkiye	94
13. İki Dil Bir Bavul			Türkiye, Hollanda	81
14. The Emperor's Club	İmparatorlar Kulübü	2002	ABD	109
15. Black	Kara	2005	Hindistan, ABD	122
16. Takhtesiah (Blackboards)	Kara Tahta	2000	İran, İtalya, Japonya	85
17. Les Choristes	Koro	2004	Fransa Almanya İsviçre	97
18. The Blind Side	Kör Nokta	2009	ABD	129
19. Être et Avoir / To Be and To Have	Olmak ve Sahip Olmak	2002	Fransa	104
20. Dead Poets Society	Ölü Ozanlar Derneği	1989	ABD	128
21. Freedom Writers	Özgürlük Yazarları	2007	ABD	123
22. Patch Adams	-	1998	ABD	115
23. Mr. Holland's Opus	Sevgili Öğretmenim	1995	ABD	143
24. EntreLesMurs (The Class)	Sınıf	2008	Fransa	128
25. Half Nelson	Tepetaklak Nelson	2006	ABD	106
26. Taare Zameen Par	Yerdeki Yıldızlar	2007	Hindistan	165

Data Collection

In order to collect data, a Film Observation Form developed by the researchers was used. Firstly, the study of the related literature was reviewed and similar studies were investigated in the preparation of the form. As a result of the investigations, certain themes were created primarily by the researchers. The main themes are as follows: the use of traditional measurement and evaluation approaches, the use of complementary evaluation approaches and the testing / examination environment. A draft form was created in accordance with the themes determined and it was checked to see in terms of functionality by the researchers and two observers selected by the researchers from the field of measurement and evaluation in the evaluation of a particular film. Later, all the evaluators examined the same film again, concomitantly. In the light of the scenes in the film, the themes were discussed and elaborated. The researchers then watched three more films in due consideration of these themes and re-examined the themes in order to determine whether they were suitable for the intended purposes or not, whether they were able to reveal the sub-texts or not, and whether they have an appropriate scope or not, and revised them. In addition to determining whether ME is used or not, the themes focus also on the issues such as the content and the type of evaluation method used, the effect of ME on the student psychology and behavior, what behaviors can be implicitly acquired through it and what messages it can give, and on the level of mental skills it addresses to. Five different films were determined by the researchers and the coding reliability was calculated by the final form. Compliance / similarity of all coding and determinations were calculated as percentage and obtained as .84. In the literature, the reliability among the raters is expected to be over .70 (Tavşancıl & Aslan, 2001). The

fact that the films can be watched again and again is a factor that increases the internal validity of the research. The Film Observation Form used in the study is presented in Appendix 1.

Data Analysis

Content analysis method was used to analyze the data collected in the study. Content analysis can be defined as the description of the basic contents of the research items and the sub-messages they contain (Cohen, Manion & Morrison, 2007). In other words, Content analysis can be defined as a reading for identifying the items that affect the individual unconsciously (Bilgin, 2006). While conducting content analysis in the research, themes were formed primarily and they were updated in the research process in accordance with the nature of qualitative research. Themes are supported by direct quotations.

RESULTS

When the films within the scope of the study have been examined, it is seen that the film 3 Idiots has been formed in a different concept as it displays a critical viewpoint for the educational activities. Leaving aside this film, it can be said that the scenes in the other films focusing on measurement and evaluation are quite few and these scenes are sloppy or arbitrary. Although the films proposed by the MoNE have pedagogical scenes / content, they do not contain a large number of items in view of the field of measurement and evaluation. Although there are many scenes related to education and training in these educational films, the lack of scenes in the field of measurement and evaluation may mean that the control mechanism is given less importance in terms of evaluation, as in real life. After examining the measurement and evaluation elements of the films with the data collection tool, 26 themes initially formed were merged into 12 main themes and discussed in line with these main themes. The 12 themes created are as follows;

1. Use of traditional and complementary measurement and evaluation approaches
2. Use of reward and punishment in ME activities
3. Use of questions to measure lower-level and higher-level skills in ME activities
4. Exam preparation and scoring process
5. Conditions of Examination Environment
6. Praising lower-level and higher mental skills
7. Use of questions that do not conform to the ME standards in the ME activities
8. Giving feedback to students at the end of the ME activities
9. The effect of me activities on student attitudes and behaviors
10. Praising the use of information in daily life
11. Use of me activities for competition
12. Cheating in ME activities

Findings on the Use of Traditional and Complementary Measurement and Evaluation Approaches

When the scenes with the ME activities was observed, it was seen that the traditional evaluation approaches were frequently used in clearly disclosed scenes, and a total of 21 scenes of this type were noted, mostly with written and verbal examinations, and the matching and true false tests were not seen in any of them (Table 2).

Table 2. Frequencies Related to the Use of Traditional and Complementary Measurement and Evaluation Approaches

	Observed	Not Observed	Number of Observed Scenes
Traditional Evaluation Approaches			
Items containing written response	9	17	16
Items containing oral response	7	19	11
Items containing multiple choice	2	24	5
Items containing short answer	3	23	6
Items with matching	0	26	0
Items containing true false	0	26	0
Complementary Evaluation Approaches	2	24	2

The sample dialogues in the films indicating the use of the traditional evaluation approach are as follows;

"Sit down. Pull out your papers. I am to give you a written test!" - Hababam Sınıfı

"Name an American composer. …. How do you know what key a concerto is in?" - Mr. Holland's Opus

"Tell me! Philosophers of the first epoch! …. Tell me my kid, who are these Balkan states?" - Hababam Sınıfı Uyanıyor

In the films, the questions in the examination papers do not appear in detail. Although traditional evaluation approaches are frequently used, only 2 films out of 26 show complementary evaluation approaches. These scenes are the entry of a student into an interview room for the acceptance to a ballet school in the film Billy Elliot, and a self-assessment of a student's own work in Freedom Writers. Since the school is located in a disadvantaged area, it is seen that the main aim of the teacher is to involve children in school life. It should expressly be noted that the complementary evaluation approaches are limited only in 2 special cases in these 2 films. Traditional approaches, such as written / oral exams, multiple-choice tests, short-response tests, and matching tests, focus more on the product and are weaker in evaluating the learning process. Complementary evaluation approach, however, is a process of evaluation that examines how the student understands and uses the information, transforms his / her existing knowledge into a product or activity or how he solves the problems in daily life, and how he uses the knowledge and skills to solve these problems (Pamukçu, 2015). What is expected in formal education is the use of different measurement and evaluation tools that will compensate for the limitations and disadvantages of traditional approaches. Excessive use of traditional measurement and evaluation approaches in films with educational themes can create a perception that student performance is always evaluated in that way. It is possible to evaluate the performance of a student, along with the traditional paper-pen tests, by following the behaviors of him in the classroom and outside the classroom, by observing his / her performance in the process, by measuring his / her interest and attitudes and by involving him in the process, from a larger perspective (Gelbal & Kelecioglu, 2007). When the studies in the related literature are examined, it is seen that teachers are rather using traditional approaches to measure and evaluate student achievement (Anıl & Acar 2008; Belet & Sağlam, 2015; Fidan & Sak 2012; Gelbal & Kelecioğlu, 2007; Yaman, 2011). When the teachers, who are accustomed to using traditional measurement and evaluation approaches, often see the similar approaches in film scenes, it may create a sense of self confidence in them to the effect that they are doing the right thing, which in turn may hinder the possibility of self-criticism.

Findings on the Use of the Reward and Punishment Elements in the ME Activities

Of the 26 films, only one of the films contains two reward scenes. This film is the Emperor's Club. The children and adults were given a crown as a reward for being the winners.

ME activities were rather used as punishment in four films in five different scenes and as a threat in some of them. For example, in one scene, the final exam was used as a threat by a teacher and the exam was prepared for the purpose of punishing the student. The teacher's line is as follows;

"You're gonna get job only when you pass in final exams. .... But this time I'm gonna set the paper for exam" - Three Idiots

Similarly, in another film, the teacher decides to apply an unscheduled examination after the students' inappropriate behaviors and uses the test as a punishment.

"You ribalds! You are playing leapfrogging in the classroom? Remove the books, pull out your sheets. I am giving you a written test" - Hababam Sınıfı

Although the use of reward and punishment in education is considered to be an old method, it is frequently used in today's education system. Class discipline should be provided for an effective learning environment where students can express themselves by respecting each other, teachers can reach out to all students in an atmosphere of cooperation and trust among individuals. What is important at this point is that the concept of discipline should not remain as a class and school rule determined by the teachers or the school administration; it should, rather, encompass an environment in which students can develop their self-discipline. Considering that the reward and punishment element is related to the behaviorist school, it is possible for individuals to develop self-discipline with only the careful use of reward and punishment. The first step to be taken in order for the measurement and evaluation studies to produce valid and reliable results is to determine the purpose for which the test will be used. Exceeding the predetermined purpose in measurement and evaluation may lead to inaccurate results, affecting validity and reliability negatively. Reward and punishment must be considered independently. It is very likely that the measurement activities which are used as punishment and threat elements in film scenes can be normalized by the viewers.

Findings on the Use of Questions to Measure Lower-Level and Higher-Level Skills in ME Activities

When examining the scenes with ME activities, it has been observed that both the written and the oral questions are mostly of a nature to measure the learning at the knowledge level. In almost all of the scenes (9 films, 17 scenes) where the exam questions are pronounced, it is observed that the teachers test their students at the knowledge level. It is obvious that the characteristics that are intended to be measured by teachers are more in the knowledge and recalling level, which is the lowest level of learning. This can lead to a perception that the examinations are used to measure only at the level of knowledge/recalling. The use of questions only at the level of knowledge/recalling in ME activities may cause learning to remain at memorization level, and may prevent learners from using knowledge in new situations, in their daily lives.

Written and oral exam questions, and questions used in competitions in Hababam Sınıfı series, the interview for entrance to a university in Black and the two quiz shows in Emperor Club are the examples of scenes where lower-level skills are measured. The sample lines taken from these films are as follows;

> Write down! Ouestion 1: Digestive system in mammals. Question 2: Give three examples of parasitic species. Question 3: The structure of ectoplasm. …. What is the date of Preveze Sea Battle? Name the parties. …. What happens if two molecules collide in the atomic reactor? …. With what forces did the Ottoman army set off for the siege of Vienna? …. How did the Patrona Halil revolt start? …. In what neighborhood and on what date did the biggest fire in Istanbul occur? …. Who became the king of Spain after Franko died? - Hababam Sınıfı Uyanıyor

"How many oceans are there in the world?" - Black

> Which emperor sought to return all power to the Senate, only to garner even greater power? …. Who introduced the modern/professional army to Rome? …. Of the first eight emperors, which name is omitted from the following list? …. What year was the

> Roman army crushed at Lake Trasimene? …. Who was the last emperor of the Western Empire? – Emperor's Club

"Name the dates of the establishment and demolition of the Anatolian Feudal States." - Hababam Sınıfı Güle Güle

"How did Marshal Ney die?" - Chorus

Contrary to the abundance of knowledge level questions observed in the scenes concerning ME activities, high level mental skills have been taken into consideration in only two scenes that question the reasoning skills of the students. It is observed that these questions are the questions that measure the interest and emotions that take individual differences into account, not the level of their learning. These questions, which aim to measure high level mental skills, were seen in the dance interview in Billy Elliot and in the interview with a student having both vision and hearing impairment for acceptance into a university. This gives a message to the audience that the questions that measure high-level mental skills can be used only in special cases. Some examples of these scenes are given below.

"Can you tell us why you first became interested in the ballet? …. Was there any particular aspect of the ballet, which caught your imagination? …. What does it feel like when you're dancing? …. What does knowledge mean to you?" - Billy Elliot

"If we are in India, on which side will America be? …. Why do you want to study?" - Black

In today's world where the means of access to knowledge have increased and developed rapidly, education is expected to raise individuals who know the ways of accessing to information, check the accuracy thereof, adapt it to new situations, interpret the events in a cause effect relationship, create a new product, and to produce solutions to problems, rather than merely memorizing the given information. The way to achieve this goal can be possible not by employing activities at the knowledge, comprehension and application levels, but by activities developing the students' ability to analyze, synthesize and evaluate the related subjects. The fact that teacher characters in films use questions about lower-level mental skills in their ME related activities can be interpreted by the viewers to the effect that teachers generally attach importance to bookish knowledge. In addition, that the low-level mental skills can be measured relatively easily may result in the misconception that teachers' task of measuring is an easy job. However, preparing a qualified exam is a difficult and painful process especially in measuring high level mental skills. The scenes in the films can create the impression that teachers' responsibility for measurement and evaluation is very easy and insignificant.

Findings Related to Exam Preparation and Scoring Process

When the films are examined, the first thing that stands out about the examination preparation and scoring process is that there are scenes which can create a perception that the teachers prepare and score exams very easily. In relation to the previous theme, teachers appear to be giving instant exams with questions mostly measuring the lower-level mental skills which can be interpreted as though exams can easily and quickly be given at any time. In the films, there is no scenes related to the preparation of the exams or the exam questions are promptly given by the teacher without use of any resource. For example, in Hababam Sınıfı, the teacher gives the exam by asking the random questions that come into his mind as soon as he enters the classroom. As it is understood from these scenes, the students have not been informed of these exams. The most common and typical of such scenes are as follows;

> Take out your papers, I will give you an exam. …. Take your papers out, I am giving you a written exam. …. Remove the books! I am giving you a written exam. …. Sit down. Take out your papers and pencils! I will give you an exam - Hababam Sınıfı Uyanıyor

As can be seen from the examples above, it is shown that the exams can be planned instantly and the questions can be easily and randomly created by the teachers. In parallel with the process of exam

preparation, there are also scenes to make the audiences think that teachers can easily and subjectively score the performances without using any scoring keys. Following line from the Hababam Sınıfı is an example to that;

"Zero to all Hababam class and your score is 10" - Hababam Sınıfı Tatilde

Another example of a scene where exam preparation and scoring are influenced by the emotions and thoughts of the teachers who act subjectively is in the Emperor's Club. In the film, it is seen that a teacher gives an undeserved high mark to a failing and mischievous student just to increase his motivation.

Developing a tool that ensures valid and reliable measurement requires preparation. A test prepared by the teacher is expected to meet the following requirements: The purpose of the examination and the desired competence should be clearly set, and an appropriate test format with a table of test specifications along with an item pool should be prepared with a view to ensure the length and the form of the test is appropriate for the desired target. Skipping the specified stages for various reasons may result in tests that are not appropriate for the purpose, which, in return, may cause miscalculation or incorrect measurements of the desired competence. The scoring of exams also requires labor. Erroneous scoring is a factor that reduces the reliability of measurement results. Well-prepared answer keys, in which the criteria are clearly expressed and the boundaries of which are clearly drawn, contribute to the objective scoring of the raters. Yet, in the films, it is seen that an objective scoring tool is not used in the scenes where students are evaluated and the exams are scored, and that the scoring process can be done with the instant decision of the teachers. This may lead to the idea in the minds of the viewers that the examinations are hastily and negligently prepared and scored.

Findings on the Application Conditions of Examinations

Nine films and 14 stages display the conditions of application. Only three of them reflect positive scenes and rest of them reflect negative scenes. When the scenes reflected on the screen are examined, the messages that the viewers can get are that the exam environments and the application conditions are sloppy, they do not match the principles of measurement and evaluation and that the students are not respected and they are treated as insignificant. Teachers disturbing the students by walking on the desks, teachers reading a newspaper or sleeping during the exam, teachers constantly shouting at and warning the students against cheating during the exam in a noisy way, students entering the classroom noisily during the exams, students distorting the attention of the teacher with a variety of stories / schemes to sabotage the effectiveness of the exam and similar other examples show how far the exam conditions are from the ideal measurement and evaluation principles.

To get valid and reliable measurement results, examination environments should be suitable and qualified for the preparation and scoring processes. Reliability, in its most general definition, is the degree of refinement of the measurement results from random errors (Crocker & Algina, 1986). Insomnia, fatigue, lack of attention, reluctance to answer questions, lack of experience, success of chance, cheating, mode of expressions of the items and directives in the exam, difficulty of the items, discriminative quality of the items, examination environment, exam duration, etc. are some of the factors leading to random errors. One of the most important of these factors is the test environment. It should be suitable for the students to be able to demonstrate the best of their performances in such a way to ensure a valid measurement. Yet, in the scenes of the films, it is seen that the results of the ME activities are performed in environments where many factors of error are likely to interfere, and that both teachers and students are frivolous. Such adversary conditions can lead to erroneous results for the exams and reduce their reliability and validity. The abundance of inappropriate scenes and the lack of appropriate scenes of examination may create a notion in the audience that the environment is insignificant in the measurement and evaluation process and that examinations can be conducted in any way under any condition.

Findings Concerning Lower and Higher Level Mental Skills

While praise for behaviors at knowledge and recall level is seen in five films, only one film contains a praise for a high level mental skill. In the seven scenes identified, the teacher / manager / the inspector praise the students for memorizing some words, a skill pertaining to knowledge level. An example of a line is in Hababam Sınıfı:

"Bravo! You have memorized the book as is." - Hababam Sınıfı

On the other hand, in an interview scene in Black a student is applauded after he has answered a series of questions requiring higher level of mental skills. This scene praising a student with "bravo" and "perfect!" is the only praise scene for a high-level mental skill.

When this theme is examined together with the themes that include the use of questions measuring lower and higher skills, it can be said that these educational films highlight only the importance of the questions that measure lower-level skills and create a perception in the audience to the effect that memorizing a book as-is is of high importance.

Findings Regarding the Use of Questions Non Conformant to ME Activities

Although some films have not examined under this theme due to the fact that the examination papers and exam questions are not clearly shown in them, in most of the ME activities of the films wherein the questions are revealed; it is seen that there are scenes which can create a misconception on teachers to the effect that they can ask a broad and unclear range of questions with no conceivable principles and purposes in their minds. The exemplary lines in the films are given below;

"Tell! Philosophers of the First Age. …. Write! Question 1. Digestive system in mammals" - Hababam Sınıfı Uyanıyor

"Tell! The philosophy and society. …. Question 1. The Era of Murat IV." - Hababam Sınıfı

"National Literary movements?" - Hababam Sınıfı Sınıfta Kaldı

While measuring the cognitive, affective and psychomotor behaviors of the students, the measurement tool is required to be suitable for the purpose, to be able to address the structure to be measured and to be able to make the valid measurements accordingly. Therefore, when preparing the items of measurement instruments used in performance measurements; maximum care should be exercised to make sure that they are clear, understandable, concrete, corresponding to and measuring a single structure, being understood by each student in the same way, having a particular and clear frame, and be answerable in sufficient time. Compliance with these principles is indispensable for a qualified measurement and evaluation while preparing both open-ended and multiple-choice items, regardless of the type of the item. It is observed that the questions asked by the teachers in the films are prepared indiscriminately by not following all of the principles of the field.

Findings Related to Providing Feedback to Students at the End of the ME Activities

The term feedback is defined as the explanation related to how much the learner learns the target, what his deficiencies are and what path he can follow to complete the missing parts in his learning (Joyce, Weil & Calhoun, 2000 as cited in Çevikbaş, 2018). Feedback can be provided in a variety of ways. According to Erişen (1997), positive feedback and correction behaviors involve checking the previous learning, correcting the missing points and mistakes, asking clear and explicit questions to ensure target behavior, simplifying the unclear questions and re-asking, replying the students with concrete, clear and comprehensible answers, giving the students sufficient time to think about the questions asked, informing the student on the accuracy or inaccuracy of his answers by showing him the missing points, giving the other students the opportunity to find the wrong or missing answers, giving the students the opportunity to give feedback and clues to each other, presenting the class inaccuracies and missing points without specifying student name and correction in the case that there is no time for

the instructor to help them individually, etc. The feedback enables the teacher to determine which points are missing in the activity and to communicate these deficiencies to the students. The feedback and corrections taken into consideration by the student improve his self-awareness, prevent him from repeating the same mistake and ensure that he gains a variety of perspectives. Considering the fact that information is a set of meaningful data built on one another, it is necessary to provide feedback for individuals with incomplete and incorrect learning in his build up process in order to realize the new learning fully. The findings in the related literature reveal that the effective oral feedback given by the teachers influences the students' academic achievement and it has a higher impact on the development of the higher cognitive awareness of the students as compared to traditional teaching (Çetin, 2014). Failure to give feedback on the results of the examination may prevent students from seeing what points they are missing or learning incorrectly and it may result in doing the same mistakes.

None of the films examined shows any feedback activity except for the score announcement. This situation does not correspond to the feedback behavior that supports the change and development of individuals and it sets a negative example to the teachers in the audience.

Findings on the Effect of ME Activities on Student Attitudes and Behaviors

In the films examined within the scope of the study -5 films and 11 scenes -ME activities are reflected as though they are vital activities, creating tension / fear and sorrow on students.

Considering examples from the scenes; students are shown to make a course to prayers and consecration to get better results in their final exams:

"Oh cow-god! .... Just, just get me passing marks. God ... God. God. I'll offer 100RS per month. Surely God! Promise!"- Three Idiots

A student who fails in his project assignment commits suicide due to having future anxiety and feeling of failure. - Three Idiots

One student flees from school because he hasn't done his math homework. Throughout the whole film, this student is reflected in fear and anxiety in every ME activity. - The Stars on the Ground

At the end of an examination, the second-best student gets upset and cries because he could not be the best. - Three Idiots

The student who gets the result of the examination gets nervous for the result in a frightened manner and when he sees that he passed it, he cries with the relief of emotional tension. - Billy Elliot

When the students find out that they passed the exam, they act as if it is of vital importance, displaying excessive happiness and excitement. - Three Idiots, Billy Elliot, Black

A student who is nervous about the exam, thinking that his performance was bad, inflicts violence on another student due to his nervousness. - Billy Elliot

As it can be understood from the scenes, the ME activities in the films are reflected as activities that cause fear and anxiety and stress for students. It has been seen that ME activities are not a tool but a goal, and are shown as the factors that affect not only the academic life but also the relationships with the family members. Another negative effect of measurement activities on the audiences is that students act as though any means to achieve a goal is justifiable. In the scenes, the students who see the evaluation activities as competition, become ambitious and develop behaviors to try all the means to get a prize. These explicit and implicit messages in film scenes have the potential to have adverse effects on the behavior of the students and teachers who are in audiences.

Findings Regarding the Praise for the Use of Information in Daily Life

None of the films shows any element of the use of the learned material in daily life in the context of the ME activities. Looking at the educational goals of today's world, students are expected to be grown

as individuals who can apply the skills they have learned in school in their daily lives and who can solve problems. In this respect, the measurements are actually expected to be made on real life situations. Considering this theme, along with the theme of praise for low-level and high-level mental skills, it can be said that the scenes in the films do not reflect high-level mental skills and to daily life skills relating to them.

Findings on the Use of ME Activities for Competition / Contest

In the films, the activities of ME have been reflected (4 films, 7 scenes) as competition elements. There are scenes and dialogues that can create the perception that the exams and exam results are a mere element of competition and that rivalry is an intrinsical quality of the exams. For example, in a scene of Three Idiots, the students were seated for a photo shoot in order of their success levels, and a student is compared to his siblings and classmates through his exam scores in Stars on the Ground. In addition, the films are abound with scenes where the students are competing to get the top place in the exams. In the scenes, the teachers / managers and the families display inciting and encouraging behaviors. Examples of the lines in these scenes are as follows:

"Rajan Damodhran always stood first in the class." - Taare-Zameen-Par / The Stars on the Ground

"Sir, Is it compulsory to sit acccording to our ranks? …. Anyone here in this batch to honour this pen? …. Nobody remembers the man who ever came second!" -Three Idiots

Film scenes reflected are in line with one of the major criticisms voiced in the education system in Turkey: Both the limitations of the traditional measurement and evaluation approaches and the high number of the population demanding education along with the low level of employment in the field seem to have created a competitive atmosphere in the educational system. Although selection as one of the objectives of the assessment involves the element of competition, it is essential that assess and improve the skills of the students' rather than having them competes with each other. The fact that the test scenes reflected in the films are of a competitive nature can cause the audience to give such a meaning into the measurement and evaluation activities.

Findings Related to Cheating in ME Activities

In 7 films, 9 scenes of cheating are seen to contain some elements which might create the perceptions to the effect that cheating is normal, that the cheater can gain prestige, and that cheating might be excused by the teacher. Particularly in the scenes belonging to the series of Hababam Sınıfı, cheating is reflected as an act that is usual for the students and it is an inseparable natural component of the exams. In the film Emperor Club, which is one of the films having a cheating scene, the student who participates in the quiz contests at both children and adults level cheats in both of the competitions and it is stated in the film that copying is normal or even necessary under the current life conditions. Obviously, cheating is a factor that reduces the validity and reliability of the measurement results. Cheating is described by Cizek & Wallack (2017) as an act to obtain an unfair gain / advantage before, during or after a test or homework. It is possible to measure the desired property according to its purpose by minimizing the negative elements such as cheating. When students are asked not to make a recourse to cheating within the scope of formal education, a moral behavior is implicitly expected from them, as well. Yet, the popular scenes of the films showing the act of cheating as excusable and pleasant can make the audiences normalize it and can create the notion that it is a behavior which is worth praising.

DISCUSSION and CONCLUSION

In the study, the scenes - related to the measurement and evaluation activities - of 26 films with educational content have been evaluated. These films have been proposed by the Ministry of National Education to the teachers. The scenes in the films have been studied both in view of measurement and

evaluation principles and of the way the measurement and evaluation activities are handled in them. Scenes have been examined under the following themes: teachers' use of traditional or complementary evaluation approaches, use of questions in measuring lower and high level mental skills; praise of these skills, the use of skills in daily life, the use of evaluation activities as reward / punishment or as competition, the use of inappropriate questions in the activities, the effects of activities on student attitudes and behaviors, the preparation of examinations, implementation and grading processes, giving feedback to students after the exam, and cheating.

A variety of studies have been conducted on the effect of visual and auditory media on the perceptions of individuals/societies (Budak, 1986; Couldry, 2000; Gerbner, 1974; Güçhan, 1993; Kaşkaya et al., 2011; Kontaş, 2016; Samsel & Perepa, 2013; Sivas, 2012; Şahin, 2011; Şentürk, 2009). In a study, Lin (2002), states that films are effective in ensuring the development of students' attitudes and motivations as well as in making learning permanent. In a similar way, according to the results of a study by Kontaş (2016), films - with educational themes - have an important contribution in the development of positive and negative attitudes for teachers. The fact that visual media appeals to more than one sensory organ and that it is easily accessible with the widespread use of technology make it attractive to individuals as it also facilitates better learning. Based on the assumption that the films recommended by MoNE to teachers are watched by the majority of teachers, it is inevitable for teachers to be affected by the scenes, content and sub-texts of these films voluntarily or involuntarily, and to develop a behavior. Another important point, however, is that these films have become favorite films also for the parents and students. So, they affect not only the teachers, but also the attitudes of the students and parents. It is seen that the teacher characters prepare the questions hastily and score them very easily with no conceivable criteria in the scenes. Teachers' acts of giving arbitrary or instantaneous exams, scoring without using a scoring key subjectively in favor of or against certain students, giving no feedback to the students after the exams can be normalized by the audiences. In a similar way, the reflections, in the scenes, of test environments which are not conformant to the principles pertaining to measurement and evaluation, of students being disturbed, of negligence of students' cognitive and affective development, and the reflections of tests for which validity and reliability factors are not taken into consideration can lead the audiences to the misconception that the students' performance can be measured under any circumstance and that the testing environments are of no importance. The scenes in the films can create the impression that teachers' responsibility for measurement and evaluation is of no significance at all.

As a result of the research, it is seen that mostly the traditional evaluation approaches have been used in the scenes involving ME activities. Relatively less use of complementary evaluation approaches in formal education reflects in the cinema as well, and the tools like structured grid, portfolio and graded scoring key are not found in the films. The types of exams that the teacher figures reflect on the screen are rather written/oral examinations and multiple choice tests which are used in conventional education system. That the traditional approach tends to evaluate the product rather than the process is reflected on the scenes as well. The learning required to be measured by the teachers remains in the recall and comprehension level and the scenes related to the measurement of high-level mental skills are insufficient. None of the films examined have any scene to reflect the use of knowledge in everyday life. The absence of these elements can create a perception in the audience that the skill which is important in the school setting is nothing but memorization. However, what is actually expected from the students in a school is that they should be able to turn the knowledge into practical skills, and their skills to abilities by using their own potential. The elements of praise for the lower-level mental skills in the scenes can reinforce the perception that these skills are more important than high-level mental skills, and that only memorization can bring success and praise. One of the striking results of the study is the use of ME activities for a goal other than their own. These activities are just a means of control mechanism for a program. It is a mechanism to check whether the program objectives have been achieved. However, in the films examined, it is seen that they are carried out in order to punish the students and they are even used as threats. There is a possibility that the viewer will normalize such behaviors upon seeing these scenes in the films. ME activities have not been implemented with the correct approaches and with their real purpose but as activities that create fear and anxiety in some scenes, affecting the individuals' own and family lives as well as their social relationships with the

environment. Since many similar examples are encountered in real life as well, this situation is very worrisome. The process of measurement and evaluation which, in fact, should be seen as a natural part of the educational process is perceived as a vital activity by students and families and thus it deviates from its real goal due to the misconceptions built in their minds. Misconception of the ME process and its goal in reality might be an outcome of these and similar other films showing it as a race process where competition prevails. Teachers' use of exams and their results as a competitive element in some scenes may also mean normalization of that misconception. It is seen in many scenes that cheating is also normalized and the students who cheat in the exams can gain prestige, and be excused by the teachers.

The research, after content analysis, has revealed the implicit messages of these films, which are suggested to teachers as part of in-service training, in terms of the concepts of Measurement and Evaluation and it has shown how these implicit messages can affect teachers, parents and students. The visual elements that the viewers are exposed to are tried to be examined rather than the intentions of the stories or scenarios. Therefore, the results of the research should not be interpreted as a critique of the films but as a set of perceptions and notions that these films can create in the audience. Also, the findings of the research on the elements in the films do not provide dependable evidence to the effect that these elements can lead to negative behavioral changes in teachers, students and families as the audience. While it is undeniable that these films would contribute to the teachers in the pedagogical sense and in many other ways, it will be appropriate to monitor / scrutinize them with an awareness of the sub-texts of the scenes in terms of the concepts of measurement and evaluation.

REFERENCES

Adıyaman, Y. (2005). *İlköğretim 4, 6 ve 8. sınıflarında Türkçe dersine giren öğretmenlerin ölçme değerlendirme düzeyleri* (Yayımlanmamış yüksek lisans tezi). Afyon Kocatepe Üniversitesi Sosyal Bilimler Enstitüsü, Afyon.

Akcan, E., & Polat, S. (2016). Eğitim konulu Türk filmlerinde öğretmen imajı: Öğretmen imajına tarihi bakış. *Kuram ve Uygulamada Eğitim Yönetimi*, *22*(3), 293-320.

Akıncı-Yüksel, N. A. (2015). Kültürel bir ürün olarak Türkiye'de sinema filmlerinde okul, öğretmen ve öğrenci temsilleri. *Global Media Journal: Turkish Edition, 6*(11), 1-17.

Anıl, D., & Acar, M. (2008). Sınıf öğretmenlerinin ölçme değerlendirme sürecinde karşılaştıkları sorunlara ilişkin görüşleri. *Yüzüncü Yıl Üniversitesi, Eğitim Fakültesi Dergisi, 5*(2), 44-61.

Anılan, H., Anagün, Ş. S., Atalay N., & Kılıç Z., (2016). Sınıf öğretmenlerinin öğrenme sürecini temel alan ölçme değerlendirme yaklaşımlarına ilişkin görüşleri. *Eğitim ve Öğretim Araştırmaları Dergisi, 5*(22), 200-221.

Armağan, İ. (1992). *Sanat toplumbilimi - demokrasi kültürüne giriş*. İzmir: İleri Kitabevi.

Bal, A. P. (2009). *İlköğretim beşinci sınıf matematik öğretiminde uygulanan ölçme ve değerlendirme yaklaşımlarının öğretmen ve öğrenci görüşleri doğrultusunda değerlendirilmesi* (Yayımlanmamış doktora tezi). Çukurova Üniversitesi Sosyal Bilimler Enstitüsü, Adana.

Beldağ, A., & Kaptan, S.Y. (2017). Arabalar filminin içerdiği değerlere ilişkin bir inceleme. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 18*(2), 487-499.

Belet, Ş. D., & Sağlam, F. (2015). Türkçe dersinde kullanılan ölçme-değerlendirme yöntem ve tekniklerinin sınıf öğretmenlerine göre değerlendirilmesi. *Anadolu Journal of Educational Sciences International, 5*(1), 115-145.

Bilgin, N. (2006). *Sosyal bilimlerde içerik analizi teknikler ve örnek çalışmalar*. Ankara: Siyasal Kitabevi.

Birkök, M. C. (2008). Bir toplumsallaştırma aracı olarak eğitimde alternatif medya kullanımı: Sinema filmleri. *Uluslararası İnsan Bilimleri Dergisi,* 5(2), 1-12.

Budak, M. (1986). *Sinema yazıları.* İstanbul: Bayrak Yayıncılık.

Cizek, G. J., & Wallack J. A. (2017). *Handboook of quantitative methods for detecting cheating on tests*. New York, NY: Routledge.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed.). New York, NY: Routledge.

Couldry, N. (2000). *The place of media power: pilgrims and witnesses of the media age*. New York, NY: Routledge.

Crocker, L., & Algina, J. (1986). *Introduction classical and modern test theory.* New York, NY: Harcourt Brace Javonovich College Publishers.

Çakan, M. (2004). Öğretmenlerin ölçme-değerlendirme uygulamaları ve yeterlik düzeyleri: İlk ve ortaöğretim. *Ankara Üniversitesi Eğitim Bilimleri Fakültesi Dergisi, 37*(2), 99-114.

Çetin, M. F. (2014). *Etkili dönütün akademik başarı, derse yönelik tutum ve üstbilişsel farkındalığa etkisi* (Yayımlanmamış yüksek lisans tezi). Çanakkale Onsekiz Mart Üniversitesi Eğitim Bilimleri Enstitüsü, Çanakkale.

Çevikbaş, M. (2018). Lise matematik öğretmenlerinin dönüt verme süreçlerinin ve dönüt algılarının incelenmesi. *Anadolu Journal of Educational Sciences International, 8*(1), 98-125.

Çoruhlu, T. Ş., Nas, S. E., & Çepni, S. (2009). Fen ve teknoloji öğretmenlerinin alternatif ölçme- değerlendirme tekniklerini kullanmada karşılaştıkları problemler: Trabzon örneği. *Yüzüncü Yıl Üniversitesi Eğitim Fakültesi Dergisi, 6*(1), 122-141.

Demirel, Ö. (2005). *Öğretme sanatı: Öğretimde planlama ve değerlendirme*. Ankara: Pegem A Yayıncılık.

Diken, B., & Laustsen. C. B. (2011). *Filmlerle sosyoloji* (Çev. S. Ertekin). İstanbul: Metis.

Doğan, C. D. (2017). "Hababam Sınıfı" filmlerinde yer alan ölçme ve değerlendirmeye ilişkin alt metinlerin incelenmesi. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 17*(1), 154-167.

Duban, N., & Küçükyılmaz, E. A. (2008). Sınıf öğretmeni adaylarının alternatif ölçme-değerlendirme yöntem ve tekniklerinin uygulama okullarında kullanımına ilişkin görüşleri. *İlköğretim Online, 7*(3), 769-784.

Ercan, E. E., & Demir, F. N. (2015). Yetiştirme kuramı: Anadolu üniversitesi fen fakültesinde yapılan araştırma. *Gümüşhane Üniversitesi İletişim Fakültesi Elektronik Dergisi, 3*(1), 127-144.

Erişen, Y. (1997). Öğretim elemanlarının dönüt ve düzeltme davranışlarını yerine getirme dereceleri. *Eğitim Yönetimi, 3*(1), 45-61.

Ertürk, S. (1972). *Eğitimde program geliştirme*. Ankara: Hacettepe Üniversitesi Basımevi.

Evin-Gencel. İ., & Özbaşı, D. (2013). Öğretmen adaylarının ölçme ve değerlendirme alanına yönelik yeterlik algılarının incelenmesi. *İlköğretim Online, 12*(1), 190-201.

Fidan, M., & Sak, İ.M. (2012). İlköğretim öğretmenlerinin tamamlayıcı ölçme değerlendirme teknikleri hakkında görüşleri. *Bartın Üniversitesi, Eğitim Fakültesi Dergisi, 1*(1), 174-189.

Fitz-Gibbon, C. T., & Morris, L. L. (1989). *How to design a program evaluation*. Los Angeles, CA: Sage Publications.

Gelbal, S., & Kelecioğlu, H. (2007). Öğretmenlerin ölçme değerlendirme yöntemleri hakkındaki yeterlik algıları ve karşılaştıkları sorunlar. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 33*, 135-145.

Gerbner, G. (1974). Communication: Society is the message. *Communication, 1*, 57-64.

Gömleksiz, M. N., & Bulut, D. (2007). Yeni fen ve teknoloji dersi öğretim programının uygulamadaki etkililiğinin değerlendirilmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 7*(1), 41-49.

Güçhan, G. (1993). Sinema-toplum ilişkileri. *Kurgu Dergisi, 12*, 51-71.

Güneş, T., Dilek, N. Ş., Hoplan, M., Çelikoğlu, M., & Demir, E. S. (2010). Öğretmenlerin alternatif değerlendirme konusundaki görüşleri ve yaptıkları uygulamalar. Z. Kaya, U. Demiray, D. Ergür, U. Tanyeri ve N. Akkuş (Ed.), *International conference on new trends in educationand their implications* (ss. 925-935). Ankara: Pegem Akademi.

Güven, S. (2008). Sınıf öğretmenlerinin yeni ilköğretim ders programlarının uygulanmasına ilişkin görüşleri. *Milli Eğitim Dergisi, 177*, 224-236.

Hamarat, D., Işıtan, S. Özcan, A., & Karaşahin, H. (2015). Okul öncesi dönem çocuklarının izledikleri çizgi filmler üzerine bir inceleme: Caillou ve sünger bob örneği. *Balıkesir University The Journal of Social Sciences Institute, 18*(33), 75-91.

Kalaycı, N. (2015). Toplumsal cinsiyet eşitliği açısından bir çizgi film çözümlemesi: Pepee. *Eğitim ve Bilim, 40* (177), 243-270.

Kaşkaya, A., Ünlü, İ., Akar, M. İ., & Özturan-Sağırlı, M. (2011). Okul ve öğretmen içerikli sinema filmlerinin öğretmen adaylarının mesleki tutumlarına ve öz yeterlik algılarına etkisi. *Kuram ve Uygulamada Eğitim Bilimleri, 11*(4), 1765-1783.

Kilmen, S., Akın Köstereliоğlu, M., & Köstereliоğlu, İ. (2007). Öğretmen adaylarının ölçme değerlendirme araç ve yaklaşımlarına ilişkin yeterlik algıları. *AİBÜ Eğitim Fakültesi Dergisi, 7*(1), 129-140.

Kontaş, H. (2016). The effect of an education-the med movie on the academic motivation of teacher candidates and their attitude towards teaching profession. *Journal of Education and Training Studies, 4*(6), 93-103.

Lin, L.Y. (2002). *The effects of feature films upon learners' motivation, listening and speaking skills: The learner-centered approach*. Taiwan: Learner Centered Instruction.

Linn, R., & Gronlund N. E. (1995). *Measurement assessment in teaching*. New Jersey, NJ: Prentice–Hall Inc.

Milli Eğitim Bakanlığı (2017). Mesleki gelişim programı. Erişim adresi: http://oygm.meb.gov.tr/www/2015-2019-mesleki-calisma-programlari/icerik/693

Özenç, M. (2013). Sınıf öğretmenlerinin alternatif ölçme ve değerlendirme bilgi düzeylerinin belirlenmesi. *Dicle Üniversitesi Ziya Gökalp Eğitim Fakültesi Dergisi, 21*, 157-178.

Özer, Ö. (2004). *Yetiştirme kuramı: Televizyonun kültürel işlevlerinin incelenmesi*. Eskişehir: Anadolu Üniversitesi Yayınları.

Pamukçu, C. (2015). *Tamamlayıcı ölçme ve değerlendirme gelişim programının coğrafya öğretmen adaylarının yeterlik algısı ve bilgi düzeyine etkisi* (Yayımlanmamış doktora tezi). Necmettin Erbakan Üniversitesi, Eğitim Bilimleri Enstitüsü, Konya.

Polat, Ç. S. (2011). *Engelli bireylere ilişkin kültürel tanımlamaların başka dilde aşk filmi üzerinden incelenmesi* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi, Sosyal Bilimler Enstitüsü, Ankara.

Polat, S., & Akcan, E. (2017). Eğitim temalı filmlerin çok kültürlü eğitim açısından analizi. *Turkish Studies International Periodical for the Languages, Literature and History of Turkishor Turkic, 12*(18), 475-504.

Samsel, M., & Perepa, P. (2013). The impact of media representation of disabilities on teachers' perceptions. *Media and Disability*, *28*(4), 138-145.

Serttaş, A. (2014). V for Vendetta filminin alımlama analizi ile sinemada televizyon. *Global Media Journal: TR Edition* 5(9), 303-321.

Sivas, A. (2012). Göstergebilim ve sinema ilişkisi üzerine bir deneme. *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi*, *11*(21), 527-538.

Sönmez, V. (2004). *Program geliştirmede öğretmen el kitabı*. Ankara: Anı Yayıncılık.

Sönmez, V., & Alacapınar, F. G. (2013). *Örneklendirilmiş bilimsel araştırma yöntemleri*. Ankara: Anı Yayıncılık.

Stiggns, R. J. (2001). *Student involved classroom assessment*. New Jersey, NJ: Prentice-Hall.

Şahin, A. (2011). *Eleştirel medya okuryazarlığı*. Ankara: Anı Yayıncılık.

Şentürk, R. (2009). Raymond Williams'ın televizyon teorisi. *Selçuk İletişim Dergisi*, *5*(4), 186-200.

Tavşancıl, E. & Aslan, E. (2001). *Sözel, yazılı ve diğer materyaller için içerik analizi ve uygulama örnekleri.* İstanbul: Epsilon Yayınevi.

Temel, A. (1991). Ortaöğretimde ölçme ve değerlendirme sorunları. *Yaşadıkça Eğitim Dergisi*, *18*, 23-27.

Tolon, B. (1978). *Toplum bilimine giriş*. Ankara: Kante Matbaası.

Turgut, F., & Baykul, Y. (2010). *Eğitimde ölçme ve değerlendirme*. Ankara: Pegem Akademi Yayıncılık.

Yakar, H. G. İ. (2013). Sinema filmlerinin eğitim amaçlı kullanımı: Tarihsel bir değerlendirme. *Hasan Ali Yücel Eğitim Fakültesi Dergisi. 19*(1), 21-36.

Yaman, S. (2011). Öğretmenlerin fen ve teknoloji dersinde ölçme ve değerlendirme uygulamalarına yönelik algıları. *İlköğretim Online, 10*(1), 244-256.

Yanpar, T. (1992). *Ankara ilkokullarındaki ikinci devre öğretmenlerinin öğretmenlik mesleği ve konu alanlarıyla ilgili eğitim ihtiyaçları* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Yapıcı, M., & Demirdelen, C. (2007). İlköğretim 4. sınıf programına ilişkin öğretmen görüşleri. *İlköğretim Online, 6*(2), 204-212.

Yıldırım, A., & Şimşek, H. (2016). *Sosyal bilimlerde nitel araştırma yöntemleri.* Ankara: Seçkin Yayınevi.

Yıldırım, C. (1999). *Eğitimde ölçme ve değerlendirme.* Ankara: ÖSYM Yayınları.

Yıldırım, N., Tüzel, E., & Yıldırım, V. Y. (2016). Aamir Khan filmlerinin eğitimsel açıdan incelenmesi: 3 Idiots (3 aptal) ve Taare Zameen Par (Her Çocuk Özeldir) üzerine nitel bir değerlendirme. A*tatürk Üniversitesi Güzel sanatlar Enstitüsü Dergisi*, *36*, 210-244.

Appendix A. Film Observation Form

Original name:				
Turkish name:				
Release year:				
Country:				
Runtime:				
In the film, the scene related to measurement and evaluation activities	☐ observed	☐ not observed		
Themes	**Y/N**	**Number of Observed Scenes**	**Time of Observation of (the Scene)**	**Annotations**
1. Use of written exams for ME				
2. Use of oral exams for ME				
3. Use of multiple-choice tests for ME				
4. Use of short-response tests for ME				
5. Use of matching tests for ME				
6. Use of true false tests for ME				
7. Use of complementary measurement tools for ME				
8. Use of interviews for ME				
9. Use of reward element for ME				
10. Use of penal clause for ME				
11. Use of questions at the recall / knowledge level in ME				
12. Use of questions to measure high-level mental processes in ME				
13. Exam preparation process (easy, difficult etc.)				
14. Exam scoring process (easy / difficult, objective / subjective, etc.)				
15. Application of examinations (environment, conditions of application, duration etc.)				
16. Use of questions that do not conform to the ME principles in examinations				
17. Giving feedback to students at the end of ME				
18. Effect of ME activities on student attitudes and behaviors				
19. Compliments on the knowledge / recall behavior				
20. Praise for high-level mental skills				
21. Compliments on the use of information in daily life				
22. Use of ME in competition				
23. Use of ME activities as a target, not as a means				
24. Number of ME activities (Number of exams faced by students, etc.)				
25. Use of ME activities for a certificate / diploma				
26. Cheating				
Quotes				
Related Theme Code	**Quote**			

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 219-234

An Example of Empirical and Model Based Methods for Performance Descriptors: English Proficiency Test *

Serkan ARIKAN ** Sevilay KİLMEN *** Mehmet ABİ **** Eda ÜSTÜNEL *****

Abstract
Great emphasis is given to the development of high-stake tests all around the world and in Turkey. However, limited emphasis is given to adequate score reporting. Too much emphasis on rankings and almost no emphasis on performance level descriptors (meaning of the scores) have leaded a "ranking culture" in Turkey. There is an immense need to raise awareness about score reporting and performance level descriptions in Turkey. This study aims to raise awareness about the use of performance level descriptors in a high-stake exam in Turkey, an English proficiency exam. The study sample is consisted of 630 undergraduate students who took the 2016-2017 English proficiency exam of a public university in the southwest of the Turkey. In order to identify the potential exemplars, two types of item mapping methods (i.e. experimental based method and model-based method) were used in the present study. Item grouping for performance level descriptors provided hierarchical and interpretable structure. Using these performance level descriptors, it is possible to give criterion referenced feedback to each student about his/her reading abilities.

Key Words: Criterion referenced assessment, performance level descriptors, empirical method, model-based method, construct map.

INTRODUCTION

Every year many exams were prepared to evaluate student performances and to give pass or fail decisions all around the world. Generally, great emphasis is given to the development of these high-stake tests. However, limited emphasis is given to adequate score reporting (Goodman & Hambleton, 2004; Karantonis, 2017). Students get their scores, but they generally do not have any idea what these scores mean. Similarly, instructors give scores to their students, but could not use these scores adequately in their instructions as these scores do not make concrete sense to them, either. In the United States, effort is given to find effective ways to report results of high-stake tests by giving meaning to scores (Karantonis, 2017). The research on standard setting is focusing on which methods are more effective (Karantonis, 2017; Karantonis & Sireci, 2006). Karantonis (2017) stated that there is still a need to examine different item-mapping methods to identify exemplar items for performance level descriptors. However, in Turkey, although exams take a crucial role in every grade level even starting from primary education, very little emphasis is given to score reporting, standard setting procedures and performance level interpretations. Each component of education is strongly affected by high-stake exams; however, stakeholders of education could not interpret and use exam results as no performance level descriptors associated with the scores are given. Students and educators are mainly interested in the normative results such as the rank of students in an exam. Criterion referenced results are very rarely used. Too much emphasis on rankings and almost no emphasis on performance level descriptors

* A part of the study was presented at 2018 EDUCCON Education Conference, Ankara University, Ankara, Turkey.
** Assist. Prof. PhD., Boğaziçi University, Faculty of Education, İstanbul-Turkey, serkan.arikan1@boun.edu.tr, ORCID ID: 0000-0001-9610-5496
*** Assoc. Prof. PhD., Bolu Abant Izzet Baysal University, Faculty of Education, Bolu-Turkey, kaplansevilay@yahoo.com, ORCID ID: 0000-0002-5432-7338
**** Lect., Muğla Sıtkı Koçman University, College of Foreign Language, Muğla-Turkey, mehmetabi@mu.edu.tr, ORCID ID: 0000-0002-4976-5173
***** Prof. PhD., Muğla Sıtkı Koçman University, College of Foreign Language, Muğla-Turkey, eustunel@mu.edu.tr, ORCID ID: 0000-0003-2137-1671

To cite this article:
Arıkan, S., Kilmen, S., Abi, M., & Üstünel, E. (2019). An example of empirical and model based methods for performance descriptors: English proficiency test. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 219-234. doi: 10.21031/epod.477857

Received: 02.11.2018
Accepted: 30.06.2019

have led a ranking culture all over the country. Additionally, there is no public or academic demand to force private and national testing companies to report test results in clear and meaningful way. Turkish teachers reported they rarely use exam results to give feedback compared to European colleagues (Demirtaşlı, 2009). Therefore, there is an immense need to raise awareness about score reporting, standard setting procedures and performance level interpretations in Turkey. As Shulman (2009) stated "assessment is a powerful tool for raising the quality of teaching and learning. It should be used diagnostically and interactively, not as a form of autopsy" (p. 237). We need to use assessment more effectively and this study aims to raise awareness about the use of performance level descriptors in a high-stake exam in Turkey by describing and exemplifying the procedures of defining performance level descriptors. This study shows how a teacher group could get performance level descriptors by using empirical method to get performance level descriptors and also shows how experts could use ConstructMap to get performance level descriptors using model-based methods.

Performance Level Descriptor Methods

There are two major methods for defining performance level descriptors: the empirical method and the model-based method. These methods are described in this part.

The empirical method

The empirical method (Zwick, Senturk, Wang, & Loomis, 2001) corresponds to direct method, defined originally by Beaton and Allen (1992). According to this method, first a few carefully dispersed scale points are determined. These points are called *anchor points* or *anchor levels* and they are defined as judgmental. Then, the student groups at anchor points are determined. But since there may be a small number of students at these points or even no student may be present, a range of points near the anchor points is determined. The items correctly answered by the majority of the students in the range are determined. These items are called exemplars. Finally, the performance represented by these items is defined (Beaton & Allen, 1992).

For example, anchor points can be defined as 10, 20, 30, and 40 on a scale scored from 0 to 50. Regarding how close a point interval to anchor points is to be determined, Beaton and Allen (1992, p. 195) recommended that "this interval should be large enough so that there will be an adequate sample in group k and yet small enough so that the score values are clearly distinguishable from the adjacent anchor points". For the anchor points in the example, near the anchor point can be specified as anchor point ±2. In this case the first anchor point interval is determined as 8 to 12 points. Other anchor intervals are determined by adding and subtracting 2 points. After the near the anchor points are identified, the correct answers are determined by the majority of the students in that range. At this point, what is meant by the majority of students is needed to be operationally defined. Different correct response probabilities (e.g. 50%, 65%, and 80%) have been used in the literature (Beaton & Allen, 1992). One of these probabilities could be selected for this method. For example, if the probability of correct response is identified to be 65%, the items correctly answered by 65% of the individuals in each anchor interval are determined. For each anchor interval, the cognitive and content related properties measured by these items are determined and the performance for each anchor interval is defined.

The model-based method

In model-based method, as in the empirical method, exemplars are chosen based on the probability of correct answer of the item. The difference of the model-based method from the empirical method is that correct response probabilities are estimated based on the item response theory model (Zwick et al., 2001). According to item response theory, ability and item parameters can be placed on the same scale. At this scale, the difficulty parameter of an item is settled at the same time as individuals who are likely to respond to that item by 50%. By utilizing this property of item response theory, it is

possible to find items with 50% probability of responding in a certain proficiency score interval. These items are the items that are likely to be correctly answered 50% by the individuals in this point range (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). For example, the items that individuals in the range of 2.20 - 3.00 points can correctly answer with 50% probability are those with difficulty parameters ranging from 2.20 - 3.00.

As mentioned above, the difference between these two methods is the way in which the response probabilities are calculated. In the empirical method, the response probability is calculated based on the classical test theory, while in the model-based method, it is calculated based on the item response theory.

Purpose of the Study

This study aimed to illustrate how performance level descriptors could be defined using a dataset of an English proficiency test. There is a need to report educational test result more efficiently by developing adequate score reporting methods, especially in Turkey. Providing verbal descriptors for related score intervals, the exam results will be more meaningful and required feedback could be given to stakeholders. An example from a high-stake English proficiency exam was used to illustrate how empirical method and model-based method using ConstructMaps could be applied in practice. With this incentive, the research question of the study is set as How can we define performance level descriptors for an English proficiency exam?

METHOD

This study is a standard setting study that aims to give a meaning to test scores. This study expected to raise awareness about the use of performance level descriptors in a high-stake exam in Turkey. In order to achieve this goal, two item-mapping methods to identify exemplar items for performance level descriptors were used. The participants, instrument and data analysis procedures were described in this section.

Participants

Total of 630 undergraduate students took the 2016-2017 English proficiency exam of a public university in the southwest of the Turkey. Sixty two percent of the students were male, and thirty two percent were females. This public university mainly has programs in Turkish but there are some programs that have the medium instruction in English. The participants of this study were the students who were registered to preparatory class of foreign language school of this university. These students were required to get overall score of 60 out of 100 to start their undergraduate programs.

Data Collection Instrument

This study used English proficiency test to define performance level descriptors. The English proficiency test has four major dimensions: Reading, Listening, Writing and Speaking. This test was developed by test development team of foreign language school of the university. The proficiency test was developed based on the assessment framework of Common European Framework and aimed to be in B1 to B2 level. This study focuses on reading part of this test. Reading part included reading paragraphs and there were 19 items in the format of matching, short answer and multiple choice.

Data Analysis

Preliminary analysis

As a preliminary analysis, internal consistency of reading test was tested using The Cronbach's Alpha reliability coefficient. According to George and Mallery (2003) Cronbach's Alpha coefficient should be higher than .700. An instrument with Cronbach's Alpha coefficient higher than .800 is considered as a good instrument as and higher than .900 is considered as a marvelous instrument. Besides, descriptive statistics related to reading test results were reported. SPSS 22.0 was used to conduct internal consistency and descriptive statistics.

Reading test was developed to measure one main reading ability. Therefore, confirmatory factor analysis was conducted to test unidimensionality of the reading test. Confirmatory factor analysis requires an assessment to establish whether or not the proposed model is a good one. A good model is a model in which the difference between covariance matrix obtained from student data and covariance matrix implied by the hypothesized model is minimum (Ullman, 2001). This difference is evaluated by using several fit indices. Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and Root Mean Square Error of Approximation (RMSEA) are widely reported fit indices to assess goodness of fit of confirmatory factor analysis. In this study, CFI and TLI values higher than .900 was considered as acceptable fit and .950 and above was considered as good fit; and RMSEA values .080 or less was considered as an acceptable fit and .060 or less was considered as a good fit (Browne & Cudeck, 1993; Hu & Bentler, 1999). Confirmatory factor analysis was conducted by MPLUS 7.4 program (Muthen & Muthen, 2015).

Differential Item Functioning (DIF) analysis was conducted to evaluate the fairness and equality of tests on item level in investigating the comparability of gender performances. Having an instrument without DIF items is an indication of a well-prepared instrument in terms of group comparisons and fairness. In the study, logistic regression (LR) and Structural Equation Modeling (SEM) DIF methods were used. In the logistic regression procedure, as a first step, only total score (model1), then total score and grouping variable (model2), and finally total score, grouping variable and their interaction (model3) were used as predictors. Significance of country and their interaction, and the change in R^2 value were taken as evidence for uniform bias and non-uniform bias, respectively (Zumbo, 1999). Jodoin and Gierl (2001) proposed ΔR^2 higher than 0.035 indicates moderate DIF and higher than 0.070 indicates large DIF. SPSS 22.0 programs were used to conduct logistic regression analysis. In the SEM procedure, a Confirmatory Factor Analysis (unifactorial, with all items as indicators of the latent variable) is conducted to assess configural and scalar invariance. The difference between incremental types of model fit is evaluated as the factor loadings and intercepts are forced to be equal for comparison groups (van de Vijver, 2017). If the difference in comparative fit index (CFI) and Tucker Lewis index (TLI) between configural and the scalar invariance model is larger than .010 modification indices are investigated to identify DIF items (Cheung & Rensvold, 2002). Mplus 7.4 program was used for SEM DIF detection procedure (Muthen & Muthen, 2015).

Defining performance level descriptors

Determination of exemplars according to the empirical method: First, the exemplar items were determined. In order to determine the potential exemplars according to empirical method using 50%, 67%, and 80% response probability, first, raw scores were converted to zero to hundred grade scale. The scores were clustered into five categories (0 - 20; 21 - 40; 41 - 60; 61 - 80; 81 - 100). The students in each score category was identified and then the proportion of correct response of each item for each score category was calculated using IBM SPSS 22. These proportions could be considered as classical test theory item difficulty indices for each item in each score category. In the present study, three different response probabilities (RP) were used to determine the exemplars: 50% RP: The items answered correctly by at least 50% of the participants in each performance level were selected as exemplar items; 67% RP: The items answered correctly by at least 67% of the participants in each

performance level were selected as exemplar items; 80% RP: The items answered correctly by at least 80% of the participants in each performance level were selected as exemplar items. For example, at the third performance level (41 - 60), the proportion of correct response for item 3 was calculated as 60.2%. This item was not chosen as an exemplar according to the empirical based method using 67% and 80%, while it was selected as an exemplar item according to empirical based method using 50%.

