Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN:1309-6575 Yaz 2019
Summer 2019
Cilt: 10- Sayı: 2
Volume: 10- Issue: 2

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.

İletişim
e-posta: epod@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

Hakem Kurulu / Referee Board

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Haziran 2019, Sayı: 10-2)

Journal of Measurement and Evaluation in Education and Psychology (June 2019, Issue: 10-2)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(2);117-132

Validity and Reliability Study of Parental Mediation for Internet Usage Scale Adolescent and Parent Forms in the Turkish Sample

Derya ATALAN ERGİN* Emine Gül KAPÇI **

Abstract

Parental mediation includes parents' attitudes and behaviors about their child's media using. Early parental mediation researches have been conducted on television. Nowadays, parental mediation researches concentrate on the Internet. The main purpose of this study is to develop assessment forms that evaluate parental mediation strategies in respect of Internet Usage. In this study, a scale has been developed including parent and adolescent forms, as the strategies used by parents could be examined based on both parents and their children's self-report. A representative sample consisting of a total of 728 parents participated in the parent form study in Mamak, Ankara (mother n=456; father n=272). A total of 718 adolescents (female n=371; male n=345) aged 11- 14 years old studying 6.-8. grades in a secondary school in Mamak participated in the adolescent form study. Exploratory factor analysis (EFA) and conformity factor analysis (CFA) were applied for the purpose of testing construct validity of the forms. EFA indicated that a two-factor model had enough fit for parent form and a three-factor model was suggested for adolescent form. Factors of parent form were named as "control/ restriction" and "active mediation", factors of adolescent form were named as "control/ restriction", "active mediation" and "monitoring". These factors have explained 63.7% and 61.7% variance on parent and adolescent scales, respectively. The results of CFA have revealed appropriateness of the factor structure (Parent form: χ^2/sd=2.08, RMSEA= .06, GFI: .91, AGFI: .88, SRMR: .03, NFI: .98, NNFI:.99, CFI:.99; Adolescent form: χ^2/sd=2.94, RMSEA= .07, GFI: .88, AGFI: .84, SRMR: .04, NFI: .98, NNFI:.98, CFI:.99). Assessed with Cronbach Alfa internal consistency reliabilities were calculated as .95 for both parent and adolescent forms. Test-retest reliabilities were .87 and .82 for parent and adolescent forms, respectively. These results have pointed out that both forms have the value of use in research on the evaluation of parental mediation on the Internet usage.

Key Words: Parental mediation, parent, adolescence, scale development.

INTRODUCTION

Nowadays, the Internet is among the most used media tools in terms of the opportunities that provides to build social networking and get information easily and quickly as well as the increase in the quantity of the information. The latest research results of TUIK (2018) in Turkey indicates that the access on the Internet at residences is 83.8%. In the same research, the rate of being regular Internet user rates are 97.3% for women and 97.6% for men. The age group that used The Internet more than the others in the last three months is between 16-24 (90.7%). The data obtained from the abroad literature shows that adolescents use the Internet more than the other age groups (Treuer, Fabian & Füredi, 2001; Widyanto & McMurran, 2004).

In studies conducted with different cultures and age groups, it is pointed out that the usage of the Internet has increased and it provides opportunities to support the adolescents` both academic and social capability developments (Lenhart, Simon & Graziano, 2001). However, the Internet also contains the risks such as accessing pornography (Sabina, Wolak & Finkelhor, 2008), exposing to exploitation

* Dr, Ministry of Education, Ankara-Turkey, deryaatalan@gmail.com, ORCID ID:0000-0002-6602-1985
** Prof.Dr., Ankara University, Eğitim Bilimleri Fakültesi, Ankara-Turkey, eminegulkapci@gmail.com, ORCID ID: 0000-0002-5094-9244

To cite this article:
Atalan Ergin, D., & Kapçı, E., G. (2019). Validity and reliability study of parental mediation strategies for internet usage scale-adolescent and parent forms in the Turkish sample. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 117-132. DOI: 10.21031/epod.457218.