Determination of exemplars according to the model-based method: In the present study, ConstructMap 4.6 (Kennedy, Wilson, Draney, Tutunciyan, & Vorp, 2010) program was used which gives the total raw score of the students, student ability estimation and item difficulty values on Wright map. The program analyzes 1-0 item scores based on the Rasch model of item response theory. The Wright map shows student ability scores and item difficulty values on the same scale. In addition, raw scores can be reported on this map. Items were given in the order related to their difficulty indices and item clusters were investigated to decide the cut scores for each performance level.

RESULTS

Psychometric Properties and Item Bias Analysis

Internal consistency analysis

The Cronbach's Alpha reliability coefficient value in the proficiency exam reading part calculated as .814 with 19 items. This value indicated a good internal consistency (George & Mallery, 2003). The corrected item-total correlation coefficient of each item was higher than .200 indicated that all items correlated with total score as expected.

Descriptive statistics

Reading test consisted of 19 items that were scored dichotomously. The reading score of students ranged from 0 to 19 (M = 10.06, SD = 4.38). Reading scores were normally distributed, with skewness of 0.15 and kurtosis of -0.86. Students were 391 men and 239 women (men: M = 9.94, SD = 4.23; women: M = 10.24, SD = 4.62). An independent-samples t-test indicated that reading scores of men and women were not significantly different ($t_{(628)} = 0.831$, $p > .05$, $d = 0.07$).

Table 1. Descriptive Statistics of the Reading Test

N	Mean	Standard Deviation	Standard Error of The Mean	Skewness	Kurtosis
630	10.06	4.38	.17	0.15	-0.86

Factor structure

Reading test aimed to measure one dimensional reading ability of students (See Figure 1). Therefore, confirmatory factor analysis was conducted to test whether 19 items reading test was unidimensional as it was proposed (see Table 2). The results showed that RMSEA, CFI and TLI values indicated an acceptable fit of the data to the unidimensional model (RMSEA = .054 < .060; CFI = .918 > .900). Thus, confirmatory factor analysis findings indicated that the proposed model was supported by the collected reading test data.

Table 2. One-dimensional Confirmatory Factor Analysis Results

χ^2/df	RMSEA	CFI	TLI
2.836***	.054	.918	.908

***p < .001.

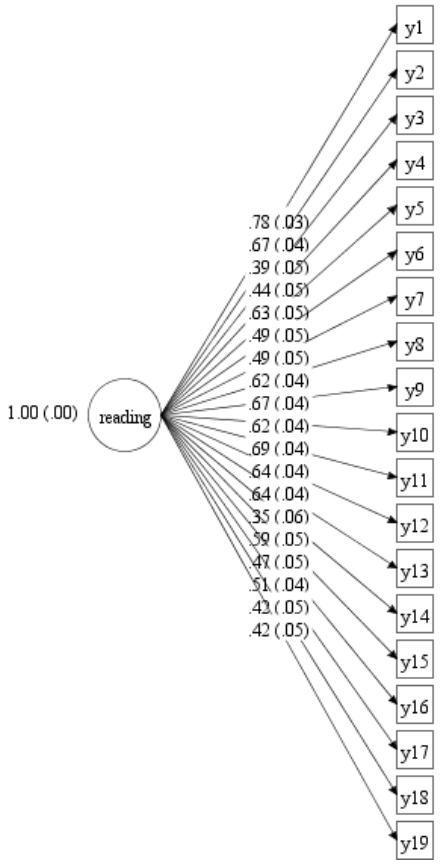


Figure 1. The Proposed Structure of Reading Test

Item bias

In this section, gender related DIF results based on Logistic Regression and Structural Equation Modeling DIF detection methods were presented. DIF results using LR method was presented in Table 3. The results indicated that none of the reading items showed DIF for gender groups. SEM DIF results are presented in Table 4. In comparing answers of girls and boys, none of the reading items showed DIF for gender groups either. Therefore, using two different DIF detection methods, it was concluded that reading test did not contain any DIF items for gender groups which was a fairness indicator of the test.

Table 3. Logistic Regression DIF Results

Item No	**Girls-Boys ΔR^2**
01	.004
02	.007
03	.002
04	.009
05	.006
06	.001
07	.001
08	.005
09	.007
10	.001
11	.001
12	.006
13	.004
14	.001
15	.004
16	.001
17	.003
18	.003
19	.002

Table 4. SEM DIF Results

Model	χ^2/df	RMSEA	CFI	ΔCFI	TLI	ΔTLI	DIF ITEMS
Configural	1.483**	.039	.956		.950		None
Scalar	1.464**	.038	.956	.000	.952	-.002	

**p < .01.

Item parameters according to classical and item response theory

In Table 5, item difficulty and item discrimination indices calculated by classical test theory and item response theory were reported. According to classical test theory item analysis statistics, the difficulty of the items were ranged from .31 to .85 with the mean value of .53; and the discrimination index was ranged from .24 to .53 with the mean value of .39. One parameter item response theory (Rasch model) results produced item difficulty indices ranging from -1.90 to 1.12 with the mean value of 0.00. These values indicated that the reading test had medium level difficulty.

Table 5. Item Parameters According to Classical and Item Response Theory

Item	Item Difficulty Index	Item Discrimination Index	b Parameter
1	.53	.53	0.03
2	.52	.43	0.04
3	.63	.28	-0.48
4	.34	.29	0.99
5	.32	.44	1.12
6	.38	.34	0.77
7	.68	.33	-0.79
8	.60	.44	-0.34
9	.59	.48	-0.30
10	.59	.44	-0.28
11	.52	.50	0.05
12	.60	.45	-0.33
13	.56	.47	-0.13
14	.31	.24	1.18
15	.85	.35	-1.90
16	.47	.34	0.29
17	.54	.37	-0.03
18	.65	.31	-0.59
19	.39	.30	0.72
Total	.53	.39	0.00

Defining Performance Level Descriptors

Identifying exemplar items using empirical method

Using RP 50, RP 67 and RP 80, exemplar items for each score interval (0 - 20; 21 - 40; etc.) were decided (see Table 6). Exemplar item grouping results were affected from chosen response probability. While an item was located to lower score intervals in RP 50, the same item was generally located to higher score intervals in RP 80. For score interval of 0 - 20, none of the items were located. This means that students who got a score between 0 and 20 in reading part could not achieve none of the items on general. In the next section how these item classifications were used to define performance level descriptors was explained. Additionally, the hierarchical structures were observed for RP 50, RP 67 and RP 80. If an item was located in one of the score interval (answered correctly by students in this score interval with required percentage) then the item was achieved by students in above score intervals with required percentage, too.

Table 6. Exemplar Items in Empirical Method

PL	n	RP 50	RP 67	RP 80*
0-20	31	-	-	-
21-40	174	15	15	-
41-60	186	3, 7, 8, 9, 10, 12, 13, 18	7	15
61-80	156	1, 2, 6, 11, 16, 17	1, 2, 3, 8, 9, 10, 11, 12, 13, 17, 18	1, 7, 8, 9, 10, 12
81-100	83	4, 5, 14, 19	4, 5, 6, 14, 16, 19	2, 3, 5, 11, 13, 16, 17, 18, 19

PL: performance level, RP: response probability. * Item 4, 6 and 14 could not be classified to any PL for RP 80.

Performance level descriptors using empirical method

In Table 6, exemplar items were reported with different response probabilities to show how each response probability affected the classification. In order to define performance level descriptors, RP 67 was selected. RP 50 was justified as the number of students at a particular score interval can do a task exceeds the number of students who cannot do the task (Zwick et al., 2001). However, RP 50 is criticized as being too low for a standard. Kolstad et al., (1998) stated that "if one is going to say that people with a particular score on an assessment can successfully perform a particular assessment task, one wants to be fairly sure that a substantial majority of them can do it" (p. 11). RP 80 could be used if the aim of the test requires higher percentage correct values. RP 80 was considered to be too stringent (Kolstad et al., 1998). In this study, three items (Item 4, 6 and 14) could not be located to any score interval for this reason. In RP 67 two third of the students were required to answer the item correctly in related score interval. RP 67 was justified as being consistent with the mastery notion (Kolstad et al., 1998) and maximizing the information of the correct response under several IRT models (Huynh, 2006). Therefore, performance level descriptors were defined using exemplar items under RP 67. The performance level descriptors were defined by three experienced scholars.

Results showed that students in score interval 0 - 20 could not show any reading ability measured in this test. Students in score interval 21 - 40 "can recognize a detail from context by using more frequently used vocabulary item (from k1 band) in the question root as an explicit clue". The ability of students in score interval 41 - 60 could be exemplified as, in addition to previously described ability, "can recognize a detail from context by using frequently vocabulary item (from k1 band) in the question root as an explicit clue". There was a small difference between these two abilities and for these groups only one item was located. For score intervals 61 - 80 and 81 - 100, there were more items. This might indicate that this test could better differentiate between score intervals of 0 - 60, 61 - 80 and 81 - 100 which is reasonable in a sense that a student should get overall score of 60 to be successful. Students in score interval 61 - 80 "can infer a detail by using an explicit clue in the text" whereas students in score interval 81 - 100 "can infer the meaning by using implicit clues in the text with less frequently used vocabulary" in addition to previously described abilities. It is also important to note that these structures are based on a probabilistic view in which a student in a score interval could have these abilities with at least 67% probability.

Cross validation of exemplar items in empirical method

As empirical method is based on percentages calculated according to classical test theory and as classical test theory is affected from different samples, the dataset was divided randomly into two to cross validate the results. In Table 8 and Table 9 these results were reported. In sample 1, for RP 50 and RP 67 only one item was located to different score interval whereas for RP 80, two items were mislocated (0.95, 0.95, and 0.89 convergence ratios, respectively). In sample 2, for RP 50 and RP 67 two items were located to different score interval whereas for RP 80, four items were located differently (0.89, 0.89, and 0.74 convergence ratios, respectively). These results showed that RP 80 was affected from sample change compared to RP 50 and RP 67. This finding also justified not selecting RP 80 for defining performance level descriptors.

Table 7. Performance Level Descriptors in Empirical Method

Level	PL	n	RP 67%	Performance Level Descriptors
1	0-20	31	-	-
2	21-40	174	15	• Can recognize a detail from context by using more frequently used vocabulary item (from k1 band) in the question root as an explicit clue.
3	41-60	186	7	• Can recognize a detail from context by using frequently vocabulary item (from k1 band) in the question root as an explicit clue.
4	61-80	156	1, 2, 3, 8, 9, 10, 11, 12, 13, 17, 18	• Can recognize a detail from context by using more frequently used vocabulary item (from k2 band) in the question root as an explicit clue. • Can follow the development of text structure and decide from where in the text each sentence is removed by using an explicit clue. • Can reach a conclusion by using an implicit clue in the text. • Can infer a detail by using an explicit clue in the text.
5	81-100	83	4, 5, 6, 14, 16, 19	• Can follow the development of text structure and can decide from where in the text each sentence is removed by using an implicit clue. • Can infer the meaning by using explicit clues in the text. • Can infer the meaning by using implicit clues in the text with less frequently used vocabulary. • Can infer writer's attitude and viewpoint.

Table 8. Cross Validation of Exemplar Items in Empirical Method-Sample 1

PL	n	RP 50	RP 67*	RP 80**
0-20	20	-	-	-
21-40	85	15	15	-
41-60	98	3, 7, 8, 9, 10, 12, 13, 18	7	15
61-80	61	1, 2, 6, 11, 16, 17, **19**	1, 2, 3, 8, 9, 10, 11, 12, 13, 17, 18	1, **2**, 7, 8, 9, 10, 12
81-100	47	4, 5, 14	4, 5, 6, 16, 19	3, **4**, 5, 11, 13, 16, 17, 18, 19

PL: performance level, RP: Response probability. * Item 14 could not be classified to any PL for RP 67. ** Item 6 and 14 could not be classified to any PL for RP 80

Table 9. Cross Validation of Exemplar Items in Empirical Method-Sample 2

PL	n	RP 50	RP 67*	RP 80
0-20	11	-	-	-
21-40	89	15	15	-
41-60	88	3, 7, 8, 9, 10, 12, **16,** 18	7, **18**	15
61-80	95	1, 2, 6, 11, **13**, 17	1, 2, 3, 8, 9, 10, 11, 12, 13, 17	1, 7, 8, 9, 10, 12, **18**
81-100	36	4, 5, 14, 19	5, 6, 14, 16, 19	2, 3, 11, 13, **14**, 17

PL: performance level, RP: Response probability. * Item 4 could not be classified to any PL for RP 67. ** Item 4, 5, 6, 16, 19 could not be classified to any PL for RP 80

Identifying exemplar items using model-based method using ConstructMap

ConstructMap 4.6.0 program was used to get Wright Map (See Figure 2). Wright Map provided ability level of students (ranging from -3 to +3), raw score associated with this ability levels, number of students in each ability level (denoted by X's) and item numbers ordered based on difficulty estimation done based on item response theory. The next step is to decide item groups by setting cut points. Among several approaches about how to decide cut points, The Construct Mapping method (Draney & Wilson, 2009) was used to identify the exemplar items. The Construct Mapping method was selected as experts defining performance level description (panelists) were given items' location and related scale scores. Panelists examined the data and items and selected the best locations for cut scores.

In the study, panelists investigated item clusters in the Wright Map and grouped items as given in Table 10. Then the scale scores intervals (theta) were reported for each level with RP67. These scale scores were estimated using the item response theory. Items were investigated in content and cognitive processes and performance level descriptors were provided. The results provide hierarchical structure for cognitive processes.

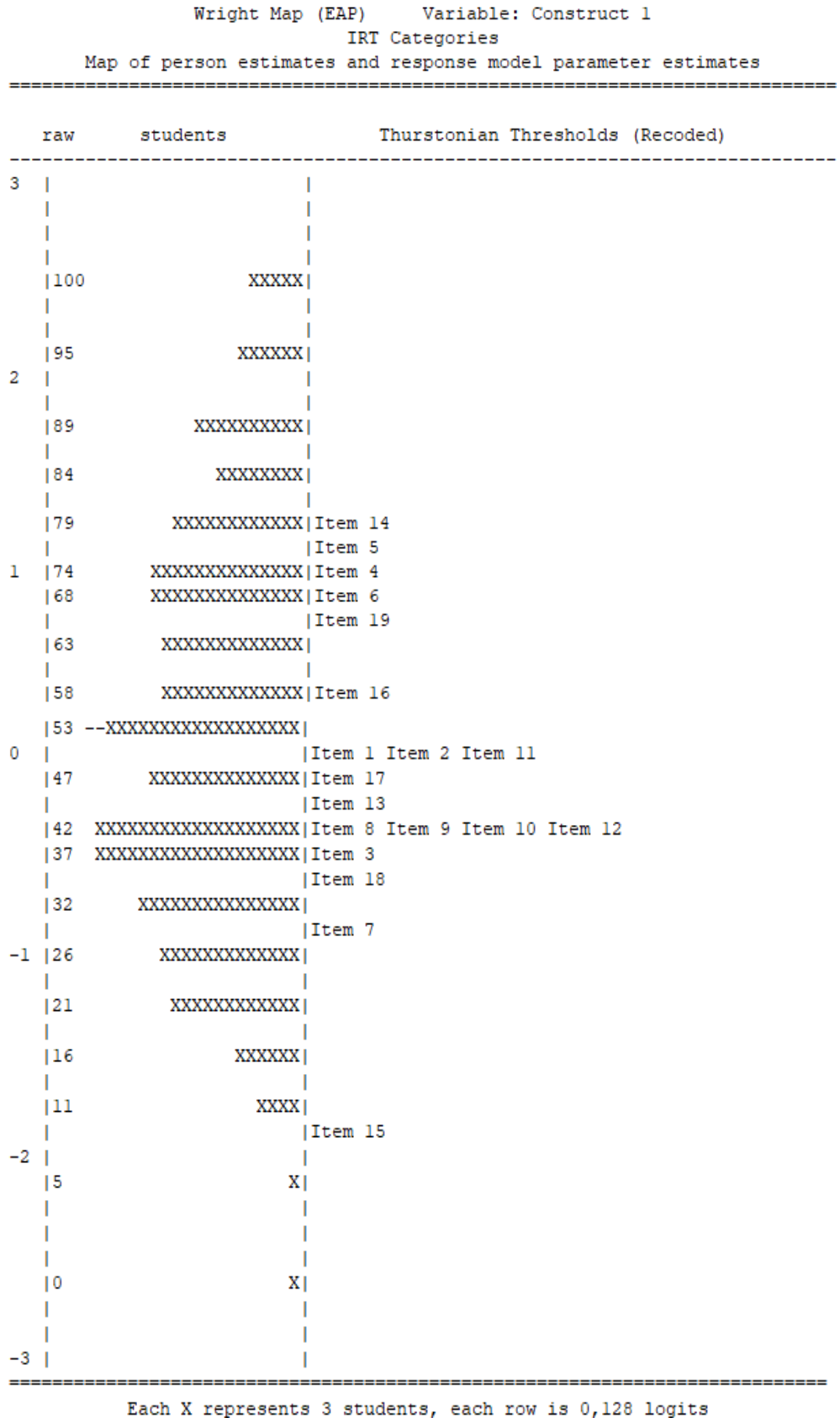


Figure 2. Wright Map Obtained by ConstructMap Program

Table 10. Item Grouping According to Construct Mapping Method

Level	Items	Theta Score Interval RP 67	Performance Level Descriptors
1	15	-0.60 and below	* Can recognize a detail from context by using more frequently used vocabulary item (from k1 band) in the question root as an explicit clue.
2	7	-0.60 and 0.00	* Can recognize a detail from context by using frequently vocabulary item (from k1 band) in the question root as an explicit clue.
3	1, 2, 3, 8, 9, 10, 11, 12, 13, 17, 18	0.00 and 0.90	* Can recognize a detail from context by using more frequently used vocabulary item (from k2 band) in the question root as an explicit clue. * Can follow the development of text structure and decide from where in the text each sentence is removed by using an explicit clue. * Can reach a conclusion by using an implicit clue in the text. * Can infer a detail by using an explicit clue in the text.
4	16	0.90 and 1.25	* Can infer writer's viewpoint.
5	4, 5, 6, 14, 19	1.25 and above	* Can follow the development of text structure and can decide from where in the text each sentence is removed by using an implicit clue. * Can infer the meaning by using explicit clues in the text. * Can infer the meaning by using implicit clues in the text with less frequently used vocabulary. * Can infer writer's attitude.

DISCUSSION and CONCLUSION

This study aimed to raise awareness about the importance of criterion referenced assessment via showing how performance level descriptors in a high-stake exam in Turkey could be defined. Giving too much emphasis on norm referenced assessment by rankings and almost no emphasis on criterion referenced assessment is continuing to harm the educational system from early years of primary school to university education. Especially national large-scale assessments that aim to select limited number of students among huge number of students to a higher educational institution focuses on norm referenced assessment in Turkey. However, there are national assessments, especially language tests, that aims to decide who are proficient or not, but even the results of these assessments are not reported with the criterion referenced perspective. Therefore, criterion referenced assessment is undervalued. There is a need to use criterion referenced assessment via providing performance level descriptors to integrate assessment results to the instructions and to provide concrete feedback to the stakeholders. Performance level descriptors could be used to follow the development of a student throughout the years of assessments. Therefore, a student who started from lower levels could increase his or her performance over years and this development could create a confidence for the student. Only ranking students is harming majority of the students as top rankings are reserved by top achievers.

One of the reasons of why assessment results based on criterion referenced assessment via performance level descriptors is not popular could be that there are very limited examples of performance level descriptors in Turkish context. Defining performance level descriptors requires more detailed effort and know how compared to providing norm referenced assessment results. This study showed how performance level descriptors could be defined using empirical method and model-based method. Empirical method is based on classical test theory and easier to implement and model-based method is based on item response theory and requires expertise on statistical software. In both methods, in the process of defining the descriptors for the score intervals, there is a hierarchical structure among the item clusters, and items that are located in higher score intervals require higher cognitive demands. As it is known, Wright maps were based on the item response theory, in which the item parameters could be estimated independently from the sample. In the study, we obtained similar results for both empirical method and model-based method. In the relevant literature, similar results were obtained in studies in different fields (e.g. mathematics). In the previous literature, it was found that the results obtained from the empirical method and wright maps were similar (e.g. Arıkan & Kilmen, 2018). As both methods produce similar item rankings and item clusters in this study, teachers could use empirical method to define performance level descriptors for their assessments and measurement experts could use model-based methods to get more stable results.

Teacher groups with limited access to the measurement experts could follow the steps described in the empirical method and could get item clusters and then could describe required abilities by the items. The study showed that with 600 students the findings were consistent with the smaller samples. With smaller number of students, the results could be more sample dependent, but the feedbacks based on performance level descriptors would be still useful for this specific group. Teachers could cooperate with other teachers to increase the number of students in their assessments and group discussion on defining performance level descriptors would be beneficial for them. Testing companies with measurement specialist and bigger schools that have measurement department are advised to use model-based method. Item statistics estimated by item response theory are sample independent which makes them more consistent (Hambleton & Jones, 1993). Cooperating with teachers and experts, Construct Mapping method is useful in defining performance level descriptors based on item analysis and item mapping.

Overall, we showed that it is possible to define performance level descriptors for an English proficiency exam. With the help of verbal descriptors for related score intervals, the exam results will be more meaningful and related feedback will be given to students, parents and school administration. Teachers and administration are expected to use this information to raise the quality of education. The student achievement outcome was defined according to what students can do and cannot do, therefore, overall success of given education throughout the year would be evaluated by these standards. When similar assessment is used for incoming proficiency exams, the outcome could also be comparable in terms of these standards. For students who could not achieve this test could be provided what they can do in addition to what they cannot do. These feedbacks are expected to help these students to shape their remedial studies.

The limitation of this study is that the number of reading items was not that high, and the items were generally loaded above score of 60. As a result, for some score intervals, one item was loaded. Defining performance level descriptors based on a limited number of items would threat the reliability of the findings. Therefore, having more items that have more equal distribution over score intervals would be preferable. Piloting items and selecting items according to pilot item analysis could be beneficial when administrating the items beforehand is possible.

REFERENCES

Arıkan, S., & Kilmen, S. (2018). Sınıf içi ölçme ve değerlendirmede puanlara anlam kazandırma: %70 doğru yanıt yöntemi. *İlköğretim Online, 17*(2), 888-908. doi: 10.17051/ilkonline.2018.419337

Beaton, A. E., & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191-204. doi: 10.3102/10769986017002191

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 137–162). Newbury Park, CA: Sage.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233–255. doi: 10.1207/S15328007SEM0902_5

Demirtaşlı, N. (2009). Eğitimde niteliği sağlamak: Ölçme ve değerlendirme sistemi örneği olarak CİTO Türkiye öğrenci izleme sistemi (ÖİS). *Cito Eğitim: Kuram ve Uygulama, 3*, 25-38.

Draney, K., & Wilson, M. (2009). Selecting cut scores with a composite of item types: The Construct Mapping procedure. In E. V. Smith Jr. & G. E. Stone (Eds.), *Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models* (pp. 276–293). Maple Grove, MN: JAM Press.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists. London: Lawrence Elbaum Associates, Publishers.

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference. 11.0 update* (4th ed.). Boston: Allyn & Bacon.

Goodman, D. P., & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145-220. doi: 10.1207/s15324818ame1702_3

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38-47. doi: 10.1111/j.1745-3992.1993.tb00543.x

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

Hu, L.-T. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. doi: 10.1080/10705519909540118

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20. doi: 10.1111/j.1745-3992.2006.00053.x

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329-349. doi: 10.1207/S15324818AME1404_2

Karantonis, A. (2017). *Using exemplar items to define performance categories: A comparison of item mapping methods* (Unpublished doctoral dissertation, University of Massachusetts). Retrieved from https://scholarworks.umass.edu/dissertations_2/1101/

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard- setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4-12. doi: 10.1111/j.1745-3992.2006.00047.x

Kennedy, C. A., Wilson, M. R., Draney, K., Tutunciyan, S., & Vorp, R. (2010). ConstructMap 4.6. [Computer software]. Berkeley, CA: BEAR Center.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Washington, DC: American Institutes for Research.

Muthen, B. O., & Muthen, L. K. (2015). Mplus (Version 7.4) [Computer software]. Los Angeles, CA: Muthen & Muthen.

Shulman, L. S. (2009). Assessment of teaching or assessment for teaching? Reflections on the invitational conference. In G. H. Gitomer (Ed.), *Measurement issues and assessment for teaching quality*. Thousand Oaks, CA: Sage Publications.

Ullman, J. B. (2001). Structural equation modeling. In B. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (4th ed.), (pp. 653-771). Boston, MA: Allyn & Bacon.

Van de Vijver, F. J. R. (2017). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis. Methods and applications* (2nd, rev. ed.). New York, NY: Routledge.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice, 20*(2), 15-25. doi: 10.1111/j.1745-3992.2001.tb00059.x

Ampirik ve Modele Dayalı Yeterlik Tanımları: İngilizce Yeterlik Sınavı Örneği

Giriş

Türkiye'de, test sonucunu daha verimli bir şekilde rapor etmek için yeterlik puan raporlama yöntemlerinin geliştirilmesine ihtiyaç duyulmaktadır. Bir testten alınabilecek puan aralıklarında tanımlanan yeterlikler sınav sonuçlarının anlamlı hale gelmesini sağlamakta ve paydaşlara gerekli geribildirimler verme konusunda yararlı olmaktadır. Bu çalışma, ampirik yöntem ve modele dayalı yöntem (ConstructMaps) kullanılarak, İngilizce yeterlilik testine ait puan aralıklarının nasıl tanımlanabileceğini göstermeyi amaçlamıştır.

Ampirik yönteme (Zwick, Senturk, Wang, & Loomis, 2001) göre yeterlik tanımlamanın ilk aşamasında, önce ölçeğe ilişkin puan aralıkları belirlenir. Ardından, bu puan aralıklarında yer alan öğrenci grupları saptanır. Her bir puan aralığındaki öğrencilerin çoğunluğu tarafından doğru olarak cevaplandırılan (örneğin %50, %65, %67 ve %80) maddeler belirlenir (Beaton & Allen, 1992). Araştırmacı belli bir doğru yanıtlama olasılığı belirleyerek bu olasılık üzerinden her bir puan

aralığındaki maddeleri belirler. Örneğin, doğru yanıt olasılığı %65 olarak belirlenmişse, her bir puan aralığında bireylerin %65'i tarafından doğru şekilde yanıtlanan maddeler bulunur. Her bir puan aralığı için, bu maddelerle ölçülen bilişsel ve içerikle ilgili özellikler belirlenir ve her bir puan aralığı için performans tanımlanır.

Modele dayalı yöntemde, ampirik yöntemde olduğu gibi, maddenin doğru yanıtlanma olasılığı esas alınarak maddeler belirlenir. Modele dayalı yöntemin ampirik yöntemden farkı, Madde Tepki Kuramı Rasch modeline göre doğru cevap olasılıklarının tahmin edilmesidir. Madde tepki kuramına göre, yetenek ve madde parametreleri aynı ölçekte yerleştirilebilir (Embretson & Reise, 2000; Hambleton, Swaminathan & Rogers, 1991). Yukarıda belirtildiği gibi, bu iki yöntem arasındaki fark, yanıt olasılıklarının hesaplanma şeklidir. Ampirik yöntemde, yanıtlanma olasılığı klasik test teorisine göre hesaplanırken modele dayalı yöntemde madde tepki kuramına göre hesaplanır.

Yöntem

Çalışma grubu

Türkiye'nin güneybatısındaki bir devlet üniversitesinin 2016-2017 İngilizce yeterlilik sınavına giren 630 lisans öğrencisi bu araştırmanın çalışma grubunu oluşturmaktadır. Öğrencilerin %68'i erkek, %32'si ise kadındır.

Veri toplama aracı

Bu çalışmada, üniversitenin yabancı dil okulu test geliştirme ekibi tarafından geliştirilen İngilizce yeterlilik testi kullanmıştır. İngilizce yeterlilik sınavının dört ana boyutu bulunmaktadır: Okuma, Dinleme, Yazma ve Konuşma. Bu çalışma, bu testin bir kısmını oluşturan okumaya odaklanmaktadır. Okuma bölümü okuma paragraflarını içermektedir. Çeşitli madde formatlarında (eşleştirme, kısa cevap ve çoktan seçmeli) 19 test maddesinden oluşmaktadır.

Verilerin analizi

Ön analiz olarak, okuma testinin iç tutarlılığı Cronbach'ın Alfa güvenilirlik katsayısı kullanılarak hesaplanmıştır. Okuma testi, okuduğunu anlama yeteneğini ölçmek için geliştirilmiştir. Bu nedenle, okuma testinin tek boyutluluğunu test etmek için doğrulayıcı faktör analizi yapılmıştır. Doğrulayıcı faktör analizi MPLUS 7.4 programı ile gerçekleştirilmiştir (Muthen & Muthen, 2015).

Maddelerin bir gruba yanlı olup olmadığını test etmek için madde yanlılığı analizi yapılmıştır. Bu çalışmada, lojistik Regresyon (LR) ve Yapısal Eşitlik Modelleme (YEM) madde yanlılığı yöntemleri kullanılmıştır.

Bu analizlerin ardından yeterlik tanımlama işlemleri yapılmıştır. Bu araştırmada yeterliklerin tanımlanmasında ampirik ve modele dayalı yöntemler kullanılmıştır. Ampirik yöntemde öğrencilerin almış oldukları puanlar beş performans seviyesine ayrılmıştır (0 - 20; 21 - 40; 41 - 60; 61 - 80; 81 - 100). Her bir puan kategorisindeki öğrenciler belirlenmiş ve daha sonra her bir puan kategorisi için her bir maddenin doğru cevaplanma oranı hesaplanmıştır. Bu çalışmada, her bir puan kategorisini temsil eden madde örneklerini belirlemek için üç farklı cevap olasılığı (RP) kullanılmıştır. %50 RP: Her bir performans seviyesinde katılımcıların en az %50'si tarafından doğru olarak cevaplanan maddeler örnek maddeler olarak seçilmiştir. %67 RP: Her bir performans seviyesinde katılımcıların en az %67'si tarafından doğru olarak cevaplanan maddeler örnek maddeler olarak seçilmiştir. %80 RP: Her performans seviyesinde katılımcıların en az %80'i tarafından doğru bir şekilde cevaplanan maddeler örnek maddeler olarak seçilmiştir. Modele dayalı yeterlik tanımlamaları ise Wright haritası üzerinde öğrencilerin toplam ham puanını, öğrenci yetenek tahminlerini ve madde güçlük indekslerini veren ConstructMap 4.6 (Kennedy, Wilson, Draney, Tutunciyan & Vorp, 2010) programı kullanılarak

yapılmıştır. Program Madde Tepki Kuramının Rasch modeline dayanarak 1-0 şeklinde puanlanan maddeleri analiz etmektedir. Wright haritası, öğrenci ölçeği puanlarını ve madde güçlük indekslerini aynı ölçekte göstermektedir.

Bulgular

Yeterlik sınavı okuma testinin Cronbach Alfa güvenirlik katsayısı .81 olarak hesaplanmıştır. Bu değer ölçekten güvenilir sonuçlar elde edildiğinin bir kanıtıdır (George & Mallery, 2003). Okuma testi öğrencilerin tek boyutlu okuduğunu anlama becerilerini ölçmeyi amaçlamıştır. Bu nedenle 19 maddelik okuma testinin tek boyutlu olup olmadığını test etmek için doğrulayıcı faktör analizi yapılmıştır. Yapılan analiz sonucunda ölçeğin tek boyutlu bir yapıda olduğu saptanmıştır (RMSEA = .054 < .060; CFI = .918 > .900). Lojistik Regresyon ve Yapısal Eşitlik Modelleme madde yanlılığı belirleme yöntemlerine dayalı analizler sonucunda okuma maddelerinin hiçbirinin cinsiyet grupları için madde yanlılığı göstermediği saptanmıştır.

Ampirik yönteme göre bulgular incelendiğinde, 0 - 20 puan aralığında öğrencilerin okuduğunu anlama becerisinin tanımlanamadığı saptanmıştır. 21 - 40 puan aralığındaki öğrencilerin soru kökündeki açık bir ipucu olarak daha sık kullanılan kelime hazinesini (k1 bandından) kullanarak içerikten bir detay tanıyabildiği belirlenmiştir. 41 - 60 puan aralığında bir puan alan öğrencilerin ise soru kökündeki açık bir ipucu olarak sık başvurulan kelime hazinesini (k1 bandından) kullanarak içerikten bir ayrıntıyı tanıyabildiği saptanmıştır. 61 - 80 puan arası bir puana sahip öğrencilerin soru kökündeki açık bir ipucu olarak daha sık kullanılan kelime hazinesini (k2 bandından) kullanarak içerikten bir detay tanıyabildiği, metin yapısının gelişimini takip edebildiği metinde açık bir ipucu kullanarak bir sonuca ve detaylara ulaşabildiği görülmüştür. En üst yeterlik düzeyi olan 81 - 100 puan arasında puan alan öğrencilerin ise metin yapısının gelişimini takip edebildiği ve bir ipucu kullanarak her cümledeki metnin nereden çıkacağına karar verilebildiği, daha az kullanılan kelime dağarcığı içeren metinde örtük ipuçlarını kullanarak anlam çıkarabildiği ve yazarın tutum ve bakış açısını yakalayabildiği saptanmıştır.

Modele dayalı bulgulara göre en alt yeterlik basamağının kesim noktası olarak -0.60 puan belirlenmiş, bu puanın altında bir puana sahip öğrenciler için yeterlik tanımları yapılabilmiştir. Ancak yapılan tanımlamalar ampirik yöntemdeki 21 - 40 puan aralığında tanımlanan yeterliklerdir. Diğer bir deyişle, ampirik yöntemde 21 - 40 puan arasında tanımlanan yeterlikler modele dayalı yöntemde en alt yeterlik basamağında tanımlanmıştır. Benzer şekilde ampirik yöntemde 41 - 60 puan aralığında belirlenen yeterlik tanımları da modele dayalı yöntemde -0.60 - 0.00 puan aralığında tanımlanmıştır. 0.00 - 0.90 arasında puan alan öğrencilerin ise soru kökündeki açık bir ipucu olarak daha sık kullanılan kelime haznesini (k2 bandından) kullanarak içerikten bir detay tanıyabildiği, metin yapısının gelişimini takip edebildiği metinde açık bir ipucu kullanarak bir sonuca ve detaylara ulaşabildiği görülmüştür. Bu yeterlik tanımı ampirik yöntemde 61 - 80 puan aralığına denk gelmektedir. 0.90 - 1.25 arasında puan alan öğrencilerin yazarın bakış açısı hakkında çıkarım yapabildiği saptanmıştır. 1.25 puan üzerinde puan alan öğrencilerin metin yapısının gelişimini takip edebildiği ve bir ipucu kullanarak her cümledeki metnin nereden çıkacağına karar verilebildiği, daha az kullanılan kelime dağarcığı içeren metinde örtük ipuçlarını kullanarak anlam çıkarabildiği ve yazarın tutumunu belirleyebildiği görülmüştür.

Sonuç ve Tartışma

Genel olarak değerlendirildiğinde, ampirik yöntem ve modele dayalı yöntem arasında yeterlik tanım basamakları açısından birtakım farklılıklar gözlense de sonuçlar yeterlik tanımlarının hiyerarşik bir şekilde sıralandığını, İngilizce yeterlilik sınavının yeterlik tanımlarının ampirik ve modele dayalı yöntemlerle tanımlanabileceğini göstermektedir. Ampirik yöntem, klasik test teorisine dayanır ve uygulanması kolaydır. Modele dayalı yöntem, madde tepki kuramına dayanır ve istatistiksel yazılım üzerinde uzmanlık gerektirir. Her iki yöntemde de puan aralıkları için tanımlayıcıların tanımlanması sürecinde, madde kümeleri arasında hiyerarşik bir yapı bulunmuş ve daha yüksek puan aralıklarında

bulunan maddeler daha yüksek bilişsel beceriler gerektirmiştir. İlgili literatürde, farklı alanlarda (örneğin matematik) yapılan çalışmalarda benzer sonuçlar elde edilmiş, literatürde, ampirik yöntem ve Wright haritalarından elde edilen sonuçların benzer olduğu bulunmuştur (Arıkan & Kilmen, 2018). Her iki yöntemde de benzer madde sıralamaları ve madde kümeleri oluşturulduğundan, öğretmenler, değerlendirmeler için performans düzeyi tanımlayıcılarını tanımlamada ampirik yöntem kullanabilir, ampirik yöntemde açıklanan adımları takip edebilir ve yeterlikleri tanımlayabilirler. Öğretmenler, diğer öğretmenlerle birlikte, öğrencilerin başarısını arttırmak için iş birliği yapabilir ve performans düzeyi tanımlayıcılarını tanımlamak için bir araya gelebilirler. Ölçme ve değerlendirme alanında uzmanlaşmış kişilerin ise modele dayalı yöntem kullanmaları tavsiye edilebilir. Çünkü madde tepki kuramı ile tahmin edilen madde istatistikleri, örneklemden bağımsızdır ve bu da parametreleri daha tutarlı hale getirir (Hambleton & Jones, 1993).

Türkiye'de geniş ölçekli testlerde bağıl ve mutlak değerlendirmeler yapılmasına rağmen daha çok bağıl değerlendirmeye vurgu yapılmaktadır. Özellikle çok sayıda öğrenci arasından sınırlı sayıda öğrenciyi yükseköğretim kurumlarına seçmeyi amaçlayan ulusal geniş ölçekli değerlendirmeler normlara odaklanmaktadır. Bununla birlikte, ulusal çapta düzenlenen mutlak değerlendirmenin kullanıldığı sınavlardan özellikle dil sınavları kimin yetkin olup olmadığına karar vermeyi amaçlamasına rağmen, kişinin yeterliklerine odaklanan bir rapor sunmamakta, sonuçlar puan ile sınırlı kalmaktadır. Oysa değerlendirme sonuçlarının puan ile sınırlı kalmayarak öğrencilere ve paydaşlara somut bir geri bildirim sağlamak için kullanılması daha yararlı olacaktır. Ayrıca, yeterlik tanımları, yıl boyunca bir öğrencinin gelişimini takip etmek için kullanılabilir. Örneğin, düşük seviyelerden başlayan bir öğrenci, yıl boyunca kendi performansını artıracak çalışmaları yeterlik göstergelerinin inceleyerek bulabilir ve kendi gelişimini başarabildiklerine ve başaramadıklarına odaklanarak kendi kendine hızlandırabilir.

Bu çalışmanın çeşitli sınırlılıkları bulunmaktadır. Sınırlı sayıda öğrenciyle elde edilen bulgular sonuçların genellenebilirliğini azaltmaktadır. Bu nedenle daha büyük örneklemlerde benzer araştırmalar yapılabilir. Okuma maddelerinin sayısının çok yüksek olmaması bazı puan aralıklarına sadece bir maddenin yerleşmesine neden olmuştur. Sınırlı sayıda maddeye dayanarak yeterliklerin tanımlanması bulguların güvenilirliğini tehdit etmektedir. Bu nedenle, daha fazla madde içeren testlerle benzer araştırmalar yapılabilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 235-248

The Influence of Using Plausible Values and Survey Weights on Multiple Regression and Hierarchical Linear Model Parameters*

Osman TAT ** İlhan KOYUNCU *** Selahattin GELBAL ****

Abstract
In large-scale assessments like Programme for International Students Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), plausible values are often used as students' ability estimations. In those studies, stratified sampling method is employed in order to draw participants, and hence, the data gathered has a hierarchical structure. In the context of large-scale assessments, plausible values refer to randomly drawn values from posterior ability distribution. It is reported that using one of plausible values or mean of those values as independent variable in linear models may lead to some estimation errors. Moreover, it is observed that sampling weights sometimes are not used during analysis of large-scale assessment data. This study aims to investigate the influence of three approaches on the parameters of linear and hierarchical linear regression models: 1) using only one plausible value, 2) using all plausible values, 3) incorporating sampling weights or not. Data used in the present study is obtained from school and student questionnaires in PISA (2015) Turkey database. Results revealed that the use of sampling weights and number of plausible values has significant effects on regression coefficients, standard errors and explained variance for both regression models. Findings of the study were discussed in details and some conclusions were drawn for practice and further research.

Key Words: Hierarchical linear modeling, multiple linear regression, plausible values, survey weights, large-scale assessments, PISA.

INTRODUCTION

When determining the group performance, large-scale assessment data are used in many countries so as to take initiatives and develop educational policies. In addition to the cognitive tests measuring the student performance, several scales are used in those applications in order to collect student-, teacher- and school-level information. Through that data, instead of individual assessment, school- and study-related student skills are taken together, and group-level inferences are made. In this type of large-scale assessments, different booklets are designed and applied to students in pairwise blocks in order to prevent the loss of time resulting from the measurement of performance in a wide range of subjects. In this case, as all students do not answer the same questions, it is incorrect and inaccurate to estimate their performance via classical statistical methods and to make a group-level comparison (Organization for Economic Cooperation and Development-OECD, 2017). Hence, such applications employ multiple values demonstrating the possible distribution of student abilities (Von Davier, Gonzales & Mislevy, 2009). The so-called *plausible values* are based on student responses to subset of tests, as well as affective features and available background information (demographic information) (Mislevy, 1991; OECD, 2009).

Plausible values refer to random values drawn from the posterior distributions of ability scores in the context of large-scale assessments (Von Davier et al., 2009). Maximum Likelihood (ML) (Rasch,

* Preliminary results of this study were presented at 6th International Conference on Education, Zagreb-Croatia, 2017.
** Res. Assist., Hacettepe University, Faculty of Education, Ankara-Turkey, osmntt@gmail.com, ORCID ID: 0000-0003-2950-9647
*** Assist. Prof. PhD., Adıyaman University, Faculty of Education, Adıyaman-Turkey, ilhankync@gmail.com ORCID ID: 0000-0002-0009-5279
**** Prof. PhD., Hacettepe University, Faculty of Education, Ankara-Turkey, sgelbal@gmail.com, ORCID ID: 0000-0001-5181-7262

To cite this article:
Tat, O., Koyuncu, İ., & Gelbal, S. (2019). The influence of using plausible values and survey weights on multiple regression and hierarchical linear model parameters. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 235-248. doi: 10.21031/epod.486999

Received: 23.11.2018
Accepted: 17.06.2019

1960), Weighted Maximum Likelihood (WML) (Warm, 1985), Joint Maximum Likelihood (JML) (Wright & Stone, 1979), and Expected A Posteriori (EAP) (Bock & Aitkin, 1981) used in estimations made through the Rasch model within the Item Response Theory are estimation methods that cover up each other's flaws. However, these methods make point estimations and do not give more than one ability estimation different from each other coming from the posterior distribution for individuals as in plausible values (Wu, 2005). The first usage of plausible values was inspired by Rubin's (1987) multiple imputation research when analyzing the US National Assessment of Educational Progress (NAEP) data in 1994. Using plausible values in large-scale tests became more common as they were also used in the next NAEP applications, the Trends in International Mathematics and Science Study (TIMSS) by OECD, as well as the Programme for International Student Assessment (PISA). In general, five plausible values are produced for each student, though there is not a strong basis for this limitation in the literature (Von Davier et al., 2009; Wu, 2005).

Plausible values correspond to the distribution of abilities a student can have depending on his / her responses to items. They are obtained by randomly drawn values out of the posterior probability distribution for θ ability values in the Item Response Theory (IRT) (Wu, 2005). The technical reports of the NAEP applications in 1983-1984 and the PISA in 2000 give detailed information about how those values are calculated and how they are drawn from the probability distribution (Adams & Wu, 2002; Beaton, 1987). Plausible values are not individual scores in the traditional sense, and should therefore not be analyzed as multiple indicators of the same score or latent variable (Mislevy, 1993). When compared to the EAP and WML methods that make point estimations, using plausible values will yield less biased results in group-level assessments, as Von Davier et al. (2009) demonstrated in their research. They point out, however, that using the averages of plausible values (PV-W) leads to more biased estimates than using the average of statistics (PV-R) derived by analyzing each value; therefore, the averages of plausible values should not be used as dependent variable (Von Davier et al., 2009). Furthermore, the simulation research by Wu (2005) shows that using any plausible value alone is enough to make highly correct estimates regarding the population parameters.

Instead of assigning point estimations of ability for each student, plausible values from the posterior ability distribution are used in large-scale assessments such as Trends in International Mathematics and Science Study (TIMSS), the PISA, and International Computer and Information Literacy Study (ICILS). The data obtained via those large-scale applications is hierarchically structured within multiple levels (student, school, regions, country, etc.). In fact, it is possible to encounter this data structure in several areas of social science research like organizational, intercultural, and developmental studies (Bryk & Raudenbush, 2002). The data in educational sciences may involve two or more levels as well, with students being nested within classes, classes within schools, and schools within cities or regions, in addition to the repeated measures for students or any unit of analysis. Over the Ghana Youth Save data, for instance, Chowa, Masa, Ramos, and Ansong (2015) examined how the properties of students and schools would affect the academic achievement of youth. Students were nested within schools in the mentioned study. By using the longitudinal data from the students participating the National Longitudinal Survey of Youth (NLSY), Stipek and Valentino (2015) investigated how well measures of short-term and working memory and attention in early childhood predicted longitudinal growth trajectories in mathematics and reading comprehension. The measures in due course were nested within the variable of student as a secondary unit. In the Sustaining Effects Study (SES), Bryk and Raudenbush (1988) used a three-level hierarchical linear model to analyze the relationship between the intensity of student and school poverty for the first to third grade students and their reading comprehension and learning mathematics.

It is common to observe two type of data use if the hierarchical data structure is not taken into consideration. Those are aggregation and disaggregation methods. Aggregation is integrating sub-units of data in upper units. Conjoining the test scores of students at the class level and obtaining school-level scores by weighting their average class-level scores can be taken as examples of aggregation. As individual differences are ruled out in this method, relationships between aggregated variables may be much stronger or lead to misinterpretations (Atar, 2010; Bryk & Raudenbush, 2002; Snijders & Bosker, 2003; Woltman, Feldstain, MacKay, & Rocchi, 2012). In disaggregation, upper units are

degraded to lower levels. Assigning the data about a school- and class-level variable to students can be an example of disaggregation. In this case, as all of the students within the same school or class have the same properties, independence of observations as a significant assumption of statistical analyses will be violated (Snijders & Bosker, 2003; Woltman et al., 2012). In conclusion, using linear regression models with aggregation and disaggregation methods will lead to related residuals, as well as to biased coefficients and standard errors on regression equations by ignoring between-group differences (Bryk & Raudenbush, 2002).

Being a way to analyze nested data, hierarchical linear models eliminate the mentioned disadvantages of aggregation and disaggregation methods. Hierarchical linear models have removed the obstacles concerning the examination of analysis unit and measurement change that were important problems in the past (Raudenbush & Bryk, 1986). Thus, estimates for variables at each level, interactions between variables at the same and different levels, as well as components of variance-covariance can be investigated through a single analysis (Bryk & Raudenbush, 2002). The advantages of using hierarchical linear models for hierarchical data include formulating within and between level relations correctly; eliminating the biases resulting from aggregation; enabling to propose more diversified and far-reaching research questions and hypotheses in empirical studies; detecting the appropriate error structures including random effects, and allowing for estimates of standard errors stemming from group effects, including the components of variance and covariance (Raudenbush, 1988). According to Goldstein (2011), hierarchical models enable statistically efficient estimates of regression coefficients, provide correct standard errors, confidence intervals, significance tests, and make it possible to examine within and between relations, as well as to compare the whole levels by taking all factors into consideration. The data analysis section of this research touches on the statistical aspects of hierarchical linear models (HLM) analyses and how they are carried out.

Ignoring the hierarchical structure in the data may lead to a considerable differentiation in the outcome. Roberts (2004) found that the relationship between urbanicity and science achievement was .77 when the hierarchical data structure was ignored, whereas the same relationship was -.88 when the students were nested within school. Likewise, a number of studies argue that using traditional linear models instead of hierarchical ones will yield biased results (Bryk & Raudenbush, 2002; Goldstein, 2011; Osborne, 2000; Raudenbush, 1988; Raudenbush & Bryk, 1986; Woltman et al., 2012). In her study which is a comparison of linear regression and hierarchical linear model, Atar (2010) found that the coefficient of *Attitude Towards Science* in linear regression differs among second level units (schools) in a range from -0.2 to 1.09. The findings shown the degree of attitude towards science significantly differs between schools and multilevel nature of the data should be taken into consideration.

According to Gelman (2006), hierarchical linear models are useful in terms of data reduction and casual inference compared to classical regression analysis. However, using hierarchical linear models do not guarantee the unbiasedness of parameter estimation in the hierarchical data, because some errors of estimates may be observed if the selected sample does not represent the number of students in the population, as it might be the case in the other linear models as well. For this reason, large-scale assessments make use of survey weights pertaining to different levels (Meinck, 2015).

The survey weights used in large-scale tests like PISA make it easier to analyze data, to calculate estimates of sampling errors appropriately, as well as to make valid estimates and inferences of the population. In this way, users are enabled to make unbiased estimates of standard errors, conduct significance tests, and create confidence intervals in consideration of the complex sample design for each participating country. The survey weights are not the same for all students in a given country, because they are to provide full representation of every selected school, to balance the participation of school populations at certain rates, to take school non-responses into consideration, to prevent larger weights in relatively small groups, and to balance the influence of additional number of students sampled for surveys in some countries (OECD, 2014). The statistical procedures underlying the survey weights in tests like TIMSS, and PISA can be found in Cochran (1977), Lohr (2010), Särndal, Swensson and Wretman (1992). The survey weights in those tests consist of the school base weight and the within-school base weight, as well as five adjustment factors. Adjustment factors are used to

consider non-participation by other schools that are somewhat similar in nature to a particular school, to balance the age and grade levels of students, to consider non-participating students within the school according to their gender, grade, and region, and to reduce the unexpected school-based and other weight factors. The detailed information on how these weight factors were calculated for PISA 2012 can be found in the technical report (OECD, 2014).

Purpose of the Study

Plausible values and weights used in large-scale assessments are grounded on conducting more precise and inclusionary measurements. Concordantly, this study aims to compare the analysis results of multiple linear regression and hierarchical linear models in predicting science literacy of students, in terms of plausible values and weights, using the PISA data in 2015. Within the frame of this general aim, we first carried out multiple linear regression and hierarchical linear model analyses which one plausible value regressed on independent variables without weights. Then, the same models repeated in such a way that whole weighted plausible values regressed on independent variables. Through this approach we could observe the impact of usage of plausible values with or without weight in both multiple linear regression and HLM. Accordingly, we investigate four research questions: how do the results of multiple regression and HLM analyses turn out in case of a) one unweighted plausible value, b) all plausible values with weights;

1. In fully unconditional model?
2. Regressed on level-1 explanatory variables (students' epistemological beliefs in science, test anxiety, motivations, and the index of economic, social, and cultural status)?
3. Regressed on level-2 explanatory variables (classroom sizes at schools, educational leadership, and shortage of educational material and staff)?
4. Regressed on both level-1 and level-2 explanatory variables?

METHOD

This study is a correlational research (Fraenkel, Wallen & Hyun, 2012) aiming to demonstrate relationship among plausible values, survey weights and few independent variables in two different analyses, with reference to the hierarchically structured data obtained from the international large-scale education research.

Working Group

The PISA 2015 dataset was used in accordance with the aim of the study. PISA is a triennial international survey conducted by OECD, mainly aiming to measure the mathematics, science, and reading performance of 15-year-old students. The first and the latest PISA surveys were conducted in 1997 and 2015, respectively. Nearly 520 thousand students from 72 countries were assessed. From Turkey, 5895 students from 187 schools in total took the PISA test.

This study incorporates two-level hierarchical data (with students being level-1, and schools being level-2), in line with the nature of hierarchical linear models. The sample of the level-2 consists of 178 schools in Turkey without any missing data. The schools with missing data were excluded from the dataset, since it is impossible to conduct analysis with missing data in level-2 units in HLM software. HLM software works with compete level-2 data. It is an obligation either to impute a value for the missing data or to delete incomplete cases. Ignoring level-2 missing observation will result in listwise deletion of incomplete level-2 units during the creation of system files (Palardy, 2011). The level-1 sample of the research consists of 5703 students receiving education in the afore-mentioned 178 schools. For hierarchical linear models, level-2 sample size of 50 or more with adequate level-1 sample

size is expected to provide unbiased estimates (Maas & Hox, 2005). Hence, the sample size of this research is appropriate enough to perform HLM-related analyses.

Data Collection Instruments

In PISA, students take mathematics, science, and reading comprehension tests. Their cognitive skills are assessed in these fields. Besides the cognitive skill tests, one of those three fields are designated as an area of focus in every application, and a student questionnaire is applied to assess affective variables related to the specified area of focus. The data related to students is gathered through cognitive tests and questionnaires in which affective variables are examined. In a similar way, a school questionnaire is applied to school principals in order to gather information in a variety of issues, such as technical infrastructure and status of educational resources at schools. In this study, the level-1 variables from the student questionnaire and the science test were used together with the level-2 variables from the school questionnaire. The variables used in the model selected via Automatic Linear Modeling (Yang, 2013) procedure. This analysis carried out with 14 index or continuous variables. Then, out of 10 most important variables eight variables (two variables exclude for having equal importance levels) decided to be used. We tried to provide a clear representation of the finding as much as possible with parsimonious models based on most important variables. The details about those variables are seen in Table 1.

Table 1. Variables and Their Properties

Level	Variable	Abbreviation	Type	Nature
Student (Level-1)	Science Literacy Scores (10 Plausible Values)	PV1SCIE (1-10)	Dependent	Continuous
	Epistemological Beliefs	EPIST	Independent	Continuous
	Test Anxiety	ANXTEST	Independent	Continuous
	Achievement Motivation	MOTIVAT	Independent	Continuous
	Index of Economic, Social and Cultural Status	ESCS	Independent	Continuous
School (Level-2)	Average Class Size of School	CLSIZE*	Independent	Continuous
	Teachers Participation	LEADTCH*	Independent	Continuous
	Shortage of Educational Material in School	EDUSHORT*	Independent	Continuous
	Shortage of Educational Staff in School	STAFFSHO*	Independent	Continuous
Weights	Final Student Weight	W_FSTUWT		Continuous
	BRR-FAY Replicate Weights (80 in number)	W_FSTURWT1-80		Continuous

*Disaggregated to the student level in multiple regression analysis

Data Analysis

The analysis of this research involves multiple regression and hierarchical linear models with the purpose of investigating the influence of plausible values and survey weights on different statistical analyses. In the multiple regression analysis, PV1SCIE1 was set as the dependent variable, and four different models were analyzed. Those models included no explanatory variables, only student-level variables, only school-level variables, and variables pertaining to both levels. In the multiple regression analysis, school-level variables were disaggregated to students. The mentioned four models were repeated while 10 plausible values (PV1SCIE1-10) were made dependent variables, and student-level weight and replications (W_FSTUWT and W_FSTURWT1-80) were used. In this way, we examined the effects of plausible values and weight use on multiple regression analysis.

Like the multiple regression analysis, the HLM analyses involved eight models in total, in four of which only the first plausible value (PV1SCIE1) was dependent variable, and in four of which all plausible models and weights were employed. For data analysis, the IDB Analyzer (International Association for the Evaluation of Educational Achievement-IEA, 2016) software was used to create the syntax that makes it possible to utilize all plausible values and weights in multiple regression. The main analyses were performed via SPSS 21.0 (International Business Machines-IBM Corp., 2012) and

HLM 7 Hierarchical Linear and Nonlinear Modelling (Bryk, Raudenbush & Congdon, 2010). .05 is significance level for all analyses.

Before the carrying out the multiple regression and HLM analyses we tested assumptions of both analyses. Firstly, we checked multiple regression assumptions in terms of linear relationship between dependent variable and independent variables, multicollinearity, independence of residuals (uncorrelated residuals), constant residual variance (homoscedasticity), normal distribution of residuals and outliers for all models (except intercept only models). By scatter plots drawn with dependent variable against independent variables for all models, we could observe linear relationship among outcome and explanatory variables. For all models, multicollinearity tested with tolerance and VIF statistics. Accordingly, it is found that none of tolerance value is smaller than 0.2 (0.7-0.9) and none of VIF is greater than 10 (1.3-1.4). The independence of residuals tested via Durbin-Watson statistics. According to the test, it is indicated that for all models mentioned statistics is in a range from one to three. For the constant residuals (homoscedasticity) we benefited from a graph of predicted standard points against standard residuals. Through the P-P graph we could decide residuals are on the diagonal line and are normally distributed. Finally, outliners tested with Cook's distance method. We did not meet any distance greater than one.

Since the first HLM model is fully unconditional model we did not check the assumptions. For all other models, homogeneity of variances and normality of residuals for each level are strictly recommended (Snijders & Bosker, 1999). For all models we created scatter plots for level-1 among level-2 units and we observed that residuals are randomly distributed among level-2 units. Finally, we drawn P-P plots of predicted standard points against standard residuals and determined that the residuals are normally distributed.

RESULTS

Findings on the First Sub-Problem

Table 2 demonstrates the details about four different models that are constructed by considering the absence of any explanatory variable. As seen in the table, the multiple regression model in which all plausible values and weights are used is the highest predictor of Turkish general science literacy score, whereas the HLM analysis in which all plausible values and weights are used is the lowest predictor. The smallest standard error estimation is obtained via multiple regression model (1.02), while the highest standard error is obtained through the HLM analysis where all plausible values and weights are used.

Table 2. Fixed Effects Pertaining to The First Model

Analysis	Fixed Effect	Coefficients	Se	t
Multiple Regression (PV1SCIE1)	Grand Mean of Science Literacy	423.19*	1.02	414.89
Multiple Regression (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy	426.22*	4.06	104.98
HLM (PV1SCIE1)	Grand Mean of Science Literacy, γ_{00}	418.48*	4.35	96.13
HLM (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy, γ_{00}	417.71*	4.90	85.29

*p < .05

The random effects from two different random effects ANOVA models are presented in Table 3. The results of both analyses indicate that the mean of student science achievement differs from a school to another. While the level-1 error term estimated in both analyses is too close, the level-2 error term estimated with the HLM analysis using all plausible values and weights is higher as compared to the first analysis.

Table 3. Random Effects Pertaining to The First Model

Analysis	Random Effect	Sd	Variance	χ^2
HLM (PV1SCIE1)	Level-2 Error Term, u_{0j}	55.43	3073.53	5499.68*
	Level-1 Error Term, r_{ij}	53.61	2873.90	
HLM (PV1SCIE1-10) Weighted	Level-2 Error Term, u_{0j}	59.46	3536.05	6332.42*
	Level-1 Error Term, r_{ij}	53.37	2848.21	

*p < .05

The intra-class correlation coefficient was used to determine the percentage of variance in science literacy explained at school level. Accordingly, the proportions obtained from both analyses are as follows:

$$\rho_1 = {}^{\tau_{00}}/_{(\tau_{00}+\sigma^2)} = 3073.53/\ (3073.53+2873.90) = 0.517 \quad (1)$$

$$\rho_2 = {}^{\tau_{00}}/_{(\tau_{00}+\sigma^2)} = 3536.05/\ (3536.05+2848.21) = 0.554 \quad (2)$$

In the first analysis, it was determined that approximately 52% (ρ_1 = 0.517) of the variance in dependent variable can be explained at school level. On the other hand, when all plausible values were used, approximately 55% (ρ_2 = 0.554) of the variance in dependent variable could be explained at level-2.

Findings on the Second Sub-Problem

Table 4 shows the coefficients pertaining to two different multiple regression analyses, in which four student-level variables were included in the model, as well as the fixed effects from two different random coefficients models.

Table 4. Fixed Effects Pertaining to The Second Model

Analysis	Fixed Effect	Coefficients	Se	t
Multiple Regression (PV1SCIE1)	Grand Mean of Science Literacy	452.97*	1.65	274.11
	EPIST	14.74*	0.83	17.80
	ANXTEST	-6.81*	0.94	-7.29
	MOTIVAT	6.05*	0.98	6.16
	ESCS	17.84*	0.82	21.66
Multiple Regression (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy	456.05*	4.56	100.08
	EPIST	15.23*	1.3	11.75
	ANXTEST	-6.27*	1.4	-4.47
	MOTIVAT	6.53*	1.38	4.74
	ESCS	18.69*	2.05	9.12
HLM (PV1SCIE1)	Grand Mean of Science Literacy, γ_{00}	423.31*	4.61	91.74
	EPIST,γ_{10}	7.37*	0.68	10.81
	ANXTEST,γ_{20}	-6.35*	0.82	-7.72
	MOTIVAT, γ_{30}	3.26*	0.90	3.63
	ESCS, γ_{40}	1.91*	0.78	2.46
HLM (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy, γ_{00}	422.10*	5.18	81.49
	EPIST, γ_{10}	7.41*	0.85	8.71
	ANXTEST, γ_{20}	-6.15*	0.93	-6.61
	MOTIVAT, γ_{30}	3.89*	1.05	3.70
	ESCS, γ_{40}	1.87*	0.91	2.03

*p < .05

In line with Table 4, it is possible to say that overall science literacy is over estimated in each analysis. The coefficients in these analyses reflect the mean science literacy when independent variables are controlled. In every analysis, all the independent variables significantly predict the dependent variable. Whereas the variable with the largest coefficient is the index of Economic, Social, and Cultural Status (ESCS) in the multiple regression analysis, this variable was estimated very lower in the HLM analyses. Students' levels of epistemological belief (EPIST) is the variable with the highest coefficient in the HLM analyses. Furthermore, it can be said that all coefficients in the multiple regression analyses were estimated higher when compared to the HLM analyses. It was seen that the standard errors pertaining to the coefficients in the unweighted multiple regression and HLM analyses were estimated low when compared to the weighted multiple linear regression. The random effects from the random coefficient models are presented in Table 5.

Table 5. Random Effects Pertaining to The Second Model

Analysis	**Random Effect**	**Sd**	**Variance**	χ^2
HLM (PV1SCIE1)	Level-2 Error Term, u_{0j}	56.50	3192.56	1982.77*
	EPIST Effect, u_{1j}	1.51	2.28	143.22
	ANXTEST Effect, u_{2j}	3.94	15.51	159.71
	MOTIVAT Effect, u_{3j}	5.26	27.69	176.79
	ESCS Effect,u_{4j}	1.68	2.82	146.86
	Level-1 Error Term, r_{ij}	52.15	2719.55	
HLM (PV1SCIE1-10) Weighted	Level-2 Error Term, u_{0j}	60.19	3622.63	2220.32*
	EPIST Effect, u_{1j}	2.47	6.09	144.92
	ANXTEST Effect, u_{2j}	2.72	7.38	146.64
	MOTIVAT Effect, u_{3j}	5.16	26.59	163.78
	ESCS Effect, u_{4j}	3.02	9.11	149.17
	Level-1 Error Term, r_{ij}	51.89	2693.07	

*p < .05

In random effects, level-1 error variances are expected to become smaller when level-1 independent variables are included in the model. Equation 3 and 4 were used to determine to what extent the level-1 variance is explained by the level-1 variables included in the model.

$$\text{Unweighted HLM } \rho_1=(\sigma^2_{ANOVA}-\sigma^2_{RIM})/\sigma^2_{ANOVA}: = (2873.90 - 2719.55) / 2873.90 = 0.05 \quad (3)$$

$$\text{Weighted HLM } \rho_2=(\sigma^2_{ANOVA}-\sigma^2_{RIM})/\sigma^2_{ANOVA} = (2848.21 - 2693.07) / 2848.21 = 0.05 \quad (4)$$

Both HLM models explained the level-1 variance to the equal extent, although the variance of level-1 error was smaller in the HLM analysis in which all plausible values and weights were used. The level-2 error variance was estimated higher when weights were used.

Findings on the Third Sub-Problem

In Table 6, the coefficients pertaining to two different multiple regression analyses, in which four school-level variables were disaggregated, as well as the fixed effects from two different HLM analyses.

Table 6. Fixed Effects Pertaining to The Third Model

Analysis	Fixed Effect	Coefficients	Se	t
Multiple Regression (PV1SCIE1)	Grand Mean of Science Literacy	411.89*	4.621	89.14
	CLSIZE	0.290*	0.09	3.07
	LEADTCH	5.00*	0.90	5.56
	EDUSHORT	-9.58*	0.94	-10.25
	STAFFSHO	-8.43*	1.01	-8.35
Multiple Regression (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy	416..84*	25.41	16.4
	CLSIZE	0.27*	0.53	0.51
	LEADTCH	4.38*	5.2	0.84
	EDUSHORT	-10.43*	3.75	-2.78
	STAFFSHO	-11.05*	4.3	-2.57
HLM (PV1SCIE1)	Grand Mean of Science Literacy, γ_{00}	399.33*	16.37	24.40
	CLSIZE,γ_{10}	0.52	0.35	1.50
	LEADTCH, γ_{20}	4.05	3.65	1.11
	EDUSHORT, γ_{30}	-9.09*	3.02	-3.01
	STAFFSHO, γ_{40}	-9.85*	4.01	-2.45
HLM (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy, γ_{00}	399.89*	20.49	19.51
	CLSIZE,γ_{10}	0.57	0.42	1.35
	LEADTCH, γ_{20}	3.38	4.42	0.77
	EDUSHORT, γ_{30}	-8.44*	3.70	-2.29
	STAFFSHO, γ_{40}	-15.39*	4.63	-3.32

*p < .05

As seen in Table 6, all the variables in both multiple regression analyses significantly predict the dependent variable, while only the shortage of educational materials (EDUSHORT) and the shortage of educational staff (STAFFSHO) remain significant in the HLM analyses. Standard errors increase with the use of weighted plausible values in both regression and HLM analyses. It is seen that some of the weighted multiple regression coefficients are slightly greater than those of the unweighted multiple regression analysis coefficients. The effect of weighting on the coefficients was not found considerable in the HLM analyses. The random effects related to the HLM analyses are presented in Table 7.

Table 7. Random Effects Pertaining to The Third Model

Analysis	Random Effect	Sd	Variance	χ^2
HLM (PV1SCIE1)	Level-2 Error Term, u_{0j}	51.64	2667.09	4700.68*
	Level-1 Error Term, r_{ij}	53.61	2873.60	
HLM (PV1SCIE1-10) Weighted	Level-2 Error Term, u_{0j}	53.46	2857.64	5243.57*
	Level-1 Error Term, r_{ij}	53.36	2847.75	

*p < .05

The variance of level-2 error term was estimated higher in the weighted HLM analysis, as demonstrated in Table 7. The level-2 variance is expected to decrease with the inclusion of level-2 variables into the completely unconditional model. Equation 5 and Equation6 were utilized to determine to what extent the level-2 variance is explained by the level-2 variables included in the model.

Unweighted HLM: $\rho_1=(\sigma^2_{ANOVA}-\sigma^2_{MAOR})/\sigma^2_{ANOVA}$ = (3073.53 - 2667.09) / 3073.53 = 0.13 (5)

Weighted HLM: $\rho_2=(\sigma^2_{ANOVA}-\sigma^2_{MAOR})/\sigma^2_{ANOVA}$ = (3536.05 - 2847.75) / 3536.05 = 0.20 (6)

Whereas 13% of the level-2 variance is explained in the unweighted HLM analysis after four level-2 independent variables are included into the model, this percentage rises to 20% in the weighted HLM

analysis. Hence, it is possible to say that weighting had a certain effect on the variance explained in the HLM analysis.

Findings on the Fourth Sub-Problem

Table 8 shows the coefficients from two different regression models, in which the level-1 variables were modelled together with the level-2 variables that were found to be significant. The table also shows the fixed effects pertaining to the model of intercepts and slopes as two different dependent variables.

Table 8. Random Effects Pertaining to The Fourth Model

Analysis	Fixed Effect	Coefficients	Se	t
Multiple Regression (PV1SCIE1)	Grand Mean of Science Literacy	453.91*	1.65	275.59
	EDUSHORT	-7.11*	0.884	-8.04
	STAFFSHO	-7.00*	0.96	-7.26
	EPIST	14.00*	0.81	17.23
	ANXTEST	-6.79*	0.92	-7.42
	MOTIVAT	5.76*	0.96	5.98
	ESCS	15.04*	0.83	18.17
Multiple Regression (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy	457.20*	4.79	95.35
	EDUSHORT	-7.8*	3.07	-2.54
	STAFFSHO	-9.07*	3.83	-2.37
	EPIST	14.16*	1.22	11.58
	ANXTEST	-6.32*	1.3	-4.86
	MOTIVAT	6.21*	1.35	4.61
	ESCS	15.63*	1.94	8.07
HLM (PV1SCIE1)	Grand Mean of Science Literacy, γ_{00}	429.33*	4.71	91.20
	EDUSHORT, γ_{01}	-8.85*	3.25	-2.72
	STAFFSHO, γ_{02}	-6.97*	3.55	-1.96
	EPIST, γ_{10}	7.43*	0.68	10.85
	ANXTEST, γ_{20}	-6.32*	0.82	-7.69
	MOTIVAT, γ_{30}	3.22*	0.89	3.61
	ESCS, γ_{40}	1.85*	0.78	2.39
HLM (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy, γ_{00}	431.37*	5.26	81.97
	EDUSHORT, γ_{01}	-8.24*	3.54	-2.33
	STAFFSHO, γ_{02}	-12.35*	3.95	-3.13
	EPIST, γ_{10}	7.43*	0.83	8.92
	ANXTEST, γ_{20}	-6.09*	0.93	-6.52
	MOTIVAT, γ_{30}	3.78*	1.05	3.59
	ESCS, γ_{40}	1.86*	0.93	2.00

*$p < .05$

According to Table 8, the significant variables in the multiple regression analyses are the epistemological beliefs (EPIST) of students and the index of economic, social, and cultural status (ESCS), both being the student-level variables, while the predictors with highest coefficients in the HLM analyses are the level of epistemological beliefs and test anxiety (ANXTEST), both being the student-level variables again. Besides, it is seen that the coefficients of regression analyses are estimated higher than those of the HLM analyses, whereas the standard errors pertaining to the coefficients are estimated lower, as it was the case in the previous models. In case of weighting, a remarkable increase is observed in standard errors of level-2 variables in the multiple regression analyses. As for the HLM analyses, weighting does not create any considerable change on the coefficients and standard errors thereof. Table 9 demonstrates the random effects pertaining to the HLM analyses.

Table 9. Random Effects Pertaining to The Fourth Model

Analysis	Random Effect	Sd	Variance	χ^2
HLM (PV1SCIE1)	Level-2 Error Term, u_{0j}	52.97	2805.70	1666.33*
	EPIST Effect,u_{1j}	1.63	2.66	143.29
	ANXTEST Effect,u_{2j}	3.95	15.59	159.68
	MOTIVAT Effect, u_{3j}	5.13	26.35	176.69
	ESCS Effect, u_{4j}	1.65	2.72	146.78
	Level-1 Error Term, r_{ij}	52.15	2719.60	
HLM (PV1SCIE1-10) Weighted	Level-2 Error Term, u_{0j}	55.15	3041.55	1786.47*
	EPIST Effect, u_{1j}	2.50	6.24	144.95
	ANXTEST Effect, u_{2j}	2.70	7.27	146.54
	MOTIVAT Effect, u_{3j}	5.10	26.99	163.38
	ESCS Effect, u_{4j}	2.96	8.79	149.15
	Level-1 Error Term, r_{ij}	51.89	2692.70	

*p < .05

In order to determine the percentages of variance explained for the models of intercepts and slopes as dependent variables, the variances obtained from these models were compared with those obtained from the random effects ANOVA model.

Level-1 variance explained:

Unweighted HLM: $\rho_1=(\sigma^2_{ANOVA}-\sigma^2_{MAOANCOVA})/\sigma^2_{ANOVA}$ = (2873.90 - 2719.60) / 2873.90 = 0.05 (7)

Weighted HLM: $\rho_2=(\sigma^2_{ANOVA}-\sigma^2_{MAOANCOVA})/\sigma^2_{ANOVA}$ = (2848.21 - 2692.70) / 2848.21 = 0.05 (8)

Level-2 variance explained:

Unweighted HLM: $\rho_1=(\sigma^2_{ANOVA}-\sigma^2_{MAOANCOVA})/\sigma^2_{ANOVA}$ = (3073.53 - 2805.70) / 3073.53 = 0.09 (9)

Weighted HLM: $\rho_2=(\sigma^2_{ANOVA}-\sigma^2_{MAOANCOVA})/\sigma^2_{ANOVA}$ = (3536.05 - 3041.55) / 3536.05 = 0.14 (10)

Accordingly, the level-1 variance explained remained the same when one plausible value was used and weighting was not applied in the analyses performed via the model of intercepts and slopes as dependent variables. Per contra, the level-2 variance explained was found higher (14%) when all plausible values and weights were used together.

DISCUSSION and CONCLUSION

This study aimed to compare the results of multiple linear regression and HLM in cases of using a plausible value and all plausible values together with survey weights as an indicator of students' science literacy. Within the scope of this aim, the estimates of those methods were compared regarding the four cases, i.e., the absence of any explanatory variable, the existence of student-level variables, school-level variables, and variables from both levels.

In the models without any explanatory variables, the highest average of science literacy was estimated through the multiple linear regression model using all plausible values and weights. In general, both multiple linear regression analyses can be said to have estimated science literacy higher than the HLM analysis did. Weighting was effective in estimating the coefficient-related standard errors in both analyses of regression and HLM. It was observed that standard errors were greater when weighting was applied in both analyses. Hence, it can be asserted that weighting has a considerable role in relation to the significance of coefficients. Students' science literacy varied from a school to another, according to the random effects of both random-effects ANOVA models. It was observed that the percentage of variance explained by schools as level-2 units was higher when weighting was applied. This means that the difference among schools further increased as a result of weighting. In this study, it was seen that approximately 55% of the variance in the dependent variable was explained by level-2 units. This

result manifests the importance of using HLM analyses as emphasized in several studies (Bryk & Raudenbush, 2002; Goldstein, 2011; Osborne, 2000; Raudenbush, 1988; Raudenbush & Bryk, 1986; Woltman et al., 2012).

For all the models, science literacy was predicted significantly by students' epistemological beliefs in science, test anxiety, motivation, and the index of economic, social, and cultural status. The literature about large scale studies such as PISA and TIMSS contain many researches that investigate economic, social and cultural development index (Acar & Öğretmen, 2012; Atar & Atar, 2012). The findings of current study related this variable is parallel with former ones. Epistemological beliefs and the index of economic, social, and cultural status were the variables with the biggest coefficients in three out of four models over which level-1 variables were examined, the coefficients related to these variables were estimated quite higher in multiple regression analyses as compared to the HLM. The standard errors estimated in those four models were quite close to each other. However, in case of weighting, the standard errors estimated were observed to be slightly higher compared to the other models. Even though all the explanatory variables were significant and the standard errors were close to each other (except for the unweighted multiple regression), it was concluded that the coefficients obtained from the HLM and multiple linear regression analyses showed remarkable differences. This result is in parallel with Roberts's (2004) observation that research findings differ significantly when the hierarchical data structure is not taken into consideration. The HLM analysis showed equal percentage of level-1 variance explained by the model in which only a plausible value was used, as well as the model in which all plausible values and weights were used together. This result may be in relation to the student-level explanatory variables versus the school-level weights. Besides, this situation is in compliance with Wu's (2005) conclusion in a simulation study that using any of the plausible values alone is enough to estimate the population parameters highly correctly.

The level-2 explanatory variables of class sizes, educational leadership, shortage of educational material and staff proved to be significant on both multiple linear regression models. However, for both HLMs, only the shortage of educational material and the shortage of educational staff were significant. This result stems from the fact that t values turn out to be higher than they must be, because the difference of level is ignored in the multiple linear regression analysis, and the level-2 variables in the nested data tend to be significant. Several other studies have also set forth that HLM is more effective in prediction and able to estimate the coefficients and related standard errors more accurately than the traditional analyses are (Gelman, 2006; Goldstein, 2011; Raudenbush, 1988). On both multiple linear regression and HLM method, using all plausible values in company with weights augmented the coefficient-related standard errors. In this case, it is possible to assert that the usage of weighting reduces the risk of type-2 errors for both analysis methods. In the HLM analysis, the use of all plausible values along with weights increased the percentage of variance explained, though they did not influence the coefficients much. Accordingly, using multiple plausible values and weights appears to enhance the performance of HLM analysis.

It was seen that all student- and school-level variables included in the model were significant factors affecting the students' overall science performance in each model. On the other hand, the variables with highest coefficients were the epistemological beliefs of students and the index of economic, social, and cultural status in the multiple regression analyses, while the level of epistemological beliefs and the shortage of educational material and staff were in the HLM analyses. The coefficients were estimated higher and the related standard errors were estimated lower in the multiple linear regression analyses than they were in the HLM analyses, even when all of the variables were included in the model. Regarding such hierarchical data, several other studies confirm that the results of HLM and those of the traditional linear models differ from each other (Bryk & Raudenbush, 2002; Gelman, 2006; Goldstein, 2011; Osborne, 2000; Raudenbush, 1988; Raudenbush & Bryk, 1986; Woltman et al., 2012). Using all coefficients in company with weights had a considerable effect on the standard errors of coefficients pertaining to the school-level variables in the multiple linear regression analyses, whereas it did not generate any remarkable effect on the HLM analyses. The use of all plausible values together with weights in the HLM analyses produced an effect, similar to that of previous models, on the percentage of variance explained at student and school levels. The percentage of student-level

variance explained did not change, while that of school-level variance increased. The inclusion of school-level variables into the model has a different impact on the results, therefore. These results support the necessity of considering the differences of level during analyses.

When all plausible values are used in concurrence with weights, model coefficients do not increase to a considerable extent, though an increase is observed in the related standard errors, in cases that student- and school-level variables are included into the models separately or together in the multiple linear regression analysis. Any increase in standard errors has a bearing on t values, from which the significance of predictor variables is affected in turn. In this study, the variables included into the model were significant despite the decrease in t values. Thus, it is possible to argue that the usage of all plausible values in company with weights does not create a remarkable change on the parameters of multiple linear regression. Although this result is in parallel with the results of a study by Wu (2005) about the use of plausible values, it shows that the way of using survey weights as proposed by OECD (2017) does not generate any change on the outcome. This finding supported by the finding of Carle's (2009) study. Carle asserts that coefficients of weighted and unweighted models are slightly different from each other. However, standard errors diverge comparably. The coefficients and the related standard errors demonstrated a similar tendency in the HLM analysis. Notwithstanding that, the models in which all plausible values and weights were used in company proved to be more conservative in terms of significance and increased the percentage of variance explained in the HLM analysis, which makes it essentially usable in precise studies.

These research results indicate that the outcomes of using HLM for hierarchically structured data are different from those of the multiple linear regression analysis. Since multiple linear regression is not appropriate and adequate for nested data, HLM analysis should be preferred for that purpose. In this way, the separate and collective effects of explanatory variables at different levels will be observed, and the explanatory variables that predict the dependent variable will be determined accurately and reliably. In this study we used just student-level weights. Under similar conditions new studies can be conducted with scholl or higher level weights.

REFERENCES

Acar, T., & Öğretmen, T. (2012). Çok düzeyli istatistiksel yöntemler ile 2006 PISA fen bilimleri performansının incelenmesi. *Eğitim ve Bilim*, *37*(163). Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/download/1040/346

Adams, R. J., & Wu, M. L. (Eds.) (2002) *PISA 2000 technical report.* Paris: OECD Publications.

Atar, B. (2010). Basit doğrusal regresyon analizi ile hiyerarşik doğrusal modeller analizinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(2), 78-84.

Atar, H. Y., & Atar, B. (2012). Examining the effects of Turkish education reform on students' TIMSS 2007 science achievements. *Educational Sciences: Theory and Practice*, *12*(4), 2632–2636.

Beaton, A.E. (1987). *Implementing the new design.* (The NAEP 1983-84 technical report, Report No. 15-TR-20). Princeton, NJ: Educational Testing Service.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika 46*, 443-459.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education 97*(1), 65-108.

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. (2010). HLM7 for Windows [Computer software]. Chicago, IL: Scientific Software International, Inc.

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, *9*(1), 1-13. doi: 10.1186/1471-2288-9-49

Chowa, G. A., Masa, R. D., Ramos, Y., & Ansong, D. (2015). How do student and school characteristics influence youth academic achievement in Ghana? A hierarchical linear modelling of Ghana Youth Save baseline data. *International Journal of Educational Development*, *45*, 129-140.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: John Wiley and Sons.

Fraenkel, J. R.; Wallen, N. E.; Hyun, H. H. (2012): *How to design and evaluate research in education* (8th Ed.). New York, NY: McGraw-Hill Humanities / Social Sciences/Languages.

Gelman, A. (2006). Multilevel (hierarchical) modelling: What it can and cannot do. *Technometrics 48*(3), 432-435.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). Oxford: John Wiley & Sons.

International Business Machines Corp. (2015). IBM SPSS Statistics for Windows (Version 23.0) [Computer software]. Armonk, NY: IBM Corp.

International Association for the Evaluation of Educational Achievement, (2016), Help Manual for the IDB Analyzer. Hamburg, Germany. Retrieved from www.iea.nl/data)

Lohr, S. (2010). *Sampling: Design and analysis* (2nd edition). Boston, MA: Brooks / Cole.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86-92. doi:10.1027/1614-2241.1.3.86

Meinck, S. (2015). Computing sampling weights in large-scale assessments in education [Special issue]. *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach*. Retrieved from https://surveyinsights.org/?p=5353

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J. (1993). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika, 58*(1), 79–85.

Organization for Economic Cooperation and Development (2009). Analyses with plausible values. In *PISA Data Analysis Manual: SPSS*, (Second Edition), OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264056275-9-en

Organization for Economic Cooperation and Development (2014). *PISA 2012 technical report*. Paris: OECD.

Organization for Economic Cooperation and Development (2017). *PISA 2015 Technical report*. Paris: OECD.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, *7*(1), 1-3.

Palardy, G. J. (2011). Review of HLM 7. *Social Science Computer Review, 29*(4), 515–520. doi: 10.1177/0894439311413437

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, *13*(2), 85-116.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*(1), 1-17.

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modelling. *Learning Disabilities: A Contemporary Journal 2*, 30-38.

Rubin, D. B. (1987). *Multiple imputations for non-response in surveys.* New York, NY: Wiley.

Särndal, C., Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer-Verlag.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* London: Sage.

Snijders, T., & Bosker, R. (2003). *Multilevel analysis: An introduction to basic and applied multilevel analysis.* Thousand Oaks, CA: Sage Publications.

Stipek, D., & Valentino, R. A. (2015). Early childhood memory and attention as predictors of academic growth trajectories. *Journal of Educational Psychology*, *107*(3), 771-788.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful. *IERI Monograph Series*, *2*, 9-36.

Warm, T. A. (1985). *Weighted maximum likelihood estimation of ability in item response theory with tests q/jinite length.* (Technical Report No. CGI-TR-85-08). Oklahoma, OK: Coast Guard Institute.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52-69.

Wright. B.D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2), 114-128.

Yang, H. (2013). The case for being automatic: Introducing the automatic linear modeling (LINEAR) procedure in SPSS statistics. *Multiple Linear Regression Viewpoints*, *39*(2), 27–37.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 249-265

Development of Selcuk Sexual Development Scale (36-72 Months) *

Ayşe ALPTEKİN ** Kezban TEPELİ ***

Abstract
This study aimed to develop valid and reliable measurement tools aiming to obtain information from the child and family to determine the sexual identity and gender behaviors of children with normal development at 36-72 months. The research was designed in the general screening model of quantitative research methods. The validity and reliability analyses of three different subscales, namely the Sexual Identity Sub-scale of the Selçuk Sexual Development Scale (SSDS, 36-72 Months), the child form, the Gender-Related Behavior Sub-Scale, the child form, and the Sexual Identity and Gender-Behavior Sub-Scale were conducted. SPSS 20, LISREL 8.80 and FACTOR software were used to analyze the data. The target population of the study consists of 36-72 months of normal development children living in the central districts of Konya between 2017-2018. As a result of the Exploratory Factor Analysis, the eight-item two-factor structure for the SSDS Sexual Identity Sub-scale child form, the eight-item single-factor structure for the Child Form of the Gender-Behavior Sub-Scale, the eight-item two-factor structure for the family form of the Sexual Identity and Gender-Behavior Scale were obtained. The results of the Confirmatory Factor Analysis of these structures showed the compatibility of the structures to the model. The reliability coefficients of the scales were calculated as .61 for the Sexual Identity Sub-Scale child form, .66 for the Child Behavior Sub-Scale, and .85 for the Sexual Identity and Gender Behavior Sub-Scale family form.

Key Words: Sexual identity, sexual development, sexual behavior, sexual development scale.

INTRODUCTION

Sexual development encompasses the growth and development of the reproductive organs of the individual's own sex and the problems and behavior changes related to this development (Ministry of National Education [Milli Eğitim Bakanlığı-MEB], 2013). Sexual development is not only related to changes in anatomical structures, but also related to emotional and cognitive developments (Tuzcuoğlu & Tuzcuoğlu, 2004). In this respect, when determining the sexual development, the situations in which the child is involved in the emotional and cognitive process should be observed.

Every child comes to the world with the anatomical structure and sexual identity that determines whether it is biologically male / female. However, the difference between the gender difference of the child is not with birth, but later in life (Gürşimşek & Günay, 2005). The acquisition of the skills required by the gender, the behavioral and self-concept of individual characteristics is defined as the process of gender discrimination (Başal & Kahraman, 2011). This is possible with the sexual education that can be given to the child and the right model in life.

The roles of men and women are considered to be defined by biological sex, although they are actually defined by the societies themselves. This view is the basis for the formation of judgments that women are different from men, that they should take different roles and that they should continue their lives in a different world than men. Hence, gender roles emphasize the qualities created by the society

* Development of Selçuk Sexual Development Scale (36-72 Months) and Examination of Sexual Development in Children with Disabilities for 48-72 Months, Prof.Dr. Kezban Tepeli, 25.06.2018; IV. International Child Development Congress Oral presentation, Hacettepe University Cultural Center, Sıhhıye- Ankara, 22-24 October 2018.

** Lect. PhD., Selcuk University, Vocational School of Health Services, Konya-Turkey, elmaliayse@hotmail.com, ORCID ID: 0000-0002-3524-5265
*** Prof. PhD., Selcuk University, Faculty of Health Sciences, Konya-Turkey, ktepeli@selcuk.edu.tr, ORCID ID: 0000-0003-3403-3890

To cite this article:
Alptekin, A. & Tepeli, K. (2019). Development of selcuk sexual development scale (36-72 months). *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 249-265. doi: 10.21031/epod.505352
Received: 30.12.2018
Accepted: 15.07.2019

related to masculinity and femininity, not physical characteristics that cause men or women to be separated from each other (Altınova & Duyan, 2013).

The Equal Opportunity Commission answers the question about what variables are the determinants of sexual identity: The word gender refers to the social character rather than the biological character. And social-cultural differences between men and women are learned over time (Skelton and Hall, 2001). The Equal Opportunity Commission emphasizes that sexual identity is based on social development and social and social perception rather than biological diversity. The child may clearly know that the gender is male or female but may feel different or may want to be a member of the opposite sex group (Zucker et al., 1993). In addition to knowing that the child is a man and a woman, being aware of the gender differences brought about by social and social perceptions is important in determining sexual development.

Childhood, which constitutes the first years of life, is gaining importance because all of this information is a period that must be acquired in a healthy way (Yurdakul, 2012). 3-6 years is one of the important periods for sexual development. It is a period in which the children's sexual curiosity is at the highest level, they acquire their own sexual identity and they acquire their sexual roles by identifying with their own gender. The social aim of sexual education in this period is to educate sexually healthy individuals. Through sexual education, children can acquire positive feelings and behaviors by learning the necessary information about sexuality (Yılmaz, 2011). Questions arise about sexuality in the age of 3-6 years. Parents cannot be sure what, how much, when and how they will give (Cole, 1998). If the child does not learn about birth and gender differences from his parents, he will start to look for answers from other sources. Then the result may not be as desired. If a person does not meet the curiosity of the child cannot be said that the problems can be solved completely (Yavuzer, 2000). Acquisition of non-age-appropriate sexual information is often claimed to be a proof indicator of sexual abuse. But very few people can define what their children know (Volbert, 2000). The questions that children ask us to help us to understand their level of knowledge. When they are not ready, the information presented is useless to confuse children (Bayrak, Başgül & Gündüz, 2011).

Children's questions about sexuality should be answered in accordance with age and developmental characteristics. If the child is not informed about sexuality and the curiosity is not resolved, the child will try to satisfy this curiosity in other ways. Sexual curiosity, which is the most innocent tool to overcome this curiosity, is likely to be used by children of later ages for their own purposes, although it is healthy (Bozer, 2009). This information emphasizes the importance of what children know, what they are curious about and what information to give, but also the necessity of measurement tools to determine these situations in order to provide healthy sexual education to children.

Purpose of the Study

When the studies examining the sexual development of children are examined, it is seen that there are many studies but the subject and scales used are generally aimed at examining gender stereotypes (see Aksoy, 1990; Aydilek-Çiftçi, 2011; Baran, 1995; Barutçu, 2002; Başal & Kahraman, 2011; Edelbrock & Sugara, 1978; Gündüz-Sentürk, 2015; Güney, 2012; Köseler, 2009; Lamb & Roopnarine, 1979; Langlois & Downs, 1980; Lobue & DeLoache, 2011; Özdemir, 2006; Özkan, 2009; Şıvgın, 2015; Şirvanlı-Özen, 1992). Considering that sexual development is related to cognitive and emotional domains, it is not sufficient to measure only gender stereotypes. Knowing whether or not the child is aware of gender roles and perceptions is not sufficient to determine the child's sexual development. Ignoring the child's emotional feelings, curiosity, and whether he / she sees the situation as a taboo makes sexual abuse possible. In the literature, the fact that the majority of the scales and sexual development are only intended to measure stereotypes, the lack of measurement tools to measure the components of sexual development, or the presence of measurement tools that provide information only from adults, is a scale that aims to obtain information from a child that includes all the components of sexual development but also provides information from an adult. the need to gain. In addition, in this study, an assessment tool consisting of child and parent forms was developed to evaluate the sexual identity and gender behaviors of children with normal development (SSDS Sexual Identity Sub-Scale

Child Form, SSDS Gender Behavior Sub-Scale Child Form, SSSS Sexual Form). Identity and Gender Behavior Subscale Family Form).

METHOD

The research was designed in the general screening model of quantitative research methods. In order to determine the indicators of sexual development during the development of the scale, theories and theories explaining sexual development were examined. Kohlberg is influenced by Piaget's views in Cognitive-Developmental Theory. Measurement tools were formed by grouping them as sexual identity based on Kohlberg, Bem and Bandura's views and gender-based behaviors based on Freud's views.

Participants

As the research group was formed to develop the Selcuk Sexual Development Scale (36-72 Months), children and their families who agreed to participate in the study were sampled and non-probability sampling methods were used. In addition, for multivariate analysis according to Kline (2013), Coşkun, Altunışık and Yıldırım (2017), attention should be paid that the number of variables used in the study is at least 10 times or more. Moreover, according to Çapık (2014), 63% of the studies in the Psych INFO database used this criterion. For sample counts, the distribution of data, the number of items, the complexity of the model should be taken into consideration criteria such as (Çapık, 2014) evaluated the opinions of 36-48 monthly 102, 49-60 monthly 113, 61-72 Children 101 per month in total 316 Children and parents of these children (the person who spends most of the time with children) were included in the study Group. In order to determine the criterion-related validity of this group, 90 children and their families whose selected gender were used were applied.

Data Collection Instruments

In the study, the development of the SESG Sexual Identity Sub-Scale Child Form, the SESG Sexual Behavior Sub-Scale Child Form, and the SESG Sexual Identity and Gender Behavior Sub-Scale Family Form were developed. In addition, in order to determine the criterion-related validity of these forms, the Gender Mold Questionnaire developed by Williams, Bennett and Deborah (1975) and adapted to Turkish by Şirvanlı-Özen (1992) was applied.

SDSS sexual identity subscale development of child form

In order to develop the scale, firstly the national and international literature on the subject was scanned, and the reference books on sexual development activities prepared for children and the gender invariance scale developed by Taylor (2004) and adapted to Turkish by Zembat and Keleş (2011). Williams et al. (1975) developed by Şirvanlı-Özen (1992) adapted to Turkish Gender Stereotype Scale, Gender Roles stereotyping Scale developed by Eren (1986), Bem Gender Role Inventory, which was developed by Bem (1974) and adapted to Turkish by Kavuncu (1987), Preschool Activity Inventory, adapted to Turkish by Ünlü (2012), developed by Golombok and Rust (1993), The Gender Role Learning Index developed by Edelbrock and Sugara (1978), the Gender Judicial Scale developed by Altınova and Duyan (2013), the Sexual Identification Scale developed by Artan (1987), the Gender Measurement Tool developed by Şıvgın (2015) were analyzed.

After forming the items in the light of literature, six experts specialized in preschool, child development, psychological counseling and assessment in education were presented. After the necessary corrections were made at the end of the expert opinion, identical cards were created for each item, and the pictures expressing the situation for the items were drawn by an illustrator specialized in the field of children's books illustration. At each stage of the drafting of the drawings and at the final stage, the opinions of the experts who examined the substances were consulted. After the necessary

corrections, a pilot study was conducted to determine whether the drawings were understood as intended by the children. Children were asked what happened in the paintings and what the children in the paintings did. One male and one female child of each age group (36-48 months, 49-60 months, 61-72 months) answered the questions. The final version of the scale was presented to the opinion of 10 experts who are experts in preschool, child development, psychological counseling and assessment in education fields, which are experts in the creation of items and pictures.

SDSS gender behavior subscale development of child form

In order to develop the scale, national and international literature was searched, and Child Sexual Behavior Inventory developed by Friedrich, Fisher, Broughton, Houston and Shafran (1997) was examined. In addition, eight preschool teachers were asked to list the sexual behaviors of their children and ten parents. Preschool teachers' behaviors indicated by their students about sexual behaviors are as follows; Trying to look at your friend entering the toilet, drawing or making sexual organs while painting or making figures with dough, addressing each other with the words of my love, darling, playing a doctor's game, playing a house game, saying that they will marry a person he knows.

Parents' sexual behaviors of their children are as follows; liking nudity while changing their tops, examining themselves naked in front of the mirror, removing dolls, trying to make up like the mother of daughters, trying to shave like the father of men, jealous of the opposite sex parent and the parent. These behaviors are accepted as sexual behavior (Friedrich et al., 1997; Kandır, 2004).

In the light of this information, an item pool was created, and two or three sentences were written for each item. Instead of asking children directly, it was found appropriate to be asked through another person who is a projective way. The statements in the direct tests allow the person performing the test to give the expected answers and mislead the test as he wishes, whereas in the indirect tests there is no possibility of such a mistake. The individual does not know the meaning and importance of his answers (Günay & Çarıkçı, 2019). Considering that the subject of sexuality is shown as a taboo to children, the indirect method is preferred to prevent the child from giving the desired and taught answer and not the situation he feels, wants and thinks, and to obtain correct results. The heroes of the stories were created from children's characters. Separate story characters were selected to facilitate identification with boys and girls. The story characters of girls are designed as girls and the story characters of boys are designed as boys. Preschool, child development, psychological counseling and assessment in education were presented to the expert opinion of six people, necessary arrangements were made in line with the opinions. Illustrations suitable for the stories were drawn by an illustrator specialized in the field of children's books illustration. At each stage of the drafting of the drawings and at the final stage, the opinion of the experts who examined the substances was sought. A pilot study was conducted to determine whether the children understood these pictures in the way they wanted to be told, and in this pilot study, children were asked what was in the pictures and what the children in the pictures were doing. One male and one female child of each age group answered the questions. The two children who participated in the study perceived the mirror in the child's picture as a door and window. The other children perceived the entire picture as intended. After these results, the pictures of children examining themselves in front of the mirror were reviewed and corrected in a way to eliminate misunderstanding. The same children were shown the pictures again and it was seen that the children perceived as they wanted to be told. At the end of these studies, the final version of the scale was presented to the opinion of 10 experts who were experts in preschool, child development, psychological counseling and assessment in education.

SDSS sexual identity and gender behavior subscale development of the family form

National and international literature was searched for the development of the scale. The 28-item family form, which was designed as a likert type including the items of the children's forms, was presented to the opinion of 10 experts specialized in preschool, child development, psychological counseling and assessment in education.

Data Collection Procedure

Sexual identity subscale of SDSS child form, gender behavior sub-scale for the implementation of the children's form was interviewed individually with the families of children of 36-72 months, the scale was shown to the family and the family was approved after the approval of the family At a time when he saw a researcher in the home environment and the child was applied to the child by the researcher. Peer cards were shown to the child, questions were asked, and the child's answers were recorded. First, the Sexual Identity Sub-Scale was administered, and the correct answer was scored as 1 and the wrong answer as 0. After that, the expected response from 36-72 months old children was scored as 1 and the other was 0. After the application of the child was completed, the family was asked to fill in the family form. The Likert-type scale was scored between 1-5 questions about sexual identity and 1-3 questions about sexual behavior. At the end of the study, 316 children and their families (the mother or father who spent more time with the child were preferred) were reached.

Data Analysis

While analyzing the collected data; Internal reliability coefficient KR-20 was used for the reliability analysis of the SSDS Sexual Identity Sub-Scale child form and the SRSG Sex-Related Behavior Sub-Scale child form. The Lawshe Scope Validity Index was calculated for the scope validity. The reliability and validity of the SRSG was developed by Williams et al. (1975). In construct validity, exploratory and confirmatory factor analyzes were performed. The Kaiser-Meyer-Olkin test and Bartlett's sphericity test were used to investigate the suitability of the data for factor analysis. As the scale was scored as 1-0, tetrachoric factor analysis was used and FACTOR software developed by Rovirai Virgili University was preferred (Aybek, 2017). As a result of the analysis, Chi-square (χ^2), χ^2/sd, RMSEA, RMR, GFI, NNFI, NFI and AGFI goodness of fit indices were examined. The Cronbach's Alpha Coefficient was used to calculate the reliability of the family form of the SRSG Sexual Identity and Gender-Behavior Sub-Scale. Lawshe Scope Validity Index was calculated for scope validity and exploratory and confirmatory factor analysis was performed for construct validity. SPSS 20 software was used for exploratory factor analysis and LISREL 8.80 software was used for confirmatory factor analysis.

RESULTS

In this section, the validity and reliability analyses of SSDS sex identity subscale child form, SSDS sex behaviour subscale child form, SSDS sexual identity and sex behaviour subscale Family form were made and the findings were studied to be explained.

Validity Analysis Result

Results on scope validity analysis

Developed SDSS sexual identity subscale child form, SSDS gender-related behavior subscale child form, SDSS sexual identity and Sexual Behavior subscale for expert evaluation of family form preschool, child development, counseling, assessment and evaluation in education expert opinion of 10 faculty members has been applied. According to Lawshe (1975) .05 coverage validity rates at the level of significance the lowest values are examined and the lowest values that the items of the tests can receive as a result of the ten expert reviews. It was determined that it was 62 and it was appropriate to eliminate test items with lower value than this. The child form, the child form of the SDSS gender-related behavior subscale, the family form of the SSDS sexual identity and Sexual Behavior subscale has been applied to the views of ten experts. SSDS sex identity sub-scale child form article when the Scope Validity Rate (SVR) of the scope of expert opinions is calculated 5, 6, 11, 12, 13, 17, 18 coverage SVR .8, other substances 1; SSDS gender-related behavior subscale child form item 1, 3, 5, 6, 7, 8, 11, 12 scope validity rate .8 and the other items were calculated as 1. As a result of the analysis,

the SSDS of no substance for child forms. Since it was not below 62, all substances were in the substance pool. It also calculated the Scope Validity Index (SVI) for all of the children's forms. SSDS sex identity subscale = child form SVI. 86, SSDS gender-related behavior subscale = child form SVI .92 have been found. The lowest scope of these values is the validity criterion (SVR = .62) was determined to be greater than the value determined for, and the scope validity of the tests was found to be statistically significant. (SVI > SVR). SDSS sex identity and gender behaviour sub-scale Family form, article when the SVR for expert opinions is calculated 5, 6, 11, 12, 13, 14, 17, 18, 19, 20, 21 coverage SVR .8, item nine .2, item ten .4 other items were calculated as 1. As a result of the analyses, the SVR of articles nine and ten. Since it remains below 62, it has been removed from the substance pool and all other substances have been placed in the substance pool. The scope validity index was also calculated for the entire test (SVI = .87). The lowest scope of this value is the validity criterion (SVR = .62) was determined to be higher than the value determined for, and the validity of the test was found to be statistically significant (SVI > SVR).

Results on criterion-related validity analyses

SSDS and, Williams et al. (1975) developed by Şirvanlı-Özen (1992) adapted to Turkish gender stereotyping scale and criteria related validity were examined and the results were explained by Table 1.

Table 1. Results on The Validity of SDSS on The Scale of Sexual Stereotyping and Criteria

	Gender Stereotype Scale					
	36- 48 month		49-60 month		61-72 month	
	n	r	n	r	n	r
SDSS sex identity sub-scale child form	30	.61	30	.74	30	.71
SDSS sex identity sub-scale child form sexual balance sub-size	30	.74	30	.71	30	.65
SDSS sex identity subtype child form sexual role subtype	30	.50	30	.63	30	.58
SDSS child form of sexual behavior subscale	30	.30	30	.21	30	.29
SDSS gender identity and Sexual Behavior subcategory Family form	30	.47	30	.52	30	.59
SDSS sex identity and Sexual Behavior subscale family Form sex identity subscale	30	.76	30	.64	30	.60
SDSS sex identity and Sexual Behavior subscale family Form gender-related behavior subtype	30	.20	30	.29	30	.31

Table 1 examined the SSDS sexual behaviour subscale between the child form and the gender stereotyping scale; SSDS sexual identity and sexual SSDS sexual identity and gender behaviour subscale family form between the gender behaviour subscale and the gender stereotyping scale of 36-48, 49-60, 61-72 months, it was determined that the correlations were not significant but there was a low level of correlation. While it can be said that there are positively significant relationships between other sub-dimensions and that there is criterion-dependent validity, this sub-scale provides low criterion-dependent validity as there is a low positive correlation between the child form and the sexual identity and sex-related behavior sub-scale of the family form and the gender-related behavior sub-dimension.

Results on construct validity analysis

Analysis of the construct validity of the scales has been made and tried to explain. The KMO value of the SSDS sex identity subscale .69 and Bartlett's globality test was $p < .05$. Hence the value of KMO. Because there is a relationship between variables greater than .60, the variables are suitable for factor analysis according to both test results.

In this study, the elimination of substances that do not measure the same construct and the determination of the number of important factors in determining Çokluk, Şekercioğlu and Büyüköztürk (2016), Kline (2013) with Coşkun et al. (2017) is based on the opinions of.

After the first factor analysis, it was seen that there were eight sub-dimensions, i.e. factors with eigenvalues above 1, and these factors explained 63.67% of the total variance. Varimax vertical rotation technique is used to explain the sub-dimensions better. Varimax rotation technique is used mostly because it can be rotated in such a way that factor variances are maximum with fewer variables (Tavşancıl, 2005). After the analysis, items that did not reach a factor load of, .30, received a high load value (overlapping) at least two factors, or were found to form a subdimension alone were excluded from the scale. After the exclusion of these items from the scale, exploratory factor analysis was performed and two sub-dimensions of eight items were obtained. After the second factor analysis, it was seen that there were two subconstruct with eigenvalues above 1 and these factors explained 43.89% of the total variance.

Table 2. SDSS Sex Identity Sub-Scale Hungry Factor Analysis Results of Two-Factor Construct Factor Load Distribution According to Varimax Rotation

Item	**1.Load Values For The Factor**	**2.Load Values For The Factor**
Sexual Balance 1	.99	
Sexual Balance 2	.68	
Sexual Balance 3	.78	
Sexual Role 4		.83
Sexual Role 7		.51
Sexual Role 8		.63
Sexual Role 9		.40
Sexual Role 10		.64
Self-worth	3.03	1.35
Described Variance Ratio	%23	%20

As can be seen in Table 2, the load values of the substances in the first construct vary between .99 and .68, while the load values of the substances in the second construct vary between .83 and .40. The first construct explains 23.73% of the variance, while the second factor explains 20.15%. All 8 items in the scale explain 43.89% of the total variance.

In the naming of the two sub-dimensions, the contents of the items in the sub-dimensions were taken into consideration. When the contents of the factors were examined, it was found that the items in the first factor (1, 2, 3) were related to the basic sexuality personality and sexual balance of the children and this factor was called sexual balance. The items in the second factor (4, 7, 8, 9, 10) were identified as expressing sexual role gains of children and therefore the factor was called sexual role.

The SDSS sex identity sub-scale is based on the results of the Two-Factor Model Verifier Factor Analysis p = .026, X^2 / sd = 3.44, RMSEA = .09, RMR = .02, GFI = .98, AGFI = .96, CFI = .99, NFI = .92, NNFI = .90 found. RMSEA is not within the generally accepted limits when other parameters are examined. Schermelleh-Engel and Moosbrugger (2003, p. 36), although the values between .08 and .10 are low, accept that the model is compatible. Based on this, the value of RMSEA (.09) is thought to adapt to the model even if it is bad. NFI, NNFI values were acceptable while the rest of the values were found to be at an excellent level. All values submitted for compliance goodness to the generally accepted criteria in the relevant field summer (Erdem, 2013; Çokluk et al., 2016; Seçer, 2015; Şimşek, 2007) is perfect and acceptable according to. From this point of view, it can be said that the alignment of the two-dimensional model to the data is confirmed.

SDSS gender-related behavior subscale predictive factor analysis KMO and Bartlett test results KMO value .75 and Bartlett's globality test result is p < determined as .05. The zero hypothesis at the level of .05 significance is rejected. In other words, there is a relationship between the variables in the main mass. Because there is a relationship between variables greater than .60, the variables are suitable for factor analysis according to both test results. The criteria used in the development of the child form of

the SSDS sex identity sub-scale were taken into consideration when analyzing the factor of explicative in the development of the scale.

After the first EFA, it was determined that there was a total of five construct with eigenvalues above 1, explaining 57.22% of the total variance of these construct. Factor load after Varimax upright rotation technique, which is used to better explain the resulting structures. Those who did not reach .30 were eliminated from the boarding and two-dimensional construct was obtained. Although the construct with two lower dimensions was obtained, these substances were removed from the scale because two substances remained in the second factor. Again, the factor analysis was performed and the seven-point single-factor construct was obtained.

Table 3. SDSS Gender-Related Behavior Sub-Scale Hungry Factor Analysis Factor Load Distribution Results for Single Factor Construct

Item No	**1.Load Values For The Factor**
Item 2	.71
Item 4	.71
Item 5	.66
Item 7	.30
Item 8	.54
Item 9	.34
Item 12	.71
Eigenvalue	2.29
Described Variance Ratio	%32.79

As shown in Table 3, the load values of the items in the first factor vary between .71 and .30. The single factor explained 32.79% of the variance. SDSS gender-related behavior subscale according to single-factor model verifier factor analysis results p = .02, X^2 / sd = 1.78, RMSEA = .06, RMR = .05, GFI = .94, AGFI = .92, CFI = .98, NFI = .92, NNFI = .98 found. When the parameters are examined, they also indicate that RMSEA (.06) while expressing that it is within acceptable limits. GFI, NFI, AGFI values are acceptable and other values are found to be excellent. From this point of view, it can be said that the alignment of the one-dimensional model to the data is confirmed.

SDSS sex identity and gender behavior subscale family form explicative factor analysis KMO and Bartlett test results KMO value .84 and Bartlett's test was $p < .05$. Since there is no relation between variables greater than .60, it is observed that variables are suitable for factor analysis according to both test results.

After the first factor analysis, it was seen that there were eight infraconstruct with eigenvalues above 1 and these infraconstruct explained 61.51% of the total variance. After the Varimax vertical rotation technique, which was used to better explain the resulting infraconstruct, the items, which had an overlapping factor, and whose factor load could not reach .30, were removed from the scale. Re-exploratory factor analysis was performed, and a two-factor construct of 18 items was obtained. After the second factor analysis, it was found that there were five factors with an eigenvalue of more than 1 and these factors explained 60.21% of the total variance. However, it was thought that the two-factor construct would be more appropriate as the five-factor construct had difficulty in naming the factors and the variables in the factors could not fully adapt to the theoretical construct. The content of the items included in the sub-dimensions was taken into account in the naming of the two sub-dimensions obtained as a result of EFA. When the contents of the sub-dimensions were examined, it was seen that the items in the first factor (2, 4, 5, 6, 7, 12, 14, 16, 20) expressed opinions about the sexual behaviors of children and this factor was called gender-related behavior. The items in the second factor (17, 19, 22, 23, 24, 25, 26, 27, 28) express the child's sexual identity acquisition and thus the factor is called sexual identity.

Table 4. SDSS Sex Identity and Gender Behavior Subscale Family Form Hungry Factor Analysis Results of Two-Factor Construct Factor Load Distribution According to Varimax Rotation

	1. Load Values for The Factor	**2. Load Values for The Factor**
Behavior 2	.63	
Behavior 4	.48	
Behavior 5	.41	
Behavior 6	.61	
Behavior 7	.54	
Behavior 12	.40	
Behavior 14	.68	
Behavior 16	.34	
Behavior 20	.31	
ID 17		.35
ID 19		.40
ID 22		.71
ID 23		.71
ID 24		.77
ID 25		.83
ID 26		.85
ID 27		.77
ID 28		.82
Self-worth	5.26	2.10
Described Variance Ratio (%)	29.22	11.69

As can be seen in Table 4, the load values of the items in the first factor vary between ".68" and ".31, and the load values of the items in the second factor vary between .85 and .35. All 18 items in the scale explain 40.91% of the total variance.

SDSS sex identity and gender behavior subscale family form two-factor model according to verifier factor analysis results p = .01, X^2 / sd= 2.98, RMSEA = .08, RMR = .08, GFI = .87, AGFI = .85, CFI = .92, NFI = .90, NNFI = .91 found. When the parameters are examined, it is also possible that RMSEA (.08) while expressing that it is within acceptable limits. All values except GFI values were deemed acceptable. All values submitted for compliance goodness (except GFI) can be deemed perfect and acceptable by adhering to generally accepted criteria in the relevant field. From this point of view, it can be said that the alignment of the two-dimensional model to the data is confirmed.

When the standardized path values were examined in Figure 1, the first factor and the variables between .56-.31, the second factor and the variables between .30-.90, the standardized path values were obtained. All t values as a result of CFA .05 it has been determined that it shows values at the level of significance and is meaningful.

To be able to use the total score of the developed scale, a second level CFA is required (Seçer, 2015). SDSS sexual identity and gender Behaviour Scale family form second level DFA compliance indices p = .01, X^2 /sd = 2.58, RMSEA = .08, RMR = .08, GFI = .87, AGFI = .84, CFI = .92, NFI = .89, NNFI = .91 found. According to the of all of the values (Büyüköztürk, Akgün, Özkahveci & Demirel, 2004; Erkorkmaz, Etikan, Demir, Özdamar & Sanisoğlu, 2013; Çokluk et al, 2016; Korucu & Usta, 2017) appears to be within acceptable limits. These values indicate that the total points of the two-dimensional model can be used and adapt to the model.

Results on Reliability Analysis

Cronbach's alpha coefficient was calculated because the Child Forms were scored between 0-1 and KR-20 was scored, and the family form was scored between 0-5 and 0-3. When the change in the KR-20 reliability coefficient was examined, the lowest value was found to be .43 and the highest value was .53 and the total value was .61. In the Sexual Balance sub-dimension, the KR-20 Reliability coefficient was .58 and .53 in the Sexual Role sub-dimension. When the change in the reliability coefficient of the KR-20 was examined, the lowest value was .57 and the highest value was .65 and the total value was .66. When the Cronbach Alpha reliability coefficient change was examined, the

lowest value was .82 and the highest value was .85 and the total value was .85. The Cronbach's alpha reliability coefficient was .65 in the Gender Related Behavior sub-dimension and .88 in the Sexual Identity sub-dimension.

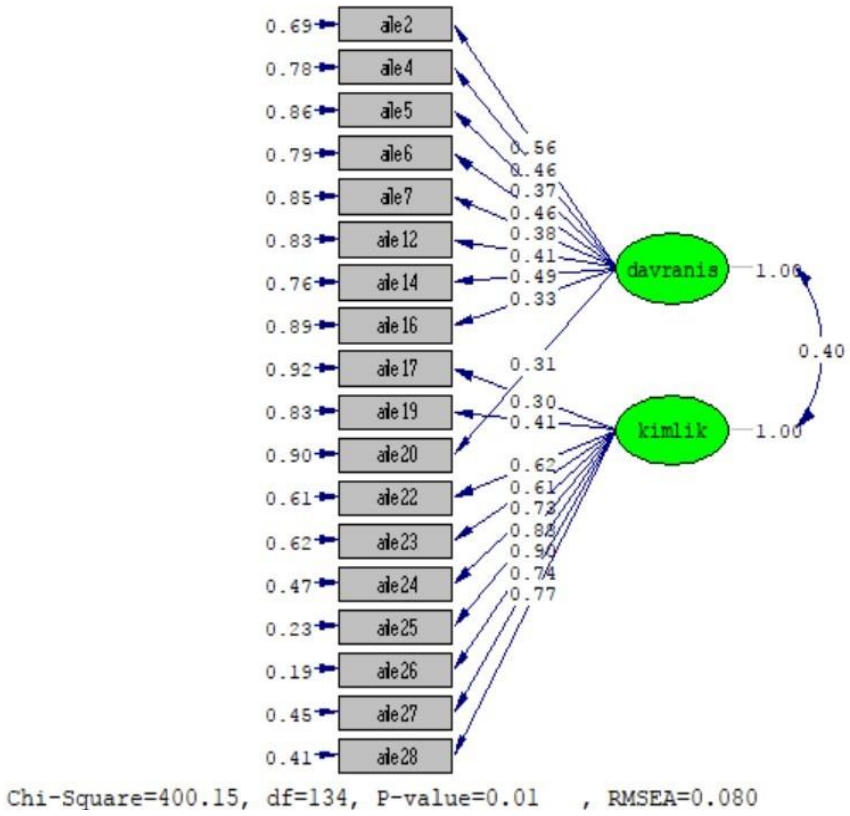


Figure 1. SDSS Sex Identity and Gender Behaviour Sub-Scale Family Form Standardized Path Diagram

DISCUSSION and CONCLUSION

If it is decided to develop a scale, the scale development process starts with the step of creating an item pool after steps such as determination of the construct to be measured, literature review, and interviews with experts (Erkuş, 2014). In this research, a pool was created, the necessary pictures were drawn, 10 experts who were experts in preschool, child development, psychological counseling, measurement and evaluation in education were presented to the opinion and then the content validity rates and content validity indices were calculated. The scale was found to provide validity.