Received: 04.09.2018
Accepted:09.03.2019

(Williams & Merten, 2011) and Internet addiction (Spada, 2014). While forming and developing functional Internet usage habits, parents are the most important people to provide adolescents to benefit from the Internet and protect them from the risks. Variables such as parents` self-sufficiency perception regarding Internet usage (Glatz, Crowe & Buchanan, 2018; Festl & Langmeyer, 2018) and the features of Internet usage (Nikken & Schols, 2015) affect the Internet usage of their children. Besides, parental mediation strategies in Internet usage (Fikkers, Piotrowski & Valkenburg, 2017) that is especially discussed in studies made abroad seems to be related to the adolescents` purpose and time of Internet usage.

Parental mediation strategies are defined as all the attitudes and behaviors of parents` who would like to increase the opportunities that children and adolescents meet in media as well as to decrease the risks of the Internet (Kirwil, 2009; Nathanson, 1999; Warren, 2001). The concepts of parental mediation or parental monitoring were first examined with the studies made for watching TV. With the proliferation of Internet usage, the concept was started to be examined for Internet usage. Although parental mediation strategies are believed to change forms regarding the differences between TV and Internet usage in terms of the abilities of the user, the user`s effectiveness level while using the media instrument and the ability level that is needed, studies show that the parental mediation strategies for Internet usage have similarities with those determined for the TV. (Sonck, Nikken & de Haan, 2013). Studies show that parents use three basic mediation strategies that are restrictive mediation, active mediation and monitoring (Valkenburg, Krcmar, Peeters & Marseille, 1999). Active mediation refers to the process of discussing certain aspects of programs with children, either during or after viewing (Valkenburg, Krcmar, Peeters & Marseille, 1999). In this mediation strategy, parents explain some surreal events or the characters` good and bad sides. In restrictive mediation, parents set up rules to limit the time or to prevent them to watch a particular content. Making use of some technologies to restrict particular channels, programs and websites are among the methods for restriction. Monitoring is defined as to monitor the Internet activities of adolescents afterwards (Cabello-Hutt, Cabello & Claro, 2017). The Internet example is parents` habits to check the children`s ``history`` of Internet usage.

The level and kind of mediation strategies in Internet usage applied by parents may vary regarding some features such as parents` communication with their children (Valkenburg, Piotrowski, Hermanns & de Leeuw, 2013), time (Fikkers, Piotrowsk & Valkenburg, 2017), consistency between parents (Mares, Stephenson, Martins & Nathanson, 2018), parents` education level (Clark, 2011; Nikken & Schols, 2015; Pasquier, Simões & Kredens, 2012; Shin & Huh, 2011), self-sufficiency perception of parents for Internet usage (Glatz, Crowe & Buchanan, 2018; Festl & Langmeyer, 2018), being a family of single parent or regular parents (Barkin, Richardson, Klinepeter, Finch & Krcmar, 2006), parental behaviors and the characteristics of the child (Padilla-Walker, Coyne & Fraser, 2012). Functional usage of parental mediation prevents Internet addiction and exposing to cyberbullying (Chang, Chiu, Miao, Chen, Lee, Chiang & Pan, 2015), and attempting the risky behaviors in the Internet (Sin & Kang, 2016) and it also decreases the time spent on the Internet (Cabello-Hutt, Cabello & Claro, 2017; Gomez Harris, Barreiro, Isorna & Rial, 2017; Shin & Kang, 2016). Besides, active mediation increases meeting with the opportunities that the Internet provides and restrictive mediation decreases meeting the risks (Livingstone, Ólafsson, Helsper, Lupiáñez-Villanueva, Veltri & Folkvord, 2017). Using restrictive and active mediation together is claimed to be the most effective method for the Internet (Valkenburg, Piotrowski, Hermanns & de Leeuw, 2013). While the risks of the Internet could be restrictive for adolescents` psycho-social and academic development, it is important to take into consideration that the opportunities that the Internet provides could support their development. Therefore, using all means of mediation strategies together may increase the benefits of the Internet.

In adolescence, the level and form of parental mediation strategies change due to the need for independence and autonomy (Chen & Chng, 2016). Parents of adolescents use less restrictive mediation than the parents who have young children (Davies & Gentile, 2012) or they decrease the level of mediation strategies they used in this period. This condition is related to the idea of parents that is older children have more self-check than the younger children (Lee, 2013) and they are more talented to cope with the negative effects of the Internet (Wang, Bianchi & Raley, 2005). Moreover, younger children spend more time at home compared to the adolescents and this provides parents to control the Internet usage of them more often. Studies show that younger children have less level of healthy Internet usage habits in comparison to adolescents (Davies & Gentile, 2012) and they spend more time before the monitor. These findings indicate that parental mediation strategies are important in terms of organizing the adolescents` Internet usage (Vaterlaus, Beckert, Tulane & Bird, 2014). This study also reflects the importance of evaluating mediation strategies in puberty for these reasons.

The findings of the cultural studies made regarding parental mediation highlight different mediation strategies. In a study conducted in Brazil, Cabello – Hutt, Cabello and Claro (2017) determine the parental mediation strategies as "active mediation", "co-using" and "restrictive mediation". A measurement tool developed by Lee and Kim (2017) for Korean adolescents is formed by sub-dimensions of "restrictive mediation", "active mediation", "co-using" and "no mediation". Livingstone, Ólafsson, Helsper, Lupiáñez-Villanueva, Veltri and Folkvord (2017) developed a measurement tool that depends on the self-statement of parents and data from eight European countries which are very different from those two measurement tools related to the adolescents` self-statement. The subdimensions obtained from the scale are determined as "active mediation for Internet usage", "child-initiated support", "active mediation of Internet safety", "technical controls", "parental monitoring" and "parental restriction". The divergence in sub-dimensions indicates that evaluations regarding culture may provide important information.

Another significant subject in measuring parental mediation strategies is that how far measurement tools related to parents` and adolescents` statements are compatible with each other as well as culture (Wang, Bianchi & Raley, 2005). In studies, half of the adolescents declare that they have parental mediation in Internet usage while this rate increases in the statements of parents (Rideout, Foehr & Roberts, 2010). Gentile, Nathanson, Rasmussen, Reimer and Walsh (2012) point out that adolescent statements may indicate real mediation level better as parents may remark more mediation for social admiration. However, it is also possible that adolescents would like to emphasize their autonomy so that they would remark the mediation strategies less. Thus, scales obtained from both adolescents and parents provide more appropriate information to reveal the real condition.

In Turkey, studies regarding parental mediation in Internet usage have been conducted by qualitative analysis (Kılınç, 2017; Sütçü, 2017). The only quantitative study that also included Turkish sample about parental mediation was conducted by Bayraktar (2017). The study of Bayraktar (2017) was conducted on the data related to the database of European Union Kids Online II Project and the risks experienced on the Internet between Turkish people living in Europe and Turkish people living in Turkey and the relation of those risks with parental mediation strategies were evaluated. The mediation strategies regarding Internet usage were examined in four dimensions that are active mediation, active mediation regarding Internet safety, restrictive mediation and parental monitoring. However, the measurement tool used in this study was not developed in Turkish sample. In Turkey, there is a measurement tool that aims to measure the close features of parental mediation strategies and to determine the parents` attitude in terms of internet usage. Internet Family Attitude Scale was developed by Eijden (2007) and was adapted to Turkish by Ayas and Horzum (2013). The scale has two sub-dimensions that are family control and family closeness. As a result of the assessment by cutoff scores, parents` attitude could be examined as laissez-failure, permissive, authoritative and authoritative. The scale was formed with reference to Baumrind`s (1991) parenting style model. In the scale, the attitude and behaviors of Internet usage are laissez-failure attitude that contains low family control and closeness, authoritative attitude that contains high family control and low family closeness, permissive attitude that contains low family control and high family closeness and authoritative control that contains high family control and closeness (Ayas & Horzum, 2013). "I determine the Internet rules with my child" or "I talk to my child about what he/she does with the Internet" can be cited as the closeness sub-dimensions of the scale adapted by Ayas and Horzum (2013). "I monitor my child while he/she surfs on the net" or "I use software to block specific Internet sites" can be cited for control. Recommended assessment style in related scale is by assigning to the groups with cutoff scores. For instance, the parents' behaviors that get lower than three in control items and higher than five in closeness items, parents are assessed as permissive. When the measurement tools in abroad literature are examined regarding parental mediation strategies in Internet usage, the items in the scale adapted by Ayas and