SSDS's, Williams et al. (1975) developed by Şirvanlı-Özen (1992) Gender stereotyping scale adapted to Turkish by examined the validity of the criteria, gender stereotyping scale and SDSS sex identity subscale the child form and its sub-dimensions, sexual identity and gender Behavior Scale the family form and sexual identity sub-dimension were found to have a positive correlation between the child form and sexual identity and gender behavior subscale the family form Sexual identity; the acceptance of the individual to the gender to which he belongs, the perception of himself within this gender, is that emotions and behaviors are appropriate to it (Barutçu, 2002). Gender stereotyping includes behaviors, attitudes, values, ways of thinking, talking, sitting or walking, dressing, and decorating one's own body (Gander and Gardiner, 2005). This information is thought to parallel sexual identity and gender stereotyping and explain the meaningful correlation. But it is assumed that gender-related behavior is not meaningful, although there is a correlation between them, as it is not very closely related to gender stereotyping. Furthermore, a scale measuring gender-related behavior adapted to Turkish has not been found in the literature.

Factor analysis can be applied to reveal the construct of the scale and many more for various purposes (Çokluk et al., 2016). In this research it has been used to determine the structure of the scale, i.e. construct validity. If the collected data is categorically scored as 1-0 and it is desired to perform an explicative factor analysis on this data, then the correlation matrix to which it should refer must be the tetrachoric correlation matrix. (Aybek, 2017; Çokluk et al., 2016; Sandal, 2015). Tetrachoric factor analysis method was preferred as SDSS sexual identity sub-scale was scored as 0-1.

Çokluk's et al. (2016), Kline's (2013), Coşkun's et al. (2017) based on reviews; If a substance is in two subconstruct, the difference between the values of these two factors is at least .10 that factors have a high variance of the common factor they explain in a substance, Kaiser criterion: the eigenvalue of each factor is at least 1, the ratio of the total variance explained by the substances on the scale .30 and more, based on the criteria for determining the number of factors according to the number of points above the point where rapid declines occur, the analysis obtained an eight-item and two-factor construct for the children's form of SDSS sex identity sub-scale and the total variance described was 43.89%. SDSS gender Behaviour Scale seven-item single-factor construct was obtained for the child form, explaining 32.79% of the variance. SDSS sex identity and gender behaviour subscale this 18-item two-factor construct was obtained for the family form, explaining 40.91% of the variance. For multi-factor scales in the Social Sciences, this ratio is expected to be between 40-60% (Çokluk et al., 2016). 30% and more of the variance described in single factor scales can be seen enough (Şekercioğlu, 2009). Therefore, it can be said that the contribution of constructs to total variance is sufficient.

With classical methods, the researcher looks at the relationship between only a few variables, and these relationships may not be sufficient to obtain a complex theory. The analyses CFA uses are advanced and advanced and can produce not just one but more results (Çapık, 2014).

All the values presented on the CFA compliance goodness of the three different sub-scales developed in the study were examined and described as excellent and acceptable. The GFI value of the family Form Two-Factor Structure was found to be .87. Büyüköztürk et al. (2004) a study of the value of GFI. They said that being equal to or greater than .80 showed that the structure was appropriate. The work of Korucu and Usta (2017), Erkorkmaz et al. (2013) also confirms this knowledge.

Büyüköztürk et al. (2004), of the reliability coefficient for a psychological test. They have stated that being .70 and over is enough. The SDSS sex identity and gender-related behavior subscale family form (Crombach alpha = .85) according to Büyüköztürk et al. (2004), it can be concluded that it is a reliable scale. However, the number of substances and the type of measuring instrument is an important factor for the coefficient of reliability, and the SSDS sexual identity sub-scale developed (KR-20 = .61) out of 8 items, the sex-related behavior subscale (KR-20 = .66) consists of 7 articles. Alpar (2014) of the value of KR-20 in measuring instruments consisting of 10-15 items. He stated that even having a value as low as 50 indicates that the test is reliable. In the light of this information, it is thought that child forms of the scales developed in the study are reliable scales.

When the findings of this study are evaluated, the norm studies of SDSS sub-scales can be performed and the sexual development levels of 36-72 months old children can be determined. When the accessible literature was examined, it was found that there was no gender-related behavioral scale directly applied to children. With this scale developed, it can be suggested that research conducted in the literature with information from families before can be repeated.

REFERENCES

Aksoy, C. (1990). *3-6 yaş arası çocukların oyuncak tercihlerinde cinsiyet faktörünün etkisinin incelenmesi* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Alpar, R. (2014). *Spor, sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik-güvenirlik* (3. Baskı). Ankara: Detay Yayıncılık

Altınova, H. H., & Duyan, V. (2013). Toplumsal cinsiyet algısı ölçeğinin geçerlik güvenirlik çalışması. *Toplum ve Sosyal Hizmet Dergisi, 24*(2), 9-22.

Artan, İ. (1987). *Annesi çalışan ve çalışmayan ilkokul birinci ve beşinci sınıf öğrencilerinin cinsel kimliklerini kazanmalarının incelenmesi* (Yayınlanmamış bilim uzmanlığı tezi). Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü, Ankara.

Aybek, E.C. (2017). Tetraktorik faktör analizi-ölçme değerlendirme. www.olcme.net adresinden edinilmiştir.

Aydilek-Çiftçi, M. (2011). *Öğretmenlerin ve farklı sosyo-ekonomik düzeye sahip anne-babaların cinsiyet rolleri algısının 60-72 ay arası çocukların oyuncak tercihleri ve akran etkileşimleri ile ilişkisinin incelenmesi* (Yayımlanmamış yüksek lisans tezi). Çukurova Üniversitesi Sosyal Bilimler Enstitüsü, Adana.

Baran, G. (1995). *Ankara'da bulunan çocuk yuvalarında kalan 7-11 yaş grubu çocuklarda cinsiyet rolleri ve cinsiyet özellikleri kalıpyargılarının gelişimi* (Yayımlanmamış doktora tezi). Ankara Üniversitesi Fen Bilimleri Enstitüsü, Ankara.

Barutçu, E. (2002). *Özel anaokullarına devam eden 6 yaş çocuklarının cinsiyet özelliklerine ilişkin kalıpyargıları ile annenin sosyal uyumları arasındaki ilişkinin incelenmesi* (Yayınlanmamış yüksek lisans tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Başal, H. A., & Kahraman, P. (2011). Anne eğitim düzeyine göre çocukların cinsiyet kalıpyargıları, ile oyun ve oyuncak tercihleri. *Journal of New World Sciences Accademy, 6*(1), 1336-1357.

Bayrak, G., Başgül, Ş. S., & Gündüz, T. (2011). *Ailede cinsel eğitim* (1. Basım). İstanbul: Timaş Yayınları.

Bem, S.I. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*(2), 155-162.

Bozer, M. (2009). *Din eğitimi açısından 0-12 yaş arası çocuklarda cinsel tutum ve davranış eğitimi* (Yayımlanmamış yüksek lisans tezi). Selçuk Üniversitesi Sosyal Bilimler Enstitüsü, Konya.

Büyüköztürk, Ş., Akgün, Ö. E., Özkahveci, Ö., Demirel, F. (2004). Güdülenme ve öğrenme stratejileri ölçeğinin türkçe formunun geçerlik ve güvenirlik çalışması. *Kuram ve Uygulamada Eğitim Bilimleri 4*(2), 207-239.

Çapık, C. (2014). Geçerlik ve güvenirlik çalışmalarında doğrulayıcı faktör analizinin kullanımı. *Anadolu Hemşirelik ve Sağlık Bilimleri Dergisi, 17*(3), 196-205.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2016). *Sosyal bilimler için çok değişkenli istatistik SPSS ve LISREL uyugulamaları* (4. Basım). Ankara: Pegem Yayıncılık.

Cole, J. (1998). *Cinsellikle ilgili merak ettikleriniz* (Çev. E. Aksay). İstanbul: Sistem Yayıncılık.

Coşkun, R., Altunışık, R., & Yıldırım, E. (2017). *Sosyal bilimlerde araştırma yöntemleri* (9.Basım). Sakarya: Sakarya Yayıncılık.

Edelbrock, C. S., & Sugara, A. I. (1978). Acquisition of sex- typed preferences in preschool-aged children. *Journal of Deveopmental Psychology, 14*(6), 614-623.

Erdem, İ. (2013). Kekeme öğrencilere ilişkin öğretmen tutumları: Bir ölçek geliştirme çalışması. *International Journal of Social Science, 6*(7), 401-416.

Eren, A. (1986). *Sex-role and sex-trait stereotypes in children* (Yayımlanmamış yüksek lisans tezi). Orta Doğu Teknik Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Erkorkmaz, Ü., Etikan, İ., Demir, O., Özdamar, K., & Sanisoğlu, S. Y. (2013). Doğrulayıcı faktör analizi ve uyum indeksleri. *Türkiye Klinikleri J Med. Sci, 33*(1), 210-223. doi: 10.5336/medsci.2011-26747

Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme-I: Temel kavramlar ve işlemler* (2. baskı). Ankara: Pegem Akademi.

Friedrich, W. N., Fisher, J., Broughton, D., Houston, M., & Shafran, C. (1997). Normative sexual behavior in children: A contemporary sample. *American Academy of Pediatrics, 101*(4), 1-8.

Gander, M. J., & Gardiner, H. W. (2005). *Çocuk ve ergen gelişimi.* (Çev. B. Onur, H. N. Çelen ve A. Dönmez). Ankara: İmge Kitabevi.

Golombok, S., & Rust, J. (1993). The pre-school activities inventory: A standardized assessment of gender role in children. *Journal of Psychological Assessment, 5*(2), 131-136. doi: 10.1037/1040-3590.5.2.131

Günay, A., & Çarıkçı, İ. H. (2019). İnsan kaynakları işe alım süreçlerinde kullanılan psikoteknik testlere ilişkin bir inceleme. *Süleyman Demirel Üniversitesi Vizyoner Dergisi, 10*(23), 178-194.

Güney, O. (2012). *5-6 yaş çocuklarında algılanan cinsiyet kslıpyargılarına ilişkin ebeveyn beklentileri ile oyuncak tercihleri arasındaki ilişki* (Yayımlanmamış yüksek lisans tezi). Maltepe Üniversitesi Sosyal Bilimler Enstitüsü, İstanbul.

Gürşimşek, I., & Günay, V. D. (2005). Çocuk kitaplarında cinsiyet rollerinin işlenişinde kullanılan dilsel ve dildışı göstergelerin değerlendirilmesi. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi, 18*, 53-63.

Kandır, A. (2004). *Gelişimde 3-6 yaş, çocuğum büyüyor* (2. Basım). İstanbul: Morpa Kültür Yayınları.

Kavuncu, N. (1987). *Bem cinsiyet rolü envanteri'nin Türk toplumuna uyarlama çalışması* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. Petscher & C. Schatsschneider, (Eds.), *Applied quantitative analysis in the social sciences* (pp. 179-185). New York, NY: Routledge.

Korucu, A. T., & Usta, E. (2017). Sosyal medya öğretmen-öğrenci etkileşimi ölçeğinin geliştirilmesi. *İlköğretim Online, 16*(1), 197-216. doi: 10.17051/io.2017.91326

Köseler, F. (2009). *Okulöncesi öykü ve masal kitaplarında toplumsal cinsiyet olgusu* (Yayımlanmamış yüksek lisans tezi). Adnan Menderes Üniversitesi Sosyal Bilimler Enstitüsü, Aydın.

Lamb, M. E., & Roopnarine, J. E. (1979). Peer influences on sex-role development in preschoolers. *Journal of Child Development, 50*(4), 1219-1222.

Langlois, H. J., & Downs, A. C. (1980). Fathers, mothers and peers as socialization agents of sex-typed play behaviors in young children. *Jornal of Child Development, 51*, 1217-1247.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Lobue, V. DeLoache, S. J. (2011). Pretty in pink: The early development of gender-stereotyped colour preferences. *British Journal of Developmental Psychology, 29*, 656-667. doi: 10.1111/j.2044-835X.2011.02027.x

Milli Eğitim Bakanlığı (2013). *Gelişim alanları*. Ankara: MEB Yayınları.

Özdemir, E. (2006). *Okulöncesi dönem çocuklarının cinsiyet özelliklerine ilişkin kalıpyargılarının incelenmesi* (Yayımlanmamış yüksek lisans tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Özkan, B. (2009). *Okulöncesi dönem 5-6 yaş çocuklarının cinsiyet özelliklerine ilişkin kalıpyargılarının bazı değişkenler açısından incelenmesi* (Yayımlanmamış yüksek lisans tezi). Marmara Üniversitesi Eğitim Bilimleri Enstitüsü, İstanbul.

Sandal, M. (2015*). Sıralayıcı ölçme düzeyi için faktör analizi ve bir uygulama* (Yayımlanmamış yüksek lisans tezi). Osman Gazi Üniversitesi Fen Bilimleri Enstitüsü, Eskişehir.

Schermelleh-Engel, K., & Moosbrugger, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit- measures. *Methods of Psychological Research Online, 8* (2), 23-74.

Seçer, İ. (2015). *Psikolojik test geliştirme ve uyarlama süreci*. Ankara: Anı Yayıncılık.

Skelton, C., & Hall, E. (2001). *The development of gender roles in young children: A review of policy and literature*. Manchester: Equal Opportunities Commission

Şekercioğlu, G. (2009). *Çocuklar için benlik algısı profilinin uyarlanması ve faktör yapısının farklı değişkenlere göre eşitliğinin test edilmesi* (Yayımlanmamış doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Şimşek, Ö. F. (2007). *Yapısal eşitlik modellemesine giriş: Temel ilkeler ve LISREL uygulamaları.* Ankara: Ekinoks Yayıncılık.

Şirvanlı-Özen, D. (1992). *Annenin çalışma durumu ve ebeveynin benimsediği cinsiyet rolü değişkenlerinin çocuğun cinsiyet özelliklerine ilişkin kalıpyargılarının gelişimi üzerindeki rolleri* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Şıvgın, N. (2015). *Cinsiyet rolleri eğitim etkinliklerinin anasınıfına devam eden 60-72 aylık çocukların toplumsal cinsiyet kalıpyargılarına etkisinin incelenmesi* (Yayımlanmamış doktora tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Taylor, L. (2004). *Gender constancy and rigidity: A cross-sectional examination of early gender development* (Unpublished doctoral dissertation, New York University). Retrieved from https://search.proquest.com/docview/305167414

Tavşancıl, E. (2005). *Tutumların ölçülmesi ve SPSS ile veri analizi* (2. Baskı), Şehir: Nobel Yayın Dağıtım.

Tuzcuoğlu, N., & Tuzcuoğlu, S. (2004). *Çocuğun cinsel eğitimi, anne ben nasıl doğdum?* (2. Basım). İstanbul: Morpa Kültür Yayınları.

Ünlü, A. (2012). *Bazı değişkenlere göre okulöncesi çocuklarının cinsiyet rolü davranışlarının incelenmesi* (Yayımlanmamış yüksek lisans tezi). Selçuk Üniversitesi Sosyal Bilimler Enstitüsü, Konya.

Volbert, R. (2000). Sexual knowledge of preschool children. In T. G. M. Sandfort, & J. Rademakers (Eds.), *Childhood sexuality* (pp. 5-26). New York, NY: Hawort Press.

Williams, J.E., Bennett, M. S., & Deborah, L. B. (1975). Awareness and expression of sex stereo- types in young children. *Developmental Psychology, 11*(5), 635-642.

Yavuzer, H. (2000). *Ana-baba ve çocuk* (13. Basım). İstanbul: Remzi Kitabevi.

Yılmaz, M. (2011). Cinsel eğitimde kütüphanelerin rolü. *Türk Kütüphaneciliği Dergisi, 25*(1), 8-34.

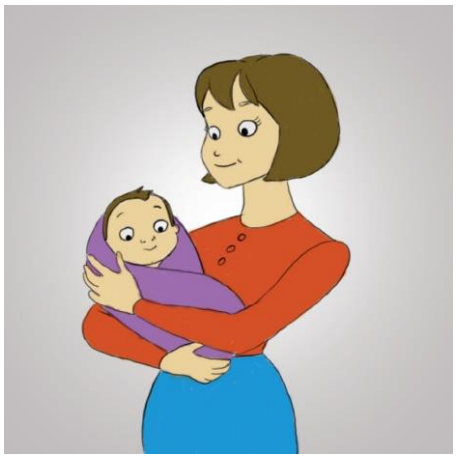
Yurdakul, R. S. (2012). *Çocuk ve cinsellik* (2. Basım). İstanbul: Kare yayınları.

Zembat, R., & Keleş, S. (2011). Erken çocuklukta toplumsal cinsiyet değişmezliği ölçeğinin türkçe formunun geçerlik ve güvenirlik çalışması. *Uluslararası İnsan Bilimleri Dergisi, 9*(1), 337-359.

Zucker, K. J., Bradley, S. J., Sullivan, C. B., Kuksis, M., Birkenfeld-Adams, A., & Mitchell, J. N. (1993). A gender identity interview for children. *Journal of Personality Assessment, 61*(3) 443-456. doi: 10.1207/s15327752jpa6103_2

Appendix A. SSDS Sexual Identity Sub-Scale Child Form Sample Item

Büyüdüğün zaman hangisi olacaksın? Anne mi, Baba mı?

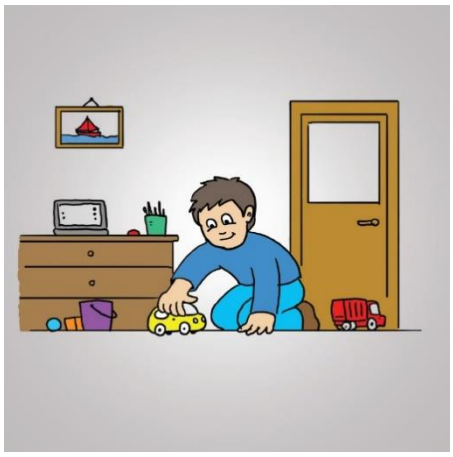


Appendix B. SDSS Gender Behavior Sub-Scale Child Form Sample Item (Girl)

Seda oyun oynarken banyonun kapısının açık olduğunu gördü. Annesi banyoda idi. Seda annesinin vücudunu merak ediyordu. Sence Seda oyununa mı devam eder yoksa annesinin nasıl yıkandığını mı izler?

Appendix C. SDSS Gender Behavior Sub-Scale Child Form Sample Item (Boy)

Can oyun oynarken banyonun kapısının açık olduğunu gördü. Babası banyoda idi. Can annesinin vücudunu merak ediyordu. Sence Can oyununa mı devam eder yoksa annesinin nasıl yıkandığını mı izler?



Appendix D. Sexual Identity and Gender Behavior Sub-Scale Family Form Sample Items

Maddeler	**Hiçbir Zaman**	**Nadiren**	**Bazen**	**Çoğu Zaman**	**Her Zaman**
1. Çıplak kişiye bakmaya çalışır.					
2. Ayna karşısında vücudunu inceler.					
22. Büyüdüğü zaman anne ya da baba olacağını bildiğini davranış ve konuşmaları ile gösterir (Evcilik oynarken baba/anne olma, geleceğe yönelik konuşmalarında anne/baba olacağı ile ilgili düşüncelerini söyleme, ilerde babası ya da annesi gibi olacağını söyleme.)					
23. Büyüdüğü zaman gelin ya da damat olacağını bildiğini davranışları ve konuşmaları ile gösterir (oyunlarında gelin/damat olma, geleceğe yönelik hayallerinde gelin/damat olacağını söyleme).					
24. Bıyık ve sakalın erkeğe özgü olduğunu konuşmalarında ve oyunlarında ifade eder.					

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 266-301

Building On-Demand Test Forms in R

Halil İbrahim SARI *

Abstract
Automated Test Assembly (ATA) plays important role in test development, especially in large scale test administrations. However, there is a lack of tutorials showing how to solve ATA problems. This tutorial aims to show to how build on-demand test forms easily for researchers and practitioners, and share the R codes for their use. The study presents the annotated R codes for thirty-nine unique examples. The examples include building one form, multiple forms and more complex ones under different constraint conditions across equal or different form lengths. All examples were solved by using "xxIRT" R package. The graphical depictions of the form-level information functions for all examples were also provided. Some important notes about the codes were also provided at the end of the paper in case one did not find a solution. The thirty-six examples were provided in the main body of the paper, the other three complex examples were given in the Supplementary material.

Key Words: On-demand test forms, automated test assembly, xxIRT, R.

INTRODUCTION

The ultimate goal of any test in the educational and psychological measurement is to estimate student's cognitive ability more accurately or precisely. However, it is quite difficult to reach this goal without a good measurement tool. This implies that the instrument or test form has to carry some certain psychometric characteristics to cover the construct of interest (e.g., math ability). It highlights the importance of the building the test forms that meet the desired features.

The topic of constructing the desired test forms is one of the most popular topics of all time. This is because regardless of the test administration type (fixed linear test, linear on the fly test, computerized multistage test or shadow test); for test security purposes, any test developer wants to build test forms that meet some certain requirements purposes, especially in large scale tests. Depending on the test administration type (e.g., linear or adaptive testing), one may want to build a single test form, two or more parallel forms that are at the same difficulty levels or multiple forms that are at the different difficulty levels. However, it is not very easy to ensure that the forms meet with both statistical (e.g., difficulty level) and non-statistical (e.g., content balancing and word count) specifications, especially when one wants to create many forms. Thus, instead of manually assembling forms, it is always better to use software to satisfy all constraints (e.g., test length, content balancing, difficulty level, word count etc.). This will help one to keep test form quality at the desired level.

Automated Test Assembly (ATA) is an integer programming approach used to solve equations that have complex constraints. In psychometrics, ATA is used to build test forms, and constraints refer to the desired test specifications. For instance, content balancing or distribution, difficulty level of form, number of test items in the form and total word count of items in a form can be thought as the constraint.

There are several integer programming software that are used to build test forms automatically. The most widely used ones are ILOG CPLEX ((International Business Machines-IBM, 2006), LINGO 12.0 (LINDO), CASTISEL (Luecht, 1998), LPSolve IDE (Berkelaar, Eikland, & Notebaert, 2004), the Premium Solver Platform 7.0 add-in for Microsoft Excel, and R packages "IpSolve" (Berkelaar et al., 2015) and "lpSolveAPI" (Konis, 2016). One can refer to Donoghue (2015) for a long list and detailed description.

* Assist. Prof., Kilis 7 Aralık University, Muallim Rıfat Faculty of Education, Kilis-Turkey, hisari87@gmail.com, ORCID ID: 0000-0001-7506-9000

To cite this article:
Sarı, H. İ. (2019). Building on-demand test forms in R. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 266-301. doi: 10.21031/epod.521330

Received: 02.02.2019
Accepted: 15.07.2019

Furthermore, there are some studies that illustrate building on-demand tests or solving ATA problems. The book written by Wim J. van der Linden (2006), "Linear Models for Optimum Test Design", details all aspects of constructing both criterion-referenced and norm-referenced test forms. It is probably the most comprehensive book written in this area. Cor, Alves and Gierl (2008, 2009) and Gierl, Daniels and Zhang (2017), in this journal, showed how to create parallel forms in Microsoft Excel. They vividly demonstrate all steps, and provided helpful screenshots. Han and Rudner (2014) showed how to build multiple parallel items with different techniques. Diao and van der Linden (2011) described how to solve complex ATA problems by using lpSolve R package in version 5.5. They presented three different ATA problems and showed how to solve them in R. Unfortunately, they provided the code for one of the problem cases only.

Purpose of the Study

The purpose of this tutorial is to show how to create on-demand test forms under different constraints for the researchers and practitioners so that they can use the codes according to their own cases. There is an abundant literature on the automated test assembly, however, a lack of R code tutorials available for the researchers. I provided the annotated R codes for thirty-nine unique examples. Due to space limit, I provided thirty-six of them in the manuscript. One can refer to the Appendix A for other three complex examples. The examples include building one form, multiple forms and more complex forms across the different conditions. I also provided test information functions for all examples.

METHOD

Item Response Theory (IRT) is integral part of ATA because IRT is used to determine the difficulty level of a form and to shape form level information function (e.g., test information function). Item parameter estimates such as difficulty, discrimination and pseudo-guessing are first computed or generated; then the best items that meet certain criteria are selected for a test form. When selecting the best items, item information function is used. This is vital because item information function allows us to see where on the theta scale an item provides the highest information or for whom an item is the best. The information function for item i (Ii(θ)) under the three parameter IRT model (Birnbaum, 1968) for an item is defined as

$$I_i(\theta)=a_i^2 \frac{(P_i(\theta)-c_i)^2}{1-c_i^2} \frac{Q_i(\theta)}{P_i(\theta)} \quad (1)$$

where *a*, *b* and *c* are the discrimination, difficulty and pseudo-guessing parameters for item *i*, *P(θ)* and *Q(θ)* are the probability of getting an item correct and incorrect for a person having θ as the ability score, respectively.

As shown in Luecht (1998), the test assembly finds a solution to maximize the Item Response Theory information function at a fixed theta point (i.e., Equation 1). Let denote θ_0 is the fixed theta point (or theta interval), and suppose we want a total of 20-item in the test. We first define a binary decision variable, x_i, (e.g., $x_i = 0$ means item *i* is not selected from the item bank, $x_i = 1$ means item *i* is selected from the item bank). The information function needs to be maximized;

$$I(\theta)=\sum_{i=1}^{N} I(\theta_0\ \xi i)x_i \quad (2)$$

where ξ_i represents the item parameters of item *i* (e.g., a, b, c parameters). Let's say one has two content areas (e.g., items measuring number properties and items measuring algebra denoted as C1 and C2, respectively), and wants to select ten items from each content area. The automated test assembly is modeled to maximize

$$I(\theta)=\sum_{i=1}^{N} I(\theta_0\ \xi i)x_i \quad (3)$$

subject to

$$\sum_{i\in C1}^{N} x_i \geq 10 \quad (4)$$

$$\sum_{i\in C2}^{N} x_i \geq 10 \quad (5)$$

$$\sum_{i=1}^{N} x_i \geq 20 \quad (6)$$

$$x_i \in (0,1),\ i=1,\ldots N \quad (7)$$

which put constraints on C1, C2, the total test length, and the range of decision variables, respectively. When the content balancing is not controlled, the constraints on the contents (Equations 4 and 5) can be removed from the model. When one wants to control other variables, he or she can add the additional constraints to the model.

It is important to note that the information function in Equation 1 can be maximized at a desired theta point (e.g., -1, 0 and 1 for easy, medium and hard test forms, respectively). Moreover, it can be maximized over a range of theta interval (e.g., -1 to 0 for an easy test form, and 0 to 1 for a hard test form). It is also possible to demand a user defined absolute amount of information either at a fixed theta point or over a range of theta interval (e.g., the amount of information is 8 at the theta point of 0 or at the interval of -1 to 1).

xxIRT PACKAGE

In this tutorial, I used "xxIRT" R package version 2.1.0 (Luo, 2018) to solve all given examples. The "xxIRT" R package is a recently released package, and uses "lpSolveAPI" package as the integer solver. The version of the package on CRAN has been recently updated. There are some core functions needed to be used when specifying and solving ATA problems. The most important function is *ata* which is used to create ATA problems. One needs to specify the information about item pool (either simulated or real), the number of test forms needs to be created, the length of form (e.g., 5 items), and maximum use of an item in the pool. The other two core functions are *ata_obj_relative* and *ata_obj_absolute*. The first is used when one wants to maximize the information at a fixed theta point or over a range of theta interval. When the theta interval is desired, the increments of theta points in the user defined theta interval can be specified. The latter is used when one wants to have an absolute amount of information for a test form. Similarly, when the target is to gather absolute information over a range of interval, the increment points can be specified by the user. The increment points are important because they may dramatically change the shape of test information function.

As discussed before, constraints are important elements of an ATA problem. The *ata_constraint* function adds the constraints to the ATA model. One can specify the number of items from each content area or total word count for a test form. Finally, *ata_solve* function solves the specified ATA problem. It is also possible to see the selected items and plot test information functions. One can refer to "xxIRT" package for more information about the codes and main functions. The annotated R codes different examples were given below.

The "xxIRT" package was primarily created to solve ATA problems for multistage testing (e.g., designing panels) so, there is no extended illustration of how to create simple or complex and single or multiple test forms. The current manual shows four simple examples but this provides shows many complex examples by using the same main functions.

Annonated Examples

I used a simulated item pool that consists of 1000 items and generated three hypothetical constraints as content area (e.g., algebra, numbers, equations), word count of each item (e.g., ranging from 30 to 150 for each item) and time required to solve the item (e.g., ranging from 100 seconds to 400 seconds for each item). I presented 12 ATA problems and for each problem, I showed how to solve the ATA by a) maximizing information function at a fixed theta point, b) maximizing information over a theta interval, and c) getting absolute amount of information for a form. For all 36 examples listed below, I always used the same simulated item pool. The number of items, distribution of item parameters, and hypothetical constraints are subject to change.

Preparing for the analysis

```
# Do not run!
#Install the "xxIRT" package first
install.packages("xxIRT",repos = "http://cran.us.r-project.org")
require("xxIRT")

# Let's generate an item pool
set.seed(10)
items=as.data.frame(cbind(
a=runif(1000, 0.5, 1.5), #a parameters from a uniform distribution. #Chan
ge accordingly!
b=runif(1000, -2, 2), #b parameters from a uniform distribution. Change #
accordingly!
c=runif(1000, 0, 0.20), #c parameters from a uniform distribution. Change
#accordingly!
content=sample(1:3,1000,replace = T), #3 content areas #(e.g., algebra, #
numbers, equations)
word_count=sample(30:150,1000,replace = T), #assigning random word counts
#for each item.
time=sample(100:400,1000,replace = T))) #assigning random time between #f
or each item.
#End
```

Problem 1: Building single forms without any constraint

Here, I created one single test form that does not have any constraints. There are three problems listed as Problem 1a, 1b and 1c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively.

```
#Problem1a: maximize the information at fixed theta point of -1
Problem1a <- ata(items, 1, #must be 1 when single form is built!
                 len=10, #test length. Change accordingly!
                 max_use=1) #Each item should be selected one time only!
Problem1a <- ata_obj_relative(Problem1a,
                              -1, # fixed theta point where we want to #
maximize the information.
                              "max")
Problem1a <- ata_solve(Problem1a, as.list=T) #Now we are ready to solve #
the ATA
plot(Problem1a) #plotting information function
#End

#Problem1b: Maximize the information at the theta interval of -1 to 1.
Problem1b <- ata(items, 1, len=10, max_use=1)
Problem1b <-  ata_obj_relative(Problem1b, seq(-1, 1, #change when a #diff
erent interval is desired
                                              0.10),  #increment of the #
theta points.
                              "max", flatten=0.10) #change accordingly!
Problem1b <- ata_solve(Problem1b, as.list=T) #Now we are ready to solve #
the ATA
plot(Problem1b) # plotting information function
```

```
# End

# Problem1c: absolute information target
theta_target=c(-1.0, -0.5 ,0, 0.5,  1.0) # target theta points (from -1 #
to 1)
tif_target= 8 # desired amount of information at the test level. Change #
accordingly!
Problem1c <- ata(items, 1, len=10, max_use=1)
Problem1c <- ata_obj_absolute(Problem1c, theta_target, tif_target) #ATA #
problem
Problem1c <- ata_solve(Problem1c, as.list=T) #Now we are ready to solve #
the ATA.
plot(Problem1c)  # plotting information function
# End
```

Problem 2: Building single form with content constraint only

Here, I created one single test form that has content constraints. There are three problems listed as Problem 2a, 2b and 2c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 2 examples, I pulled 10 items as 2, 3, and 5 items from the content1, content2 and content3, respectively.

```
# Problem2a: maximize the information at fixed theta point of -1.
Problem2a <- ata(items, 1, #single test form. This must be 1 when single
#form is created!
                 len=10, #test length of 10. Change accordingly!
                 max_use=1) #Each item should be selected one time only!
Problem2a <-  ata_obj_relative(Problem2a, -1, "max") #specifying ATA #pro
blem
#Now let's add content distribution constraints before solving the ATA.
Problem2a <-  ata_constraint(Problem2a, "content", min=2, max=2, level=1)
# 2 items from Content 1
Problem2a <-  ata_constraint(Problem2a,"content", min=3, max=3, level=2)
# 3 items from Content 2
Problem2a <-  ata_constraint(Problem2a, "content", min=5, max=5, level=3)
# 5 items from Content 3
Problem2a <- ata_solve(Problem2a, as.list=T) #Now, ATA is ready to solve!
plot(Problem2a) # plotting information function
# End

# Problem2b: maximize the information at theta interval of -1 to 1.
Problem2b <- ata(items, 1, len=10, max_use=1) # Test length of 10. Change
#accordingly!
Problem2b <-  ata_obj_relative(Problem2b, seq(-1, 1, 0.10), "max", flatte
n=0.10)
#Now let's add content constraints before solve the ATA.
Problem2b <- ata_constraint(Problem2b, "content", min=2, max=2, level=1)
# 2 items from C 1
Problem2b <- ata_constraint(Problem2b,"content", min=3, max=3, level=2)
# 3 items from C 2
Problem2b <- ata_constraint(Problem2b, "content", min=5, max=5, level=3)
# 5 items from C 3
Problem2b <- ata_solve(Problem2b, as.list=T) #Now, ATA is ready to solve!
```

```
plot(Problem2b) # plotting information function
# End

#Problem2c: absolute information target
theta_target=c(-1.0, -0.5 ,0, 0.5,  1.0) # target theta points where you
#want to maximize information.
tif_target= 8 # desired amount of information we want. Change #accordingl
y!
Problem2c <- ata(items, 1, len=10, max_use=1) # Test length of 10. Change
#accordingly!
Problem2c <- ata_obj_absolute(Problem2c, theta_target, tif_target) #speci
fying ATA problem
#Now let's add content constraints before solve the ATA.
Problem2c <-  ata_constraint(Problem2c, "content", min=2, max=2, level=1)
# 2 items from C 1
Problem2c <-  ata_constraint(Problem2c,"content", min=3, max=3, level=2)
# 3 items from C 2
Problem2c <-  ata_constraint(Problem2c, "content", min=5, max=5, level=3)
# 5 items from C 3
Problem2c <- ata_solve(Problem2c, as.list=T) #Now, ATA is ready to solve!
Problem2c$items #see selected items
plot(Problem2c) # plotting information function
# End
```

Problem 3: Building single form with two constraints

Here, I created one single test form that has content and word count constraints. There are three problems listed as Problem 3a, 3b and 3c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 3 examples, I pulled 10 items as 2, 3, and 5 items from the content1, content2 and content3, respectively. The average word count of the items in the forms is between 60 and 70.

```
#Problem3a: maximize the information at the fixed theta point of -1.
Problem3a <- ata(items, 1, # Building single form. Change accordingly!
                 len=10, #Test length of 10. Change accordingly!
                 max_use=1) # Each item should be selected one time only!
Problem3a <- ata_obj_relative(Problem3a, -1, "max") #specifying ATA #prob
lem
#Now let's add content constraints before solve the ATA.  Change #accordi
ngly!
Problem3a <- ata_constraint(Problem3a, "content", min=2, max=2, level=1)
# 2 items from C 1
Problem3a <- ata_constraint(Problem3a,"content", min=3, max=3, level=2) #
3 items from C 2
Problem3a <- ata_constraint(Problem3a, "content", min=5, max=5, level=3)
# 5 items from C 3
#Now let's add word count constraint before solve the ATA. Change #accord
ingly!
Problem3a <-  ata_constraint(Problem3a, "word_count", min=60*10, max=70*1
0)
Problem3a <- ata_solve(Problem3a, as.list=T) #Now, ATA is ready to solve!
Problem3a$items #see selected items
plot(Problem3a) # plotting information function
```

```
# End

# Problem3b:  maximize the information at theta interval of -1 to 1.
Problem3b <- ata(items, 1, len=10, max_use=1)
Problem3b <-  ata_obj_relative(Problem3b, seq(-1, 1, 0.10),  "max", flatt
en=0.10)
#Now let's add content constraints before solve the ATA.  Change #accordi
ngly!
Problem3b <-  ata_constraint(Problem3b, "content", min=2, max=2, level=1)
# 2 items from Content 1
Problem3b <-  ata_constraint(Problem3b,"content", min=3, max=3, level=2)
# 3 items from Content 2
Problem3b <-  ata_constraint(Problem3b, "content", min=5, max=5, level=3)
# 5 items from Content 3
#Now let's add word count constraint before solve the ATA.  Change #accor
dingly!
Problem3b <-  ata_constraint(Problem3b, "word_count", min=60*10, max=70*1
0)
Problem3b <- ata_solve(Problem3b, as.list=T) #Now, ATA is ready to solve!
Problem3b$items #see selected items
plot(Problem3b) # plotting information function
# End

# Problem3c: absolute information target
theta_target=c(-1.0, -0.5 ,0, 0.5,  1.0) # target theta points where you
#want to maximize
#information. One can also specify fixed theta point! Change accordingly!
tif_target= 8 # desired amount of information. Change accordingly!
Problem3c <- ata(items, 1, len=10, max_use=1)
Problem3c <- ata_obj_absolute(Problem3c, theta_target, tif_target) #speci
fying ATA problem
#Now let's add content constraints before solve the ATA.  Change #accordi
ngly!
Problem3c <-  ata_constraint(Problem3c, "content", min=2, max=2, level=1)
# 2 items from C 1
Problem3c <-  ata_constraint(Problem3c,"content", min=3, max=3, level=2)
# 3 items from C 2
Problem3c <-  ata_constraint(Problem3c, "content", min=5, max=5, level=3)
# 5 items from C 3
#Now let's add word count constraint before solve the ATA.  Change #accor
dingly!
Problem3c <-  ata_constraint(Problem3c, "word_count", min=60*10, max=70*1
0)
Problem3c <- ata_solve(Problem3c, as.list=T) #Now, ATA is ready to solve!
plot(Problem3c) # plotting information function
# End
```

Problem 4: Building single form with three constraints

Here, I created one single test form that has content, word count and time constraints. There are three problems listed as Problem 4a, 4b and 4c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 4 examples,

I pulled 10 items as 2, 3, and 5 items from the content1, content2 and content3, respectively. The average word count of the items in the forms is between 60 and 70. The average time to solve an item is between 200 and 300 seconds.

```
#Problem4a: maximize the information at the fixed theta point of -1.
Problem4a <- ata(items, 1, # Building single form. Change accordingly!
                 len=10, #Test length of 10. Change accordingly!
                 max_use=1) # Each item should be selected one time only!
Problem4a <-  ata_obj_relative(Problem4a, -1, "max")
#Now let's add content constraints before solve the ATA.  Change #accordi
ngly!
Problem4a <-  ata_constraint(Problem4a, "content", min=2, max=2, level=1)
# 2 items from C 1
Problem4a <-  ata_constraint(Problem4a,"content", min=3, max=3, level=2)
# 3 items from C 2
Problem4a <-  ata_constraint(Problem4a, "content", min=5, max=5, level=3)
# 5 items from C 3
#Now let's add word count constraint before solve the ATA.  Change #accor
dingly!
Problem4a <-  ata_constraint(Problem4a, "word_count", min=60*10, max=70*1
0)
#Now let's add time constraint before solve the ATA.  Change accordingly!
Problem4a <-  ata_constraint(Problem4a, "time", min=200*10, max=300*10)
Problem4a <- ata_solve(Problem4a, as.list=T) #Now, ATA is ready to solve!
Problem4a$items #see selected items
plot(Problem4a) # plotting information function
# End

# Problem4b: maximize the information at theta interval of -1 to 1.
Problem4b <- ata(items, 1, len=10, max_use=1) #A total of 10 items. #Chan
ge accordingly!
Problem4b <-  ata_obj_relative(Problem4b, seq(-1, 1, 0.10), "max", flatte
n=0.10)
#Now let's add content constraints before solve the ATA.  Change #accordi
ngly!
Problem4b <-  ata_constraint(Problem4b, "content", min=2, max=2, level=1)
# 2 items from C 1
Problem4b <-  ata_constraint(Problem4b,"content", min=3, max=3, level=2)
# 3 items from C 2
Problem4b <-  ata_constraint(Problem4b, "content", min=5, max=5, level=3)
# 5 items from C 3
#Now let's add word count constraints before solve the ATA.  Change #acco
rdingly!
Problem4b <-  ata_constraint(Problem4b, "word_count", min=60*10, max=70*1
0)
#Now let's add time constraint before solve the ATA.  Change accordingly!
Problem4b <-  ata_constraint(Problem4b, "time", min=200*10, max=300*10)
Problem4b <- ata_solve(Problem4b, as.list=T) #Now, ATA is ready to solve!
Problem4b$items #see selected items
plot(Problem4b) # plotting information function
# End

# Problem4c:  absolute information target
theta_target=c(-1.0, -0.5 ,0, 0.5,  1.0) # target theta points where you
```

```
#to maximize #information. One can also specify fixed theta point! Change
#accordingly!
tif_target= 8 # desired amount of information. Change accordingly!
Problem4c <- ata(items, 1, len=10, max_use=1) #A total of 10 items. #Chan
ge accordingly!
Problem4c <- ata_obj_absolute(Problem4c, theta_target, tif_target) #speci
fying ATA problem
#Now let's add content constraints before solve the ATA.  Change #accordi
ngly!
Problem4c <-  ata_constraint(Problem4c, "content", min=2, max=2, level=1)
#2 items from C 1
Problem4c <-  ata_constraint(Problem4c,"content", min=3, max=3, level=2)
#3 items from C 2
Problem4c <-  ata_constraint(Problem4c, "content", min=5, max=5, level=3)
#5 items from C 3
#Now let's add word count constraints before solve the ATA.  Change #acco
rdingly!
Problem4c <-  ata_constraint(Problem4c, "word_count", min=60*10, max=70*1
0)
#Now let's add time constraints before solve the ATA.  Change #accordingl
y!
Problem4c <-  ata_constraint(Problem4c, "time", min=200*10, max=300*10)
Problem4c <- ata_solve(Problem4c, as.list=T) #Now, ATA is ready to solve!
Problem4c$items #see selected items
plot(Problem4c) # plotting information function
#End
```

Problem 5: Building two equal-length forms with no constraints

Here, I created two test forms that have any constraints. There are three problems listed as Problem 5a, 5b and 5c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 5 examples, I pulled equal test lengths (10 items) and content was not controlled.

```
# Problem5a: maximize the information at the fixed theta point of 0 for #
all forms
Problem5a <- ata(items, 2, #Building 2 forms. Change accordingly!
                 len=10, # Test length of 10. Change accordingly!
                 max_use=1) #We don't want item overlapping. Change #acco
rdingly!
Problem5a <-  ata_obj_relative(Problem5a, 0, # change when a different #f
ixed theta point is desired
                               "max")
Problem5a <- ata_solve(Problem5a, as.list=T) # "as.list=F" gives all #sel
ected items together
Problem5a$items #see selected items
plot(Problem5a) # plotting information function
#End

# Problem5b: maximize the information at theta interval of -1 to 1.
Problem5b <- ata(items, 2, len=10, max_use=1)
Problem5b <-  ata_obj_relative(Problem5b, seq(-1, 1, 0.50), # change #acc
ordingly!
```

```
                                "max", flatten=0.10) # change accordingly!
Problem5b <- ata_solve(Problem5b,as.list=T) #Now let's solve the ATA!
Problem5b$items #see selected items
plot(Problem5b) # plotting information function
#End

# Problem5c: absolute information target
theta_target=c(-1.0, -0.5 ,0, 0.5,  1.0) # target theta points. Change #a
ccordingly!
tif_target= 8 # desired amount of information. Change accordingly!
Problem5c <- ata(items, 2, len=10, max_use=1)
Problem5c <- ata_obj_absolute(Problem5c, theta_target, tif_target) # #Spe
cifying the ATA.
Problem5c <- ata_solve(Problem5c, as.list=T) # Let's solve the ATA #probl
em.
Problem5c$items #see selected items
plot(Problem5c) # plotting information function
# End
```

Problem 6: Building unequal-length two forms with no constraints

Here, I created two test forms that have any constraints same in problem 5a, 5b and 5c. However, in problem 6 examples, the test lengths are not equal. There are three problems listed as Problem 6a, 6b and 6c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 6 examples, I pulled 5 items for form 1 and 8 items for form 2 but content distribution was not controlled for both forms.

```
# Problem6a:  maximize the information at the fixed theta point of 0 for
#all forms
Problem6a <- ata(items, 2, # don't specify form length "len=10" because #
of unequal test lengths
                 max_use=1) #we do not want overlapping items.
Problem6a <-  ata_obj_relative(Problem6a, 0, #fixed theta point for both
#forms. Change accordingly!
                                "max")
Problem6a <- ata_constraint(Problem6a,1, min=5, max=5, forms=1) #5 items
#in form 1
Problem6a <- ata_constraint(Problem6a,1,  min=8, max=8, forms=2) #8 items
#in form 2
Problem6a <- ata_solve(Problem6a, as.list=T) # Let's solve the ATA #probl
em.
Problem6a$items #see selected items
plot(Problem6a) # plotting information function
#End

# Problem6b: maximize the information at theta interval of -1 to 1.
Problem6b <- ata(items, 2,  # two forms
                 max_use=1) #we do not want overlapping items. Change #ac
cordingly!
Problem6b <-  ata_obj_relative(Problem6b, seq(-1, 1, 0.50),  #Change the
interval accordingly!
                                "max", flatten=0.10)
Problem6b <- ata_constraint(Problem6b,1, min=5, max=5, forms=1) #5 items
```

```
#in form 1
Problem6b <- ata_constraint(Problem6b,1,  min=8, max=8, forms=2) #8 items
#in form 2
Problem6b <- ata_solve(Problem6b,as.list=T) #Now Let's solve the ATA
Problem6b$items #see selected items
plot(Problem6b) # plotting information function
#End

# Problem6c: absolute information target
theta_target=0 #the theta point where we want to maximize the #informatio
n. Change accordingly!
tif_target= 5 # desired amount of information. Change accordingly!
Problem6c <- ata(items, 2, max_use=1)
Problem6c <- ata_constraint(Problem6c,1, min=5, max=5, forms=1) #5 items
#in form 1
Problem6c <- ata_constraint(Problem6c,1,  min=6, max=6, forms=2) #8 items
#in form 2
Problem6c <- ata_obj_absolute(Problem6c, theta_target, tif_target, forms
= 1) # ATA for form 1
Problem6c <- ata_obj_absolute(Problem6c, theta_target, tif_target, forms
= 2) # ATA for form 2
Problem6c <- ata_solve(Problem6c, as.list=T) #Now Let's solve the ATA
plot(Problem6c) # plotting information function
#End
```

Problem 7: Building equal-length two forms with content constraint

Here, I created two test forms with controlling content distribution. There are three problems listed as Problem 7a, 7b and 7c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all problem 7 examples, for both forms, the test length is 10, and I pulled 2, 3 and 5 items content2 and content3, respectively.

```
# Problem7a: Maximize the information at the fixed theta point of 0 for #
all forms.
Problem7a <- ata(items, 2, len=10, #Equal test length of 0 for both #form
s. Change accordingly!
                 max_use=1) #I want non-overlapping forms.
Problem7a <- ata_obj_relative(Problem7a, 0, # fixed theta point. Change #
accordingly!
                              "max")
Problem7a <- ata_constraint(Problem7a, "content", min=2, max=2, level=1)
#2 items from C 1
Problem7a <- ata_constraint(Problem7a,"content", min=3, max=3, level=2) #
3 items from C 2
Problem7a <- ata_constraint(Problem7a, "content", min=5, max=5, level=3)
#5 items from C 3
Problem7a <- ata_solve(Problem7a, as.list=T) #Now Let's solve the ATA
plot(Problem7a) # plotting information function
#End

# Problem7b: Maximize the information at theta interval of -1 to 1.
Problem7b <- ata(items, 2, len=10, max_use=1)
Problem7b <-  ata_obj_relative(Problem7b, seq(-1, 1, 0.50), #Change the #
```

```
interval accordingly!
                                "max", flatten=0.50)
Problem7b <- ata_constraint(Problem7b, "content", min=2, max=2, level=1)
#2 items from content 1
Problem7b <- ata_constraint(Problem7b,"content", min=3, max=3, level=2) #
3 items from content 2
Problem7b <- ata_constraint(Problem7b, "content", min=5, max=5, level=3)
#5 items from content 3
Problem7b <- ata_solve(Problem7b,as.list=T) #Now Let's solve the ATA
Problem7b$items #see selected items
plot(Problem7b) # plotting information function
#End

# Problem7c: absolute information target
theta_target=c(-1.0, -0.5 ,0, 0.5,  1.0)
tif_target= 8 # desired amount of information
Problem7c <- ata(items, 2, len=10, max_use=1)
Problem7c <- ata_constraint(Problem7c, "content", min=2, max=2, level=1)
#2 items from C 1
Problem7c <- ata_constraint(Problem7c,"content", min=3, max=3, level=2) #
3 items from C 2
Problem7c <- ata_constraint(Problem7c, "content", min=5, max=5, level=3)
#5 items from C 3
Problem7c <- ata_obj_absolute(Problem7c, theta_target, tif_target) # ATA
#for both forms
Problem7c <- ata_solve(Problem7c, as.list=T) #Now Let's solve the ATA
Problem7c$items #see selected items
plot(Problem7c) # plotting information function
#End
```

Problem 8: Building unequal-length two forms with content constraint

Here, I created two unequal-length test forms with controlling content distribution. There are three problems listed as Problem 8a, 8b and 8c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 8 examples in below, for form 1, I pulled 5 items as 1, 2 and 2 from content 1, 2 and 3, respectively. For form 2, I pulled 8 items as 2, 3 and 3 from content 1, 2 and 3, respectively.

```
# Problem8a: Maximize the information at the fixed theta point
Problem8a <- ata(items, 2, #two test forms
                 max_use=1) #I don't want item overlapping across the #fo
rms
Problem8a <-  ata_obj_relative(Problem8a, 0, "max")
#Now let's specify total test lengths for both forms.
Problem8a <- ata_constraint(Problem8a,1, min=5, max=5, forms=1) #5 items
#in form 1
Problem8a <- ata_constraint(Problem8a,1,  min=8, max=8, forms=2) #8 items
#in form 2
#Now let's add content constraints for form 1.  Change accordingly!
Problem8a <- ata_constraint(Problem8a, "content", min=1, max=1, level=1,f
orms = 1)
Problem8a <- ata_constraint(Problem8a,"content", min=2, max=2, level=2, f
orms = 1)
```

```
Problem8a <- ata_constraint(Problem8a, "content", min=2, max=2, level=3,f
orms = 1)
#Now let's add content constraints for form 2.  Change accordingly!
Problem8a <- ata_constraint(Problem8a, "content", min=2, max=2, level=1,f
orms = 2)
Problem8a <- ata_constraint(Problem8a,"content", min=3, max=3, level=2, f
orms = 2)
Problem8a <- ata_constraint(Problem8a, "content", min=3, max=3, level=3,f
orms = 2)
Problem8a <- ata_solve(Problem8a, as.list=T) # ATA is ready to solve
Problem8a$items #see selected items
plot(Problem8a) # plotting information function
#End

# Problem8b: maximize the information at theta interval of -1 to 1.
Problem8b <- ata(items, 2, max_use=1)
Problem8b <-  ata_obj_relative(Problem8b, seq(-1, 1, 0.50), #theta #inter
val. Change accordingly!
                               "max", flatten=0.10)
Problem8b <- ata_constraint(Problem8b,1, min=5, max=5, forms=1) #5 items
#in form 1
Problem8b <- ata_constraint(Problem8b,1,  min=8, max=8, forms=2) #8 items
#in form 2
#Now let's add content constraints for form 1.  Change accordingly!
Problem8b <- ata_constraint(Problem8b, "content", min=1, max=1, level=1,f
orms = 1)
Problem8b <- ata_constraint(Problem8b,"content", min=2, max=2, level=2, f
orms = 1)
Problem8b <- ata_constraint(Problem8b, "content", min=2, max=2, level=3,f
orms = 1)
#Now let's add content constraints for form 2.  Change accordingly!
Problem8b <- ata_constraint(Problem8b, "content", min=2, max=2, level=1,f
orms = 2)
Problem8b <- ata_constraint(Problem8b,"content", min=3, max=3, level=2, f
orms = 2)
Problem8b <- ata_constraint(Problem8b, "content", min=3, max=3, level=3,f
orms = 2)
Problem8b <- ata_solve(Problem8b,as.list=T) # ATA is ready to solve
Problem8b$items #see selected items
plot(Problem8b) # plotting information function
#End

# Problem8c:  absolute information target
theta_target=c(-1.0, -0.5 ,0, 0.5,  1.0) #theta interval I want to #maxim
ize the information
tif_target= 5 # desired amount of information. Change accordingly!
Problem8c <- ata(items, 2, max_use=1)
Problem8c <- ata_constraint(Problem8c,1, min=5, max=5, forms=1) #5 items
#in form 1
Problem8c <- ata_constraint(Problem8c,1,  min=8, max=8, forms=2) #8 items
#in form 2
#Now let's add content constraints for form 1.  Change accordingly!
Problem8c <- ata_constraint(Problem8c, "content", min=1, max=1, level=1,f
```

```
orms = 1)
Problem8c <- ata_constraint(Problem8c,"content", min=2, max=2, level=2, f
orms = 1)
Problem8c <- ata_constraint(Problem8c, "content", min=2, max=2, level=3,f
orms = 1)
#Now let's add content constraints for form 2.  Change accordingly!
Problem8c <- ata_constraint(Problem8c, "content", min=2, max=2, level=1,f
orms = 2)
Problem8c <- ata_constraint(Problem8c,"content", min=3, max=3, level=2, f
orms = 2)
Problem8c <- ata_constraint(Problem8c, "content", min=3, max=3, level=3,f
orms = 2)
Problem8c <- ata_obj_absolute(Problem8c, theta_target, tif_target, forms
= 1) #ATA for form 1
Problem8c <- ata_obj_absolute(Problem8c, theta_target, tif_target, forms
= 2) #ATA for form 2
Problem8c <- ata_solve(Problem8c, as.list=T) #Now ATA is ready to solve
Problem8c$items #see selected items
plot(Problem8c) # plotting information function
#End
```

Problem 9: Building equal-length two forms with two constraints

Here, I create two equal-length test forms with controlling content distribution and word count. There are three problems listed as Problem 9a, 9b and 9c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 9 examples, I pulled 10 items for both forms, and in both forms, there are 2, 3, and 5 items from the content1, content2 and content3, respectively. The average word count of the items in both forms is between 60 and 70.

```
# Problem9a: maximize the information at the fixed theta point of 0 for #
both forms.
Problem9a <- ata(items, 2, # number  of forms
        len=10, # Equal test length of 10 for both forms. Change #accordi
ngly!
max_use=1) #I don't want item overlapping across the forms.
Problem9a <- ata_obj_relative(Problem9a, 0, # fixed theta point. Change #
accordingly!
                        "max")
#Now let's add content constraints for the forms. Change accordingly!
Problem9a <- ata_constraint(Problem9a, "content", min=2, max=2, level=1)
#2 items from content 1
Problem9a <- ata_constraint(Problem9a,"content", min=3, max=3, level=2) #
3 items from content 2
Problem9a <- ata_constraint(Problem9a, "content", min=5, max=5, level=3)
#5 items from content 3
#Now let's add word count constraints before solve the ATA.  Change #acco
rdingly!
Problem9a <-  ata_constraint(Problem9a, "word_count", min=60*10, max=70*1
0)
Problem9a <- ata_solve(Problem9a, as.list=T) #Let's solve the ATA
Problem9a$items #see selected items
plot(Problem9a) # plotting information function
```

```
#End

# Problem9b: maximize the information at theta interval of -1 to 1.
Problem9b <- ata(items, 2, len=10, max_use=1)
Problem9b <-  ata_obj_relative(Problem9b, seq(-1, 1, 0.50), #theta #inter
val. Change accordingly!
                           "max", flatten=0.50)
#Now let's add content constraints for the forms. Change accordingly!
Problem9b <- ata_constraint(Problem9b, "content", min=2, max=2, level=1)
#2 items from content 1
Problem9b <- ata_constraint(Problem9b,"content", min=3, max=3, level=2) #
3 items from content 2
Problem9b <- ata_constraint(Problem9b, "content", min=5, max=5, level=3)
#5 items from content 3
#Now let's add word count constraints before solve the ATA.  Change #acco
rdingly!
Problem9b <-  ata_constraint(Problem9b, "word_count", min=60*10, max=70*1
0)
Problem9b <- ata_solve(Problem9b,as.list=T) #Let's solve the ATA
Problem9b$items #see selected items
plot(Problem9b) # plotting information function
#End

# Problem9c: Absolute information target.
theta_target=0 #the point where we want the absolute information. One can
#specify interval as well!
tif_target= 8 # desired amount of information. Change accordingly!
Problem9c <- ata(items, 2, len=10, max_use=1)
#Now let's add content constraints for the forms. Change accordingly!
Problem9c <- ata_constraint(Problem9c, "content", min=2, max=2, level=1)
#2 items from content 1
Problem9c <- ata_constraint(Problem9c,"content", min=3, max=3, level=2) #
3 items from content 2
Problem9c <- ata_constraint(Problem9c, "content", min=5, max=5, level=3)
#5 items from content 3
#Now let's add word count constraints before solve the ATA.  Change #acco
rdingly!
Problem9c <-  ata_constraint(Problem9c, "word_count", min=60*10, max=70*1
0)
Problem9c <- ata_obj_absolute(Problem9c, theta_target, tif_target) #Speci
fy the ATA
Problem9c <- ata_solve(Problem9c, as.list=T) #Let's solve the ATA
Problem9c$items #see selected items
plot(Problem9c) # plotting information function
#End
```

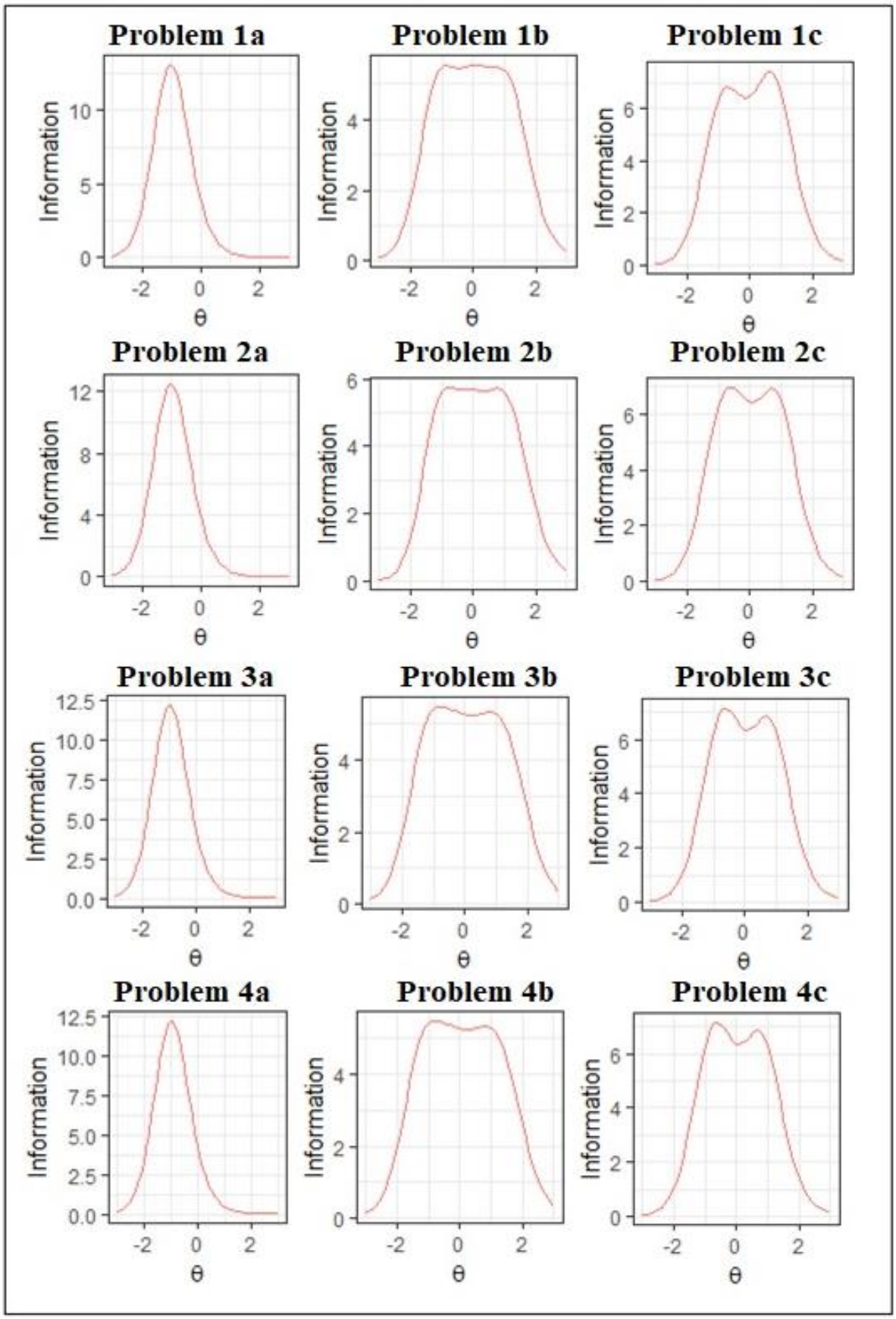


Figure 1. Plots for The Solutions in Examples from 1a to 4c.

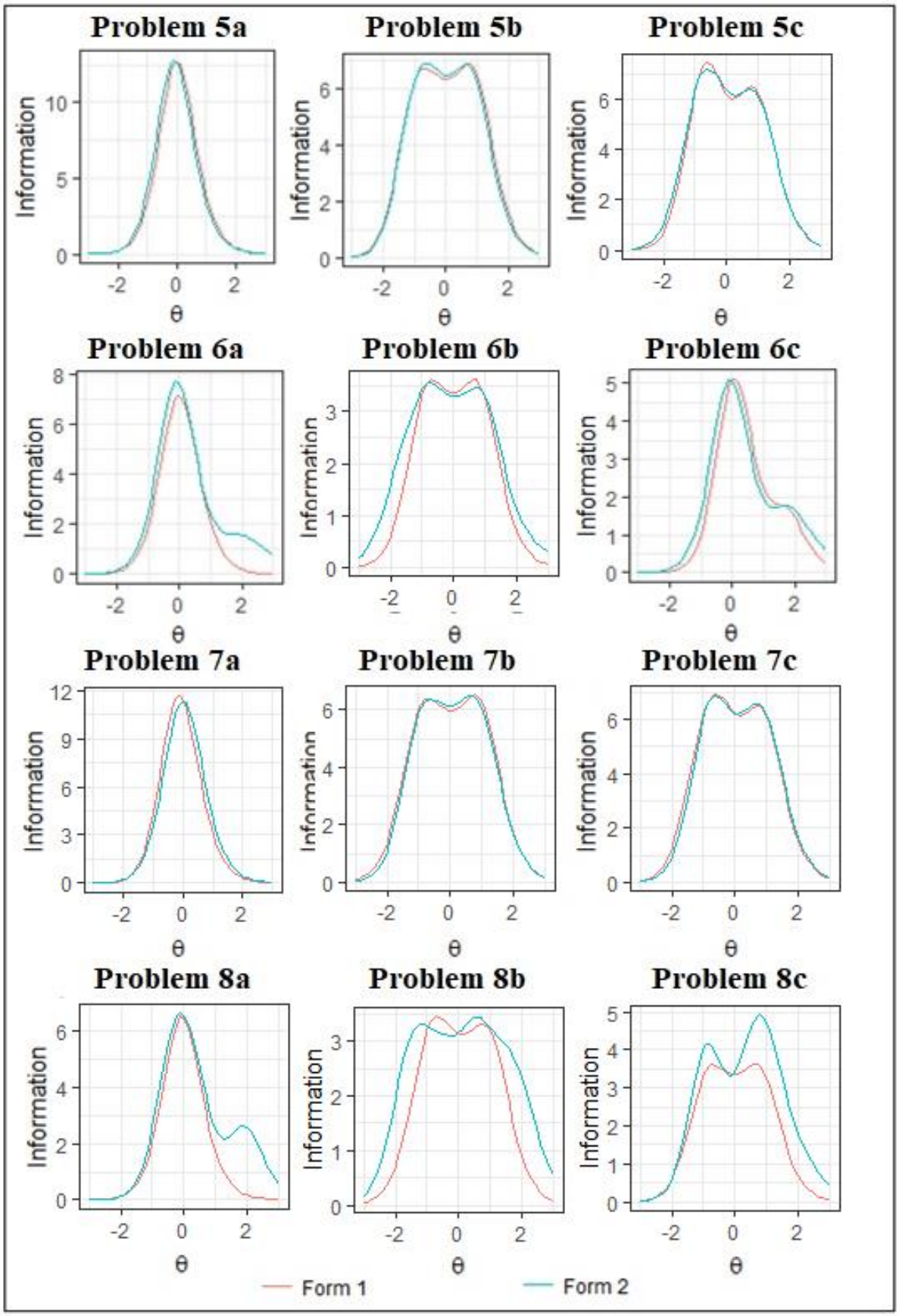


Figure 2. Plots for The Solutions in Examples from 5a to 8c.

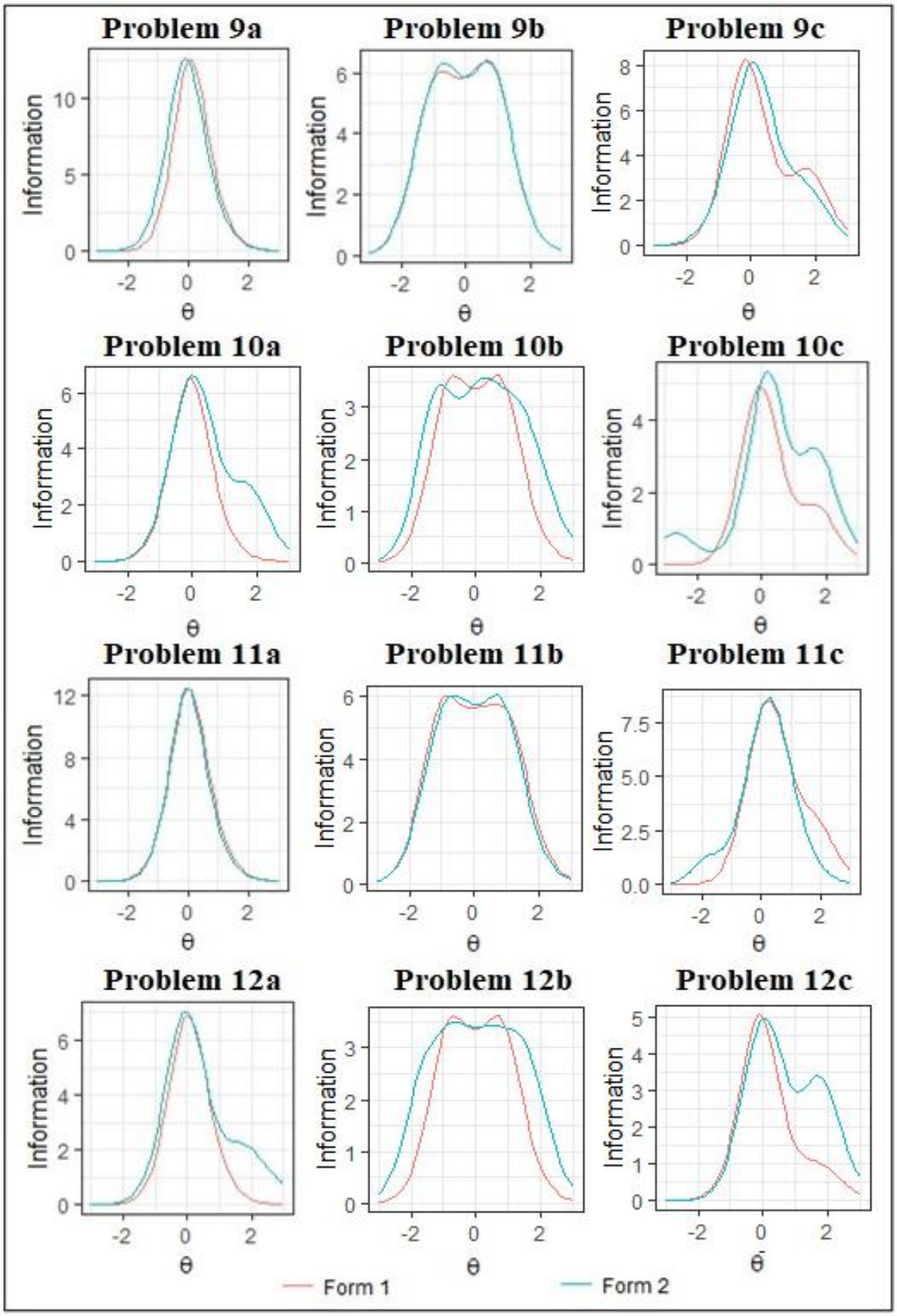


Figure 3. Plots for The Solutions in Examples from 9a to 12c.

Problem 10: Building unequal-length two forms with two constraints

Here, I create two unequal-length test forms with controlling content distribution and word count. There are three problems listed as Problem 10a, 10b and 10c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 10 examples, I pulled 5 items for form 1 and 8 items for form 2. In form 1, there are 1, 2, and 2 items from the content1, content2 and content3, respectively. In form 2, there are 2, 3, and 3 items from the content1, content2 and content3, respectively. The average word count of the items in both forms is between 30 and 80.

```
# Problem10a: maximize the information at the fixed theta point
Problem10a <- ata(items, 2, #Two test forms. Change accordingly!
max_use=1) #we don't want item-overlapping
Problem10a <-  ata_obj_relative(Problem10a, 0, #fixed theta point. Change
#accordingly!
"max")
Problem10a <- ata_constraint(Problem10a,1, min=5, max=5, forms=1) #5 #ite
ms in form 1
Problem10a <- ata_constraint(Problem10a,1,  min=8, max=8, forms=2) #8 #it
ems in form 2
#Now let's add content constraints for form 1. Change accordingly!
Problem10a <- ata_constraint(Problem10a, "content", min=1, max=1, level=1
,forms = 1)
Problem10a <- ata_constraint(Problem10a,"content", min=2, max=2, level=2,
forms = 1)
Problem10a <- ata_constraint(Problem10a, "content", min=2, max=2, level=3
,forms = 1)
#Now let's add content constraints for form 2. Change accordingly!
Problem10a <- ata_constraint(Problem10a, "content", min=2, max=2, level=1
,forms = 2)
Problem10a <- ata_constraint(Problem10a,"content", min=3, max=3, level=2,
forms = 2)
Problem10a <- ata_constraint(Problem10a, "content", min=3, max=3, level=3
,forms = 2)
#Now let's add word count constraints before solve the ATA. Change #accor
dingly!
Problem10a <-  ata_constraint(Problem10a, "word_count", min=30*10, max=80
*10)
Problem10a <- ata_solve(Problem10a, as.list=T) # Now let's solve the ATA
Problem10a$items #see selected items
plot(Problem10a) # plotting information function
#End

# Problem10b: maximize the information at theta interval of -1 to 1.
Problem10b <- ata(items, 2, max_use=1)
Problem10b <-  ata_obj_relative(Problem10b, seq(-1, 1, 0.50), #theta #int
erval. Change accordingly!
"max", flatten=0.10)
Problem10b <- ata_constraint(Problem10b,1, min=5, max=5, forms=1) #5 #ite
ms in form 1
Problem10b <- ata_constraint(Problem10b,1,  min=8, max=8, forms=2) #8 #it
ems in form 2
#Now let's add content constraints for form 1. Change accordingly!
Problem10b <- ata_constraint(Problem10b, "content", min=1, max=1, level=1
```

```
,forms = 1)
Problem10b <- ata_constraint(Problem10b,"content", min=2, max=2, level=2,
forms = 1)
Problem10b <- ata_constraint(Problem10b, "content", min=2, max=2, level=3
,forms = 1)
#Now let's add content constraints for form 2. Change accordingly!
Problem10b <- ata_constraint(Problem10b, "content", min=2, max=2, level=1
,forms = 2)
Problem10b <- ata_constraint(Problem10b,"content", min=3, max=3, level=2,
forms = 2)
Problem10b <- ata_constraint(Problem10b, "content", min=3, max=3, level=3
,forms = 2)
#Now let's add word count constraints before solve the ATA. Change #accor
dingly!
Problem10b <-  ata_constraint(Problem10b, "word_count", min=30*10, max=80
*10)
Problem10b <- ata_solve(Problem10b,as.list=T) #Solve the ATA
Problem10b$items #see selected items
plot(Problem10b) # plotting information function
#End

# Problem10c: absolute information target
theta_target=0 #One can specify theta interval as well. Change #according
ly!
tif_target= 5 # desired amount of information
Problem10c <- ata(items, 2, max_use=1)
Problem10c <- ata_constraint(Problem10c,1, min=5, max=5, forms=1) #5 #ite
ms in form 1
Problem10c <- ata_constraint(Problem10c,1,  min=8, max=8, forms=2) #8 #it
ems in form 2
#Now let's add content constraints for form 1. Change accordingly!
Problem10c <- ata_constraint(Problem10c, "content", min=1, max=1, level=1
,forms = 1)
Problem10c <- ata_constraint(Problem10c,"content", min=2, max=2, level=2,
forms = 1)
Problem10c <- ata_constraint(Problem10c, "content", min=2, max=2, level=3
,forms = 1)
#Now let's add content constraints for form 2. Change accordingly!
Problem10c <- ata_constraint(Problem10c, "content", min=2, max=2, level=1
,forms = 2)
Problem10c <- ata_constraint(Problem10c,"content", min=3, max=3, level=2,
forms = 2)
Problem10c <- ata_constraint(Problem10c, "content", min=3, max=3, level=3
,forms = 2)
#Now let's add word count constraints before solve the ATA.  Change #acco
rdingly!
Problem10c <-  ata_constraint(Problem10c, "word_count", min=30*10, max=80
*10)
Problem10c <- ata_obj_absolute(Problem10c, theta_target, tif_target, form
s = 1) # ATA for form 1
Problem10c <- ata_obj_absolute(Problem10c, theta_target, tif_target, form
s = 2) # ATA for form 2
Problem10c <- ata_solve(Problem10c, as.list=T) #Solve the ATA
```

```
Problem10c$items #see selected items
plot(Problem10c) # plotting information function
#End
```

Problem 11: Building equal-length two forms with three constraints

Here, I create two equal-length test forms with controlling content distribution, word count and time. There are three problems listed as Problem 11a, 11b and 11c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 11 examples, I pulled 10 items for both forms, and in both forms, there are 2, 3, and 5 items from the content1, content2 and content3, respectively. The average word count of the items in both forms is between 60 and 70. The average time to solve the item is between 200 and 300 seconds.

```
#Problem11a: maximize the information at the fixed theta point
Problem11a <- ata(items, 2, #we are building two forms
len=10, #total test length for a form. Change accordingly!
max_use=1) #Each item can be selected for a form only. Change #accordingl
y!
Problem11a <- ata_obj_relative(Problem11a, 0, # fixed theta point.
                          "max")
#Now let's add content constraints for the forms. Change accordingly!
Problem11a <- ata_constraint(Problem11a, "content", min=2, max=2, level=1
)
Problem11a <- ata_constraint(Problem11a,"content", min=3, max=3, level=2)
Problem11a <- ata_constraint(Problem11a, "content", min=5, max=5, level=3
)
#Now let's add word count constraints before solve the ATA. Change #accor
dingly!
Problem11a <-  ata_constraint(Problem11a, "word_count", min=60*10, max=70
*10)
#Now let's add time constraints before solve the ATA. Change #accordingly
!
Problem11a <-  ata_constraint(Problem11a, "time", min=200*10, max=300*10)
Problem11a <- ata_solve(Problem11a, as.list=T) #Let's solve the ATA
Problem11a$items #see selected items
plot(Problem11a) # plotting information function
#End

# Problem11b: Maximize the information at theta interval of -1 to 1.
Problem11b <- ata(items, 2, len=10, max_use=1)
Problem11b <-  ata_obj_relative(Problem11b, seq(-1, 1, 0.50), #Change #ac
cordingly!
                          "max", flatten=0.50)
#Let's add content distribution constraints. Change accordingly!
Problem11b <- ata_constraint(Problem11b, "content", min=2, max=2, level=1
)
Problem11b <- ata_constraint(Problem11b,"content", min=3, max=3, level=2)
Problem11b <- ata_constraint(Problem11b, "content", min=5, max=5, level=3
)
#Let's add word count constraints. Change accordingly!
Problem11b <-  ata_constraint(Problem11b, "word_count", min=60*10, max=70
*10)
#Let's add time constraints. Change accordingly!
```

```
Problem11b <-  ata_constraint(Problem11b, "time", min=200*10, max=300*10)
Problem11b <- ata_solve(Problem11b,as.list=T) #Now let's solve the ATA
Problem11b$items #see selected items
plot(Problem11b) # plotting information function
#End

# Problem11c:  Absolute information target
theta_target=c(0, 0.5) # either specify fixed theta point or theta #inter
val
tif_target= 8 # desired amount of information. Change accordingly!
Problem11c <- ata(items, 2, len=10, max_use=1)
#Let's add content distribution constraints. Change accordingly!
Problem11c <- ata_constraint(Problem11c, "content", min=2, max=2, level=1
)
Problem11c <- ata_constraint(Problem11c,"content", min=3, max=3, level=2)
Problem11c <- ata_constraint(Problem11c, "content", min=5, max=5, level=3
)
#Let's add word count constraints. Change accordingly!
Problem11c <-  ata_constraint(Problem11c, "word_count", min=60*10, max=70
*10)
#Let's add time constraints. Change accordingly!
Problem11c <-  ata_constraint(Problem11c, "time", min=200*10, max=300*10)
Problem11c <- ata_obj_absolute(Problem11c, theta_target, tif_target) #Spe
cify the ATA problem
Problem11c <- ata_solve(Problem11c, as.list=T) #Let's solve the ATA
Problem11c$items #see selected items
plot(Problem11c) # plotting information function
#End
```

Problem 12: Building unequal-length two forms with three constraints

Here, I create two unequal-length test forms with controlling content distribution, word count and time. There are three problems listed as Problem 12a, 12b and 12c. The codes for these problems are written for a fixed theta point, over a range and absolute amount of information cases, respectively. In all Problem 12 examples, I pulled 5 items for form 1 and 8 items for form 2. In form 1, there are 1, 2, and 2 items from the content1, content2 and content3, respectively. In form 2, there are 2, 3, and 3 items from the content1, content2 and content3, respectively. The average word count of the items in both forms is between 30 and 80. The average time to solve the item is between 100 and 400 seconds.

```
#Problem12a : Maximize the information at the fixed theta point of 0.
Problem12a <- ata(items, 2, #Building two forms
max_use=1) #we don't want item overlapping. Change accordingly!
Problem12a <-  ata_obj_relative(Problem12a, 0, # Change accordingly!
"max")
#Let's specify test lengths for the two forms. Change accordingly!
Problem12a <- ata_constraint(Problem12a,1, min=5, max=5, forms=1) #5 #ite
ms in form 1
Problem12a <- ata_constraint(Problem12a,1,  min=8, max=8, forms=2) #8 #it
ems in form 2
#Let's add content distribution constraints for form 1. Change #according
ly!
Problem12a <- ata_constraint(Problem12a, "content", min=1, max=1, level=1
,forms = 1)
```

```
Problem12a <- ata_constraint(Problem12a,"content", min=2, max=2, level=2,
forms = 1)
Problem12a <- ata_constraint(Problem12a, "content", min=2, max=2, level=3
,forms = 1)
#Let's add content distribution constraints for form 2. Change #according
ly!
Problem12a <- ata_constraint(Problem12a, "content", min=2, max=2, level=1
,forms = 2)
Problem12a <- ata_constraint(Problem12a,"content", min=3, max=3, level=2,
forms = 2)
Problem12a <- ata_constraint(Problem12a, "content", min=3, max=3, level=3
,forms = 2)
#Let's add word count constraints. Change accordingly!
Problem12a <-  ata_constraint(Problem12a, "word_count", min=30*10, max=80
*10)
#Let's add time constraints. Change accordingly!
Problem12a <-  ata_constraint(Problem12a, "time", min=100*10, max=400*10)
Problem12a <- ata_solve(Problem12a, as.list=T) #Now, let's solve the ATA
Problem12a$items #see selected items
plot(Problem12a) # plotting information function
#End

#Problem12b: maximize the information at theta interval of -1 to 1.
Problem12b <- ata(items, 2, max_use=1)
Problem12b <-  ata_obj_relative(Problem12b, seq(-1, 1, 0.50),  # theta #i
nterval. Change accordingly!
"max", flatten=0.10)
Problem12b <- ata_constraint(Problem12b,1, min=5, max=5, forms=1) #5 #ite
ms in form 1
Problem12b <- ata_constraint(Problem12b,1,  min=8, max=8, forms=2) #8 #it
ems in form 2
Problem12b <- ata_constraint(Problem12b, "content", min=1, max=1, level=1
,forms = 1) #Form1C1
Problem12b <- ata_constraint(Problem12b,"content", min=2, max=2, level=2,
forms = 1) #Form1C2
Problem12b <- ata_constraint(Problem12b, "content", min=2, max=2, level=3
,forms = 1) #Form1C3
Problem12b <- ata_constraint(Problem12b, "content", min=2, max=2, level=1
,forms = 2) #Form2C1
Problem12b <- ata_constraint(Problem12b,"content", min=3, max=3, level=2,
forms = 2) #Form2C2
Problem12b <- ata_constraint(Problem12b, "content", min=3, max=3, level=3
,forms = 2) #Form2C3
Problem12b <-  ata_constraint(Problem12b, "word_count", min=30*10, max=80
*10) #word counts
Problem12b <-  ata_constraint(Problem12b, "time", min=100*10, max=400*10)
#time constraints
Problem12b <- ata_solve(Problem12b,as.list=T) #Now, let's solve the ATA
Problem12b$items #see selected items
plot(Problem12b) # plotting information function
#End
```

```
#Problem12c: Absolute information target
theta_target=0 #the theta point where you want the absolute amount of #in
formation
tif_target= 5 # desired amount of information. Change accordingly!
Problem12c <- ata(items, 2, max_use=1)
Problem12c <- ata_constraint(Problem12c,1, min=5, max=5, forms=1) #5 #ite
ms in form 1
Problem12c <- ata_constraint(Problem12c,1,  min=8, max=8, forms=2) #8 #it
ems in form 2
Problem12c <- ata_constraint(Problem12c, "content", min=1, max=1, level=1
,forms = 1) #Form1 C1
Problem12c <- ata_constraint(Problem12c,"content", min=2, max=2, level=2,
forms = 1) #Form1 C2
Problem12c <- ata_constraint(Problem12c, "content", min=2, max=2, level=3
,forms = 1) #Form1 C3
Problem12c <- ata_constraint(Problem12c, "content", min=2, max=2, level=1
,forms = 2) #Form2 C1
Problem12c <- ata_constraint(Problem12c,"content", min=3, max=3, level=2,
forms = 2) #Form2 C2
Problem12c <- ata_constraint(Problem12c, "content", min=3, max=3, level=3
,forms = 2) #Form2 C3
Problem12c <-  ata_constraint(Problem12c, "word_count", min=30*10, max=80
*10) #word counts
Problem12c <-  ata_constraint(Problem12c, "time", min=100*10, max=400*10)
#time constraints
Problem12c <- ata_obj_absolute(Problem12c, theta_target, tif_target, form
s = 1) # ATA for form 1
Problem12c <- ata_obj_absolute(Problem12c, theta_target, tif_target, form
s = 2) # ATA for form 2
Problem12c <- ata_solve(Problem12c, as.list=T) #Now, ATA is ready to #sol
ve!
Problem12c$items #see selected items
plot(Problem12c) # plotting information function
#End of the tutorial
```

Important Notes

1. It is important to note that finding a solution in any example depends on the psychometric characteristics of the items in the pool.
2. In this paper, I used a simulated item pool. Thus, when you replicate the item pool, you may or may not find the same solutions.
3. The item pool was generated based on the 3PL Item Response Theory Model. In case you use different model than the 3PL, you should change the item parameters accordingly. For example, when you use Rasch model, you should fix all discrimination parameters at 1 and all pseudo-guessing parameters at 0.
4. All of the codes were carefully written, and their functionality was checked again and again by the author. In case you have problems to run any example, you can try the following steps first. If you still cannot find a solution, please do not hesitate contacting the author.
    a. If you use simulated item pool, you may want to re-generate the item pool, and try again.

b. In case you do not find a solution for your own cases, you can try relaxing the constraints. For example, you can specify lower amount of absolute information (see Problems 1c, 2c, 3c and 4c) or maximize the information in a narrower theta interval (see Problems 1b, 2b, 3b and 4b).

c. In same cases, you may want to allow item overlapping, especially when you have limited number of total items.

5. Finding a solution also depends on the constraints you specify. The likelihood of finding a solution becomes difficult as you use strict restrictions.

6. For the demonstration purposes, the constraints used in this study are the hypothetical constraints (e.g., content area, word count and time). You can use your own constraints or more logical ones.

REFERENCES

Berkelaar, M., & others (2015). Package 'lpSolve'. Retrieved from https://cran.r-project.org/web/packages/lpSolve/lpSolve.pdf

Berkelaar, M., Eikland, K., & Notebaert, P. (2004). lp_solve reference guide menu. Retrieved from http://lpsolve.sourceforge.net/5.5/

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Cor, K., Alves, C., & Gierl, M. (2008). Conducting automated test assembly using the Premium Solver Platform Version 7.0 with Microsoft Excel and the large-Scale LP/QP solver engine add-in. *Applied Psychological Measurement*, *32*, 652-663. doi: 10.1177/0146621608316603

Cor, K., Alves, C., & Gierl, M. (2009). Three applications of automated test assembly within a user-friendly modeling environment. *Practical Assessment, Research, & Evaluation*, *14*(14), 1-23. Retrieved from http://pareonline.net/getvn.asp?v=14%26n=14

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5), 398-409. doi: 10.1177/0146621610392211

Donoghue, J. R. (2015). *Comparison of integer programming (IP) solvers for automated test assembly (ATA)* (Research Report No. ETS RR-15-05). Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/ets2.12051

Gierl, M. J., Daniels, L., & Zhang, X. (2017). Creating parallel forms to support on-demand testing for undergraduate students in psychology. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 288-302. doi: 10.21031/epod.305350

Han, K. T., & Rudner, L. M. (2014). Item *pool construction using mixed integer quadratic programming* (MIQP) (Research Report No. RR-14-01). Retrieved from https://files.eric.ed.gov/fulltext/ED558455.pdf

International Business Machines (2006). *ILOG CPLEX 10.0 user's manual*. Paris: ILOG SA.

Konis, K. (2016). Package 'lpSolveAPI'. Retrieved from https://cran.r-project.org/web/packages/lpSolveAPI/lpSolveAPI.pdf

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*(3), 224-236. doi: 10.1177/01466216980223003

Luo, X. (2018). Package 'xxIRT'. Retrieved from https://cran.r-project.org/web/packages/xxIRT/xxIRT.pdf

Van der Linden, W. J. (2006). *Linear models for optimal test design*. New York, NY: Springer Science & Business Media.

Test Formları Oluşturma Üzerine Öğretici R Çalışması

Giriş

Eğitimde ve psikolojide ölçmenin temel amaçlarından biri öğrenci yeteneğini veya bilgisini en doğru veya en az hata ile ölçmektir. Ancak iyi bir ölçme aracı olmadan bunu başarabilmek oldukça zordur. Geliştirdiğimiz test formu veya ölçme aracının zorluk derecesi, içerik alanı, maddelerin kelime sayısı gibi psikometrik özelliklerinin önceden belirlenmesi gerekir.

İstenilen özelliklerde test formu oluşturmak özellikle bilgisayar ortamında bireye uyarlanmış testler, bireye uyarlanmış çok aşamalı testler, kâğıt-kalem testleri için büyük öneme sahiptir. Bireye uyarlanmış testlerde aynı maddeleri çok fazla kişinin almasını engellemek için, birden fazla test formu oluşturmak bir gerekliliktir. Ancak oluşturulan test formlarının birbirine birçok açıdan benzer olması bir zorunluluktur. Örneğin birbirine paralel formlar oluşturulmak istendiğinde test formaların zorluk derecesi (kolay, zor form gibi), maddelerin konu alanı (doğal sayılar, kümeler, fonksiyonlar gibi), formlardaki maddelerin uzunlukları (kelime sayısı) birbiriyle aynı veya benzer olmalıdır. Bunu sağlama işlemine *otomatik test derleme* adı verilmektedir.

Otomatik Test Derleme (OTD) çok bilinmeyenli karmaşık denklemleri veya kısıtlamaları çözmek için kullanılan tamsayı (integer) programlama metodudur. OTD psikometride istenilen kriterlere sahip test formu oluşturmada kullanılmaktadır. İstenilen özellik veya kısıtlama ile kastedilen örneğin test formunun zorluk derecesi, maddelerin içerik veya konu alanları, maddelerin uzunluları olabilir.

Literatürde otomatik test derleme ile alakalı birçok faydalı kaynak bulunmaktadır. Wim J. Van der Linden (2005) tarafından yazılan kitap bu alandaki en faydalı kaynakların başında gelmektedir. Yazar kitabında eşik-dayanaklı ve norm dayanaklı test formlarının nasıl oluşturulduğunu detaylıca anlatmaktadır. Cor, Alves ve Gierl (2008, 2009) ve Gierl, Daniels ve Zhang (2017) Microsoft Excel'de isteğe bağlı testlerin nasıl hazırlanabileceğini göstermişlerdir. Diao ve van der Linden (2011) karmaşık otomatik test derlemenin R'da nasıl yapılabileceğini anlatmışlardır. Ancak yazarlar yanlızca üç farklı problem üzerinde durmuş ve sadece bir problem durumuna ait R kodunu okuyucuyla paylaşmışlardır.

Otomatik test derleme yapmak için kullanılabilecek çok sayıda bilgisayar programı bulunmaktadır. Bunlardan bazıları ILOG CPLEX, LINGO 12.0, LPSolve IDE, "IpSolve ve "lpSolveAPI" R paketleri.

Çalışmanın amacı

Bu çalışmanın amacı, araştırmacılar ve uygulayıcılar için farklı kısıtlamalar altında isteğe bağlı test formlarını nasıl oluşturabileceklerini göstermektir. Otomatik test derleme hakkında literatürde bol miktarda çalışma olmasına rağmen araştırmacıların kullanabileceği ücretsiz R kodları sınırlı miktarda bulunmaktadır. Çalışmada birbirinden farklı otuz dokuz farklı problem için açıklamalı R kodu verilmiştir. Kelime sınırı nedeniyle otuz altı tanesi bu belgede verilmiştir. Geriye kalan üç problem daha karmaşık durumlar altında otomatik test derlemenin nasıl çözülebileceğini göstermekte olup, Ek A'da verilmiştir. Örnekler, istenilen koşullar altında tek bir test formu, çoklu test formları ve daha karmaşık test formları oluşturmayı içermektedir. Açıklamalı R kodlarının yanı sıra her bir problem için form bilgi fonksiyonu da verilmiştir.

"xxIRT" R Paketi

Bu çalışmada, verilen tüm örnekleri çözmek için "xxIRT" R paketi versiyon 2.1.0 (Luo, 2018) kullanılmıştır. Bu paket yeni yayımlanmış olup, OTD problemlerini çözmek için "lpSolveAPI" R paketini kullanmaktadır. Paket içerisinde bulunan ve kullanıcı tarafından bilinmesi gereken önemli fonksiyonlardan bazıları *ata, ata_obj_relative, ata_obj_absolute, ata_constraint, ata_solve* fonksiyonlarıdır.

Fonksiyon *ata* kaç tane test formu oluşturulacağı (örneğin iki), formun uzunluğu (örneğin 5 madde) ve madde havuzu olarak neyi kullanacağı bilgilerini kullanarak OTD için problemi tanımlar. Diğer iki temel fonksiyondan *ata_obj_relative* fonksiyonu test bilgi fonksiyonunu sabit bir yetenek seviyesi noktasında maksimum hale getirirken, *ata_obj_absolute* ise test bilgi fonksiyonunu bir yetenek seviyesi aralığında maksimum hale getirir. Test bilgi fonksiyonunun verilen bir yetenek seviyesi aralığında maksimum hale getirilmesi istendiğinde, kullanıcı tarafından tanımlanan yetenek aralığının yanı sıra, yetenek seviyesi artış miktarları da belirlenebilir. Herhangi bir yetenek seviyesinde veya yetenek aralığında mutlak test bilgi fonksiyonuna sahip bir test formu oluşturulmak istendiğinde de istendiğinde *ata_obj_absolute* fonksiyonu kullanılır. Bu durumda da artış noktaları kullanıcı tarafından belirlenebilir. Yetenek seviyesi için sabit bir nokta değil de aralık belirlenmesi durumunda artış noktaları önemlidir, çünkü bu test bilgisi fonksiyonunun şeklini önemli ölçüde değiştirebilir.

Daha önce tartışıldığı gibi, kısıtlamalar bir OTD probleminin önemli unsurlarıdır. Kısıtlamaları oluşturmak için kullanılacak fonksiyon *ata_constraint* fonksiyonudur. Bu fonksiyon belirtilen kısıtlamayı OTD probleminin modeline ekler. Formdaki konu alanlarından kaçar tane madde seçileceği, formdaki maddelerin toplam sayılarının hangi aralıkta olacağı gibi birçok kısıtlama bu fonksiyon kullanılarak modele eklenebilir. Tüm kısıtlamalar da girildikten sonra OTD modeli çözmek için *ata_solve* fonksiyonu kullanılır. Bu fonksiyon kullanılıp, test formları için uygun maddeler seçildikten sonra hangi maddelerin seçildiği ve oluşturulan formların bilgi fonksiyonlarının grafikleri de çizilebilir. Temel fonksiyonlar veya komutlar hakkında daha fazla bilgi için "xxIRT" paketine başvurabilir. Bu fonksiyonlarının kullanımı gösteren açıklamalı R kodları aşağıda vermiştir.

Örnek R Kodları

Çalışmada verilen problem durumlarını çözmek için öncelikli olarak 1000 maddeden oluşan 3 parametreli madde tepki kuramına göre madde havuzu oluşturulmuştur. Bununla birlikte her bir madde için rastgele a) içerik alanı (örneğin cebirsel ifadeler, sayılar, denklemler gibi), b) kelime sayısı (örneğin, her bir madde için 30 ile 150 arasında değişen) ve c) maddeyi çözmek için gerekli zaman (örneğin, her bir madde için 100 saniye ile 400 saniye arasında) atanmıştır. Rastgele atanan bu değerler problem durumuna bağlı olarak çözülmesi istenen otomatik test derleme işleminde kullanılmak içindir.

Çalışmanın ana metninde 12 farklı OTD problemi sunulmuş olup, her bir problem 3 farklı durum altında çözülmüştür. Her bir problemin a şıkkı sabit bir yetenek seviyesi noktasında bilgi fonksiyonu maksimum hale çıkarılmak istendiğinde, b şıkkı test bilgi fonksiyonu yetenek seviyesi istenilen yetenek seviyesi aralığında maksimum hale getirilmek istendiğinde, c şıkkı ise istenilen miktarda test bilgisi elde edilmek istendiğinde OTD'nin nasıl çözüleceğini göstermektedir. Çalışmada listelenen toplamda 36 örnek için, simülasyonla üretilmiş aynı madde havuzu kullanılmıştır. Madde sayısı, madde parametrelerinin dağılımı ve varsayımsal kısıtlamalar kullanıcı tarafından değiştirilebilir. Aşağıda her bir problem durumunda çözülen OTD problemindeki test formlarının özellikleri verilmiştir.

Problem 1: Herhangi bir kısıtlama olmaksızın tek bir test formunun nasıl oluşturulacağını göstermektedir. Formda 10 madde yer almaktadır.

Problem 2: İçerik ağırlıklandırılması göz önünde bulundurularak tek bir test formunun nasıl oluşturulacağını göstermektedir. Test formu için 3 farklı içerikten sırasıyla 2, 3 ve 5 madde olmak üzere 10 madde seçilmiştir.

Problem 3: İçerik ağırlıklandırma ve kelime sayısı göz önünde bulundurularak iki farklı kısıtlamayla tek bir test formunun nasıl oluşturulabileceğini göstermektedir. Test formu için 3 farklı içerikten sırasıyla 2, 3 ve 5 madde olmak üzere 10 madde seçilmiştir. Formdaki maddelerin ortalama kelime sayısının 60 ile 70 arasında olması istenmiştir.

Problem 4: İçerik ağırlıklandırma, kelime sayısı ve zaman göz önünde bulundurularak üç farklı kısıtlamayla tek bir test formunun nasıl oluşturulabileceğini göstermektedir. Test formu için 3 farklı içerikten sırasıyla 2, 3 ve 5 madde olmak üzere 10 madde seçilmiştir. Formdaki maddelerin ortalama

kelime sayısının 60 ile 70 arasında olması istenmiştir. Maddeleri çözmek için gereken ortalama zaman 200 saniye ile 300 saniye arasında olacak şekilde ayarlanmıştır.

Problem 5: Herhangi bir kısıtlama olmaksızın eşit test uzunluğuna sahip iki test formunun nasıl oluşturulacağını göstermektedir. Her iki formda da 10 madde yer almaktadır.

Problem 6: Herhangi bir kısıtlama olmaksızın farklı test uzunluğuna sahip iki test formunun nasıl oluşturulacağını göstermektedir. Birinci form için 5, ikinci form için 8 madde seçilmiştir.

Problem 7: İçerik ağırlıklandırılması göz önünde bulundurularak eşit test uzunluğuna sahip iki test formunun nasıl oluşturulacağını göstermektedir. Her iki test formu için 3 farklı içerikten sırasıyla 2, 3 ve 5 madde olmak üzere 10 madde seçilmiştir.

Problem 8: İçerik ağırlıklandırılması göz önünde bulundurularak farklı test uzunluğuna sahip iki test formunun nasıl oluşturulacağını göstermektedir. Birinci test formu için üç farklı içerik alanından sırasıyla 1, 2 ve 2 olmak üzere toplam 5 madde, ikinci test formu için üç farklı içerik alanından sırasıyla 2, 3 ve 3 olmak üzere toplam 8 madde seçilmiştir.

Problem 9: İçerik ağırlıklandırma ve kelime sayısı göz önünde bulundurularak iki farklı kısıtlamayla eşit test uzunluğuna sahip iki test formunun nasıl oluşturulabileceğini göstermektedir. Her iki test formu için 3 farklı içerikten sırasıyla 2, 3 ve 5 madde olmak üzere 10 madde seçilmiştir. Formlardaki maddelerin ortalama kelime sayısının 60 ile 70 arasında olması istenmiştir.

Problem 10: İçerik ağırlıklandırma ve kelime sayısı göz önünde bulundurularak iki farklı kısıtlamayla farklı test uzunluğuna sahip iki test formunun nasıl oluşturulabileceğini göstermektedir. Birinci test formu için üç farklı içerik alanından sırasıyla 1, 2 ve 2 olmak üzere toplam 5 madde, ikinci test formu için üç farklı içerik alanından sırasıyla 2, 3 ve 3 olmak üzere toplam 8 madde seçilmiştir. Formlardaki maddelerin ortalama kelime sayısının 30 ile 80 arasında olması istenmiştir.

Problem 11: İçerik ağırlıklandırma, kelime sayısı ve zaman göz önünde bulundurularak üç farklı kısıtlamayla eşit test uzunluğuna sahip iki test formunun nasıl oluşturulabileceğini göstermektedir. Her iki test formu için 3 farklı içerikten sırasıyla 2, 3 ve 5 madde olmak üzere 10 madde seçilmiştir. Formlardaki maddelerin ortalama kelime sayısının 60 ile 70 arasında olması istenmiştir. Maddeleri çözmek için gereken ortalama zaman 200 saniye ile 300 saniye arasında olacak şekilde ayarlanmıştır.

Problem 12: İçerik ağırlıklandırma, kelime sayısı ve zaman göz önünde bulundurularak üç farklı kısıtlamayla farklı test uzunluğuna sahip iki test formunun nasıl oluşturulabileceğini göstermektedir. Birinci form için 3 farklı içerikten sırasıyla 1, 2 ve 2 olmak üzere toplam 5 madde, ikinci form için 3 farklı içerikten sırasıyla 2, 3 ve 3 olmak üzere toplam 8 madde seçilmiştir. Formlardaki maddelerin ortalama kelime sayısının 30 ile 80 arasında olması istenmiştir. Maddeleri çözmek için gereken ortalama zaman 100 saniye ile 400 saniye arasında olacak şekilde ayarlanmıştır.

Önemli Notlar

1. Herhangi bir Otomatik Test derleme probleminin çözümü madde havuzundaki maddelerin kalitesine bağlıdır.
2. Bu çalışmada simülasyonla üretilmiş madde havuzu kullanılmıştır. Dolayısıyla bir başkası kodları çalıştırdığında aynı sonuçlara ulaşamayabilir veya daha iyi sonuçlar elde edebilir.
3. Bu çalışmada test bilgi fonksiyonlarının hesaplanması için 3 parametreli Madde Tepki Kuramı Modeli kullanılmıştır. Bir başkası isteğe göre farklı modeller kullanabilir.
4. Çalışmada herhangi bir problemin çözülememesi durumunda aşağıdaki yöntemler denenebilir. Hâlâ problem yaşanması durumunda yazar ile irtibata geçmekten çekinmeyiniz.
    a. Simülasyonla üretilmiş madde havuzu kullanılmışsa, madde havuzu tekrardan üretilebilir.

b. Bazı kısıtlamalar gevşetilebilir. Örneğin yetenek seviyesi aralığı daraltılabilir veya mutlak bir bilgi seviyesi miktarı istenildiğinde, istenilen bilgi miktarı azaltılabilir.

c. Bazı durumlarda formlardaki ortak madde olmasına izin verilebilir.

5. Unutulmamalıdır ki herhangi bir OTD problemini çözmek belirtilen kısıtlamalara bağlıdır. Kısıtlamaların zorluğu veya miktarı arttıkça OTD probleminin çözülme imkânı azalır.
6. Bu çalışmada gösterim amacıyla içerik ağırlıklandırma, kelime sayısı ve zaman olmak üzere varsayımsal kısıtlamalar kullanılmıştır. Kullanıcılar kendi durumlarına göre farklı kısıtlamalar kullanabilirler.

Appendix A. Three Complex Examples

Preparing for the analysis

```
# Do not run!
#Install the "xxIRT" package first
install.packages("xxIRT",repos = "http://cran.us.r-project.org")
require("xxIRT")

# Let's generate an item pool
set.seed(10)
items=as.data.frame(cbind(
a=runif(1000, 0.5, 1.5), #a parameters from a uniform distribution. #Chan
ge
#accordingly!
b=runif(1000, -2, 2), #b parameters from a uniform distribution. Change
#accordingly!
c=runif(1000, 0, 0.20), #c parameters from a uniform distribution. Change
#accordingly!
content=sample(1:3,1000,replace = T), #3 content areas (e.g., algebra,num
bers#equations)
word_count=sample(30:150,1000,replace = T), #assigning random word counts
for #each item.
time=sample(100:400,1000,replace = T))) #assigning random time between fo
r #e#ach item.
#End

# DO NOT RUN
#BUILDING MORE COMPLEX TEST FORMS

#Example 1: maximize the information at the different fixed theta points

#Pulling five sets of item (5 test forms)
#For Form 1 and 2 maximize the information at the fixed theta point of -1
(two easy forms)
#For Form 3 maximize the information at the fixed theta point of 0 (1 med
ium #form)
#For Form 4 and 5 maximize the information at the fixed theta point of 1
(two #hard forms)
#Test length for form 1 and 2 is 10 (2, 3, 5 items from Contents 1, 2 and
3, #respectively)
#Test length for form 3 is 15 (5, 6, 4 items from Contents 1, 2 and 3,res
pect#ively)
#Test length for form 4 and 5 is 20 (4, 7, 9 items from Contents 1, 2 and
3, #respectively)
#For Forms 1 and 2, average word count across the items in the forms is
#between 30 and 80
#For Form3, average word count across the items in the forms is between
# 50 and 90
#For Forms 4 and 5, average word count across the items in the forms is
#between 20 and 100
#For Form 1 and 2, the average time to solve the item is between 200 and
250 #seconds
#For Form 3, average time to solve the item is between 200 and 300 second
```

```
s
#For Form 4 and 5, the average time to solve the item is between 200 and
400 #seconds
Ex1 <- ata(items, 5, #building 5 test forms at the same time!
#Change accordingly!
           max_use=1) #we don't want item overlapping!
Ex1 <-  ata_obj_relative(Ex1, -1, #fixed theta point for forms 1 and 2.
# Change accordingly!
                         "max",
                         forms=c(1,2) #the specified theta point of -1 wa
s
#for both forms 1 and 2 only!
)
Ex1 <-  ata_obj_relative(Ex1, 0, #fixed theta point for form 3.
#Change accordingly!
                         "max",
                         forms=3 #the specified theta point of 0 was for
both #form 3 only!
)
Ex1 <-  ata_obj_relative(Ex1, 1, #fixed theta point for forms 4 and 5. Ch
ange #accordingly!
                         "max",
                         forms=c(4,5) #the specified theta point of 1 was
for #both forms 4 and 5 only!
)
Ex1 <- ata_constraint(Ex1,1, min=10, max=10, forms=c(1,2)) #Test length f
or
#forms 1 & 2
Ex1 <- ata_constraint(Ex1,1,  min=15, max=15, forms=3) #Test length for f
orm #3
Ex1 <- ata_constraint(Ex1,1,  min=20, max=20, forms=c(4,5)) #Test length
for #forms 4 & 5
#For forms 1 and 2, specify content distributions for content 1, 2, and 3
,
#respectively.
Ex1 <- ata_constraint(Ex1, "content", min=2, max=2, level=1,forms=c(1,2))
Ex1 <- ata_constraint(Ex1,"content", min=3, max=3, level=2, forms=c(1,2))
Ex1 <- ata_constraint(Ex1, "content", min=5, max=5, level=3,forms=c(1,2))
#For form 3, specify content distributions for content 1, 2, and 3,
# respectively.
Ex1 <- ata_constraint(Ex1, "content", min=5, max=5, level=1,forms = 3)
Ex1 <- ata_constraint(Ex1,"content", min=6, max=6, level=2, forms = 3)
Ex1 <- ata_constraint(Ex1, "content", min=4, max=4, level=3,forms = 3)
#For forms 4 and 5, specify content distributions for content 1, 2, and 3
,
#respectively.
Ex1 <- ata_constraint(Ex1, "content", min=4, max=4, level=1,forms=c(4,5))
Ex1 <- ata_constraint(Ex1,"content", min=7, max=7, level=2, forms=c(4,5))
Ex1 <- ata_constraint(Ex1, "content", min=9, max=9, level=3,forms=c(4,5))
#For forms 1 and 2, specify word counts.
Ex1 <-  ata_constraint(Ex1, "word_count", min=30*10, max=80*10,forms=c(1,
2))
#For form 3, specify word counts.
```

```
Ex1 <-  ata_constraint(Ex1, "word_count", min=50*10, max=90*10,forms=3)
#For forms 4 and 5, specify word counts.
Ex1 <-  ata_constraint(Ex1, "word_count", min=40*10, max=100*10,forms=c(4
,5))
#For forms 1 and 2, specify time as seconds.
Ex1 <-  ata_constraint(Ex1, "time", min=250*10, max=400*10,forms=c(1,2))
#For form 3, specify time as seconds.
Ex1 <-  ata_constraint(Ex1, "time", min=200*10, max=300*10,forms=3)
#For forms 4 and 5, specify time as seconds.
Ex1 <-  ata_constraint(Ex1, "time", min=200*10, max=400*10,forms=c(4,5))
Ex1 <- ata_solve(Ex1, as.list=T) # Now, solve the ATA
Ex1$items #see selected items
plot(Ex1) # plotting information function
#End
```

```
#Example 2: maximize the information at the different theta intervals

#Pulling five sets of item (5 test forms)
#For forms 1 and 2 maximize the information at theta interval of -1.5 to
-0.5
#For form 3 maximize the information at theta interval of 0 to 0.5
#For forms 4 and 5 maximize the information at theta interval of 0.5 to 1
.5
#Test length for forms 1 and 2 is 10 (2, 3, 5 items from Contents 1, 2 an
d 3, #respectively)
#Test length for form 3 is 15 (5, 6, 4 items from Contents 1, 2 and 3,
#respectively)
#Test length for forms 4 and 5 is 20 (4, 7, 9 items from Contents 1, 2 an
d 3, #respectively)
#For forms 1 and 2, average word count across the items in the forms is
# between 30 and 80
#For form 3, average word count across the items in the forms is between
#50 and 90
#For forms 4 and 5, average word count across the items in the forms is
#between 20 and 100
#For forms 1 and 2 average time to solve the item is between 200 and 250
#seconds
#For form 3 average time to solve the item is between 200 and 300 seconds
#For forms 4 and 5 average time to solve the item is between 200 and 400
#seconds
Ex2 <- ata(items, 5, max_use=1)
Ex2 <-  ata_obj_relative(Ex2, seq(-1.5, -0.5, 0.10), #theta interval of -
1.5 #to -0.5
                         "max", flatten=0.50,
                         forms=c(1,2) #the specified interval is for Form
s 1 #and 2. Change accordingly!
)
Ex2 <-  ata_obj_relative(Ex2, seq(0, 0.5, 0.10), #theta interval of 0 to
0.5
                         "max", flatten=0.50,
                         forms=3) #the specified interval is for Form 3.
#Change accordingly!
Ex2 <-  ata_obj_relative(Ex2, seq(0.5, 1.5, 0.10), # interval of 0 to 1.5
```

```
                          "max", flatten=0.50,
forms=c(4,5) #the specified interval is for Forms 4&5. Change accordingly
!
)
Ex2 <- ata_constraint(Ex2,1, min=10, max=10, forms=c(1,2)) #Test length f
or
#forms 1 & 2
Ex2 <- ata_constraint(Ex2,1,  min=15, max=15, forms=3) #Test length for f
orm 3
Ex2 <- ata_constraint(Ex2,1,  min=20, max=20, forms=c(4,5)) #Test length
for #forms 4 & 5
#For forms 1 and 2, specify content distributions for content 1, 2, and 3
,
#respectively.
Ex2 <- ata_constraint(Ex2, "content", min=2, max=2, level=1,forms=c(1,2))
Ex2 <- ata_constraint(Ex2,"content", min=3, max=3, level=2, forms=c(1,2))
Ex2 <- ata_constraint(Ex2, "content", min=5, max=5, level=3,forms=c(1,2))
#For form 3, specify content distributions for content 1, 2, and 3,
#respectively.
Ex2 <- ata_constraint(Ex2, "content", min=5, max=5, level=1,forms = 3)
Ex2 <- ata_constraint(Ex2,"content", min=6, max=6, level=2, forms = 3)
Ex2 <- ata_constraint(Ex2, "content", min=4, max=4, level=3,forms = 3)
#For forms 4 and 5, specify content distributions for content 1, 2, and 3
,
#respectively.
Ex2 <- ata_constraint(Ex2, "content", min=4, max=4, level=1,forms=c(4,5))
Ex2 <- ata_constraint(Ex2,"content", min=7, max=7, level=2, forms=c(4,5))
Ex2 <- ata_constraint(Ex2, "content", min=9, max=9, level=3,forms=c(4,5))
#For forms 1 and 2, specify word counts.
Ex2 <-  ata_constraint(Ex2, "word_count", min=30*10, max=80*10,forms=c(1,
2))
#For form 3, specify word counts.
Ex2 <-  ata_constraint(Ex2, "word_count", min=50*10, max=90*10,forms=3)
#For forms 4 and 5, specify word counts.
Ex2 <-  ata_constraint(Ex2, "word_count", min=40*10, max=100*10,forms=c(4
,5))
#For forms 1 and 2, specify time as seconds.
Ex2 <-  ata_constraint(Ex2, "time", min=250*10, max=400*10,forms=c(1,2))
#For form 3, specify time as seconds.
Ex2 <-  ata_constraint(Ex2, "time", min=200*10, max=300*10,forms=3)
#For forms 4 and 5, specify time as seconds.
Ex2 <-  ata_constraint(Ex2, "time", min=200*10, max=400*10,forms=c(4,5))
Ex2 <- ata_solve(Ex2, as.list=T) # Now, solve the ATA
Ex2$items #see selected items
plot(Ex2) # plotting information function
#End
```

```
#Example 3: Specifying different absolute amount of information for diffe
rent forms
```

```
# pulling five sets of item (5 test forms)
# for forms 1 and 2 target theta is -1 and target information is 5
# for form 3 target theta is 0 and target information is 10
# for forms 4 and 5 target theta is 1 and target information is 15
# Test length for forms 1 and 2 is 10 (2, 3, 5 items from Contents 1, 2 a
nd 3#respectively)
#Test length for form 3 is 15 (5, 6, 4 items from Contents 1, 2 and 3,
#respectively)
#Test length for forms 4 and 5 is 20 (4, 7, 9 items from Contents 1, 2 an
d 3, #respectively)
#For forms 1 and 2, average word count across the items in the forms is
#between 30 and 80
#For form 3, average word count across the items in the forms is between
50
#and 90
#For forms 4 and 5, average word count across the items in the forms is
#between 20 and 100
#For forms 1 and 2 average time to solve the item is between 200 and 250
#seconds
#For form 3 average time to solve the item is between 200 and 300 seconds
#For forms 4 and 5 average time to solve the item is between 200 and 400
theta_target1=-1 #theta point (or it can be interval) where you want the
#absolute information.
theta_target2=0 #theta point (or it can be interval) where you want the
# absolute information.
theta_target3=1 #theta point (or it can be interval) where you want the a
bsolute information.
tif_target1= 5 #The amount of information for forms 1 and 2. Change
#accordingly!
tif_target2= 10 #The amount of information for form 3. Change accordingly
!
tif_target3= 15 #The amount of information for forms 4 and 5. Change
#accordingly!
Ex3 <- ata(items, 5, #we are building 5 forms at the same time. Change
#accordingly!
           max_use=1) #we don't want item overlapping. Change accordingly
!
#Specify ATA for forms 1&2
Ex3 <- ata_obj_absolute(Ex3, theta_target1, tif_target1, forms = c(1,2))
#Specify ATA for forms 3
Ex3 <- ata_obj_absolute(Ex3, theta_target2, tif_target2, forms = 3)
#Specify ATA for forms 4&5
Ex3 <- ata_obj_absolute(Ex3, theta_target3, tif_target3, forms = c(4,5))
Ex3 <- ata_constraint(Ex3,1, min=10, max=10, forms=c(1,2)) #Test length f
orms #1&2
Ex3 <- ata_constraint(Ex3,1,  min=15, max=15, forms=3) #Test length forms
3
Ex3 <- ata_constraint(Ex3,1,  min=20, max=20, forms=c(4,5)) #Test length
#forms 4&5
#For forms 1 and 2, specify content distributions for content 1, 2, and 3
,
# respectively.
Ex3 <- ata_constraint(Ex3, "content", min=2, max=2, level=1,forms=c(1,2))
```

```
Ex3 <- ata_constraint(Ex3,"content", min=3, max=3, level=2, forms=c(1,2))
Ex3 <- ata_constraint(Ex3, "content", min=5, max=5, level=3,forms=c(1,2))
#For form 3, specify content distributions for content 1, 2, and 3,
#respectively.
Ex3 <- ata_constraint(Ex3, "content", min=5, max=5, level=1,forms = 3)
Ex3 <- ata_constraint(Ex3,"content", min=6, max=6, level=2, forms = 3)
Ex3 <- ata_constraint(Ex3, "content", min=4, max=4, level=3,forms = 3)
#For forms 4 and 5, specify content distributions for content 1, 2, and 3
,
#respectively.
Ex3 <- ata_constraint(Ex3, "content", min=4, max=4, level=1,forms=c(4,5))
Ex3 <- ata_constraint(Ex3,"content", min=7, max=7, level=2, forms=c(4,5))
Ex3 <- ata_constraint(Ex3, "content", min=9, max=9, level=3,forms=c(4,5))
#For forms 1 and 2, specify word counts.
Ex3 <-  ata_constraint(Ex3, "word_count", min=30*10, max=80*10,forms=c(1,
2))
#For form 3, specify word counts.
Ex3 <-  ata_constraint(Ex3, "word_count", min=50*10, max=90*10,forms=3)
#For forms 4 and 5, specify word counts.
Ex3 <-  ata_constraint(Ex3, "word_count", min=40*10, max=100*10,forms=c(4
,5))
#For forms 1 and 2, specify time as seconds.
Ex3 <-  ata_constraint(Ex3, "time", min=250*10, max=400*10,forms=c(1,2))
#For form 3, specify time as seconds.
Ex3 <-  ata_constraint(Ex3, "time", min=200*10, max=300*10,forms=3)
#For forms 4 and 5, specify time as seconds.
Ex3 <-  ata_constraint(Ex3, "time", min=200*10, max=400*10,forms=c(4,5))
Ex3 <- ata_solve(Ex3, as.list=T) #Now, let's solve the ATA!
Ex3$items #see selected items
plot(Ex3) # plotting information function

# END OF THE CODE #
```

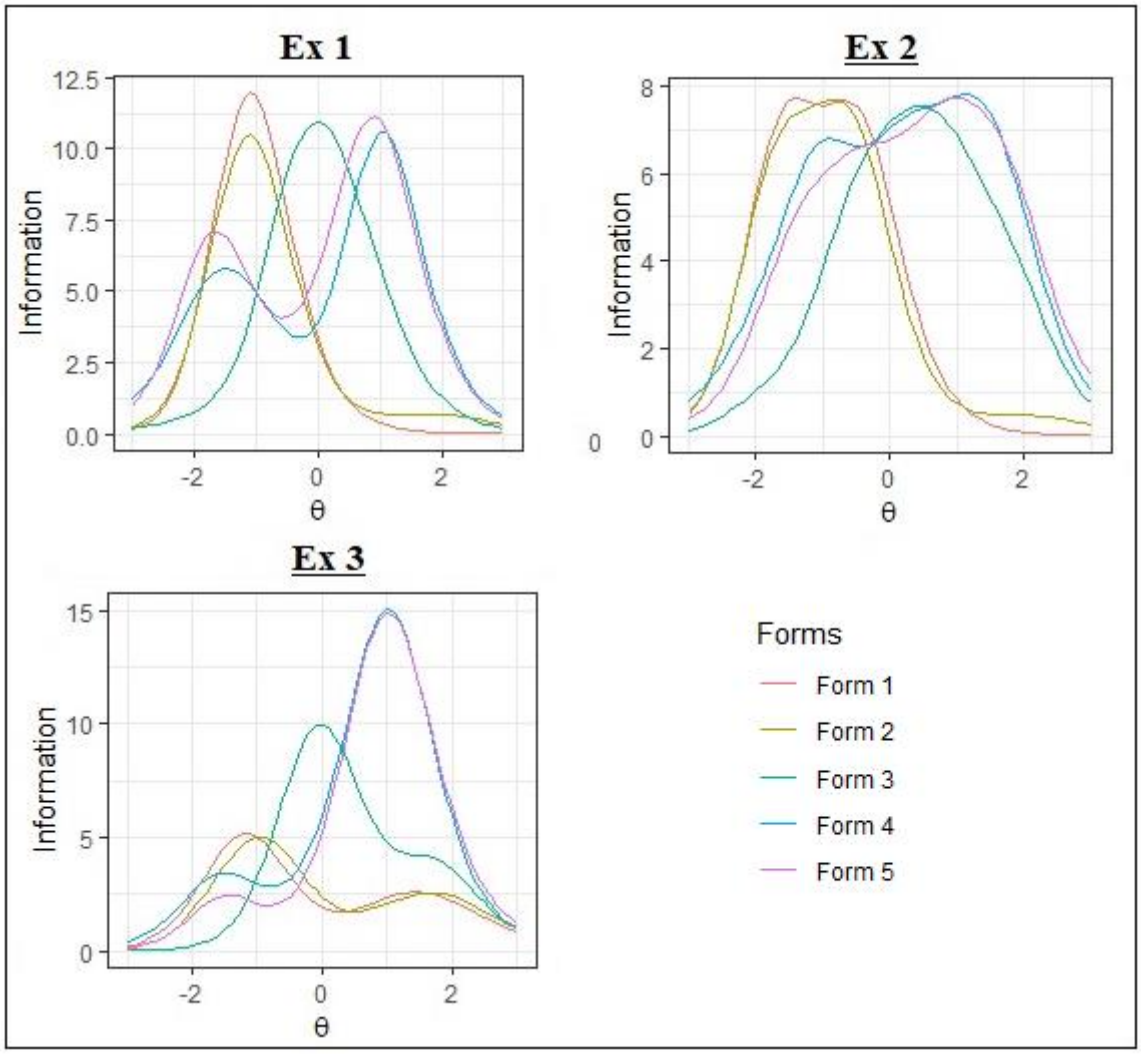


Figure A. 1. Plots for the solutions in Examples from 1 to 3.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 302-314

Inadvertent Use of ANOVA in Educational Research: ANOVA is not A Surrogate for MANOVA *

Lokman AKBAY ** Tuncer AKBAY *** Osman EROL **** Mustafa KILINÇ *****

Abstract
ANOVA and MANOVA address different research questions and decision on conducting one or the other of these tests relies on the research purpose. One prominent illegitimate analysis of multivariate data is developed out of conducting multiple ANOVAs rather than conducting a MANOVA. Another common mistake about MANOVA applications is the use of improper post hoc procedure. Post hoc procedures are needed to determine why the null hypothesis was rejected. Although the correct post hoc procedure for MANOVA is descriptive discriminant analysis (DDA), many researchers fail to conduct DDA to interpret their MANOVA results. The purpose of this study is two-fold; (1) we aim to emphasize the theory behind the MANOVA and its appropriate post hoc procedure and make clear distinction between surrogate statistical procedures such as ANOVA; and (2) this study also investigates the extent of incorrect analysis of multivariate dependent variables in educational research in Turkey. First, we provided a small simulation study to demonstrate the extent to which multiple ANOVAs yields contradictory results when they are inadvertently used to test group mean differences on multiple dependent variables. Results of the simulations indicated that MANOVA and multiple ANOVAs had severe disagreements under many conditions. Disagreement rate is elevated under the conditions where MANOVA retains the null hypothesis. Then, we systematically reviewed the archives of three education journals, which are classified as higher-, medium, and lower quality journals. Results indicated that correct use of MANOVA with its proper post hoc procedure is not common practice across educational researchers who publish in Turkish education journals.