Horzum (2013) are discussed in sub-dimensions of active mediation, control and monitoring (Hutt & Cabello, 2017; Lee & Kim, 2017; Livingstone & Olafsson, 2017). Assessing the measurement tools concerning mediation strategies is done from the total score which is different from Internet Family Attitude Scale. The Internet Family Attitude Scale adapted by Ayas and Horzum and the differences between the dimensions and forms of the evaluation of the scales for parental mediation strategies developed in the literature of the abroad studies are indicative of the differentiation in the theoretical foundations. In Internet Family Attitude Scale, based on Baumrind`s classification, authoritative parental attitudes are the desired behaviors. However, all of the forms (control, active mediation, co-use) in parental mediation strategies are assessed as positive strategies. Therefore, bringing in an instrument to literature that is related to the assessment of parental mediation strategies in Internet usage is significant. Besides, two forms are aimed to be developed by gathering data from two sources in order to take into consideration the differences in adolescents` and parents` statements.

The importance of protecting adolescents from the risks of the Internet and the awareness about doing studies in this issue gradually increases in Turkey. The Ministry of Education conducts various studies in education institutions about Internet addiction and functional usage of information and communication technologies. It seems both important and necessary to consider the cultural differences while determining the parental mediation strategies. The main purpose of this study is to develop two forms that would reveal possible cultural differences in assessing the parental mediation strategies and that is based on both adolescents` and parents` statements.

METHOD

The model of the research is survey model that aims to describe the existing situation. The purposive sampling method was used to select participants. Thus, a sample was determined regarding the previous theoretical information about the universe, its own information and the special purpose of the research (Fraenkel & Wallen, 1993). In this study, the purposive sampling method is preferred because adolescents need a social media or an e-mail account to fill in the evaluation instruments.

Working Group

The study groups are named as study group 1 for Explanatory Factor Analysis (EFA), study group 2 for Confirmatory Factor Analysis (CFA) and study group 3 for Test-Retest. In study groups, adolescent form is indicated by the letter A and parent form is indicated by letter P to emphasize the difference between the adolescent forms and parent forms. For parent form EFA study group is named 1P, for CFA study group is named 2P; and for the adolescent form EFA study group is named 1A, for CFA study group is named 2A; for Test-Retest adolescent study group is named 3A, for parents' study group is named 3P. In the next parts, study groups will be referred with those names.

In the study, adolescents between the ages 11-14, having their secondary education in Ankara province, Mamak district and parents` whose children are at the same age range at the same school were contacted. For study group 1P 432 parents (n_{mother}=272, n_{father} = 160), and for study group 1A 361 adolescents ($n_{6th\ grade}$=159, $n_{7th\ grade}$ 115, $n_{8th\ grade}$= 81) were contacted. In study group 1P, 29.17% of them are primary school graduate (n =126), 28.24% of them are secondary school graduate (n=122), %33.79 of them are high school graduate (n=146) and %8.80 of them are university graduate (n=38). For study group 2P 296 parents (n_{mother} =184, n_{father}=112) and for the study group 2A 355 adolescents ($_{n6th\ grade}$=124, $_{n7th\ grade}$=147, $n_{8th\ grade}$= 84) were contacted. In study group 2P, 29.39% of the parents are primary school graduate (n=87), 31.08% of them are secondary school graduate (n=92), 30.07% of them are high school graduate (n=89) and %9.46 of them are university graduate (n=28). The study group 3 was formed from the randomly chosen and volunteered people in study group 1 and 2 that are 49 parents (n_{mother}=34; n_{father}=15) and 51 adolescents (n_{girls}=29; n_{boys}=22) to calculate the reliability of the test-retest method.