Key Words: Multivariate data analysis, multivariate dependent variable, ANOVA, MANOVA.

INTRODUCTION

Univariate and multivariate data analysis are the two distinct statistical approaches. Univariate analysis involves only one variable at a time while two or more variables are involved in multivariate analysis. The analysis on group mean differences on a single outcome variable is referred to as Analysis of Variance (ANOVA); yet when multiple outcome variables are involved, we speak of Multivariate Analysis of Variance (MANOVA) (Fish, 1988; Stevens, 2002). Primary purpose of conducting both analyses is to determine treatment variable effect. MANOVA can be considered as a more general procedure of ANOVA. Although MANOVA is the most commonly used multivariate data analysis procedure (Kieffer, Reese & Thompson, 2001; Zientek & Thompson, 2009); literature indicates that MANOVA and its accompanying post hoc procedures are not properly understood by a considerable amount of social science researchers (Tonidandel & LeBreton, 2013; Warne, 2014; Warne, Lazo, Ramos & Ritter, 2012).

* Preliminary results of this work were presented at the 27th International Conference on Educational Sciences, Antalya, Turkey 2016.

** Assist. Prof., Burdur Mehmet Akif Ersoy University, Faculty of Education, Burdur-Turkey, lokmanakbay@gmail.com, ORCID ID: 0000-0003-4026-5241
*** PhD., Burdur Mehmet Akif Ersoy University, Faculty of Education, Burdur-Turkey, tuncerakbay@mehmetakif.edu.tr, ORCID ID: 0000-0003-3938-1026
**** Assist. Prof. Burdur Mehmet Akif Ersoy University, Faculty of Education, Burdur-Turkey, oerol@mehmetakif.edu.tr, ORCID ID: 0000-0002-9920-5211
***** Assist. Prof., Burdur Mehmet Akif Ersoy University, Faculty of Education, Burdur-Turkey, mkilinc@mehmetakif.edu.tr, ORCID ID: 0000-0002-2759-4916

To cite this article:
Akbay, L., Akbay, T., Erol, O., & Kılınç, M. (2019). Inadvertent use of ANOVA in educational research: ANOVA is not a surrogate for MANOVA. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 302-314. doi: 10.21031/epod.524511

Received: 08.02.2019
Accepted: 30.05.2019

ANOVA and MANOVA address different research questions so that decision on conducting one or the other of these analyses must be determined by the purpose of the research. One prominent inadvertent analysis of multivariate data is derived from conducting multiple ANOVAs rather than conducting a MANOVA. Conducting multiple ANOVAs fundamentally differs from MANOVA in two ways: (1) Multiple ANOVAs yield increase in the likelihood of committing Type I error. In a series of ANOVA, experiment-wise error can be as high as $1-(1-\alpha)^t$, where α is the Type I error rate and t is the number of ANOVAs conducted. For instance, the experiment-wise error will be .185 (i.e., $1-(1-.05)^4$) for $\alpha = .05$ and $t = 4$. Of course, this is the extreme case where dependent variables are uncorrelated. It should be noted that Type I error rate inflation depends on the correlation between the dependent variables (Hummel & Sligo, 1971). Therefore, Bonferroni correction (i.e., α/t) cannot overcome this problem unless dependent variables are truly uncorrelated.

Second fundamental difference (2) relies on the fact that ANOVA and MANOVA tend to answer to distinct empirical questions. Former statistical procedure is used to test the group mean differences on an observed variable, whereas the latter is used to test the group mean differences on underlying latent variables (Zientek & Thompson, 2009). Multiple ANOVAs fail to determine relationship between the independent variable(s) and combination of dependent variables (Warne, 2014). Notice that we are not interested in the possible group mean differenced on indicators (i.e., observed variables) of a latent dependent variable; yet we would like to detect the group mean difference on the latent variable that may be determined by a linear combination of the indicator variables. For example, from the statistical point of view, there might be no statistically significant difference in each of the dependent variables, yet a significant difference might be suggested by combination of them.

Another common mistake that is made in conducting MANOVA is related to use of improper post hoc procedure. Post hoc procedures are generally needed when the null hypothesis (H_0) is rejected in MANOVA (Stevens, 2002) to determine why the H_0 was rejected. Although the proper post hoc procedure for MANOVA is descriptive discriminant analysis (DDA) (Warne, 2014), most researchers do not conduct DDA to interpret their MANOVA results (Huberty & Morris, 1989; Warne et al., 2012). This is mainly because many researchers use SPSS for MANOVA and it automatically conducts an ANOVA for each dependent variable. However, some researchers claim that because ANOVA is only concerned with observed variable, use of ANOVA as a follow-up procedure to significant MANOVA result is against the nature of MANOVA (Kieffer et al., 2001; Zientek & Thompson, 2009). Underlying rationale to this claim relies on the difference in the empirical questions ANOVA and MANOVA are exposed to (i.e., ANOVA tests the mean differences on the observed variable whereas MANOVA tests the mean differences on the underlying latent variables).

Purpose of the Study

The purpose of this study is two-fold. (1) We aim to emphasize the theory behind the MANOVA and its appropriate post hoc procedure (i.e., DDA) and make clear distinction between surrogate statistical procedures such as ANOVA. (2) This study also investigates the extent of inadvertent analysis of multivariate dependent variables in educational research in Turkey. In other words, this study aims to determine to the extent to which educational researchers conduct MANOVA when it is the most appropriate way of analyzing the data to answer their empirical question.

Univariate and Multivariate Hypothesis Testing

To find out whether the mean score on a dependent variable is equal across two or more groups, ANOVA test is conducted and an F-statistic is computed. To test the null hypothesis (i.e., group means are equal) observed F-statistic compared against the sampling distribution. The null hypothesis is rejected when observed statistic fall beyond a predetermined critical value; otherwise the null hypothesis is retained. When multiple dependent variables are employed in the analysis, each of them may or may not fall in the rejection region. Furthermore, linear combinations of the dependent variables may or may not fall in the rejection region. Imagine a case where two perfectly uncorrelated

dependent variables are tested; as can be seen in Figure 1, rejection region becomes the outside of the circle. Further assume that these two uncorrelated observed dependent variables equally contribute to the underlying latent variable. Then one of the four possible cases may be observed.

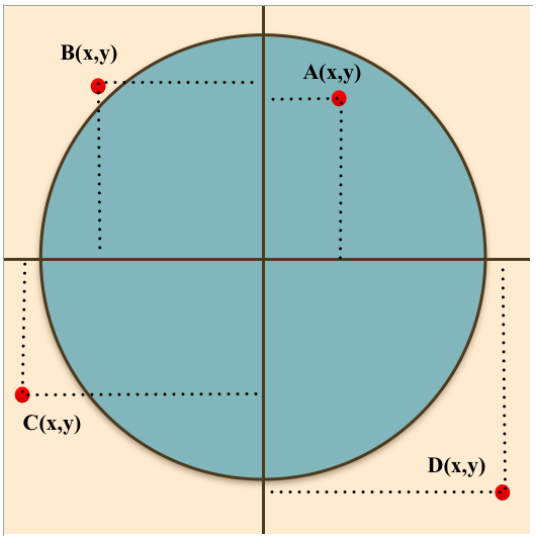


Figure 1. Possible Hypothesis Testing Results for Two Perfectly Uncorrelated Dependent Variables

In the first case, both of the observed variables (i.e., x and y) and the latent variable (i.e., A) do not fall outside the circle so that all of the hypotheses testing the group mean differences in the observed and latent variables are retained. In other words, neither the ANOVAs nor the MANOVA suggest any significant difference. In the second case (i.e., latent variable B), although both ANOVAs fail to reject the null hypotheses, MANOVA rejects the null hypothesis. In case of latent variable C, MANOVA and ANOVA testing the difference on observed variable x yield significant difference; whereas ANOVA for the observed variable y suggests no significant difference. In the last case, all tests reject the null hypotheses. As shown in the Figure 1, MANOVA and multiple ANOVAs may result in contradicting results.

When a MANOVA test result rejects the null hypothesis of equality of group means we need to go ahead and identify how one or more groups of observations differ by interrelated multiple dependent variables. Difference can be in anywhere: in one variable or in a combination of multiple variables. DDA should be run to find the source of the difference. Although we have no intention to explain DDA in details, several reminders might be noted here. DDA provides us with discriminant functions, which are created by the linear combination of the dependent variables to maximize group differences (Sherry, 2006). DDA treats outcome variables as the linear combination of the dependent variables that maximizes group differences. DDA, in general, help us determine how much each of the dependent variable contribute to group difference on the outcome variable.

METHOD

This study can be regarded a documentary survey, which is a type of survey research under the descriptive research method. Documentary surveys are akin to content analysis or document analysis. The term content analysis is used to define the process of summarizing and reporting written data (Cohen, Manion & Morrison, 2002). Document analysis is defined as a systematic procedure for evaluating or reviewing printed and/or electronic materials (Bowen, 2009). With this documentary survey, we aim to ascertain whether use of MANOVA with its proper post hoc procedure is common practice across educational researchers who publish in Turkish education journals.

Data Collection Procedure

We have obtained our data through screening the archives of three education journals. We specifically reviewed all issues published in these journals in the last four years (i.e., 2015-2018). These three journals may represent the higher, medium, and lower quality journals based on where they are indexed. Based on our classification, Thompson Reuters Social Sciences Citation Index indexes the higher quality journal. The medium and lower quality journals are indexed by the Thompson Reuters Emerging Sources Citation index and ULAKBIM Social and Human Sciences Database (Sosyal ve Beşeri Bilimler Veri Tabanı), respectively. Detailed information on the journals may be provided upon request.

Researchers reviewed the articles published in these three Turkish education journals and reported the counts as well as the type of analyses used to test intervention effect or group mean differences in multiple outcome variables. Furthermore, counts and the types of post-hoc procedures are also reported. We considered the following types of multivariate dependent variable analyses:

1. Use of MANOVA to test group mean differences in multivariate data
   - followed by DDA
   - followed by ANOVA
   - followed by other procedures or no post hoc
2. Use of ANOVA with sum scores to test group mean difference in multivariate data
3. Use of multiple ANOVAs to test group mean difference in multivariate data

Moreover, we provided a small simulation study to demonstrate the extent to which multiple ANOVAs yields incorrect results when they are inadvertently used to test group mean differences on multiple dependent variables. This simulation is also designed to determine to what extent the results of multiple ANOVAs agree to the results obtained from MANOVA. For the simulation conditions, data were generated from a standard multivariate normal distribution. Sample size is fixed to 100 for each group. Number of groups and number of dependent variables are fixed to two, and three, respectively. Correlation between the dependent variables, difference in the population means, and distribution variance are the three variables considered to create the simulation conditions. Correlation had two levels, which specifies lower- and higher-correlation conditions. More specifically, in lower correlation condition, the correlations between the dependent variables are drawn from a uniform distribution with minimum of .2 and maximum of .4. Likewise, correlations for the higher correlation condition are drawn from a uniform distribution with minimum value of .6 and maximum value of .8. Note that the mean of these distributions (i.e, .3 and .7) are the cutoff scores for describing the magnitude of a relationship in social sciences. As argued by Köklü, Büyüköztürk and Çokluk (2007), a correlation coefficient smaller than .3 represent a low relationship and one larger than .7 represents a high relationship.

Table 1. Variables Used in Simulation

Corr	**Δμ**	**σ^2**
Lower = U(.2, .4)	Small = 0.2 standard deviation	Lower = .5
Higher = U(.6, .8)	Medium = 0.4 standard deviation	Medium = 1.0
	Large = 0.6 standard deviation	Higher = 1.5

Note: Corr is the correlations between the dependent variables; Δμ is the population mean differences; σ^2 is the distribution variance.

Population mean difference had three levels, which are labeled as small-, medium-, and large-difference conditions. These three levels were fixed to 0.2, 0.4, and 0.6 standard deviations. Here we have no intention to define what is a small or a large difference is; rather, we are just using these arbitrary differences to demonstrate the impact of the size of mean differences. More specifically, one

group is generated from a multivariate normal distribution MVN(0,Σ), where Σ is the variance-covariance matrix determined by the variance of and correlations specified for each conditions. Then, 0.2, 0.4, or 0.6 is added to the mean vector of the second group for the small, medium, and large mean difference conditions, respectively. Last variable is the distribution variance for which 0.5, 1.0, and 1.5 were used to represent lower-, medium-, and higher-variance conditions. These three variables and their levels are summarized in Table 1. Combination of two correlation levels, three mean difference levels, and three variance levels yield 18 conditions. Number of replication for each condition is fixed to 500.

Table 2. The Extent to which Multiple ANOVAs are in Conformity with MANOVA

			MANOVA	**ANOVAs**		**MANOVA**	**ANOVAs**	
Corr	**Δμ**	**σ^2**	**p ≥ .05**	**p ≥ .05**	**p < .05**	**p < .05**	**p < .05**	**p ≥ .05**
Lower	Small	Lower	.368	.174	.194	.632	.632	.000
		Medium	.676	.456	.220	.324	.322	.002
		Higher	.796	.558	.238	.204	.202	.002
	Medium	Lower	.004	.002	.002	.996	.996	.000
		Medium	.108	.026	.082	.892	.892	.000
		Higher	.258	.116	.142	.742	.742	.000
	Large	Lower	.000	.000	.000	1.000	1.000	.000
		Medium	.000	.000	.000	1.000	1.000	.000
		Higher	.024	.000	.024	.976	.976	.000
Higher	Small	Lower	.566	.322	.244	.434	.432	.002
		Medium	.798	.564	.234	.202	.196	.006
		Higher	.828	.626	.202	.172	.166	.006
	Medium	Lower	.032	.006	.026	.968	.968	.000
		Medium	.236	.056	.180	.764	.764	.000
		Higher	.446	.188	.258	.554	.554	.000
	Large	Lower	.000	.000	.000	1.000	1.000	.000
		Medium	.016	.002	.014	.984	.984	.000
		Higher	.096	.026	.070	.904	.904	.000

Note: Corr is the correlations between the dependent variables; Δμ is the population mean differences; σ^2 is the distribution variance; and p is the type I error rate of the test.

RESULTS

Simulation Results

Data generation and the analyses of the generated data are conducted in R language and statistical computing environment (R core team) using R-package "MASS" (Venables & Ripley, 2002). R code used for data generation and analyses is given in the Appendix A. Simulation results are summarized in Tables 2 and 3. These tables present the conformity on test results of MANOVA and multiple ANOVAs without and with Bonferroni correction, respectively. It should be noted here that, under the (multiple) ANOVAs condition, retain refers to the conditions where all three tests corresponding to three dependent variables are retained; whereas, reject refers to the conditions where at least one hypothesis out of the three is rejected. In the MANOVA tests, we used the Pillai's trace as rejection criterion because it is more robust to MANOVA violation of test assumptions (Olson, 1974).

First of all, result tables present two expected results: (1) Increase in the sample variance yields increase in the number of retained null hypotheses when the mean difference is tested by either multiple ANOVAs or by a MANOVA. For example, under the lower correlation and small mean difference cases, MANOVA retains about 37% to 80% of the null hypothesis as the variance increases from 0.5 to 1.5. Similarly, when we conduct multiple ANOVAs without Bonferroni correction, approximately 17% to 56% of the null hypotheses are retained as the sample variance increases from 0.5 to 1.5. Under the same conditions, when we conduct multiple ANOVAs with Bonforreni correction, these percentages become 34% (i.e., .318+.022) to 77% (i.e., .750+.016).

Table 3. The Extent to which Multiple Bonferroni Corrected ANOVAs Agree with MANOVA

			MANOVA	ANOVAs		MANOVA	ANOVAs	
Corr	**Δμ**	**σ^2**	**p ≥ .05**	**p ≥ .0167**	**p < .0167**	**p < .05**	**p < .0167**	**p ≥ .0167**
Lower	Small	Lower	.368	.318	.050	.632	.610	.022
		Medium	.676	.634	.042	.324	.290	.034
		Higher	.796	.750	.046	.204	.188	.016
	Medium	Lower	.004	.002	.002	.996	.996	.000
		Medium	.108	.088	.020	.892	.876	.016
		Higher	.258	.228	.030	.742	.710	.032
	Large	Lower	.000	.000	.000	1.000	1.000	.000
		Medium	.024	.000	.000	1.000	1.000	.000
		Higher	.024	.020	.004	.976	.976	.000
Higher	Small	Lower	.566	.458	.108	.434	.424	.010
		Medium	.798	.724	.074	.202	.180	.022
		Higher	.828	.766	.062	.172	.154	.018
	Medium	Lower	.032	.020	.012	.968	.966	.002
		Medium	.236	.152	.084	.764	.758	.006
		Higher	.446	.340	.106	.554	.548	.006
	Large	Lower	.000	.000	.000	1.000	1.000	.000
		Medium	.016	.008	.008	.984	.982	.002
		Higher	.096	.060	.036	.904	.902	.002

Note: Corr is the correlations between the dependent variables; Δμ is the population mean differences;σ^2 is the distribution variance; and p is the type I error rate of the test.

Another expected result is (2) the increase in the rejection rates of the tests along with the increase in the sample mean differences. For example, under the lower correlation and higher variance conditions, rejection rates of MANOVA varied from .204 to .976 as the sample mean differences increases from 0.2 standard deviation to 0.6 standard deviation. Rejection rates of multiple ANOVAs without Bonferroni correction vary between .440 (i.e., .238+.202) to 1.000 (i.e., .024+.976) for the same conditions. When ANOVAs are conducted with Bonferroni correction, rejection rates of multiple ANOVAs vary between .234 (i.e., .046+.188) to .980 (i.e., .004+.976). Although these are the expected results, we are more interested in the agreement between the MANOVA and multiple ANOVAs in terms of hypothesis test results. Remember that this simulation study only considers the similarity of the test results from a statistical point of view. We do not have any intention to downgrade the importance of theoretical considerations on choosing one or the other analysis.

When we look at the results obtained under lower and higher correlation conditions, MANOVA tend to fail to reject the null hypothesis as the correlation between the dependent variables increases. For example, when sample variance is higher and correlation between the dependent variables is lower, MANOVA retains the null hypothesis .796, .258, and .024 of the time for the small-, medium-, and large mean difference cases; whereas these rates rise up to .828, .446, and .096 under the higher correlation cases. As long as the simulation results concerned, we are mainly interested in the agreement rates of the two types of dependent variable analysis results. Looking at the retain rates, we observed a great quantity of disagreement under certain conditions. For instance, MANOVA retains the null hypotheses with a rate of .368 (i.e., 184 out of 500) under the lower correlation, small mean difference, and lower sample variance case. Multiple ANOVAs, however, only retain 87 out of the 184 null hypotheses, which are already retained by MANOVA (i.e., agreement on retaining the null hypotheses is .174). When Bonferroni correction is applied to ANOVA tests, this agreement rate is reported to be 159 out of 184 times (i.e., .318).

Tables 2 and 3 suggest that multiple ANOVAs procedure rejects a great deal of the null hypotheses that are already rejected by MANOVA. The highest disagreement rates for the ANOVAs are observed under small mean difference cases when Bonferroni correction is applied to ANOVAs (i.e., up to .034 and .022 under the lower and higher correlation conditions, respectively). In general, these results indicate that application of multiple ANOVAs rather than a single MANOVA yields higher rejection rates.

Results on Document Analysis

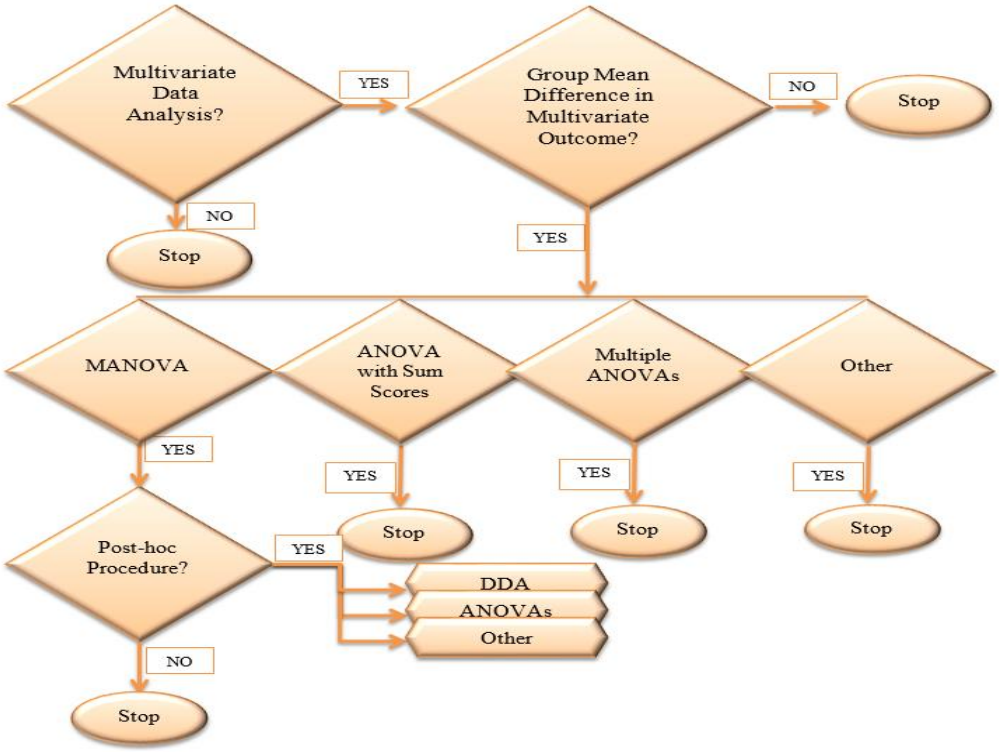


Figure 2. Flow Chart Used to Categorize the Reviewed Research.

We have gathered our archival data by screening the archives of three education journals (e.g., higher, medium, and lower quality). We have found 144 studies investigating the mean difference of multivariate dependent variables as we have viewed a total of 767 articles. We looked at the data analysis technique used for testing the group mean differences. To categorize reviewed works, we have used the flow chart given in Figure 2. In our archival survey, we have come across multiple t-tests applied to test the mean differences across two groups on multiple dependent variables. These studies were counted toward multiple ANOVAs category.

Table 4. Results on The Archival Survey

Journal Quality	**Years**	**Number of Articles**	**Multivariate Mean Difference**	**MANOVA**			**Sum Score ANOVA**	**Multiple ANOVAs**	
				No post hoc	ANOVA	DDA		No post hoc	Post hoc
HQ-J	2015	88	15	1	3	0	5	6	0
	2016	62	11	1	1	0	1	7	1
	2017	80	14	2	2	0	6	4	0
	2018	60	11	0	0	0	2	9	0
MQ-J	2015	61	13	0	2	0	4	7	0
	2016	50	12	1	2	0	2	7	0
	2017	60	3	0	0	0	0	3	0
	2018	45	10	0	2	0	4	4	0
LQ-J	2015	38	12	0	1	0	3	10	0
	2016	65	11	0	0	0	5	5	0
	2017	76	18	0	0	0	6	10	0
	2018	82	14	0	1	0	4	9	0
All-3-J	15-18	767	144	5	14	0	42	81	1

Note: HQ-J = higher quality journal; MQ-J = medium quality journal; LQ-J = lower quality journal; All-3-J = all three journals.

Results on the archival survey are summarized in Table 4. Rate of articles investigating treatment variable effect on the multivariate variables are about 18% (i.e., 51/290 and 38/216) for the higher and medium quality journals, while this rate is slightly higher for the lower quality journal (i.e., 21% or 55/261). Rate of MANOVA test use for detecting treatment effect is quite low: 10/51; 7/38; and 2/55 for the higher, medium, and lower quality journal publications, respectively. Although the maximum number of studies investigating mean differences on multivariate data is reported to be published in the lower quality journal, use of MANOVA to test the mean difference is only about 4% (i.e., 2 out of 55). Within the rare use of MANOVA, employment of ANOVA as post hoc tests is quite common (i.e., 14 out of 19). This may be mainly due to the fact that ANOVA tests are readily available when MANOVA test is run by the statistical package for the social sciences (SPSS). Moreover, although the most accurate inferences can be made when DDA is run as a follow up test for MANOVA, we have not come across any study that used DDA to interpret MANOVA results.

It is obvious from the results summarized in Table 4 that many researchers do not use MANOVA when it is the most appropriate way to test effect of independent variable(s) on the multivariate dependent variables. Rather than using MANOVA, many educational researchers who published in Turkish educational journals run a single ANOVA on the sum score obtained from multivariate dependent variables or they run multiple ANOVAs to test the effect on each of the dependent variables separately. Figure 3 displays these results based on the three types of journals as well as the results obtained from all three journals altogether. This figure shows that employment of MANOVA is quite rare across all, especially for the lower, quality journal publications. At least more than half of the studies run multiple ANOVAs rather than running a single MANOVA to test group mean differences on the multivariate dependent variables. Furthermore, approximately 30% of the studies used a single ANOVA test on a dependent variable, which is obtained by summing all the scores on multiple dependent variables.

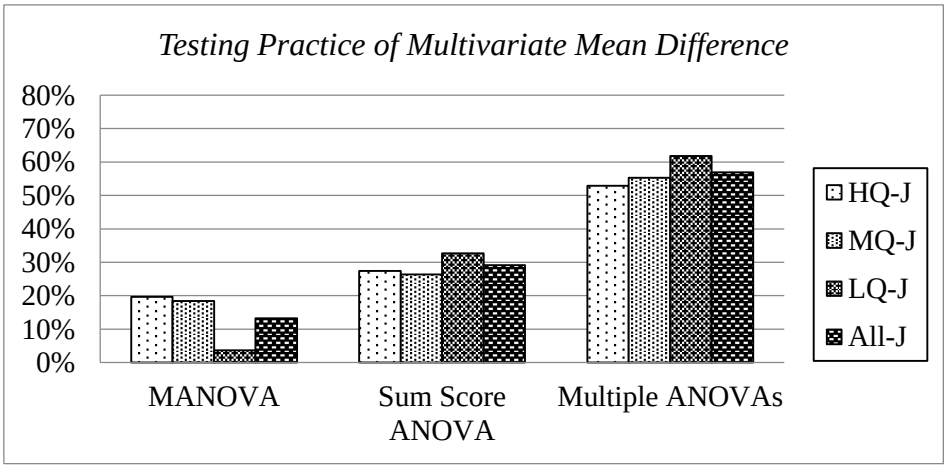


Figure 3. Rate of Analyses Used to Test Multivariate Mean Differences.

DISCUSSION and CONCLUSION

Primary purpose of conducting univariate or multivariate analysis of variance is to determine treatment variable effects. Although MANOVA can be considered as a more general procedure of ANOVA, it is not just a statistical extension of ANOVA because they address different research questions. ANOVA is used to test the group mean differences on an observed variable whereas MANOVA is used to test the group difference on an underlying latent variables. By conducting a MANOVA we basically test the group mean differences on a linear combination of the dependent variables. Because we are not interested in the mean difference of any single dependent variable when we conduct MANOVA, conducting multiple ANOVAs (i.e., an ANOVA for each dependent variable) would not

be the same as conducting a single MANOVA. To do so would not address the empirical questions researchers begins with and yield different statistical test results.

With this study, we aimed to emphasize the theory behind the MANOVA and to make clear distinction between surrogate statistical procedures such as ANOVA. We not only focused on the theoretical difference between the two; through a small simulation study, we also demonstrated the discrepancy between obtained statistical test results. Then, we further investigated the extent of incorrect analysis of multivariate data in educational studies that are published in Turkish education journals. We specifically focused on the analysis of multivariate data for treatment variable effects and the post hoc procedures used for follow up. Results indicated that correct use of MANOVA with its proper post hoc procedure is not common practice across educational researchers who publish in Turkish education journals.

Although the courses given in the graduate level include the analysis of multivariate data, it is observed that, at least in case of MANOVA, the areas of application are not properly understood. The underlying reason for this may be the presentation of practical information on how to analyze data at hand with specific statistical package programs (eg., SPSS) rather than presentation of the theoretical background of these statistical data analysis techniques. In order to eliminate such deficiencies and misunderstandings of individuals who are conducting research in education, it is useful to take steps to gain theoretical knowledge on the basis of statistical analysis in the graduate education programs. We also suggest researchers to co-operate with the experts of the related fields if they deem necessary.

REFERENCES

Bowen, G. A., (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27-40.

Cohen, L., Manion, L., & Morrison, K. (2002). *Research methods in education.* London; Routledge.

Fish, L. J. (1988). Why multivariate methods are usually vital. *Measurement and Evaluation in Counseling and Development, 21*, 130-137.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analysis. *Psychological Bulletin, 105*, 302-308.

Hummel, T. J., & Sligo, J. R. (1971). Empirical comparison of univariate and multivariate analysis of variance procedures. *Psychological Bulletin, 76*, 49-57.

Kieffer, K. M., Reese, R. J., & Thompson, B. (2001). Statistical techniques employed in *AERJ* and *JCP* articles from 1988 to 1997: A methodological review. *Journal of Experimental Education, 69*, 280-309.

Köklü, N., Büyüköztürk, Ş., & Çokluk, Ö. (2007). *Sosyal bilimler için istatistik.* Ankara: Pegem Yayınları.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of The American Statistical Association, 69*(348), 894-908.

R Core Team (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Retrieved from http:www.R-project.org/

Sherry, A. (2006). Discriminant analysis in counseling psychology research. *The Counseling Psychologist, 34*, 661–683. doi:10.1177/0011000006287103

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences.* Mahwah, NJ: Lawrence Erlbaum Associates.

Tonidandel, S., & LeBreton, J. M. (2013). Beyond step-down analysis: A new test for decomposing the importance of dependent variables in MANOVA. *Journal of Applied Psychology, 98*, 469-477.

Venables, W. N. & Ripley, B. D. (2002) *Modern applied statistics with S* (Fourth Edition). New York, NY: Springer.

Warne, R. T. (2014). A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. *Practical Assessment, Research & Evaluation, 19*(17), 1-10.

Warne, R. T., Lazo, M., Ramos, T., & Ritter, N. (2012). Statistical methods used in gifted education journals, 2006-2010. *Gifted Child Quarterly, 56*, 134-149.

Zientek, L. R., & Thompson, B. (2009). Matrix summaries improve research reports: Secondary analyses using published literature. *Educational Researcher, 38*, 343-352.

ANOVA'nın Eğitim Araştırmalarında Dikkatsizce Kullanımı: ANOVA, MANOVA için Yer Tutucu Değildir

Giriş

Tek değişkenli varyans analizi (ANOVA) ve çok değişkenli varyans analizi (MANOVA) farklı araştırma sorularına cevap arayan iki farklı istatistiksel yöntemdir. Bu ikisi arasındaki seçim araştırmanın amacına bağlı olarak yapılır; tek bir bağımlı değişken için gruplar arası fark bakılırken ANOVA, birden fazla bağımlı değişken için gruplar arası fark bakılırken MANOVA'dan bahsediyoruzdur (Fish, 1988; Stevens, 2002). MANOVA istatistiksel olarak ANOVA'nın daha genel bir prosedürü olarak düşünülebilir. MANOVA en sık kullanılan çok değişkenli data analiz prosedürlerinden biri olsa da (Kieffer, Reese & Thompson, 2001; Zientek & Thompson, 2009); alan yazın incelendiğinde bu prosedür ve analize eşlik etmesi gereken doğru post hoc prosedürünün azımsanmayacak sayıda sosyal bilimler araştırmacısı tarafından doğru anlaşılmadığı görülmektedir (Tonidandel & LeBreton, 2013; Warne, 2014; Warne, Lazo, Ramos & Ritter, 2012).

MANOVA testinin kullanılması gereken yerlerde en sık karşımıza çıkan yanlış kullanım her bir bağımlı değişkeni ayrı ayrı test eden ANOVA testleri serisinin tercih edilmesidir. Ancak, birden fazla ANOVA testinin uygulanması bir tek MANOVA testinin uygulanmasından iki şekilde farklılık arz eder: (1) birden fazla ANOVA uygulaması *birinci tip hatasının* yapılma olasılığını artırır. Bu hatanın artış oranı bağımlı değişkenler arasındaki korelasyonun büyüklüğü ile değişmekte olup kolayca kontrol altına alınamaz. Tamamen bağımsız yani korelasyonun sıfır olduğu durumlar için Bonferroni düzeltmesi uygulamak bu hata oranının ancak kontrol altına alınmasını sağlayabilir (Hummel & Sligo, 1971) ki sosyal bilimlerdeki çoklu bağımsız değişkenler arasındaki korelasyonun sıfır olduğu durum (eğer varsa) sınırlıdır.

Çoklu ANOVA ve MANOVA arasındaki ikinci temel fark ise (2) bu testlerin farklı ampirik sorulara cevap verebilir olmasıyla ilgilidir. ANOVA gözlenen değişkenlerden elde edilen veriler için uygun bir test iken; MANOVA gözlenmeyen (gizil) değişkenler üzerinden gruplar arası farklılık olup olmadığını anlamak için yapılabilecek uygun bir testtir (Zientek & Thompson, 2009). Birden fazla ANOVA testinden elde edilen sonuçlar bağımsız değişken(ler) ile bağımlı değişkenlerin kombinasyonu arasında anlamlı bir ilişki olup olmadığını test etmede yetersiz kalır (Warne, 2014). MANOVA testinin kullanımında araştırmacılar gözlenmeyen değişkenlerin gözlenen gösterge (indicator) değişkenleri açısından gruplar arasında fark olup olmadığını değil, bu gösterge değişkenlerin lineer bir kombinasyonundan oluşan gözlenemeyen değişken açısından gruplar arasında anlamlı bir farklılık olup olmadığını araştırmaktadır.

MANOVA yerine yanlışlıkla ANOVA kullanımının bir diğer şekli ise bağımlı değişkenlerden elde edilen skorların toplamı üzerinden bir tek ANOVA testinin yapılmasıdır. Bu çalışmanın iki temel amacı vardır. (1) MANOVA'nın ve devamında uygulanması gereken post hoc testinin alt yapısını oluşturan teoriyi vurgulayarak ANOVA ve MANOVA arasındaki farklılıkların anlaşılmasına yardımcı olmak; (2) Türkiye'de yayınlanan eğitim dergilerinde basılmış makalelerde MANOVA testinin ve doğru post hoc testinin kullanılması gerektiği durumlarda bunların kullanılmış olma oranını ortaya koymaktır.

Yöntem

Çalışmanın yöntemi betimsel araştırma yöntemlerinden doküman analizidir. Doküman analizi içerik analizine yakın bir veri analizi yöntemidir. Bu yöntem basılı ya da elektronik materyallerin sistematik bir şekilde incelenmesinin ve değerlendirilmesinin yapılması şeklinde tanımlanabilir (Bowen, 2009). Bu doküman analiziyle araştırmacılar, Türkiye'de yayın yapan eğitim dergilerinde basılmış makalelerde, MANOVA testinin ne ölçüde doğru kullanıldığının tespitini yapmayı amaçlamaktadırlar. Üç eğitim dergisinin arşivlerinden son dört yılda (2015-2018) yayınlanan tüm sayıları incelemek

kaydıyla veriler elde edilmiştir. Bu üç dergi, endekslendikleri yerlere göre yüksek, orta ve düşük kaliteli dergileri temsil edecek şekilde seçilmiştir. Bu sınıflandırma dergilerin tarandıkları veri tabanları göz ününde bulundurularak yapılmıştır (ör. Thompson Reuters Sosyal Bilimler Atıf Dizini yüksek kaliteli dergiyi endekslemektedir). Araştırmacılar bu üç Türk eğitim dergisinde yayınlanan makaleleri incelemiş çok değişkenli bağımlı değişken ile bağımsız değişken(ler) arasındaki ilişkiye bağlı olarak gruplar arası anlamlı farklılıkları test etmek için kullanılan analiz türlerini raporlaştırmışlardır.

Tablo 1. Similasyonda Manipüle Edilen Değişkenler

Corr	**Δμ**	**σ^2**
Düşük = U(.2, .4)	Küçük = 0.2 standart sapma	Düşük = .5
Yüksek = U(.6, .8)	Orta = 0.4 standart sapma	Orta = 1.0
	Büyük = 0.6 standart sapma	Yüksek = 1.5

Not: Corr, bağımlı değişkenler arasındaki korelasyon; Δμ, popülasyon ortalamalarındaki fark; σ^2, dağılımların varyansı.

Ayrıca, çoklu ANOVA'ların yanlışlıkla çok değişkenli bağımlı değişkenler üzerindeki grup ortalama farklarını test etmek için kullanıldığında istatistiksel olarak ne ölçüde tutarlı sonuçlar verdiğini gösteren küçük bir simülasyon çalışması yaptık. Simülasyon koşulları için, standart çok değişkenli normal dağılımdan veriler üretilmiştir. Örneklem büyüklüğü her grup için 100'e sabitlenmiştir. Grup sayısı ve bağımlı değişken sayısı sırasıyla iki ve üçe sabitlenmiştir. Bağımlı değişkenler arasındaki korelasyon, popülasyon ortalamaları arasındaki fark ve dağılım varyansı, simülasyon koşullarını oluşturmak için manipüle edilen değişkenlerdir. Bu üç değişken ve değişkenlerin düzeyleri Tablo 1'de özetlenmiştir. İki korelasyon düzeyi, üç ortalama fark düzeyi ve üç varyans düzeyinin çaprazlanmasıyla toplam 18 simülasyon durumu oluşturulmuştur. Her durum için replikasyon sayısı 500 olarak belirlenmiştir.

Sonuç ve Tartışma

Düşük ve yüksek korelasyon koşulları altında elde edilen simülasyon sonuçlarına baktığımızda, MANOVA bağımlı değişkenler arasındaki korelasyon arttıkça yokluk hipotezini daha sıklıkla reddetme eğilimindedir. Bu simülasyon sonuçları içinden biz bağımlı değişken analizinde kullanılan iki tür testin (MANOVA ve Çoklu ANOVA) sonucunun mutabakat oranlarıyla daha çok ilgileniyoruz. Mutabakat oranlarına bakıldığında, belirli koşullar altında büyük miktarda anlaşmazlık olduğunu gözlemleyebiliriz. Örneğin, MANOVA yokluk hipotezini düşük korelasyon, küçük popülasyon ortalama farkı ve düşük dağılım varyansı durumunda .368 oranında reddedemektedir. Bununla birlikte, çoklu ANOVA'lar, reddedilemeyen yokluk hipotezlerinin en az yarısını reddetmektedir. Bonferroni düzeltmesi ANOVA testlerine uygulandığında, MANOVA ve çoklu ANOVA arasındaki yokluk hipotezlerini reddedememe mutabakatlarının oldukça yükseldiği gözlenmiştir. Simülasyon sonuçları çoklu ANOVA ve MANOVA'nın yokluk hipotezini reddetme mutabakatlarının oldukça yüksek olduğu sonucunu ortaya koymaktadır. Birkaç istisna dışında, üç ANOVA'dan en az biri, MANOVA tarafından zaten reddedilmiş yokluk hipotezlerini reddetmektedir. Genel olarak, bu sonuçlar tek bir MANOVA yerine birden fazla ANOVA uygulamasının daha yüksek oranda yokluk hipotezi reddetme eğilimi gösterdiğini ortaya koymaktadır.

Üç eğitim dergisinin arşivleri taranarak çok değişkenli bağımlı değişkenlerin grup ortalama farkını araştıran 144 çalışma bulunmuştur. Çok değişkenli bağımlı değişkenler üzerinde bağımsız değişkeninin etkisini araştıran makalelerin oranı, yüksek ve orta kaliteli dergiler için yaklaşık %18 (yani, 51/290 ve 38/216) iken, düşük kaliteli dergi için %21 (55/261) olarak bulunmuştur. Bağımsız değişken etkisinin saptanmasında MANOVA testi kullanım oranının oldukça düşük olduğu görülmüştür: 10/51; 7/38; ve 2/55 sırasıyla yüksek, orta ve düşük kaliteli dergiler için. MANOVA'nın nadir kullanımı içinde, ANOVA'nın post hoc testi olarak kullanımının oldukça yaygın olduğu görülmüştür (14/19). Bu durum MANOVA testinin sosyal bilimler için istatistiksel paket (SPSS) programı tarafından gerçekleştirildiğinde, ANOVA testlerinin otomatik olarak uygulanıyor

olmasından kaynaklanıyor olabilir. Bununla birlikte, DDA MANOVA için en doğru post hoc prosedürü olmasına rağmen, MANOVA sonuçlarını yorumlayabilmek için post hoc olarak DDA kullanılan herhangi bir çalışmaya rastlanmamıştır.

MANOVA'yı kullanmak yerine, Türk eğitim dergilerinde yayınlanan birçok eğitim araştırmacısı, çok değişkenli bağımlı değişkenlerden elde edilen toplam puan üzerinde tek bir ANOVA testini uygulamakta veya bağımlı değişkenlerin her biri üzerindeki bağımsız değişken etkisini ayrı ayrı test etmek için birden fazla ANOVA testi kullanmaktadır. Sonuçlar, MANOVA'nın uygulamasının bütün dergi türlerinde oldukça nadir olduğunu göstermektedir. Çalışmaların yarısından fazlası, çok değişkenli bağımlı değişkenlerdeki grup ortalama farklarını test etmek için tek bir MANOVA çalıştırmak yerine birden fazla ANOVA kullanıyor. Ayrıca, çalışmaların yaklaşık %30'u, çoklu bağımlı değişkenlerden elde edilen toplam puanlar üzerinden tek bir ANOVA testi yaparak bağımsız değişkenlerin etkisini ortaya çıkarmaya çalışmaktadır. Bütün bu sonuçlar bize MANOVA'nın teorisinin ve uygulamasının ülkemizdeki eğitim dergilerinde yayın yapan eğitim araştırmacılarınca yeterince anlaşılmadığını göstermektedir.

Lisansüstü eğitim dönemlerinde her ne kadar çok değişkenli verilerin analizini içeren dersler veriliyor olsa da MANOVA açısından bakıldığında, en azından uygulama alanlarının yeterince iyi anlaşılmadığı görülmektedir. Bunun altında yatan temel sebep, istatistiksel veri analizi yöntemlerinin teorik alt yapısından ziyade, belirli istatistiksel paket programlar (ör. SPSS) ile nasıl analiz yapılacağına ilişkin pratik bilgilerin sunuluyor olması olabilir. Eğitimde araştırma yapan bireylerin bu tür eksik ve yanlışlarının giderilmesi için lisansüstü eğitim programlarının istatistiksel analizlerin dayandığı teorik bilgileri kazandırmaya yönelik adımlar atması ve eğitim araştırmacılarının da gerekli gördükleri durumlarda ilgili alanların uzmanlarıyla iş birliğine yönelmeleri faydalı olabilir.

Appendix A. R Code Used for Data Generation and Analyses

```
library(MASS)
d=3   ### dimensions....fixed
N=100    #### sample size....fixed
M_fark=.2   #### group mean differences....variable
variance=.5   #### distribution variance....variable
mincorr=.2   ### minimum correlation....variable
maxcorr=.4   ### maximum correlation....variable
var=matrix(variance,d)
std=sqrt(var)
R=500    ### number of replication....fixed

Pvalues<-matrix(NA,R,d+1)
for (r in 1:R){
    corr=matrix(runif(d,mincorr,maxcorr))
    corMat=matrix(c(1,corr[1,],corr[2,],corr[1,],1,corr[3,],corr[2,],corr[3,],1),ncol=d)
    covMat=std%*%t(std)*corMat
    Mean1=matrix(0,d)
    Mean2=Mean1+M_fark
    data1=mvrnorm(N,Mean1,covMat)
    data2=mvrnorm(N,Mean2,covMat)
    data=rbind(data1,data2)
    ind=rbind(matrix(1,N),matrix(2,N))
    ##### MANOVA
    fit <-summary(manova(data ~ ind) , test="Pillai")
    p<-fit$stats[1:1, "Pr(>F)"]
    p<- matrix(unlist(p), nrow=length(p))
    ##### ANOVAs
    values1=data[,1]
    values2=data[,2]
    values3=data[,3]
    aov_sum1=summary(aov(values1~ind))
    p1=lapply(aov_sum1, function(aov_sum1){aov_sum1$'Pr(>F)' })
    p1<- matrix( unlist(p1), nrow=length(p1) )
    aov_sum2=summary(aov(values2~ind))
    p2=lapply(aov_sum2, function(aov_sum2){aov_sum2$'Pr(>F)' })
    p2<- matrix( unlist(p2), nrow=length(p2) )
    aov_sum3=summary(aov(values3~ind))
    p3=lapply(aov_sum3, function(aov_sum3){aov_sum3$'Pr(>F)' })
    p3<- matrix( unlist(p3), nrow=length(p3) )
    pval=round(cbind(p[,1],p1[,1],p2[,1],p3[,1]), digits=3) ### manova + 3 anovas
    Pvalues[r,]<-pval
    }
ret<-ifelse(Pvalues>=.05,1,0)   #### Retained hypotheses rates of MANOVA and ANOVAs
manret<-ifelse(ret[,1]==1,3,9)
anoret<-ifelse(rowSums(ret[,2:4])==3,3,99)
agree<-as.numeric(sum(ifelse(manret==anoret,1,0)))
disagree<-as.numeric((sum(ret[,1])-agree))
match<-as.matrix(cbind(agree,disagree))
rej<-ifelse(Pvalues<.05,1,0)
anorej<-ifelse(rowSums(rej[,2:4])>0,1,9)
agreement<-as.numeric(sum(ifelse(anorej==rej[,1],1,0)))   ### Not manova but at least one anova is significant
disagreement<-as.numeric((sum(rej[,1]))-agreement)
matching<-as.matrix(cbind(agreement,disagreement ))   #### retained manova while rejected anova
results<-cbind(match,matching)

#Bonferroni corrected
retB<-ifelse(Pvalues>=.0167,1,0)   ### Retained hypotheses rates of MANOVA and ANOVAs
anoret<-ifelse(rowSums(retB[,2:4])==3,3,99)
agree<-as.numeric(sum(ifelse(manret==anoret,1,0)))
disagree<-as.numeric((sum(ret[,1])-agree))
match<-as.matrix(cbind(agree,disagree))
rejB<-ifelse(Pvalues<.0167,1,0)
anorejB<-ifelse(rowSums(rejB[,2:4])>0,1,9)
agreement<-as.numeric(sum(ifelse(anorejB==rej[,1],1,0)))   ### Not manova but at least one anova is significant
disagreement<-as.numeric((sum(rej[,1]))-agreement)
matching<-as.matrix(cbind(agreement,disagreement ))   ### retained manova while rejected anova
resultsB<-cbind(match,matching)
cbind(results,resultsB)
```

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 315-326

Investigation of Item Selection Methods According to Test Termination Rules in CAT Applications *

Sema SULAK ** Hülya KELECİOĞLU ***

Abstract
In this research, computerized adaptive testing item selection methods were investigated in regard to ability estimation methods and test termination rules. For this purpose, an item pool including 250 items and 2000 people were simulated (M = 0, SD = 1). A total of thirty computerized adaptive testing (CAT) conditions were created according to item selection methods (Maximum Fisher Information, a-stratification, Likelihood Weight Information Criterion, Gradual Information Ratio, and Kullback-Leibler), ability estimation methods (Maximum Likelihood Estimation, Expected a Posteriori Distribution), and test termination rules (40 items, SE < .20 and SE < .40). According to the fixed test-length stopping rule, the SE values that were obtained by using the Maximum Likelihood Estimation method were found to be higher than the SE values that were obtained by using the Expected a Posteriori Distribution ability estimation method. When ability estimation was Maximum Likelihood, the highest SE value was obtained from a-stratification item selection method when the test length is smaller then 30. Whereas, Kullback-Leibler item selection method yielded the highest SE value when the test length is larger then 30. According to Expected a Posteriori ability estimation method, the highest SE value was obtained from a-stratification item selection method in all test lengths. In the conditions where test termination rule was SE < .20, and Maximum Likelihood Ability Estimation method was used, the lowest and highest average number of items were obtained from the Gradual Information Ratio and Maximum Fisher Information item selection method, respectively. Furthermore, when the SE is lower than .20 and Expected a Posteriori ability estimation method was utilized, the lowest average number of items was obtained through Kullback-Leibler, and the highest was obtained through Likelihood Weight Information Criterion item selection method. In the conditions where the test termination rule was SE < .40, and ability estimation method was Maximum Likelihood Estimation, the maximum and minimum number of items were obtained by using Maximum Fisher Information and Kullback-Leibler item selection methods respectively. Additionally, when Expected a Posteriori ability estimation was used, the maximum and minimum number of items were obtained via Maximum Fisher Information and a-stratification item selection methods. For the cases where the stopping rule was SE < .20 and SE < .40 and Maximum Likelihood Estimation method was used, the average number of items were found to be highest in all item selection methods.

Key Words: Computerized adaptive testing, maximum fisher information, a-stratification, likelihood weight information criterion, gradual information ratio, kullback-leibler.

INTRODUCTION

Computerized Adaptive Test (CAT) algorithm consists of applying selected items to the examinee in computer environment, estimating examinee ability level through given responses, selecting new items according to the most recent estimated ability, and administrating test until the specified test termination rule is conducted (Orcutt, 2002; Thissen & Mislevy, 2000; Wainer, 2000; Weiss, 1983).

The key questions for CAT are (Wainer, 2000);

- How is the first item selected to start the test?

* The present study is a part of PhD Thesis entitled "Bireyselleştirilmiş Bilgisayarlı Test Uygulamalarında Kullanılan Madde Seçme Yöntemlerinin Karşılaştırılması" completed within Hacettepe University Graduate School of Educational Sciences.

** Assist. Prof. PhD., Bartin University, Faculty of Education, Bartın-Turkey, semasulak@gmail.com, ORCID ID: 0000-0002-2849-321X
*** Prof. PhD., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Sulak, S. & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 315-326. doi: 10.21031/epod.530528

Received: 21.02.2019
Accepted: 06.07.2019

- How are the subsequent items selected from the item pool based on examinee responses, and how is the examinee ability predicted based on given responses?
- How is the test terminated?

There are different methods for selecting the first item to start testing. Either relevant information about examinees (i.e., previous test scores, grades, etc.) are used or a set of items, which do not impact examinees' final scores, are applied to all examinees to determine the first item. (Slater, 2001; Sireci, 2003). The most commonly used ability estimation methods in CAT applications are Maximum Likelihood and Bayesian Based Estimation. The major item selection methods used in CAT applications are Maximum Fisher Information (MFI), a-stratification, Likelihood Weight Information Criterion (LWIC), Gradual Information Ratio (GIR) and Kullback-Leibler (KL). The methods used in this study are explained below.

Maximum Fisher Information

The MFI item selection method aims to find the maximal interim ability to estimate regarding every previously administered item. MFI item selection investigate the i^{th} item that results in the largest value of,

$$I_i[\hat{\theta}_{m-1}]= \frac{(Da_i)^2(1-c_i)}{\left[c_i+e^{Da_i(\hat{\theta}_{m-1}-b_i)}\right]\left[1+e^{-Da_i(\hat{\theta}_{m-1}-b_i)}\right]^2} \quad (1)$$

In the Equation 1, a_i, b_i, and c_i; represent the discrimination, difficulty, and pseudo-guessing parameters in 3PLM respectively, and D stands for the scaling constant, 1.702. (Han, 2010).

Kullback-Leibler

The KL information selection method was developed by Chang and Ying (1996) based on the global knowledge approach. KL information for an item is defined as Equation 2.

$$K_i(\theta||\theta_0)= P_i(\theta_0)\log\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right]+[1-P_i(\theta_0)]\log\left[\frac{1-P_i(\theta_0)}{1-P_i(\theta)}\right] \quad (2)$$

KL information is a function of two variables (θ and θ_0) and is a surface in three-dimensional space. As a function of these two θ levels, KL information characterizes the change capacity of an item between two θ levels.

Likelihood Weight Information Criterion

LWIC item selection method was developed by Veerkamp and Berger (1997). In this method, the information function is collected along the θ scale and weighted by the likelihood function after the administration of the item.

The item to be selected in the LWIC criterion is determined by selecting the item that will maximize the value of the Equation 3.

$$\int_{\theta=-\infty}^{\infty} L(\theta;x_{m-1})I_i[\theta]d\theta \quad (3)$$

a-Stratification

The method of a-stratification item selection is constituted with the suggestion of layering according to the a parameter values in the item pool by Chang and Ying (1999). In this method, items are stratified into K strata based on their a values. Accordingly, the item selection process is divided into K stages. In the first stage, items are selected from the first stratum, which corresponds to the items with the lowest a values. In the second stage, items are selected from the second stratum. In the K^{th} stage, items

are selected within the K^{th} level (Chang, Qian, & Ying, 2001). This method utilizes low a-items at early stages of the test. By doing so, the test precision and efficiency are maintained (Chang & Ying, 1996).

Gradual Information Ratio

The GIR item selection method was developed by Han (2009). Han proposed an alternative method based on the expected item effectiveness to improve the use of item pool instead of MFI method.

Han (2009) proposed to take the item efficacy (expected item information) into account on the item adequacy. Thus, this method looks for the item that makes the following criteria maximum,

$$\frac{I_i[\hat{\theta}_{m-1}]}{I_i[\theta_i^*]}\left(1-\frac{m}{M}\right)+I_i[\hat{\theta}_{m-1}]\frac{m}{M} \quad (4)$$

In Equation 4, M shows the length of the test, and m denotes the number of administered items +1.

There are two test stopping methods in CAT applications; fixed-length tests and standard error termination (Sireci, 2003; Wainer, 2000; Weiss & Kingsbury, 1984). Fixed-length termination rules continue until an examinee takes a predetermined number of items. According to the standard error (SE) termination rule, the exam continues until the estimate of the θ reaches a certain level.

CAT applications have numerous advantages. The most important advantage provided by CAT applications is that the test can be tailored to the examinees' ability level. In order to obtain valid results from CAT applications, it is critical to select the item that maximizes the test information about the examinee. MFI is widely used in CAT applications; however, this method tends to use items with a high a parameter and is insufficient in the ability estimation at the beginning of the test (Van der Linden & Glas, 2010; Wainer, 2000; Weiss, 1983). Veldkamp (2012) stated that it is important to investigate different item selection methods in order to eliminate the aforementioned (proposed) limitations of MFI item selection method. There are researches indicating a-stratification item selection method is preferred to MFI due to selecting high a parameter items (Chang & Ying, 1999; Deng, Ansley, & Chang, 2010; Deng & Chang, 2001). Additionaly, Eggen (1999) found that KL item selection method provides more accurate ability estimation in comparison to MFI. Weissman (2003) stated in his study that ability estimation methods affect item selection methods. Bock and Mislevy (1982) indicated that Expected a Posteriori (EAP) ability estimation method was better than Maximum Likelihood Estimation (MLE) methods; while Wang and Visposel (1998) proposed that EAP ability estimation method was more biased. There are additional researches regarding the relationship between the test termination rules and item selection methods (Han, 2009; Weissman, 2003).

Purpose of the Study

The key point of the item selection process in the CAT applications is to match the ability of the respondent with the difficulty of the item. Namely, in CAT, ability estimation is reperformed after each item is answered, and the most recent ability estimation is used in the selection of subsequent items. MLE and EAP which are among the ability estimation methods were included in the research, and it was attempted to determine how ability estimation methods affect the item selection methods. There are studies suggesting that item selection methods are inadequate (especially when the test length is smaller than five items) at the beginning of CAT applications (Han, 2009; Linda, 1996; Van der Linden & Glas, 2010). According to the literature when the CAT has more than 20 items, the difference in the performance of a newly proposed method and MFI turns out to be trivial (Passos, Berger & Tan, 2007; as cited in Şahin & Özbaşı, 2017). Chen, Ankenmann and Chang (2000) conducted a simulation study to compare item selection methods, and they found that for CATs with more than 10 items, there is no difference between item selection methods. Veerkamp and Berger (1997) conducted a simulation study according to 60 items termination rule and found that item selection performances vary over 20 items. One of the advantages of CAT applications is to shorten the test. An item pool of 60 items was not selected, and an item pool of more than 20 items was used.

Thus, different test lengths (5, 10, 20, 30, and 40 items) were also taken as a variable to determine how the item selection methods differ depending on the test length. In order to compare the item selection methods in CAT applications where the test stopping rule was determined based on a fixed standard error, conditions were established in which the standard error was .20 and .40.

This study aims to answer the following questions:

1) How do standard errors in relation to the methods used in item selection (Maximum Fisher Information, a-stratification, Likelihood Weight Information Criterion, Gradual Information Ratio, and Kullback-Leibler) differ in terms of
   a) test length (5, 10, 20, 30 and 40 items)
   b) ability estimation (Maximum Likelihood and Expected a Posteriori) methods?
2) How do the average number of items utilized in item selection methods (Maximum Fisher Information, a-stratification, Likelihood Weight Information Criterion, Gradual Information Ratio, and Kullback-Leibler) differ in terms of
   a) test termination rules (SE < .20 and SE < .40)
   b) ability estimation methods (Maximum Likelihood and Expected a Posteriori)?

When the literature regarding the current study is reviewed, the following results are found:

In their study, Veerkamp and Berger (1997) compared the Interval Information Criteria and LWIC methods with MFI method, and the authors concluded that these methods did not have a substantial superiority to MFI. Eggen (1999) have compared KL and MFI item selection methods. According to the results of this study, KL item selection method performed better than the MFI. In a simulation study, Wen, Chang and Hau (2001) compared a-stratification item selection method and MFI item selection method. They concluded that MFI item selection method yielded more effective results than a-stratification item selection method. Weissman (2003) investigated the effectiveness of item selection methods in CAT applications. According to the findings, the ability estimation method impacted the effectiveness of item selection more than item selection method. Han (2009) explored random selective MFI, fade-away selective MFI, GIR, and fade-away selective GIR item selection methods in CAT application. It was concluded that MFI and GIR item selection methods exhibited lowest SE through theta criteria. Costa, Karino, Moura and Andrade (2009) evaluated the performance of MFI, KL, and Maximum Expected Information item selection methods. They concluded that all methods performed similarly to estimate examinees' θs by means of bias and mean square error.

Deng et al. (2010) compared MFI, a-stratification, and refined a-stratification item selection methods. The study findings yielded that MFI was more effective in predicting ability in comparison to other methods. Han (2010) compared five different item selection methods, which are a-stratification, Interval Information Criteria, Likelihood Weighted Information Criterion (LWIC), Kullback-Leibler Information, and Gradual Information Ratio (GIR). The study results showed that SE values decreased in all item selection methods due to test length. Low SE values were calculated under MFI, KL and GIR item selection methods, whereas high SE values were calculated under a-stratification item selection methods.

Research findings related to different item selection methods in the literature indicated that item selection methods have strengths as well as weaknesses in different conditions (Deng et al., 2010; Eggen, 2009; Wen, et al., 2001; Yi & Chang, 2003) and two-item selection method were compared. In the studies investigating more than two item selection methods (Han, 2010; Weissman, 2003), stopping rules and ability estimation methods were not elaborated together.

METHOD

The data of the study were simulated by SimulCAT computer program, which was developed by Han (2012). In data collection stages, first, the group where the research was to be carried out, then the item pool and CAT conditions were formed.

Participants

2000 hypothetical examinee were simulated. Examinee ability parameters (N = 2000) were randomly drawn from a normal distribution ~N(0, 1). Dichotomous item responses for the entire item bank were generated using the SimulCAT program (Han, 2012).

Data Collection Instruments

Item pool

An item pool with 250 dichotomously-scored items was created using the three-parameter logistic (3PL) item response model. In his research, Urry (1977) found that an item pool of at least 100 items is adequate to estimate ability. Kingsbury and Zara (1989) indicated that item pool size for adaptive tests should always be -more is better-. Stocking (1992) determined that an item pool size should be 6 to 12 times more than the item number.

Item discrimination parameters were randomly drawn from a uniform distribution ~U(0.8, 1.5); item difficulty parameters were randomly drawn a uniform distribution ~U(-3, 3); guessing parameters were randomly drawn from a uniform distribution ~U(.05, .15). Following the suggestions from previous studies regarding data simulation for the 3PL model, the simulation was conducted. Feinberg and Rubright (2016) indicated 3PL IRT model parameters are often simulated as uniform. Ree and Jensen (1983) said that "a values below 0.5 are insufficiently discriminating for most testing purposes, and a values above 2.0 are infrequently found … most test items have c parameters less than or equal to .30" (pp. 135-146).

Process

The data collection process was simulated using the SimulCAT computer program. As the first step, examinee and item pool files were created and uploaded to the computer program. In the second step, item selection and stopping rules were specified, and in the final step, ability estimation methods, test initiation rule, number of replications and output files were selected. The test initiation rule was determined as $\theta = 0.5$, and 100 replications were performed for all simulation conditions. A crossed-factorial design resulted in a total of 30 simulation conditions; 5 item selection methods * 2 ability estimation methods * 3 stopping rules. For each crossed condition, 100 replications were conducted. The number of replications depends on the research question. However, with too many replications simulation may be more complex and might take a long time to complete (Bulut & Önder, 2017; Feinberg & Rubright, 2016). Because of the 30 conditions, the researcher decided to make 100 replications. Harwell, Stone, Hsu and Kirisci (1996) suggested a minimum of 25 replications and indicated that "aggregating results over replications produces more stable and reliable results" (p. 110). Thus, the simulation study was ended after 100 replications and interim, and final θ values were aggregated over the 100 replications.

Data Analysis

In order to determine how item selection methods differ according to the test length in the CAT conditions, where the stopping rule was specified as 40 items, interim θ and standard error (SE) of the

estimation were calculated for 5, 10, 20, 30 and 40 items. The standard error of the estimation is calculated via the Equation 5.

$$SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}} \quad (5)$$

In the conditions with test stopping rule of SE < .20 and SE < .40, item selection methods were evaluated according to the average number of items. Since CAT administration would terminate at a specific standard error value, the average number of items used until reaching this standard error value was investigated.

RESULTS

To determine how standard error associated with different item selection methods (MFI, GIR, LWIC, a-stratification, KL), the test length (5, 10, 20, 30 and 40 items) and ability estimation methods (MLE and EAP), mean of the interim ability estimations ($\hat{\theta}$) were used in the analysis of the results. Item selection methods were compared according to SE values, and the results are presented Table 1.

Table 1. Statistics Regarding the Item Selection Methods According to the Test Length (For a 40-Item Fixed-Length CAT Administration Where MLE Ability Estimation is Used)

	Test Length									
Item Selection	**5**		**10**		**20**		**30**		**40**	
Methods	**$\hat{\theta}$**	**SE**	**$\hat{\theta}$**	**SE**	**$\hat{\theta}$**	**SE**	**$\hat{\theta}$**	**SE**	**$\hat{\theta}$**	**SE**
MFI	0.12	.55	0.05	.36	0.03	.25	0.02	.20	0.02	.18
a-stratification	-1.55	.78	-1.57	.52	-1.39	.31	-1.29	.23	-1.19	.19
LWIC	-0.60	.74	-0.28	.38	-0.11	.25	-0.62	.21	-0.04	.18
GIR	-1.52	.50	-1.28	.35	-1.26	.25	-1.06	.21	-0.68	.19
KL	-1.6	.67	-1.20	.37	-1.10	.25	-0.57	.22	-0.21	.22

When Table 1 is examined, it is observed that the method of a-stratification item selection shows high SE value in cases where the test length is less than 30 items (n < 30), while the method of KL item selection shows high SE value in cases where the test length is greater than thirty items (n > 30). While the highest SE value that was obtained from the a-stratification item selection method is similar to the results of Han's (2009) research, it differs from Linda's (1996) study which shows that KL item selection method is better than the MFI item selection method.

Considering all item selection methods according to test lengths, it was determined that there was a great difference between the SE values of the item selection methods after administering five items. However, the difference between SE values was decreased after administering ten items. When the inadequacy of MFI item selection method in the predictive estimation at the beginning of the CAT applications (n < 5) was examined, it was found that only the GIR item selection method showed a lower SE value than MFI. These two findings indicated that all of the item selection methods included in this study were limited in their ability estimation at the beginning of CAT applications and that they did not have a significant advantage over MFI item selection method.

Table 2. Statistics on the Methods of Item Selection According to the Test Length in the CAT Conditions Where the Test Stopping Rule is Determined as 40 Items and the EAP Ability Estimation is Used

	Test Length									
Item Selection	**5**		**10**		**20**		**30**		**40**	
Method	**$\hat{\theta}$**	**SE**	**$\hat{\theta}$**	**SE**	**$\hat{\theta}$**	**SE**	**$\hat{\theta}$**	**SE**	**$\hat{\theta}$**	**SE**
MFI	0.01	.47	0.02	.33	0.02	.23	0.02	.20	0.02	.18
a-stratification	0.01	.70	0.01	.49	0.02	.31	0.02	.23	0.02	.18
LWIC	0.01	.55	0.02	.35	0.02	.24	0.02	.20	0.02	.18
GIR	0.01	.49	0.01	.33	0.02	.24	0.02	.20	0.02	.18
KL	0.01	0.47	0.02	0.33	0.02	0.24	0.02	0.20	0.02	0.18

As shown in Table 2, a-stratification item selection method had the highest SE value among all test lengths. When the findings were examined, it was found that there was a substantial difference between the SE values of the item selection methods while the test length was 5 items, but the difference between the SE values was decreased in the CAT conditions where the test length was specified as 10 items and higher. The differences decreased as test length increased, and the results were close to each other. In addition, when the test length reached 40 items, the SE values of the item selection methods were found equal to each other. This significant decrease in all item selection methods at the beginning of the CAT applications (n < 5) was interpreted as the absence of a significant superiority of other item selection methods except for the KL item selection method in the problem of MFI item selection method in terms of ability estimation.

When MLE and EAP ability estimation methods were examined, the highest SE value was obtained from the a-stratification item selection method in both MLE and EAP ability estimation methods. In general, the SE values obtained when the MLE ability estimation was used were found higher than the SE values obtained when EAP ability estimation was used.

The most important difference was detected when the test length was 5 items. For example, the SE value of the KL item selection method was .67 for MLE ability estimation, whereas the SE value was calculated as .47 for EAP ability estimation. Wang and Visposel (1998) found that EAP ability estimation showed a lower SE value compared to MLE ability estimation method.

The findings obtained in the present study align with these results. This finding may indicate that EAP ability estimation method should be primarily preferred especially at the beginning of the test in the application of CAT. In both cases where MLE and EAP ability estimation were used, a sharp decrease in SE values was observed when the test length reached to 10 items from 5 items.

To be able to determine how the average number of items related to item selection methods (MFB, GIR, LWIC, a-stratification, KL) changes according to test stopping rule (SE < .20 and SE < .40) and ability estimation methods (MLE and EAP), the mean number of items was calculated. The findings were presented in Table 3.

Table 3. Statistics of Ability Estimation and Item Selection Methods in CAT Conditions Where the Test Stopping Rule is Based on Fixed Standard Error

	Stopping rule						
Ability Estimation Method	**Item Selection Method**	**SE < .20**			**SE < .40**		
		Minimum Item	**Maximum Item**	**Average Item**	**Minimum Item**	**Maximum Item**	**Average Item**
MLE	MFI	26	95	40.71	7	9	8.72
	a-stratification	-	-	-	13	16	14.65
	LWIC	27	88	32.85	8	13	9.54
	GIR	12	41	31.75	7	10	8.96
	KL	13	38	32.63	8	12	9.72
EAP	MFI	18	124	30.07	6	11	7.07
	a-stratification	-	-	-	12	17	12.54
	LWIC	26	78	31.18	8	9	8.41
	GIR	18	43	30.23	7	12	7.46
	KL	27	48	30.13	6	11	7.16

According to the results on Table 3, the lowest and highest number of items were obtained from GIR and MFI item selection methods respectively in the CAT conditions where the standard error was less than .20 and the MLE ability estimation was used. In the CAT applications where EAP ability estimation was used, the average of the lowest and highest number of items was obtained from KL and LWIC item selection methods. The a-stratification item selection method did not function as expected in both MLE and EAP ability estimates. The computer program could not complete the simulation because no suitable item was found in the item pool. This situation was interpreted as the

insufficiency of the item pool or the small size of the a-parameter range. In the method of a-stratification item selection, the item pool is stratified according to the a parameters, and in the present research, the item pool is divided into three layers. In the literature, studies have been carried out for the various size of item pools.

Wen et al. (2001) determined four layers for an item pool of 360 items and a-stratification item selection method in their research where the parameter value ranges from 0.40 to 2. On the other hand, Costa et al. (2009) were able to use the a-stratification item selection method for a standard error value of .20 using a pool of 246 items. When the existing research was examined, it was considered that keeping a parameter value between 0.80 and 1.5 could be the reason why a-stratification method has not been realized under the condition that the standard error is less than .20 as well as the effect of item pool size.

The average number of items was examined for each ability estimation methods. The mean number of items obtained from CAT conditions using MLE ability estimation was found to be higher than the mean number of items from CAT conditions using EAP ability estimation.

This was interpreted as the ability to estimate EAP ability to obtain shorter tests in CAT applications. In CAT conditions test stopping rule, where standard error is defined as less than .40 and MLE ability estimation is used, the lowest and highest number of items were obtained from MFI and a-stratification item selection methods, respectively regarding the mean number of items. In CAT conditions using EAP ability estimation, MFI and KL item selection method had the lowest value while the method of a-stratification item selection was found to be the highest in terms of the average number of items.

The lowest test length was obtained from MFI, and the highest test length was obtained from a-stratification item selection method in cases where both of the ability estimation methods were used. The a-stratification item selection method requires the highest number of items to achieve the standard error value of .40 may be related that this method selects items by stratification of the item pool.

DISCUSSION and CONCLUSION

In the beginning of CAT conditions, where MLE ability estimation method used, the lowest SE value was obtained from the GIR item selection method after five items administered ($n < 5$). a-stratification item selection method showed the highest SE value while the test length is shorter than 30 items ($n < 30$), and KL showed the highest SE value while the test length is longer than 30 items ($n > 30$). In the beginning of CAT conditions, where MLE ability estimation method used and the number of items was less than 10 ($n < 10$), it was seen that there was a great difference between the SE values of the item selection methods investigated, but this difference decreased as the test length increased.

When using EAP ability estimation, the highest SE values were obtained from a-stratification item selection method for all different test lengths included in the study. At the beginning of CAT conditions where EAP ability estimation method used and the number of items was less than 10 ($n < 10$), it was seen that there was a great difference between the SE values of the item selection methods investigated, but this difference decreased as the test length increased. When the test length was set to 40 items, the SE values of all the item selection methods yielded equal results. The SE values observed when MLE ability estimation was used were found to be higher than the SE values obtained when EAP ability estimation was used.

The lowest item number was obtained from GIR item selection method, and the highest item number was obtained from MFI item selection method when MLE ability estimation was used in the CAT conditions where SE was accepted as $SE < .20$. When EAP ability estimation is used, the lowest mean of the item number is obtained from KL item selection method, and the highest mean of the item number is obtained from the item selection method. In both cases where MLE and EAP ability estimations were used, a-stratification item selection method did not yield meaningful results. It was concluded that this finding was due to insufficient pool size and low level of the parameter value. When the average of the number of items was examined in terms of ability estimation method, it was

concluded that the conditions in which MLE ability estimation was used were higher than those in which EAP ability estimation is used.

When MLE ability estimation was used in the CAT conditions where SE < .40 was used, the lowest average of item number was obtained from MFI item selection method, and the highest average of item number was obtained from KL item selection method. When EAP ability estimation was used, the lowest average of item number was obtained from MFI and KL item selection methods, and the highest average of item number was obtained from a-stratification item selection method. For all of the item selection methods included in the study, the average test length obtained from MLE ability estimation was higher than the average test length obtained from EAP ability estimation. It was concluded that EAP ability estimation shorten the test length. SE values for item selection methods were lower when EAP ability estimation was used. EAP ability estimation is recommended for operational CAT applications. One of the most important advantages of CAT applications is that it produces a shorter test length than paper-based tests. When the results are investigated, it is recommended that EAP ability estimation method can be preferred in CAT applications.

The method of a-stratification item selection did not yield meaningful result in the condition that the test stop rule was SE < .20. This finding shows that further research is needed. It is recommended that future studies may be conducted by determining different item pool sizes and a-parameter values. In addition, the relationship between the number of layers used in the method of a-stratification item selection method may be studied.

Future studies should be carried out to investigate what would happen if there were more constraints placed on the items in the pool, such as, content constraints which may differ how the item pool is conducted. Also, the effect of b parameter value (b-blocking, etc.) on item selection methods can be investigated. In this research, a parameter value range is narrow, and this research can be repeated according to different a parameter range. Different item pool sizes and ability estimation methods can be examined for the same simulative conditions of research. How different item selection methods work in an item pool weighted according to content can be examined. In this study, one-dimensional item response theory is used, in the future studies multi-dimensional item response theory can be used. The present study has been done on the simulation data, and the operational CAT applications can be investigated in future studies.

REFERENCES

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431-444. doi: 10.1177/014662168200600405

Bulut, O., & Sünbül, Ö. (2017). R programlama dili ile madde tepki kuramında monte carlo simülasyon çalışmaları. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266-287. doi: 10.21031/epod.305821

Chang, H.-H, Qian, J., & Ying, Z. (2001). a-stratified multistage adaptive testing with b blocking. *Applied Psychological Measurement, 25*(4), 333-341. doi: 10.1177/01466210122032181

Chang, H.-H, & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213-229. doi: 10.1177/014662169602000303

Chang, H. H, & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222. doi: 10.1177/01466219922031338

Chen, S.-Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*(3), 241-255. doi: 10.1177/01466210022031705

Costa, D., Karino, C., Moura, F., & Andrade, D. (2009, June). *A comparision of three methods of item selection for computerized adaptive testing*. Paper session presented at the meeting of 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Deng, H., & Chang, H. H. (2001, April). *A-stratified computerized adaptive testing with unequal item exposure across strata*. Paper session presented at the American Educational Research Association Annual Meeting 2001. Retrieved from https://www.learntechlib.org/p/93050/

Deng, H., Ansley, T., & Chang, H. H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement, 47*(2), 202-226. doi: 10.1111/j.1745-3984.2010.00109.x

Eggen, T. H. J. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*(3), 249-261. doi: 10.1177/01466219922031365

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49. doi: 10.1111/emip.12111

Han, K. (2009). *Gradual maximum information ratio approach to item selection in computerized adaptive testing* (Graduate Management Admission Council Research Reports No. 09-07). USA.

Han, K. (2010). *Comparision of non-fisher information item selection criteria in fixed length computerized adaptive testing*. Paper session presented at the annual meeting of the National Council on Measurement in Education. Denver, CO.

Han, K. (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement, 36*(1), 64-66. doi: 10.1177/0146621611414407

Harwell, M. R., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125. doi: 10.1177/014662169602000201

Kingsbury, G. G., Zara, A. R. (1989). Procedures for selecting ıtems for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359-375.

Linda, T. (1996, April). *A comparision of the traditional maximum information method and the global information method in cat item selection.* Paper session at annual meeting of the National Council on Measurement in Education, New York, NY.

Orcutt, V. L. (2002, February). *Computerized adaptive testing: Some issues in development.* Paper session at the annual meeting of the Educational Research Exchange. Denton, TX.

Ree, M. J., & Jensen, H. E. (1983). Effects of sample size on linear equating of item characteristic curve parameters, In Weiss, D. (Ed). *New horizons in testing latent trait test theory and computerized adaptive testing* (pp.135-146). London: Academic Press. doi: 10.1016/B978-0-12-742780-5.50017-2

Şahin, A., & Özbaşı, D. (2017). Effects of content balancing and item selection method on ability estimation in computerized adaptive testing. *Eurasian Journal of Educational Research*, *17*(69), 21-36. Retrieved from http://dergipark.org.tr/ejer/issue/42462/511414

Sireci, S. (2003). Computerized adaptive testing: An introduction. In Wall, & Walz (Eds), *Measuring up: Assessment issues for teachers, counselors and administrators* (pp. 684-694). USA: CAPS Press.

Slater, S. C. (2001). *Pretest ıtem calibration within the computerized adaptive testing environment* (Unpublished doctoral dissertation, Graduate School of the University Massachusetts). Retrieved from https://elibrary.ru/item.asp?id=5337539. Amherst.

Stocking, M. L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report No. 93-2). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1993.tb01513.x

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, (Ed.), *Computerized adaptive testing: A primer* (pp. 101-133). Mahwah, NH: Lawrence Erlbaum Associates, Inc.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*(2), 181-196.

Van Der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing, statistics for social and behaviorel sciences*, New York, NY: Springer.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*(2), 203-226. doi: 10.3102/10769986022002203

Veldkamp, B. P. (2012). Ensuring the future of computerized adaptive testing. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.). *Psychometrics in practice at RCEC* (pp. 35-46). Netherlands: RCEC, Cito.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T., & Visposel, W. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109-135.

Weiss, D. J. (1983). Latent trait theory and adaptive testing. In D. J. Weiss (Ed.). *New horizons in testing: latent trait test theory and computerized adaptive testing* (pp. 5-7). New York, NY: Academic Press.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Weissman, A. (2003, April). *Assessing the efficiency of item selection in computerized adaptive testing*. Paper session presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Wen, H., Chang, H., Hau, K. (2001, April). *Adaption of a-stratified method in variable length computerized adaptive testing.* Paper session at the American Educational Research Association Annual Meeting, Seattle, WA.

Yi, Q., Chang, H. (2003). a-stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 359-378. doi: 10.1348/000711003770480084

Bireyselleştirilmiş Bilgisayarlı Test Uygulamalarında Madde Seçme Yöntemlerinin Test Durdurma Kurallarına Göre İncelenmesi

Giriş

Bireyselleştirilmiş Bilgisayarlı Test (BBT) algoritması, seçilen maddelerin bilgisayar ortamında cevaplayıcıya sunulması, verilen cevaplar yoluyla yetenek düzeyinin kestirilmesi, hesaplanan yetenek düzeyine göre yeni maddelerin seçilmesi ve testin durdurma kuralı yerine gelinceye kadar test etme sürecine devam edilmesine göre gerçekleşir (Orcutt, 2002; Thissen & Mislevy, 2000; Wainer, 2000; Weiss, 1983).

Teste başlamak için ilk maddenin seçilmesinde farklı yöntemler vardır. Cevaplayıcı hakkında önceden sahip olunan bilgi (önceki testlerden aldığı puanlar, karne notu vb.) veya BBT uygulamalarına başlamadan önce cevaplayıcıların nihai test puanlarına etki etmeyecek madde setleri, tüm cevaplayıcılara uygulanır ve elde edilen yetenek düzeyi ilk maddenin seçilmesinde kullanılabilir (Sireci, 2003; Slater, 2001). BBT uygulamalarında yaygın olarak kullanılan yetenek kestirim yöntemleri, En Çok Olabilirlik ve Bayes kestirimine dayalı olan yöntemlerdir. BBT uygulamalarında kullanılan belli başlı madde seçme yöntemleri ise, Maksimum Fisher Bilgisi (MFB), Kullback-Leibler Bilgisi (KL), Aralık Bilgisi Ölçütü (ABÖ), Olabilirlik Ağırlıklı Bilgi Ölçütü (OAB), a-tabakalama, Aşamalı Maksimum Bilgi Oranıdır (AMBO). BBT uygulamalarında testi durdurmak için; sabit test uzunluğu ve değişken test uzunluğu olmak üzere iki yöntem vardır (Sireci, 2003; Wainer, 2000; Weiss & Kingsbury, 1984). BBT uygulamalarında MFB yaygın olarak kullanılır; ancak, bu yöntem yüksek a parametresine sahip maddeleri kullanmaya meyillidir ve özellikle testin başlangıcında yetenek kestiriminde yetersiz kalmaktadır (Van der Linden & Glas,2010; Wainer,2000; Weiss, 1984). Bu araştırmada, MFB madde seçme yönteminin yüksek a parametresine sahip maddeleri seçme özelliğinin farklı madde seçme yöntemleri ile karşılaştırılması yapılmıştır.

BBT uygulamalarında madde seçme sürecinin anahtar noktası, cevaplayıcının yeteneği ile madde güçlüğünü eşleştirmektir. Şöyle ki; BBT uygulamalarında her madde cevaplandıktan sonra yetenek kestirimi yapılmaktadır ve bu yetenek kestiriminin sonucu madde seçiminde kullanılmaktadır. Yetenek kestirim yöntemlerinden En Çok Olabilirlik Tahmini (EOT) ve Beklenen Sonsal Dağılım (BSD) araştırmaya dahil edilerek madde seçme yöntemlerini nasıl etkilediği belirlenmeye çalışılmıştır. BBT uygulamalarının başında (özellikle test uzunluğu beş maddeden küçük olduğunda) madde seçme yöntemlerinin yetersiz kaldığı yönünde araştırmalar mevcuttur. Test uzunluğuna bağlı olarak madde seçme yöntemlerinin nasıl farklılaştığını belirlemek için farklı test uzunlukları (5, 10, 20, 30 ve 40 madde) da bir değişken olarak alınmıştır. Testi durdurma kuralının sabit standart hataya bağlı olarak belirlendiği BBT uygulamalarında madde seçme yöntemlerini karşılaştırmak için ise, standart hatanın .20 ve .40 olduğu koşullar oluşturulmuştur. Eldeki araştırmanın amacı yetenek kestirim yöntemi, sabit madde sayısı ve standart hataya dayalı durdurma kuralının madde seçme yöntemlerini nasıl etkilediğini belirlemektir.

Yöntem

Bu araştırma simülatif olarak gerçekleştirilmiştir. 250 maddelik bir madde havuzu, ortalaması 0 ve standart sapması 1 olacak şekilde normal dağılım gösteren 2000 kişi simülatif olarak oluşturulmuştur. BBT koşulları, madde seçme yöntemleri (MFB, KL, OAB, a-tabakalama, AMBO), yetenek kestirim yöntemleri (EOT, BSD) test durdurma kuralları (40 madde, SH < .20 ve SH < .40) olmak üzere toplam otuz koşuldan oluşturulmuştur. Test durdurma kuralı 40 madde olarak belirlenen BBT koşullarında, test uzunluğuna göre madde seçme yöntemlerinin nasıl farklılaştığını bulmak amacıyla interim θ ve tahminin standart hatası (SH) hesaplanmıştır. Test durdurma kuralı SH < .20 ve SH < .40 olan BBT koşullarında, madde seçme yöntemleri, madde sayısına göre değerlendirilmiştir.

Sonuç ve Tartışma

Test uzunluğu 5, 10, 20, 30 ve 40 madde olarak belirlendiği ve EOT yetenek kestiriminin kullanıldığı BBT koşullarında; ilk beş madde kullanıldıktan sonra ($n < 5$) en düşük SH değeri AMBO madde seçme yönteminden elde edilmiştir. Test uzunluğu $n < 30$ iken, a-tabakalama; $n > 30$ iken KL madde seçme yöntemi en yüksek SH değerini göstermiştir. BBT koşullarının başında ($n < 10$), araştırmaya alınan madde seçme yöntemlerinin SH değerleri arasında büyük farklar olduğu, ancak test uzunluğu arttıkça bu farkın azaldığı görülmüştür. BSD yetenek kestirimi kullanıldığında ise; araştırmaya alınan bütün farklı test uzunluklarında en yüksek SH değeri a-tabakalama madde seçme yönteminden elde edilmiştir.

Test uzunluğu 40 madde olduğunda bütün madde seçme yöntemlerinin SH değerleri birbirine eşit sonuçlar vermiştir. EOT yetenek kestirimi kullanıldığında elde edilen SH değerleri, BSD yetenek kestirimi kullanıldığında elde edilen SH değerlerinden daha yüksek bulunmuştur.

$SH < .20$ olduğu BBT koşullarında EOT yetenek kestirimi kullanıldığında en düşük madde sayısı ortalaması AMBO madde seçme yönteminden, en yüksek madde sayısı ortalaması MFB madde seçme yönteminden elde edilmiştir. EOT ve BSD yetenek kestirimlerinin kullanıldığı her iki durumda da a-tabakalama madde seçme yöntemi sonuç vermemiştir. Bu durumun madde havuzu büyüklüğünün yetersiz kalmasından ve araştırmaya alınan a parametre değeri ranjının düşük olmasından kaynaklandığı sonucuna varılmıştır. Madde sayısı ortalamaları, yetenek kestirim yöntemleri bakımından incelendiğinde; EOT yetenek kestiriminin kullanıldığı koşulların, BSD yetenek kestiriminin kullanıldığı koşullardan daha yüksek olduğu sonucuna varılmıştır.

$SH < .40$ olduğu BBT koşullarında EOT yetenek kestirimi kullanıldığında en düşük madde sayısı ortalaması MFB madde seçme yönteminden, en yüksek madde sayısı ortalaması KL madde seçme yönteminden elde edilmiştir. BSD yetenek kestirimi kullanıldığında en düşük madde sayısı ortalaması MFB ve KL madde seçme yöntemlerinden, en yüksek madde sayısı ortalaması a-tabakalama madde seçme yönteminden elde edilmiştir. Araştırmaya alınan bütün madde seçme yöntemleri için, EOT yetenek kestiriminden elde edilen ortalama test uzunluğu, BSD yetenek kestiriminden elde edilen ortalama test uzunluğundan yüksek bulunmuştur. BSD yetenek kestiriminin kullanıldığı BBT uygulamalarında daha kısa testler elde edileceği sonucuna varılmıştır. Madde seçme yöntemlerine ait SH değerleri, BSD yetenek kestirimi kullanıldığında daha düşük sonuç vermiştir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 327-347

Examination of the Reliability of the Measurements Regarding the Written Expression Skills According to Different Test Theories *

Merve YILDIRIM SEHERYELİ ** Şeref TAN ***

Abstract
The aim of the study is to examine the reliability estimations of written expression skills analytical rubric based on the Classical Test Theory (CTT), Generalizability Theory (GT) and Item Response Theory (IRT) which differ in their field of study. In this descriptive study, the stories of the 523 students in the study group were scored by seven raters. CTT results showed that Eta coefficient revealed that there was no difference between the scoring of the raters ($\eta = .926$); Cronbach Alpha coefficients were over .88. GT results showed that G and Phi coefficients were over .97. The students' expected differentiation emerged, the difficulty levels of the criteria did not change from one student to another, and the consistency between the scores among raters was excellent. In the Item Response Theory, parameters were estimated according to Samejima's (1969) Graded Response Model and item discrimination differed according to the different raters. According to b parameters, for all the raters; individuals are expected to be at least -2.35, -0.80, 0.41 ability level in order to be scored higher than 0, 1 or 2 categories respectively with .50 probability. Marginal reliability coefficients were quite high (around .93). The Fisher Z' statistic was calculated for the significance of the difference between all reliability estimates. GT revealed more detailed information than CTT in the explanation of error variance sources and determination of reliability; while IRT provided more detailed information than CTT in determining the item-level error estimations and the ability level. There was a significant difference between the estimated parameters of CTT and GT in interrater reliability ($p < .05$); there was no significant difference between the parameters predicted according to CTT and IRT ($p > .05$).

Key Words: Classical test theory, generalizability theory, item response theory, interrater reliability, reliability, rubric.

INTRODUCTION

Nowadays, the aim of education is to educate individuals as producers of knowledge in line with the needs of society. Individuals who produce knowledge are, at the same time, critical thinkers, problem-solvers, and researchers. In this respect, changing education policies require a change in the measurement and evaluation methods as well. These changes increased the use of evaluation materials and studies related to the higher level of thinking skills (Kutlu, Doğan & Karakaya, 2014).

There are many ways that enable individuals to demonstrate their high-level skills. However, the most important means of transforming abstract thoughts into concrete form is writing or writing skills. Writing is defined as thinking on thinking. It also allows individuals to expand their thoughts by organizing information (Karatay, 2015).

* This article was written by Merve YILDIRIM SEHERYELİ based on her M.A. thesis in May 2018 at the Institute of Educational Sciences at Gazi University under the supervision of Şeref TAN. In addition, a part of the study is presented as a paper: The examination of the reliability of written expression skills rubric according to classical test and generalizability theories. III. INES Education and Social Science Congress, 21st April - 01st May 2018, Antalya.

** Res. Assist., Hasan Kalyoncu University, Faculty of Education, Gaziantep-Turkey, yldrm.mrv.7806@gmail.com, ORCID ID: 0000-0002-1106-5358
*** Prof. PhD., Gazi University, Institute of Educational Sciences, Ankara-Turkey, sereftan4@yahoo.com, ORCID ID: 0000-0002-9892-3369

To cite this article:
Yıldırım-Seheryeli, M. & Tan, Ş. (2019). Examination of the reliability of the measurements regarding the written expression skills according to different test theories. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 327-347. doi: 10.21031/epod.559470

Received: 30.04.2019
Accepted: 22.07.2019

While studies are carried out to measure written expression skills in developed countries in detail, a common study is not carried out on determining the deficiencies of students in this field in our country. Moreover, the lack of a common writing approach in the teaching process also makes it difficult to follow the development of the students' written expression skills. Therefore, it is not possible to identify learning deficiencies and provide constructive feedback regarding these deficiencies (Karatay, 2015). Therefore, the present study investigates the evaluation of the storytelling, which is one of the written expression skills.

It is necessary to obtain a valid and reliable measurement as well as the suitability of the criterion to make a correct decision about the students. As the errors involved in the measurement process decrease, the reliability of the measurement process increases, and therefore, the accuracy of the decision we make about the individual trait measured increases (Köse, 2014). Therefore, measurement theories somehow differ depending on the purpose of use, limits, and how to use the results of measurement, just as Classical Test Theory (CTT), Generalizability Theory (GT) and Item Response Theory (IRT) differ from one another.

According to the CTT, the score a person receives from any test is the observed score, and this score indicates the degree of presence of the property measured by the test. In addition, when some assumptions are met, the observed score is estimated by the sum of a person's true score and the error score. In CTT, this error score is only one score, which is the sum of random errors that are caused by the individual, measurement expert, the environment, the rater, etc. In the GT, which is an extension of CTT and variance analysis, these error sources are included in the measurement processes in order to control them. The greatest advantage of the GT is that it can partition the variances into different error sources. While CTT is concerned with the reliability of measurements obtained from a group of individuals; GT is concerned with generalizing measurements beyond the measurements, materials, and raters obtained from a group of individuals. Thus, with a single analysis, a single reliability estimation can be made in CTT, and the data can be interpreted under reliability and generalizability. The results obtained by the generalizability study in GT prepare the basis for decision studies so that the effect of the changes in raters, number of items, etc. on reliability estimations can be determined. The accurate estimation of the population where all observation conditions and sources of variability take place provides a new perspective on the difference between reliability and validity. However, validity and reliability studies in CTT require different analyzes. While CTT gives us the variability from all error sources as a single estimate, GT provides the opportunity to examine the error sources such as students, items, and raters together and if there is a variation between students they are called *measurement object*. Measurement object may change depending on the purpose of the study, and it can be item or rater. The way that variation sources (fixed or random facet) are chosen determines the generalizability of the source. The source of a fixed variable is limited to the measurement situation. Therefore, it will be difficult to comment on the generalization of the measurement results even if the source of error decreases, and the measurement accuracy increases. In addition, a single reliability coefficient can be estimated in CTT when relative and absolute assessments are to be taken, while two different reliability coefficients can be estimated in GT according to the fact that individuals are compared to other individuals or treated free from the group. Different patterns can be used depending on whether a source of variability is observed in all conditions of the other source of variability in GT. It is possible to make estimations for all sources of variability when using the crossed design only (Brennan, 2000; Cardinet, Johnson & Pini, 2010; Gulliksen, 1950; Güler, Kaya-Uyanık & Taşdelen-Teker, 2012; Shavelson & Webb, 1991).

In the IRT whole-test and item-level analyses are performed with the relationship between ability estimations and response patterns. In IRT the degree of the latent trait in individuals can be calculated with ability estimations independent from items and with item parameters independent from the sample (Atılgan, 2005; Baykul, 2010; Erkuş, Sünbül, Ömür-Sünbül, Aşiret & Yormaz, 2017). IRT estimates item-based error using the response patterns given to each item. For reliability and validity of three parameter model, the parameters of a, b, c, and θ are examined, and the marginal reliability coefficients are estimated (Baker, 2016; DeMars, 2016). IRT, which is based on fixed variability source, has no purpose of generalizing differently from GT. The difficulty of providing IRT with the

assumptions of unidimensionality and local independence also makes it difficult to use this theory (Ayala, 2009; Hambleton & Jones, 1993; Hambleton, Swaminathan & Rogers, 1991; Ostini & Nering, 2006).

Purpose of the Study

The purpose of the present study is to compare the reliability estimation methods based on CTT, GT, and IRT by using written expression skill scores which are one of the high-level thinking skills of the students and to provide a theoretical contribution to the field by determining their superiorities and differences, limitations and assumptions.

This study is also important in terms of providing the assumptions for the three theories and revealing the findings and interpretations about the difficulties and solutions that the researchers may face concerning the applicability of these theories.

The literature shows that the studies comparing the two theories are more in number than the studies comparing the three theories (Brennan, 2011; Güler, 2008). In the studies which CTT and GT have compared the reliability in terms of internal consistency scores that were obtained from the scales, Kendall's concordance coefficient for non-parametric tests in the occurrence of more than two measures, and G and Phi coefficients obtained by using crossed design in GT were calculated. In general, the results showed that the GT has more detailed results than CTT, and when the number of items and raters increased, the generalizability and reliability coefficients increased as well. For future studies, it is suggested that different items, raters or designs may be used for the same analyses and that the results may be compared by doing analyses in IRT (Bağcı, 2015; Büyükkıdık, 2012; Deliceoğlu, 2009; Güler, 2011; Öztürk, 2011; Şalgam, 2016; Yelboğa & Tavşancıl, 2010).

In studies that compare CTT with IRT, it is generally aimed to compare the item parameters, and it has been observed that large-scale study groups were used with the tests with two-category items. Although they are generally similar in item parameters, it is concluded that IRT provides more detailed results than CTT; CTT is useful in pass-fail decisions; IRT is superior in item invariance or individualized test. Although there is not much research based on reliability comparison, the a and b parameters have been examined on the basis of the item, and it has been seen that reliability interpretations are made only on the item and test functions (Çelen & Aybek, 2013; Doğan & Tezbaşaran, 2003; Gelbal, 1994; İlhan, 2016; Kan, 2006; Kelecioğlu, 2001; Kim & Feldt, 2010; Koch, 1983; Köse, 2015; Lee, Torre & Park, 2012; Morales, 2009; Nartgün, 2002; Özdemir, 2004; Özer-Özkan, 2012; Sebille et al., 2010; Sünbül, 2011).

In the studies comparing the GT and IRT, many facet Rasch measurement model (MFRMM) is generally used. While GT is used to obtain the group and general information, MFRMM is used to obtain information about the sources of variability of items. Apart from examining the sources of variability, the estimation of the reliability coefficients for IRT was not mentioned (Arsan, 2012; Kim & Wilson, 2009; Ure, 2011).

The theories to be used vary depending on the purpose of the researchers, the measurement tool, the data collection method, the measurements obtained, the distribution of measurements, the sampling, the purpose for which the measurements are used and the limitations of the theories. However, a common point of view is that using at least two theories together yields more reliable results. This study compares the CTT, GT, and IRT in the reliability estimation of the scores obtained from a scale which is scored polytomously in line with the suggestions of the studies in the literature.

METHOD

In this study, the techniques used in estimations of reliability in CTT, GT, and IRT methods will be compared by using the story writing skill rubric. This study is a descriptive study, as it just presents

the results as it is without questioning causality or making comparisons and without the effort of determining the relationship or the difference (Erkuş, 2017).

Study Group

The study group consisted of 523 primary and secondary school students. The data were collected in the spring of 2017. One school was in Karabük and the other was in Gaziantep. The distribution of students across province and class levels is as follows:

Table 1. Distribution of Students in the Study Group Across Province and Class

	3rd Grade	4th Grade	5th Grade	6th Grade	7th Grade	Total
Karabük	50	28	18	36	26	158
Gaziantep	52	58	98	74	83	365
Total	102	86	116	110	109	523

Two teachers from Bursa, three from Karabük, one from Gaziantep and one from Ankara volunteered for scoring the data. Work experience of teachers varies between two and ten years. One of them is Turkish teacher, five of them are elementary school teachers, and the last one is an assessment expert.

Data Collection Instruments and Procedure

In this study, the students were asked to write a story according to the criteria given in the determined subjects. Since this practice was done within the class hour, the students and the teachers were chosen voluntarily. The themes of the forms were unanimously voted by three academicians who work in the fields of Elementary School Teaching, Turkish Education and Curriculum Development in Education. The theme for 3rd grade is *forest*, for 4th grade is *colors*, for 5th grade is *books*, for the 6th grade is *teacher*, and for 7th grade is *discrimination*.

Written stories were scored by seven raters according to the written expression skill (analytical) rubric. Each of the raters is provided with the necessary training on how to use the rubric. Scoring range is 0-3, and the highest score that can be obtained from the rubric for 11 criteria is 33, and the lowest score is 0.

Data Analysis

IBM SPSS 22 was used for Eta correlation and Cronbach's Alpha (α) coefficients for CTT, Edu-G 6.1e were used for G and Phi (ϕ) coefficients for GT, and Multilog 7.03 was used for a, b_1, b_2, b_3 (b: parameters of step functions) parameters, and information functions for IRT analysis. In order to compare the reliability coefficients, t-test was performed for the significance of the difference between the two correlation coefficients using Fisher's Z transformation in Microsoft Office Excel 2016 program. For normality assumptions, graphs in IBM SPSS 22, skewness and kurtosis coefficients in Microsoft Office Excel 2016 program were examined. The principal components analysis in SPSS 22 for the assumptions of unidimensionality and local independence were calculated. For model-data fit, the differences between observed and expected ratios in Multilog 7.03 program were investigated.

RESULTS

The skewness and kurtosis coefficients were calculated with Microsoft Excel 2016 before starting analysis under CTT. The skewness coefficients of all grade levels are between -0.612 and 0.873. The kurtosis coefficients are between -1.491 and 0.735. In this case, it can be said that the distribution of data is not skewed and that the kurtosis is acceptable. The results reveal a normal distribution.

Before moving on to the sub-problems of the research, descriptive statistics of the total scores which were scored by seven raters are given below.

Table 2. Descriptive Statistics of the Total Scores Across Grade Levels which were Scored by Seven Raters

Grade Levels	Raters	Min	Max	Mean	Std. Deviation	Grade Levels	Raters	Min	Max	Mean	Std. Deviation
3	1	2	32	11.51	6.456	6	1	0	33	17.84	9.059
N = 102	2	2	32	11.51	6.565	N = 110	2	0	33	17.60	9.201
	3	1	32	11.80	6.236		3	0	33	17.70	9.095
	4	2	32	11.83	6.456		4	0	33	17.95	8.931
	5	2	33	12.28	6.692		5	0	33	18.16	9.064
	6	5	31	**16.45**	5.538		6	4	32	**21.01**	6.905
	7	4	32	14.11	6.038		7	4	33	19.89	8.275
4	1	1	33	12.70	8.889	7	1	4	33	23.28	7.277
N = 86	2	1	33	12.42	8.982	N = 109	2	4	33	22.89	7.288
	3	1	33	12.19	9.145		3	4	33	22.89	6.915
	4	1	33	12.83	9.131		4	4	33	23.05	7.099
	5	1	33	12.51	9.176		5	4	33	22.96	6.987
	6	1	31	**17.06**	6.886		6	3	33	**23.22**	7.186
	7	3	33	15.04	6.364		7	3	33	21.31	7.267
5	1	1	32	14.26	8.012	Total	1	0	33	16.10	9.018
N = 116	2	1	33	14.24	8.285	N = 523	2	0	33	15.92	9.080
	3	1	33	13.90	8.092		3	0	33	15.88	8.948
	4	1	33	14.05	8.325		4	0	33	16.11	9.002
	5	1	33	13.92	8.305		5	0	33	16.15	9.032
	6	6	31	**19.80**	6.264		6	1	33	**19.66**	7.007
	7	4	32	18.37	7.198		7	3	33	17.92	7.601

Table 2 shows that the scores given to the students range between 1.00 and 33.00 in the 3rd, 4th, 5th and 7th-grade levels; however, for the 6th grade, it is between .00 and 33.00. In all levels, the 6th rater scored with a higher mean than the other raters, and as the grade levels increased, the means of the scores given by each rater increased as expected. The most homogeneous scoring was done by 6th rater for the 3rd, 5th, 6th grades and all students (total); by 7th rater for the 4th grade; by 3rd rater for the 7th grade.

Results of Classical Test Theory

The Eta correlation coefficient was calculated using the random block ANOVA results for the consistency of the scores of the seven raters. As a result, it was observed the degree of agreement between the raters who scored the story writing skill of each student was $\eta = .926$ for 11 items, which shows us that the fit among the raters is high. However, this correlation coefficient does not provide us whether each rater scored correctly. For this reason, Cronbach's Alpha (α) internal consistency coefficient was calculated for the reliability of the scores given by the seven raters to the writings of 523 students.

Table 3. Cronbach α Internal Consistency Coefficients for Each Rater

Raters						
1	2	3	4	5	6	7
.936	.936	.934	.937	.938	.880	.901

Table 3 reveals that the scores of each rater are quite high (over .88). In particular, the reliability coefficients of the scores of the first five raters and the seventh rater are considerably high (.90 and above).

Results of Generalizability Theory

In order to calculate the variance and percentages obtained by the G study, seven raters (p) were asked to rate the writings of 523 students (b) using 11 criteria (o), and a completely crossed pattern (b×o×p) was applied. The main effects of b, o and p in this pattern and the effects of bo, bp, op, bop are presented in the table below.

Table 4. Estimated Variances in G study and their Percentages in Total Variance

Source of Variance	df	Sum of Squares	Mean Square	Variance	Percentage
b	522	21401.956	41.000	.525	44.5
o	10	1028.432	102.843	.015	1.3
p	6	72.255	12.043	-.006	.0
bo	5220	4715.283	.903	.059	5.0
bp	3132	565.278	.181	-.028	.0
op	60	2808.594	46.810	.089	7.5
bop	31320	15447.874	.493	.493	41.8
Total					100

It was found that the variance value (.525) estimated for the main effect of the student variable (b) explained 44.5% of the total variance. This variance component for the population score shows how the students differ from each other in a systematic way. The highest (the first rank) value of the variance component is the desired outcome.

The percentage of the estimated error variance (.015) for the main effect of the criterion variable (o) is 1.3%. The low value indicates that there is not much variation among item difficulties.

The percentage of total variance estimation of the predicted error variance value for the main effect of the rater (p) was 0% (-.006, negative values are rounded to zero since the variance cannot be negative). This value gives the degree of variation among the scores of the raters. Because this value is zero, it is an indication of the excellent consistency between the scores of the raters.

The error variance component resulting from the student-criterion (bo) interaction is the difference in students' responses from one criterion to another. The estimated variance value (.059) for this interaction accounted for 5% of the total variance. Accordingly, the difficulty levels of the criteria do not differ much from one student to another.

The error variance component (-.028) resulting from the interaction of the student-rater (bp) explains 0% of the total variance. This value indicates that if a rater gave a high score to a student, other raters gave a high score to that student as well.

The error variance value (.089) resulting from the criterion-scoring (op) interaction accounts for 7.5% of the total variance. This value implies the extent to which a rater is strict when scoring a criterion and flexible when scoring another.

The student-criterion-rater (bop) (residual) variability source is the variability caused by the interaction of the student, the criterion, and the raters and by the random errors. This error variance value (.493), which is the second highest, accounts for 41.8% of the total variance. This value is an indicator of the existence of systematic or random variability sources that cannot be measured in this study by the interaction between students, criteria, and raters.

G and Phi coefficients which are estimated as a result of the decision studies performed by doubling the number of criteria and decreasing it by 2, 6; decreasing the number of raters by 2, 4, 5 and increasing it by 1 are given in the table below.

Table 5. G and Phi (ϕ) Coefficients Obtained from the D Study on Measurement of Written Expression Skills of Students

	Number of the Raters									
	2		3		5		7		8	
Number of criteria	G	Phi	G	Phi	G	Phi	G	Phi	G	Phi
5	.896	.878	.922	.907	.943	.932	.953	.944	.956	.947
9	.939	.928	.955	.946	.968	.961	.973	.968	.975	.970
11	.950	.941	.963	.956	.973	.968	**.978**	**.974**	.980	.975
22	.974	.969	.981	.977	.987	.984	.989	.987	.990	.987

Table 5 shows the result of the real application where 11 criteria were scored by seven raters in which the G coefficient is .978, and ϕ coefficient is .974. The table also reveals that ϕ coefficient is smaller than the G coefficient under similar conditions. Due to the high value of the obtained results, instead of examining the increase in the criteria and raters in D studies, it was tried to obtain values closer to .80 to ensure practicality.

While the smallest G and ϕ coefficients were .896 and .878 respectively when there were five criteria and two raters. The biggest G and ϕ coefficients were .990 and .987, respectively when there were 22 criteria and eight raters. G and ϕ coefficients decreased when the number of raters was decreased, and the number of criteria was fixed. G and ϕ coefficients increased when the number of raters was increased. However, G and ϕ coefficients decreasingly increased after a certain number of items and the raters.

Results of Item Response Theory

One of the polychotomous IRT models: Samejima's graded response model (GRM)

First of all, it is necessary to check the assumptions of IRT. The normality distribution of the data was shown in the CTT analyses. In IRT, the assumptions of unidimensionality and local independence were examined.

In order to check the unidimensionality assumption, Principal Components Analysis (PCA) was performed for each of the seven raters. Eigenvalues, lowest factor loads, and explained variance rates are given in the table below.

Table 6. PCA Results for Unidimensionality Assumption regarding the data of Seven Raters

Rater	The eigenvalue of factor 1	The eigenvalue of factor 2	Proportions of eigenvalues	Assumption of unidimensionality	The lowest factor load	The variance explained by a unidimensional model (%)
1	6.726	1.525	4.41	Provided.	.627	61.144
2	6.726	1.502	4.48	Provided.	.605	61.144
3	6.651	1.588	4.19	Provided.	.615	60.466
4	6.757	1.498	4.51	Provided.	.608	61.426
5	6.792	1.462	4.65	Provided.	.605	61.750
6	5.148	2.293	2.25	Not provided.		
7	5.627	2.028	2.77	Not provided.		

Table 6 indicates that the structure has a dominant dimension for the first five raters since the first eigenvalues are more than four times the second eigenvalues (Çokluk, Şekercioğlu & Büyüköztürk, 2014). Data for 6th and 7th raters could not be included in GRM analysis because they did not meet the assumption of unidimensionality.

If the scale shows the unidimensionality, the assumption of local independence is met as well (Crocker & Algina, 2006), which means that the assumption of local independence is met for the first five raters.

When the observed and expected ratios of each item scored by five raters for model data fit were examined, it was found that the maximum residual value was .0321. Uyar, Öztürk-Gübeş and Kelecioğlu (2013) state that the differences between observed rates and expected rates are named as *residual*. Also, they mention that when the residues approach zero the model – data fit is achieved. Table 7 presents the estimated item parameters and their standard errors according to GRM in measuring the written expression skills.

Table 7. Step-Function Parameters and Standard Errors with the Discrimination of the Items of Written Expressions Rubric

	Raters																			
Items	**1**				**2**				**3**				**4**				**5**			
	a	**b_1**	**b_2**	**b_3**	**a**	**b_1**	**b_2**	**b_3**	**a**	**b_1**	**b_2**	**b_3**	**a**	**b_1**	**b_2**	**b_3**	**a**	**b_1**	**b_2**	**b_3**
1	1.45	-1.15	0.40	1.36	1.40	-0.97	0.29	1.44	1.30	-1.29	0.17	1.45	1.41	-1.15	0.25	1.48	1.52	-1.11	0.29	1.42
SE	0.16	0.14	0.11	0.15	0.15	0.13	0.11	0.16	0.14	0.15	0.11	0.18	0.15	0.14	0.11	0.16	0.16	0.14	0.10	0.14
2	1.54	-1.32	0.06	1.21	1.55	-1.28	-0.02	1.18	1.53	-1.25	-0.05	1.06	1.73	-1.31	0.02	1.05	1.67	-1.23	0.00	1.20
SE	0.16	0.15	0.10	0.13	0.17	0.14	0.10	0.13	0.15	0.13	0.09	0.13	0.17	0.13	0.09	0.12	0.16	0.13	0.09	0.12
3	1.71	-1.33	-0.28	0.68	1.60	-1.42	0.24	0.79	1.74	-1.50	-0.41	0.59	1.88	-1.39	-0.30	0.66	1.88	-1.34	-0.28	0.62
SE	0.18	0.14	0.09	0.11	0.16	0.15	0.09	0.11	0.17	0.14	0.09	0.10	0.18	0.13	0.08	0.09	0.18	0.13	0.08	0.09
4	1.57	-2.17	-0.61	0.44	1.36	-2.04	-0.65	0.52	1.37	-2.35	-0.80	0.41	1.41	-2.22	-0.79	0.45	1.53	-2.15	-0.72	0.46
SE	0.18	0.25	0.10	0.10	0.16	0.24	0.12	0.12	0.16	0.25	0.12	0.11	0.16	0.25	0.12	0.11	0.16	0.23	0.11	0.10
5	1.67	-1.47	-0.14	0.71	1.58	-1.47	-0.20	0.77	1.67	-1.44	-0.35	0.63	1.83	-1.43	-0.31	0.70	1.80	-1.41	-0.27	0.70
SE	0.18	0.16	0.09	0.11	0.17	0.16	0.09	0.11	0.17	0.14	0.09	0.11	0.19	0.14	0.08	0.10	0.18	0.14	0.08	0.10
6	3.05	-0.44	0.44	1.33	2.81	-0.46	0.56	1.51	3.05	-0.56	0.32	1.31	3.17	-0.48	0.35	1.24	2.96	-0.47	0.42	1.28
SE	0.25	0.06	0.06	0.08	0.24	0.07	0.07	0.09	0.25	0.06	0.06	0.09	0.28	0.06	0.06	0.07	0.24	0.07	0.06	0.08
7	4.14	-0.46	0.35	1.11	3.81	-0.47	0.31	1.17	3.74	-0.58	0.20	1.05	4.04	-0.50	0.28	1.04	3.85	-0.50	0.32	1.15
SE	0.34	0.05	0.05	0.06	0.31	0.05	0.06	0.06	0.30	0.05	0.05	0.06	0.31	0.05	0.05	0.06	0.31	0.05	0.05	0.06
8	5.98	-0.44	0.39	0.95	5.89	-0.41	0.34	0.98	5.48	-0.52	0.29	0.94	5.30	-0.43	0.38	0.93	5.52	-0.41	0.36	0.97
SE	0.54	0.04	0.05	0.04	0.50	0.04	0.04	0.04	0.48	0.04	0.05	0.05	0.46	0.04	0.05	0.05	0.49	0.04	0.05	0.04
9	5.73	-0.44	0.32	0.92	6.44	-0.39	0.33	0.97	4.94	-0.50	0.20	0.94	6.20	-0.45	0.22	0.90	6.14	-0.42	0.29	0.93
SE	0.47	0.04	0.05	0.04	0.57	0.04	0.05	0.05	0.44	0.05	0.04	0.05	0.57	0.04	0.04	0.04	0.56	0.04	0.04	0.04
10	5.05	-0.44	0.39	0.95	5.17	-0.46	0.37	0.98	4.81	-0.64	0.26	0.90	4.33	-0.59	0.29	0.95	4.91	-0.54	0.33	0.91
SE	0.42	0.04	0.05	0.05	0.46	0.04	0.05	0.05	0.38	0.04	0.05	0.05	0.35	0.05	0.05	0.05	0.41	0.05	0.05	0.05
11	1.38	-1.74	0.70	1.60	1.26	-1.90	0.72	1.82	1.28	-2.00	0.57	1.87	1.33	-1.88	0.64	1.90	1.24	-1.93	0.74	1.87
SE	0.15	0.22	0.12	0.17	0.15	0.25	0.13	0.19	0.14	0.23	0.12	0.21	0.15	0.23	0.12	0.20	0.15	0.25	0.13	0.20

Table 7 demonstrates that a parameters of the 11 items ranged from 1.24 to 6.44 in all raters. Baker (2016) classified the discrimination parameters as very low (0.01-0.34), low (0.35-0.64), moderate (0.65-1.34), high (1.35-1.69), and very high (1.70 and above). As it is given in Table 7, a parameters in all items are high and very high with regard to the five raters. a parameters show the the slope of item characteristic curve in dichotomous IRT. In polychotomous IRT it additionally shows the item information (DeMars, 2010). According to the 1st and 3rd raters, the most informative item is the 8th one. According to the 2nd, 4th, and 5th raters, the 9th criterion is the most informative one. The least informative item was the 11th criterion according to all the raters.

Table 7 shows the parameters of location related to the step functions (the threshold values for the categories). The b parameters indicate the ability levels of individuals who have been scored into the relevant category by the raters with the probability of .50. Individuals need a lower ability level to be scored into a lower category, while a higher level of skills requires higher categories. For all raters, individuals must have a minimum score of -2.35 ability level to score higher than 0 with .50 probability, and a minimum of -0.80 ability level to score higher than category 1, and a minimum of 0.41 ability level to score higher than category 2.

Characteristic curves and information functions are also used to examine the statistical analysis of items. Item characteristic curves and information functions of the 1st rater are given in appendix A as an example.

Appendix A shows the graphs of items 3 and 8. Since the curves of item 8 are higher (5.98) than the curves of item 3, the discrimination for item 8 is higher; the curves of item 3 are more skewed; therefore, the discrimination for item 3 is lower (1.71).

b parameters of item 8 show that individuals are expected to be at the ability levels of (-∞, about -0.60), [-0.60, about 0.40), [0.40, about 1.20) or [1.20, + ∞) in order to be scored into 0, 1, 2 or 3 categories respectively with .50 probability. Item information function of item 8 revealed that the ability levels in which the item gives the most information are approximately between -0.60 and 1.20.

b parameters of item 3 show that individuals are expected to be at the ability levels of (-∞, about -1.30), [-1.30, about -0.30), [-0.30, about 0.50) or [0.50, + ∞) in order to be scored into 0, 1, 2 or 3 categories respectively with .50 probability. Item information function of item 3 revealed that the ability levels in which the item gives the most information are approximately between -1.50 and 1.00.

In addition to the parameters, the test information function, which is the sum of the contribution of each item to the test, and the marginal reliability coefficient are calculated under IRT. The test information functions of the five raters are given in appendix B.

In appendix B, the figures indicate that even though the test information functions for each rater changes depending on the ability levels, they are relatively higher for individuals with varying ability levels between -1.00 and 1.50. As the amount of information in test information functions increases, the standard error decreases. Then, for individuals who have the ability between -1.00 and 1.50, the measurement results are estimated with fewer errors. As the test information increases, the error level decrease and vice versa.

The marginal reliability coefficient is the coefficient of reliability that is estimated for the whole scale. It takes a value between 0-1; as you get closer to 1, the reliability of the scores obtained from the scale increases. The marginal reliability coefficients of the five raters are given below.

Table 8. Marginal Reliability Coefficients of Five Raters

Raters				
1	2	3	4	5
.9313	.9304	.9313	.9330	.9330

Table 8 shows that all coefficients are around .93 and the reliability is quite high. The Cronbach's Alpha coefficients in CTT for each rater were compared with the marginal reliability coefficients in IRT. In order to compare the Cronbach's Alpha coefficients in the CTT with the coefficients in the GT, the median of the Cronbach Alpha (α) coefficients of the seven scores was calculated.

Table 9. α (median), Eta, G and Phi Coefficients for All Students

α (median)	Eta	G	Phi
.936	.926	.978	.974

The four coefficients in Table 9 are compared in pairs with Fisher's Z test, with .95 probability (.05 significance level). The Z test statistic results for Fisher's Z values and their significance (p) levels are given in the table below:

Table 10. Z Test Results for Fisher's Z Values for Four Coefficients

Fisher Z Coefficients		***α* (median)**	**Eta**	**G**
		1.705	1.35	2.249
Eta	1.35	1.20		
G	2.25	-8.78*	-9.98*	
Phi	2.17	-7.42*	-8.62*	1.36

*p < .05

Table 10 shows that there is no significant difference between the G and Phi coefficients (Z = 1.36, p > .05), and α and Eta correlation coefficients (Z = 1.20, p > .05), while there were significant differences at .05 level between α and G, α and Phi, G and Eta, and finally Phi and Eta correlation coefficients.

According to Table 3 and 8, when the two coefficients were compared with the Z test statistic performed by Fisher's Z conversion, the results obtained with .95 confidence (.05 significance level) are given in the table below:

Table 11. The Results of the Stability Test of the Fisher Z Values of Two Coefficients

Coefficients	**Raters**				
	1	**2**	**3**	**4**	**5**
Fisher Z (α)	1.7047	1.7047	1.6888	1.713	1.721
Fisher Z (Marginal reliability)	1.6681	1.6614	1.6681	1.681	1.681
Z test statistics	.5909	.6996	.3345	.513	.646
p values	.55	.48	.74	.61	.52

Table 11 shows that there is no significant difference between the stability coefficients at .05 level. In this case, the same results were obtained for inter-rater reliability in both CTT and IRT.

DISCUSSION and CONCLUSION

According to the CTT, the Eta correlation coefficient was estimated for the seven raters, and it was seen that the raters' scoring consistency were high. Cronbach α reliability coefficients were high in the internal consistency of the test scores of seven raters. These findings yielded similar results with Cronbach's internal consistency coefficients calculated over .77 in the studies of Bağcı (2015), Büyükkıdık (2012), Deliceoğlu (2009), Güler (2008), Öztürk (2011) and Yelboğa (2007). However, in Güler's (2011) study, the coefficient was very low. Güler (2011) stated that the reason for this result was the purpose of the study and that random data with low validity and reliability were used.

The estimated parameter values in the measurement of written expression skill under GT are explained below.

The error variances and the percentage of total variance estimations that were estimated as a result of the G study of the b×o×p design, in which student (b), criterion (o) and rater (p) variability sources were crossed, were examined.

- It is possible to say that the scoring revealed the variability between the students.
- The criteria do not differ too much from each other as easy, medium, and difficult.
- The consistency between the scoring of the raters is excellent.
- It can be said that the difficulty levels of the criteria do not differ very much from one student to another.
- Students who got high scores from one rater got high scores from others as well.

- Raters can be very strict when scoring a criterion and can be very generous when scoring another. In this study, it was revealed that there are unexplained systematic or random variability sources by design.

These results comply with the results of the studies of Arsan (2012), Brennan (2011), Büyükkıdık (2012), Güler (2008), Deliceoğlu (2009), Şalgam (2016) and Yelboğa (2007). These studies were conducted with a completely crossed design; the number of participants ranged from 72 to 397 and the number of raters ranged from 2 to 9, and data were obtained using likert scales, holistic and analytical rubrics.

G and Phi coefficients obtained as a result of the decision study (D) by increasing and decreasing the number of scoring and criterion in the b×o×p design were examined.

As a result of the real implementation of 11 criteria scored by seven raters, the coefficients G are over .96, and the ϕ coefficients are estimated to be over .95. At the same time, ϕ coefficient was found to be smaller than the G coefficient under similar circumstances as it should be theoretically. Due to the high value of the obtained results, instead of examining the increase in the criteria and raters in D studies, it was tried to obtain values closer to .80 to ensure practicality. These results differ with the studies of Güler (2011) and Öztürk (2011), which had low values of G and Phi coefficients. The reasons for this difference are the fact that Öztürk (2011) used observation form and Güler (2011) used the random data which had low level of reliability and validity.

In this case, GT yields more detailed results than CTT by separating the sources of variability and providing both separate (main) and interactive results including students, criteria, and raters. The literature shows that Çelen and Aybek (2013), Doğan and Tezbaşaran (2003), Gelbal (1994), Kan (2006), Kelecioğlu (2001), Lee et al. (2012), Morales (2009), Nartgün (2002), Özdemir (2004) estimated parameters using dichotomous IRT models with achievement tests or simulated data. Arsan (2012), Ilhan (2016), Kim and Wilson (2009), Özer-Özkan (2012), Sünbül (2011), Ure (2011) estimated parameters using polychotomous-based Rasch model.

According to GRM, a parameters of 11 items ranged from 1.24 to 6.44 in all raters and discriminations of items for each rater and information that items provide are high. According to the 1st and 3rd raters, the item that gives the most information is the 8th criterion, and according to the 2nd, 4th and 5th raters, it is 9th criterion that gives the most information. The least informative item was the 11th criterion in all the raters. Of Koch's (1983), Köse (2015), Nartgün's (2002) and Özdemir's (2004) studies, which are based on polychotomous IRT model, the highest value of item discrimination is 3.34, estimated for the first item of the sample consisted of all males in Nartgün's (2002) study. In this study, the reason for the fact that discrimination value is 6.44 can be because of the academic achievement levels of the schools in the sample, the familiarity of the students to the written expression studies and the inclusion of all students between the 3rd and 7th grades.

Although the difficulty levels of the items do not differ much according to the GT, b parameters according to IRT vary between -2.35 and 1.90 and θ levels vary from -0.50 to 1.20.

It was revealed that marginal reliability coefficients were quite high (around .93). This finding is very close to the marginal reliability coefficients obtained by Köse (2015) and Nartgün (2002) (.97 and .93), whereas it differs from the coefficients (between .65 to .73) obtained by Özdemir (2004). Apart from the marginal reliability coefficient, a single coefficient for reliability in IRT was calculated in Morales (2009) (Person reliability .95) and Çelen and Aybek (2013) (Empirical reliability .80).

Similar to CTT and GT, the reliability of the scores of the five raters in the IRT was high. As a result, the reliability estimates obtained from the three reliability theories used for our measurements were all very high.

The reliability estimates of the three measurement theories used in this study were examined in two ways. There was a significant difference between α and G, α and Phi, G and Eta, Phi and Eta coefficients ($p < .05$) at each grade level and all students in favor of GT. In this case, CTT and GT coefficients differed in reliability estimation. The literature in this field shows that there has been no

analysis for the significance of the difference between the correlation coefficients in the studies comparing CTT to GT.

In the second part, Cronbach Alpha coefficients in CTT and marginal reliability coefficients in IRT were compared. There was no significant difference between the coefficients ($p < .05$). Similar results were obtained for inter-rater reliability in CTT and IRT. Nartgün (2002) examined the difference between Cronbach's Alpha internal consistency coefficient and the marginal reliability coefficient with Fisher Z transformation and found no significant difference. In contrast to this study, Doğan and Tezbaşaran (2003) examined the significance of the difference between item discriminations and difficulties in CTT and IRT with Fisher's Z transformation and concluded that there was no significant difference.

As a result of the present study which aimed to estimate the reliability of the measurements, it was revealed that when the number of samples is at least 500 and the unidimensionality-local independence assumptions are met, making item-level error estimations with Samejima's (1969) Graded Response model and making reliability estimates through item and test information functions in IRT provide more detailed information than those provided by CTT. Unlike CTT, when the number of samples is less than 500 and the variability sources are more than two, it is possible to calculate the generalizability and reliability coefficients, which differ based on the absolute and relative decisions, by examining the error variances separately and together using GT. In studies in which there is a single source of variability, the use of CTT is more useful if there are pass-fail decisions or when the researcher has a purpose of ranking.

REFERENCES

Arsan, N. (2012). *Buz pateninde hakem değerlendirmelerinin genellenebilirlik kuramı ve Rasch modeli ile incelenmesi* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Atılgan, H. (2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama, 4*(7), 95-108.

Ayala, R. J. (2009). *The theory and practice of item response theory*. USA: The Guildford Press.

Bağcı, V. (2015). *Matematiksel muhakeme becerisinin ölçülmesinde klasik test kuramı ile genellenebilirlik kuramındaki farklı desenlerin karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Baker, F. B. (2016). *Madde tepki kuramının temelleri* (Çev. N. Güler ve M. İlhan). Ankara: Pegem Akademi.

Baykul, Y. (2010). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* Ankara: Pegem Akademi.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement And Education, 24*, 1-21. doi: 10.1080/08957347.2011.532417

Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339–353.

Büyükkıdık, S. (2012). *Problem çözme becerisinin değerlendirilmesinde puanlayıcılar arası güvenirliğin klasik test kuramı ve genellenebilirlik kuramına göre karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. USA: Routledge-Taylor & Francis Group.

Çelen, Ü., & Aybek, E. C. (2013). Öğrenci başarısının öğretmen yapımı bir testle klasik test kuramı ve madde tepki kuramı yöntemleriyle elde edilen puanlara göre karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 4*(2), 64-75.

Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2014). *Sosyal bilimler için çok değişkenli istatistik spss ve lisrel uyglamaları.* Ankara: Pegem Akademi.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* USA: Cengage Learning.

Deliceoğlu, G. (2009). *Futbol yetilerine ilişkin dereceleme ölçeğinin genellenebilirlik ve klasik test kuramına dayalı güvenirliklerinin karşılaştırılması* (Yayımlanmış doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

DeMars, C. (2016). *Madde tepki kuramı* (Çev. E. H. Özberk ve H. Kelecioğlu). Ankara: Nobel Akademi.

Doğan, N., & Tezbaşaran, A. A. (2003). Klasik test kuramının ve örtük özellikler kuramının örneklemler bağlamında karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 25*, 58-67.

Erkuş, A. (2017). *Bilimsel araştırma süreci.* Ankara: Seçkin.

Erkuş, A., Sünbül, Ö., Ömür Sünbül, S., Aşiret, S., & Yormaz, S. (2017). *Psikolojide ölçme ve ölçek geliştirme II.* Ankara: Pegem Akademi.

Gelbal, S. (1994). p madde güçlük indeksi ile Rasch modelinin b parametresi ve bunlara dayalı yetenek ölçüleri üzerine bir karşılaştırma. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 10*, 85-94.

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve Rasch modeli üzerine bir araştırma* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Güler, N. (2011). Rastgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması. *Eğitim ve Bilim, 36*(162), 225-234.

Güler, N., Kaya-Uyanık, G., & Taşdelen-Teker, G. (2012). *Genellenebilirlik kuramı.* Ankara: Pegem Akademi.

Gulliksen, H. (1950). *Theory of mental tests.* USA: John Wiley & Sons.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement Issues and Practice, 12*(3), 38-47. doi: 10.1111/j.1745-3992.1993.tb00543.x

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

İlhan, M. (2016). Açık uçlu sorularla yapılan ölçmelerde klasik test kuramı ve çok yüzeyli Rasch modeline göre hesaplanan yetenek kestirimlerinin karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 31*(2), 346-368. doi: 10.16986/HUJE.2016015182

Kan, A. (2006). Klasik test teorisine ve örtük özellikler teorisine göre kestirilen madde parametrelerinin karşılaştırılması üzerine ampirik bir çalışma. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 2*(2), 227-235.

Karatay, H. (2015). Süreç temelli yazma modelleri: 4+1 Planlı yazma ve değerlendirme modeli. M. Özbay (Ed.), *Yazma eğitimi* içinde (s. 21-48). Ankara: Pegem Akademi.

Kelecioğlu, H. (2001). Örtük özellikler teorisindeki b ve a parametreleri ile klasik test teorisindeki p ve r istatistikleri arasındaki ilişki. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 20*, 104-110.

Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review Journal, 11*(2), 179-188. doi: 10.1007/s12564-009-9062-8

Kim, S., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and many-facet rasch measurement. *Journal of Applied Measurement, 10*(4), 408-422.

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7*(1), 15-32. doi: 10.1177/014662168300700104

Köse, A. (2014). Ölçmede güvenirlik. R. N. Demirtaşlı (Ed.), *Eğitimde ölçme ve değerlendirme* içinde (s. 86-109). Ankara: Edge Akademi.

Köse, A. (2015). Aşamalı tepki modeli ve klasik test kuramı altında elde edilen test ve madde parametrelerinin karşılaştırılması. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi, 15*(2), 184-197.

Kutlu, Ö., Doğan, C., & Karakaya, İ. (2014). *Ölçme ve değerlendirme performansa ve portfolyoya dayalı durum belirleme.* Ankara: Pegem Akademi.

Lee, Y.-S., Torre, J. d., & Park, Y. S. (2012). Relationships between cognitive diagnosis, CTT, and IRT indices: An empirical investigation. *Asia Pacific Educ. Rev. 13*(2), 333–345. doi: 10.1007/s12564-011-9196-3

Morales, R. A. (2009). Evaluation of mathematics achievement test: A Comparison between CTT and IRT. *The International Journal of Educational and Psychological Assessment, 1*(1), 19-26.

Nartgün, Z. (2002). *Aynı tutumu ölçmeye yönelik likert tipi ölçek ile metrik ölçeğin madde ve ölçek özelliklerinin klasik test kuramı ve örtük özellikler kuramına göre incelenmesi* (Yayımlanmamış doktora tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models.* USA: Sage Publication.

Özdemir, D. (2004). Çoktan seçmeli testlerin klasik test teorisi ve örtük özellikler teorisine göre hesaplanan psikometrik özelliklerinin iki kategorili ve ağırlıklandırılmış puanlanması yönünden karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi, 26*, 117-123.

Özer-Özkan, Y. (2012). *Öğrenci başarılarının belirlenmesi sınavından (ÖBBS) klasik test kuramı, tek boyutlu ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması* (Yayımlanmamış doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Öztürk, M. E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların genellenebilirlik ve klasik test kuramına göre karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Sosyal Bilimler Enstitüsü, Ankara.

Şalgam, A. (2016). *Kısa cevaplı matematik yazılı sınavının genellenebilirlik kuramı ve test tekrar test yöntemiyle güvenirliğinin kıyaslanması.* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores [Monograph]. *Psychometrika, 34*(4, Pt. 1). doi: 10.1007/BF03372160

Sebille, V., Hardouin, J.-B., Neel, T. L., Kubis, G., Boyer, F., Guillemin, F., & Falissard, B. (2010). Methodological issues regarding power of classical test theory (CTT) and item response theory (IRT)-based approaches for the comparison of patient-reported outcomes in two groups of patients - A simulation study. *BMC Medical Research Methodology, 10*. doi: 10.1186/1471-2288-10-24

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A Primer.* USA: Sage Publications.

Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda, madde parametrelerinin değişmezliğinin klasik test teorisi, tek boyutlu madde tepki kuramı ve çok boyutlu madde tepki kuramı çerçevesinde incelenmesi* (Yayımlanmamış doktora tezi). Mersin Üniversitesi Eğitim Bilimleri Enstitüsü, Mersin.

Ure, A. C. (2011). *The effect of raters and rating conditions on the reliability of the missionary teaching assessment* (Unpublished master thesis). University of Brigham Young, USA.

Uyar, Ş., Öztürk-Gübeş, N., & Kelecioğlu, H. (2013). PISA 2009 tutum anketi madde puanlarının aşamalı tepki modeli ile incelenmesi. *Eğitim ve Öğretim Araştırmaları Dergisi, 2*(4), 125-134.

Yelboğa, A. (2007). *Klasik test kuramı ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi* (Yayımlanmış doktora tezi). Ankara Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Yelboğa, A., & Tavşancıl, E. (2010). Klasik test ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi. *Kuram ve Uygulamada Eğitim Bilimleri, 10*(3), 1825-1854.

Yazılı Anlatım Becerisine İlişkin Ölçümlerin Güvenirliğinin Farklı Test Kuramlarına Göre İncelenmesi

Giriş

Günümüzde eğitimin amacı kişileri, toplumun ihtiyaçları doğrultusunda, tüketen değil bilgiyi üreten bireyler olarak yetiştirmektir. Bilgi üretebilen nitelikteki kişilerin sorunları çözebilen, sorgulayan, üst düzey ve eleştirel düşünebilen, araştırma-geliştirme becerisine sahip ve yaratıcı bireyler olması gerekmektedir. Bireylerin üst düzey becerilerini ortaya koymaları sağlayan birçok araç vardır. Fakat soyut durumdaki düşünceleri somutlaştırarak incelenebilir hâle dönüştüren en önemli araç yazma ya da yazılı anlatım becerisidir. Yazma, düşünme üzerine düşünme olarak tanımlanmaktadır. Ayrıca bireylerin bilgiyi düzenleyerek düşüncelerini genişletmelerini sağlamaktadır (Karatay, 2015).

Gelişmiş ülkelerde yazılı anlatım becerilerinin detaylı olarak ölçülmesi için çalışmalar yapılırken ülkemizde henüz öğrencilerin bu yöndeki eksikliklerinin belirlenmesi üzerine ortak bir çalışma yapılmamaktadır. Ölçmenin yanında öğretim sürecinde de ortak bir yazma yaklaşımının olmaması, öğrencilerinin yazılı anlatım becerilerinin gelişimlerinin takip edilmesini de güçleştirmektedir. (Karatay, 2015). Bu nedenle, bu çalışmada da yazılı anlatım becerilerinden biri olan hikâye yazma becerilerinin değerlendirilmesi konu edinilmiştir.

Öğrenciler hakkında doğru karar verebilmek (değerlendirme yapmak) için ölçütün uygunluğunun yanı sıra geçerli ve güvenilir bir ölçüm de elde etmek gerekmektedir. Ölçme işlemine karışan hatalar azaldıkça ölçme işleminin güvenirliği dolayısıyla da bireyde ölçülen özellik hakkında verdiğimiz kararın doğruluğu artmaktadır (Köse, 2014). Bu nedenle hata teorileri (kuramlar); kullanım amacına, sınırlıklarına, ölçme sonuçlarının ne şekilde kullanılacağına göre Klasik Test Kuramı (KTK), Genellenebilirlik Kuramı (GK) ve Madde Tepki Kuramı (MTK) gibi farklılaşmıştır.

Araştırmacıların, araştırmanın amacına, kullanılan ölçme aracına, veri toplama yöntemine, elde edilen ölçümlere, ölçümlerin dağılımına, örnekleme, ölçümlerin hangi amaçla kullanılacağına, kuramların sınırlıklarına bağlı olarak kullanılması önerilen kuramlar da değişmektedir. Ortak bir bakış açısı, en az iki kuramın birlikte kullanılmasının daha güvenilir sonuçlar ortaya koyduğu yönündedir. Bu araştırma ile öğrencilerin üst düzey düşünme becerilerinden biri olan yazılı anlatım becerisi puanları kullanılarak KTK, GK ve MTK'ye dayalı güvenirlik kestirme yöntemlerinin karşılaştırılması, birbirlerine göre üstünlükleri ve farkları, sınırlıkları ve sayıltıları belirlenerek alana kuramsal bir katkı

sağlanması hedeflenmektedir. Bu çalışma, incelenen üç kuram için sayıltıların sağlanması ve bu kuramların uygulanabilirliğine yönelik olarak araştırmacıların karşılaşabileceği güçlükler ve çözüm yollarına yönelik bulgu ve yorumların yapılması bakımından da önem taşımaktadır.

Yöntem

Araştırmanın çalışma grubunu 2017 yılı bahar döneminde Karabük ve Gaziantep'te bulunan birer okulda öğrenim gören toplam 523 ilkokul ve ortaokul öğrencisi oluşturmaktadır. Bu öğrencilerin 102'si 3. sınıfta, 86'sı 4. sınıft, 116'sı 5. sınıfta, 110'u 6. Sınıfta ve 109'u 7. sınıfta öğrenim görmektedir.

Çalışma grubunda verileri puanlamak için Bursa'dan 2, Karabük'ten 1, Gaziantep'ten 1 ve Ankara'dan 1 kişi olmak üzere toplam 7 öğretmen gönüllü olmuştur. Öğretmenlerin iş tecrübesi 2 ile 10 yıl arasında farklılaşmaktadır. Öğretmenlerimizden biri Türkçe, beşi sınıf öğretmeni ve biri ölçme değerlendirme uzmanı olarak görev yapmaktadır.

Veri toplama araçları

Bu çalışmada öncelikle öğrencilerden belirlenen konularda verilen ölçütlere göre hikâye yazmaları istenmiştir. Bu uygulama ders saati içinde yapıldığından, öğrencilerin ve öğretmenlerin seçilmesinde gönüllülük esas alınmıştır. Formların temaları Sınıf Öğretmenliği, Türkçe Eğitimi ve Eğitimde Program Geliştirme alanlarında çalışmalar yapan üç akademisyen tarafından oy birliği ile 3. sınıf için *orman*, 4. sınıf için *renkler*, 5. sınıf için *kitaplar*, 6. sınıf için *öğretmen*, 7. sınıf için *ayrımcılık* olarak belirlenmiştir.

Yazılan hikâyeler, yazılı anlatım becerisi (analitik) puanlama anahtarına göre yedi puanlayıcı tarafından puanlanmıştır. Puanlayıcıların her birine puanlama anahtarının nasıl kullanılacağı ile ilgili gerekli eğitimler verilmiştir. 0-3 arasında yapılan puanlamada 11 ölçüt için puanlama anahtarından alınabilecek en yüksek puan 33 en düşük puan 0 olarak belirlenmiştir.

Veri analizi

Güvenirlik belirlemede KTK'de Eta korelasyon ve Cronbach Alfa (α) katsayıları için SPSS 22; GK'de G ve Phi (ϕ) katsayıları için Edu-G 6.1e ve MTK'de a, b_1, b_2, b_3 (b: adım fonksiyonlarının parametreleri) ve θ parametreleri ile bilgi fonksiyonları için Multilog 7.03 programları kullanılmıştır. Elde edilen güvenirlik katsayılarının karşılaştırılması için ise Microsoft Ofis Excel 2016 programında Fisher'in Z dönüştürmesi kullanılarak iki korelasyon katsayısı arasındaki farkın manidarlığı için t testi yapılmıştır. Normallik sayıltısı için SPSS 22'de grafikler, Microsoft Ofis Excel 2016 programında çarpıklık ve basıklık katsayıları; tek boyutluluk ve yerel bağımsızlık sayıltıları için yine SPSS 22'de temel bileşenler analizi; model-veri uyumu için ise Multilog 7.03 programında gözlenen ve beklenen oranlar arasındaki farklar incelenmiştir.

Sonuç ve Tartışma

KTK'ye göre, yedi puanlayıcı için puanlayıcılar arasındaki Eta korelasyon katsayısı hesaplanmıştır ve puanlayıcıların öğrencileri puanlamadaki uyumlarının yüksek olduğu görülmüştür. Yedi puanlayıcının da test puanlarının iç tutarlılık olarak Cronbach α güvenirlik katsayıları yüksek bulunmuştur. Bu bulgular Bağcı (2015), Büyükkıdık (2012), Deliceoğlu (2009), Güler (2008), Öztürk (2011) ve Yelboğa' nın (2007) çalışmalarında .77'nin üzerinde hesapladıkları Cronbach α iç tutarlılık katsayıları ile benzer sonuçlar vermiş, farklı olarak Güler'in (2011) rastgele veriler üreterek yaptığı çalışmada çok düşük düzeyde bulunmuştur. Güler (2011) bu durumun sebebinin çalışmanın amacından kaynaklandığını, düşük geçerlik ve güvenirliğe sahip rastgele verilerin kullanıldığını belirtmiştir.

GK'ye göre yazılı anlatım becerisinin ölçülmesinde kestirilen parametre değerleri aşağıda açıklanmıştır.

Öğrenci (b), ölçüt (ö) ve puanlayıcı (p) değişkenlik kaynaklarının tümüyle çaprazlandığı b×o×p deseninin G çalışması sonucunda kestirilen hata varyansları ve toplam varyansı açıklama yüzdeleri incelenmiştir.

- Yapılan puanlamaların öğrenciler arasındaki farklılaşmayı ortaya çıkardığını söylemek mümkündür.
- Ölçütler kolay, orta ve zor gibi birbirinden güçlük bakımından çok fazla farklılaşmamaktadır.
- Puanlayıcıların puanlamaları arasındaki tutarlılık mükemmel düzeydedir.
- Ölçütlerin güçlük düzeylerinin bir öğrenciden diğerine çok büyük farklılıklar göstermediği söylenebilir.
- Bir puanlayıcının yüksek puan verdiği öğrenciler diğer puanlayıcılardan da yüksek puan almıştır.
- Puanlayıcıların bir ölçütü puanlarken çok katı, diğer ölçütte ise cömert olabildikleri görülmektedir. Bu çalışmada ölçülemeyen sistematik ya da tesadüfi değişkenlik kaynaklarının bulunduğu saptanmıştır.

Bu sonuçlar Arsan (2012), Brennan (2011), Büyükkıdık (2012), Güler (2008), Deliceoğlu (2009), Şalgam (2016) ve Yelboğa'nın (2007) tümüyle çaprazlanmış desende 72 ile 397 arasında birey, 2 ile 9 arasında puanlayıcı, likert ölçekler, bütüncül ve analitik rubrik kullanarak elde ettikleri veriler ile örtüşmektedir.

b×o×p deseninde puanlayıcı ve ölçüt sayılarının arttırılıp azaltılmasıyla yapılan karar çalışması (K) sonucunda elde edilen G ve Phi katsayıları incelenmiştir.

11 ölçütün yedi puanlayıcı tarafından puanlandığı asıl uygulama sonucunda G katsayılarının .96'nın üzerinde, ϕ katsayılarının .95'in üzerinde kestirildiği görülmektedir. Aynı zamanda teorik olarak olması gerektiği gibi benzer durumlar altında her ϕ katsayısı, G katsayısından küçük bulunmuştur. Elde edilen sonuçların yüksek değerlerde olması sebebi ile K çalışmalarında ölçüt ve puanlayıcı sayılarının artışlarını incelemek yerine kullanışlılık (ekonomiklik) sağlanması adına daha az puanlayıcı ve ölçüt ile .80'e yakın değerler elde edilmeye çalışılmıştır. Bu sonuçlar ise G ve Phi katsayıları düşük düzeyde elde edilen Güler (2011), Öztürk'ün (2011) çalışmaları ile farklı durumlar ortaya koymuştur. Bu durumun sebebi ise Öztürk'ün (2011) çalışmasında gözlem formu, Güler'in (2011) çalışmasında geçerlik ve güvenirliği düşük olması istenen rastgele veriler olarak belirtilmiştir.

Bu durumda GK, değişkenlik kaynaklarını ayrıştırarak öğrenciler, ölçütler ve puanlayıcıları ayrı ayrı ve etkileşimlerini içeren sonuçlarla KTK'ye göre daha ayrıntılı sonuçlar vermiştir.

Alanyazın incelendiğinde Özdemir (2002), Nartgün (2002), Doğan ve Tezbaşaran (2003), Kan (2006), Morales (2009), Gelbal (1994), Kelecioğlu (2001), Lee, Torre ve Park. (2012), Çelen ve Aybek'in (2013) araştırmalarında başarı testleri ya da simülasyon ile üretilmiş veri kullanarak iki kategorili MTK modelleri; Özer-Özkan (2012), Sünbül'ün (2011) çok boyutlu MTK modelleri; Arsan (2012), İlhan (2016), Kim ve Wilson (2009), Ure'nin (2011) çok değişkenlik kaynaklı Rasch modeli kullanarak parametre kestirimleri yaptıkları görülmüştür.

Derecelendirilmiş (Aşamalı) Tepki Modeli'ne (DTM) göre, 11 maddenin a parametrelerinin tüm puanlayıcılarda 1.24 ile 6.44 arasında değiştiğinden her puanlayıcı için maddelerin ayırt ediciliklerinin ve verdikleri bilgilerin yüksek düzeyde olduğu görülmüştür. 1 ve 3. puanlayıcılara göre en fazla bilgiyi veren madde 8. ölçüt iken 2, 4 ve 5. puanlayıcılara göre en fazla bilgiyi 9. ölçüt vermektedir. En az bilgi veren madde ise tüm puanlayıcılara göre 11. ölçüt olarak bulunmuştur. Çok kategorili MTK modellerinin kullanıldığı Koch (1983), Köse (2015), Nartgün (2002) ve Özdemir (2002)'nin çalışmalarında madde ayırt edicilik değerleri en yüksek Nartgün'ün (2002) çalışmasında erkek örnekleminden birinci madde için kestirilen 3.34 değeridir. Bu çalışmada ise ayırt edicilik değerinin 6.44 bulunması örneklemdeki okulların eğitim düzeyleri, öğrencilerin yazılı anlatım çalışmalarına

aşinalığı, 3 ile 7. sınıflar arasındaki tüm öğrencilerin örnekleme dâhil edilmesinin olabileceği düşünülmektedir.

GK'ye göre maddelerin güçlük düzeyleri çok fazla farklılaşmıyor olarak bulunmasına rağmen MTK'ye göre b parametreleri -2.35 ile 1.90 ve θ düzeyleri -0.50 ile 1.20 arasında farklılaşmaktadır.

Marjinal güvenirlik katsayıları incelendiğinde ise güvenirliğin oldukça yüksek (.93 civarında) olduğu görülmüştür. Bu bulgu Köse (2015) ve Nartgün'ün (2002) elde ettikleri marjinal güvenirlik katsayıları ile çok yakınken (.97 ve .93) Özdemir'in (2002) elde ettiği katsayılardan (.65 ile .73 arasında) farklılaşmaktadır. Marjinal güvenirlik katsayısı dışında MTK'de güvenirlik için tek bir katsayıya Morales (2009) -Person reliability (kişi güvenirliği) .95- ve Çelen ve Aybek'in (2013) -Empirical reliability (Görgül güvenirlik) .80- çalışmalarında rastlanmıştır.

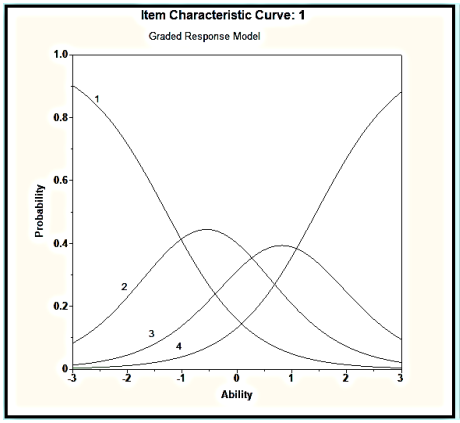
KTK ve GK ile benzer şekilde MTK'de da beş puanlayıcının öğrencilerin yazılı anlatım becerilerini puanlamaları sonucu elde edilen puanların güvenirliği yüksek düzeyde bulunmuştur. Sonuçta ölçümlerimiz için kullanılan üç güvenirlik kuramlarından elde edilen güvenirlik kestirimlerinin hepsi oldukça yüksek bulunmuştur.

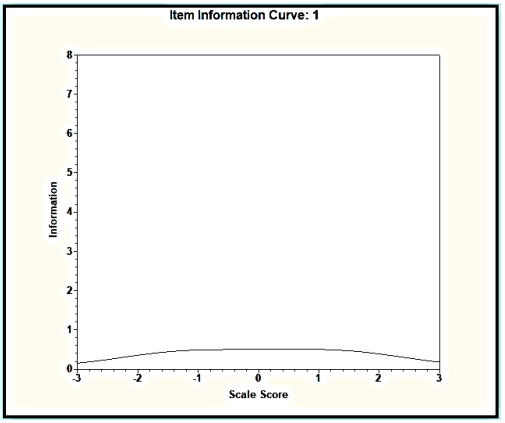
Bu çalışmada kullanılan üç ölçme teorisinden elde edilen güvenirlik kestirimleri arasında manidar bir farklılık olup olmadığı iki şekilde incelenmiştir. İlk kısımda tüm öğrencilere ve tek uyum puanına göre elde edilen KTK'deki Cronbach Alpha katsayıları, Eta korelasyon katsayıları ile GK'deki G ve Phi katsayıları karşılaştırılmıştır. Bu işlem için yedi puanlayıcıya ait Cronbach Alfa (α) katsayılarının ortancası alınmıştır. Her sınıf düzeyinde ve tüm öğrencilerde α ile G, α ile Phi, G ve Eta, Phi ve Eta katsayıları arasında .05 düzeyinde Genellenebilirlik Kuramı katsayıları lehine anlamlı bir fark bulunmuştur. Bu durumda güvenirlik kestirimi için KTK ile GK katsayılarının farklılaştığı görülmüştür. Alanyazın incelendiğinde KTK ile GK'yi karşılaştıran çalışmalarda korelasyon katsayılarının arasındaki farkın manidarlığı için yapılan bir analizle karşılaşılmamıştır.

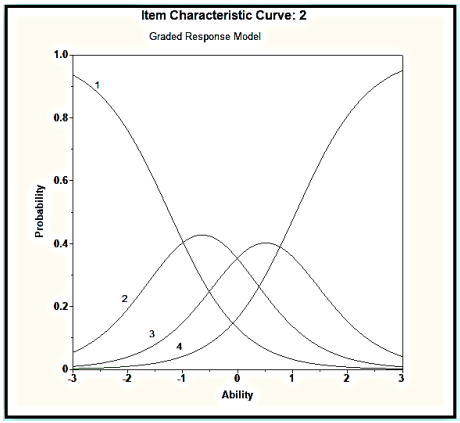
İkinci kısımda ise puanlayıcılara göre elde edilen KTK'deki Cronbach Alpha katsayıları ile MTK'deki marjinal güvenirlik katsayıları karşılaştırılmıştır. .05 anlamlılık düzeyinde katsayılar arasında anlamlı bir fark olmadığı tespit edilmiştir. Bu durumda KTK ve MTK'ye göre puanlayıcılar arası güvenirlik için benzer sonuçlar elde edilmiştir. Alanyazın incelendiğinde KTK ile MTK'yi karşılaştıran çalışmalarda Nartgün (2002) bu çalışma ile benzer olarak Cronbach Alfa iç tutarlılık katsayısı ile marjinal güvenirlik katsayısı arasındaki farkı Fisher'in Z dönüşümü ile inceleyerek manidar bir fark olmadığı sonucuna ulaşmıştır. Doğan ve Tezbaşaran (2003) ise bu çalışmadan farklı olarak KTK ve MTK'deki madde ayırt edicilikleri ve güçlükleri arasındaki farkın manidarlığını Fisher'in Z dönüşümü ile incelemiş, manidar bir fark olmadığı sonucuna ulaşmıştır.

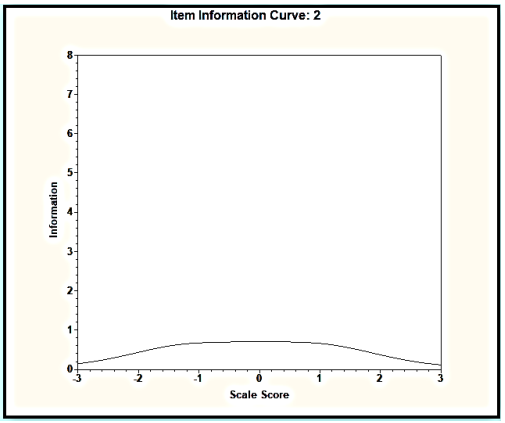
Sonuç olarak ölçümlerin güvenirliğini kestirmeye yönelik olan bu çalışmaya göre, örneklem sayısı en az 500 olduğunda ve tek boyutluluk-yerel bağımsızlık varsayımları karşılandığında MTK'de Samejima'nın (1969) Derecelendirilmiş Tepki modeli ile madde düzeyinde hata kestirimleri yapmak madde ve test bilgi fonksiyonları aracılığıyla güvenirlik kestirimleri yapmak KTK'ye göre daha ayrıntılı bilgiler sunmaktadır. Örneklem sayısı 500'den az, değişkenlik kaynakları ikiden fazla olduğunda, GK kullanılarak hata varyanslarının ayrı ayrı ve birlikte ele alınması ile mutlak ve bağıl kararlara göre farklılaşan genellenebilirlik ve güvenirlik katsayıları hesaplamak KTK'den farklı olarak mümkün olmaktadır. Değişkenlik kaynağının tek olduğu çalışmalarda, geçti-kaldı kararları ya da araştırmacının sıralama yapma amacı olduğunda ise KTK'nin kullanılması daha kullanışlıdır.

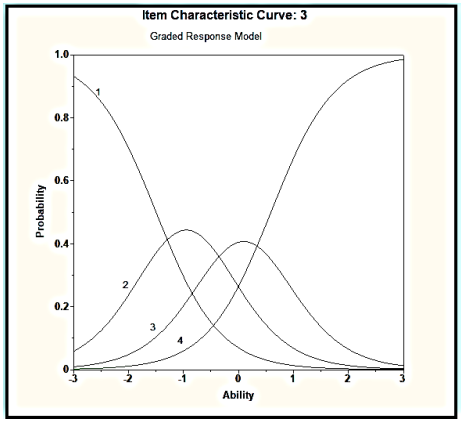
Appendix A. Characteristic Curves and Information Functions of Items-1st Rater

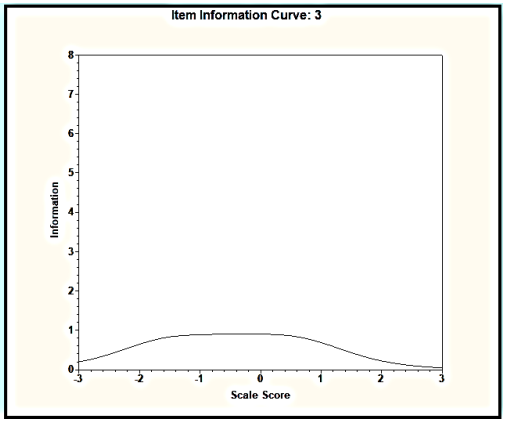


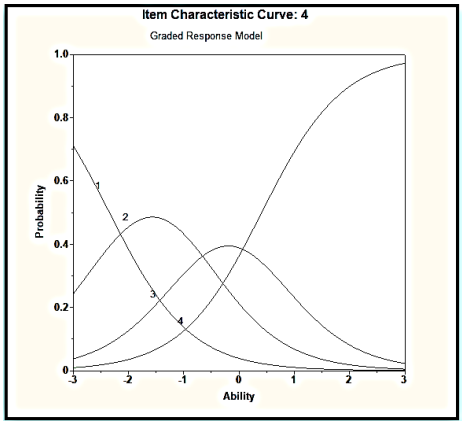


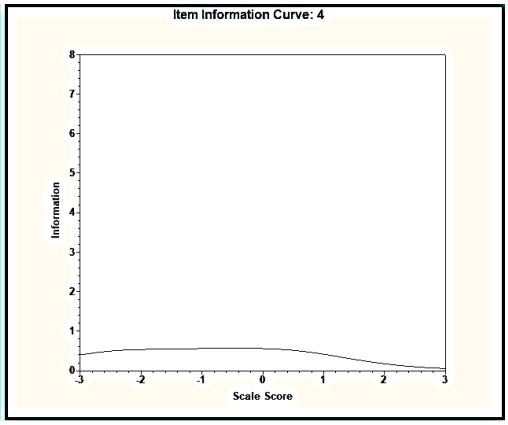


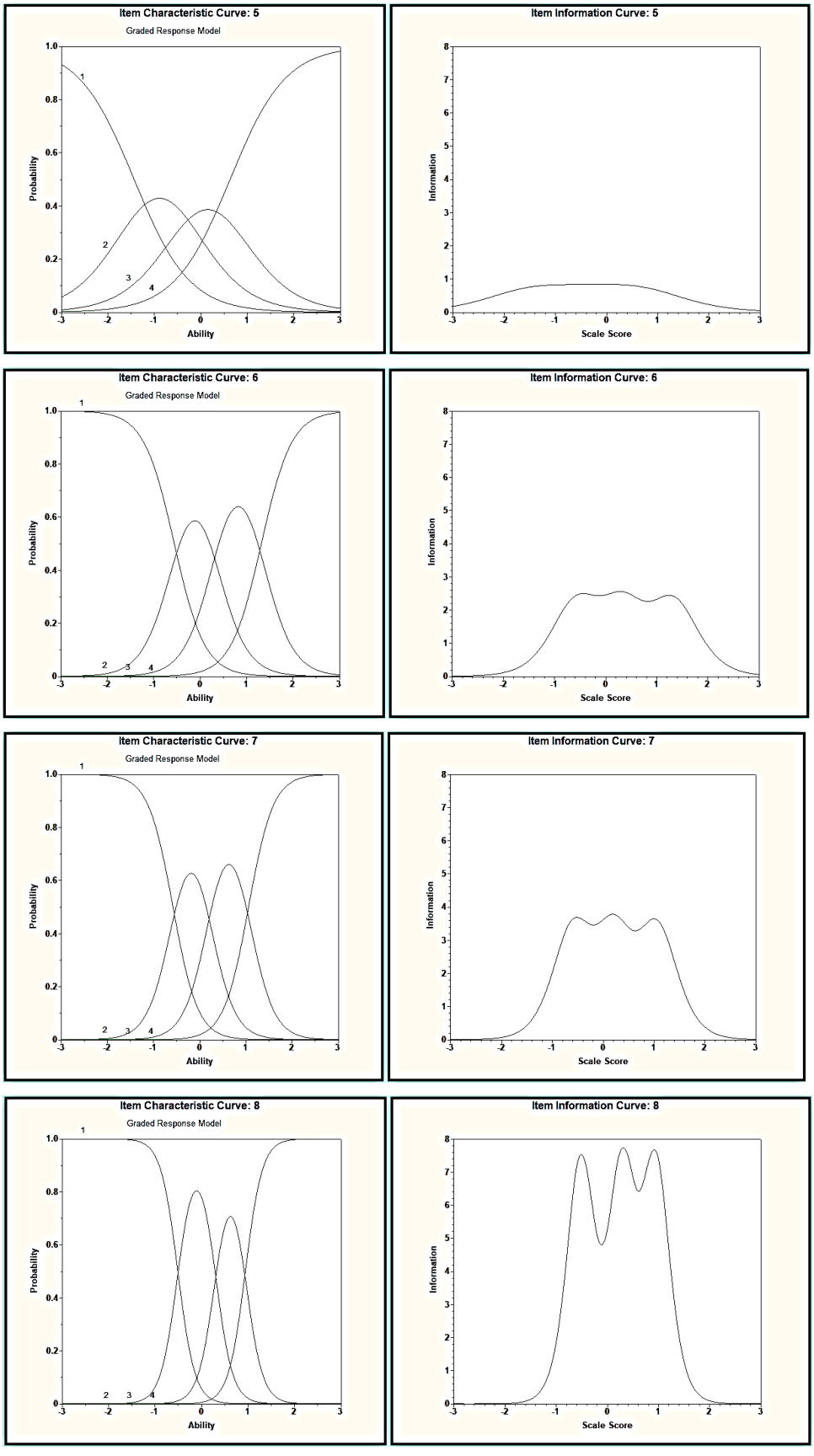


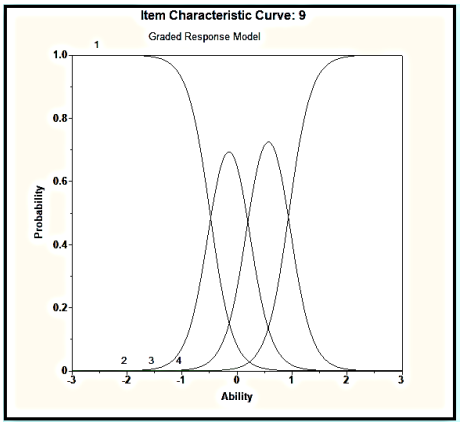


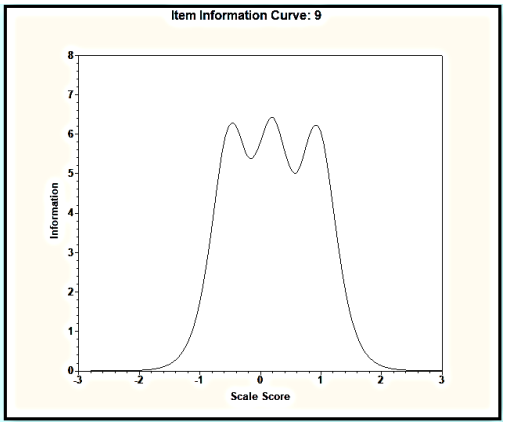


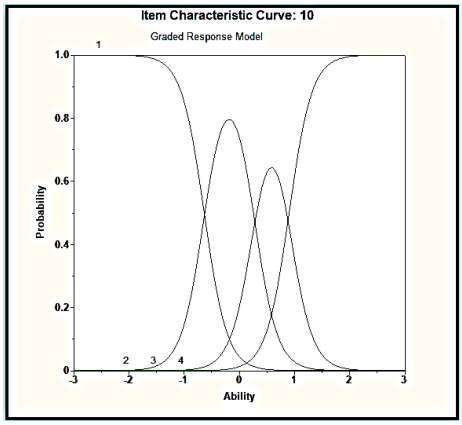


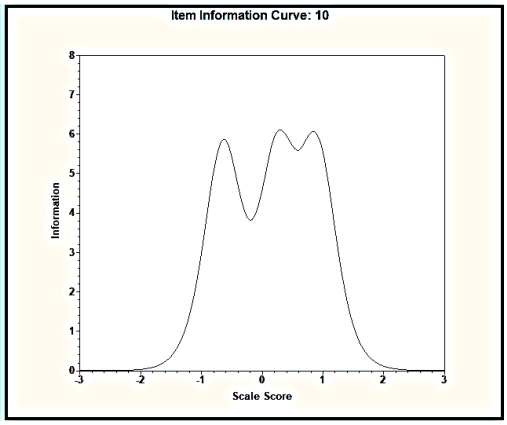


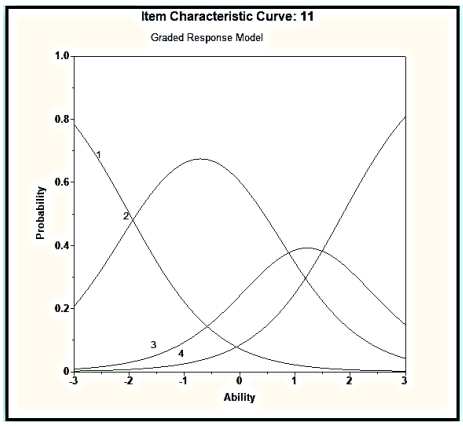

Item Characteristic Curve: 5
Graded Response Model
Probability
Ability
Item Information Curve: 5
Information
Scale Score
Item Characteristic Curve: 6
Graded Response Model
Probability
Ability
Item Information Curve: 6
Information
Scale Score
Item Characteristic Curve: 7
Graded Response Model
Probability
Ability
Item Information Curve: 7
Information
Scale Score
Item Characteristic Curve: 8
Graded Response Model
Probability
Ability
Item Information Curve: 8
Information
Scale Score

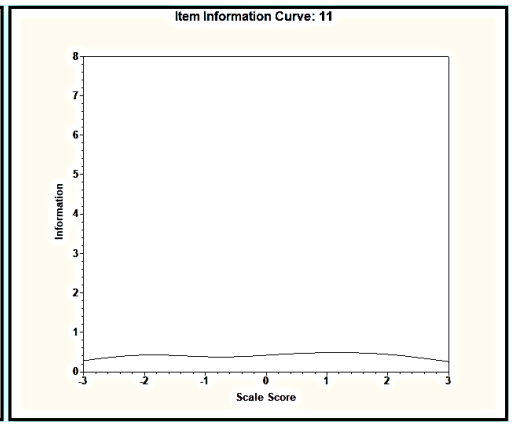

Item Characteristic Curve: 9
Graded Response Model
Probability
Ability
1
2
3
4


Item Information Curve: 9
Information
Scale Score

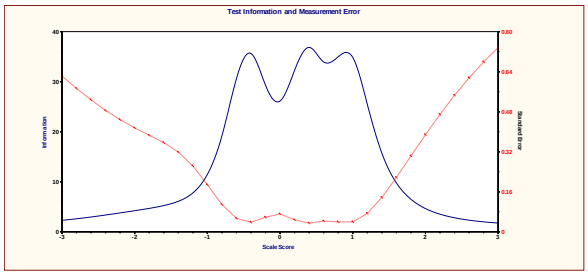

Item Characteristic Curve: 10
Graded Response Model
Probability
Ability
1
2
3
4


Item Information Curve: 10
Information
Scale Score

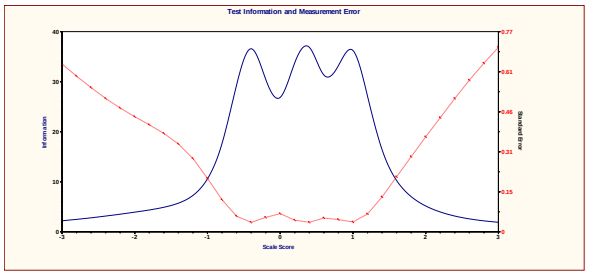

Item Characteristic Curve: 11
Graded Response Model
Probability
Ability
1
2
3
4


Item Information Curve: 11
Information
Scale Score

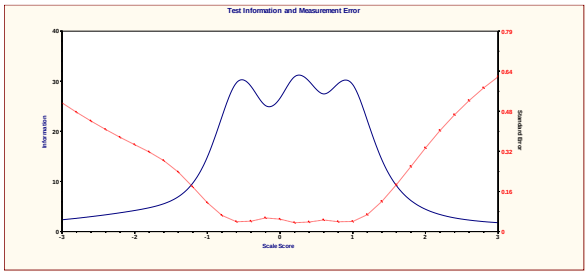
Appendix B. Test Information Functions of Five Raters



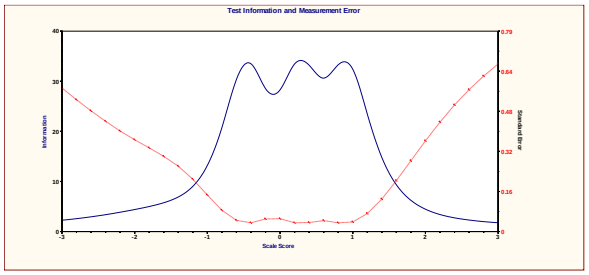
Test information function of the first rater



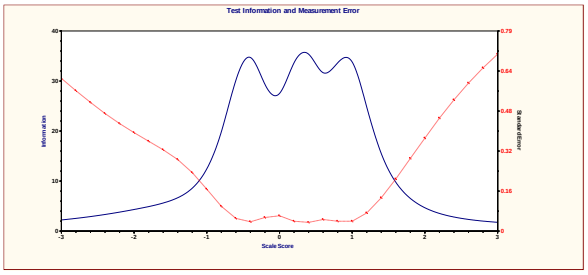
Test information function of the second rater



Test information function of the third rater



Test information function of the fourth rater



Test information function of the fifth rater