Data Collection Instruments

Parental mediation for Internet usage scale – adolescent form (PMS-A)

In order to measure the theoretical basis and parental mediation strategies, a pool of 54 items, each of which was evaluated in 5-point Likert type, was prepared considering the scales previously developed in abroad literature studies. When preparing the item pool, two sentences considered to measure the same feature were written. After the preparation of the item pool, opinions were received from five adolescents for the comprehension of the items through individual interviews and the items which were more understandable than the two items were taken into the measurement tool and the others were excluded from the measurement tool. After the arrangements, 25 items were remained and an "Expert Opinion Form" was prepared to assess the appropriateness and comprehensibility of the items by the experts. In the form that aims to have experts` opinion with 3-point Likert type, there is a part in which experts would point out their opinion and correction points for each item. The form was reached out by a specialist clinic psychologist, a psychological counsellor and guidance specialist, two academicians of education psychology, a Turkish language specialist and an evaluation and assessment specialist. With specialists` suggestions, the measurement tool was determined to have 22 items. After the pre-application of the scale, the scale had the last arrangements before the main application.

Parental mediation for Internet usage scale- parent form (PMS – P)

In order to measure the theoretical basis and parental mediation strategies, a pool of 54 items, each of which was evaluated in 5-point Likert type and was formed of pairs, was prepared considering the scales previously developed in abroad literature studies. When preparing the item pool, two sentences considered to measure the same feature were written. After preparing the item pool, the assessment of comprehensibility for the items was done by three mothers and two fathers by individual interviews. One of the two items that measure the same feature was taken into the measurement tool and the other one was excluded. After the arrangements done on the items by parents` evaluations, in order to assess the appropriateness and comprehensibility of 25 items, an ``Expert Opinion Form`` was sent to a specialist clinic psychologist, a psychological counsellor and guidance specialist, two academicians of education psychology, a Turkish language specialist and an evaluation and assessment specialist. With specialists` suggestions, the scale had 23 items and pre-applications were done for the main application.

Personal information form

In addition to the adolescents` scale, the form regarding information about a nickname, gender, grade or whether having an e-mail or social media account or not was given to the participants. For parent form, the participant of each parent was given a form regarding the information about a nickname, closeness degree (mother or father), education level, having knowledge about whether his/her child has an e-mail or a social media account or not. Information about e-mail and social media account are necessary to answer the questions about the related accounts in the measurement tool. Participants who don`t have any e-mail or social media account were excluded from the study.

Data Collection Procedure

Before the data collection, legal permission regarding the application was obtained from the Ministry of Education. Afterwards, the appropriate days and hours of practice for the institution and the grades of the study group were determined. The parents of study group students were sent "Informed Consent Form" one week before the application. Related form contains the purpose of the research, by whom it is going to be done, the duration of time to fill in the scales, privacy policy, the communication information of the researcher, and consent parts. In this stage, the parents of all 6th, 7th and 8th-grade students were given PMS– P by the students (who have a social media and an e-mail account). This process was carried out one week before the application to be made to the adolescents by considering that the return period of the forms would be completed by the parents. As a result, a total of 987 parents were sent the forms and 728 returns were provided . No data loss was experienced by the data returned and the return rate was stated as 73.75%. After the expected week for the return of the consent forms, adolescents had the application. The participants of each class level who had parental consent form were met in an empty class arranged by the school administration. Before collecting the data, the participants were informed about the purpose of the study, privacy and volunteering policy. Next, they were asked whether they have an e-mail and a social media account which is necessary to participate in the research. Students who have not at least one of the related accounts were excluded from the study. After this process, the participants filled the measurement tool and personal information form. There was no one who did not want to participate in the study or who left the form undone. Applications lasted about 20 minutes. After four weeks of all the applications were done, responses of randomly and voluntarily selected parents and adolescents were used to calculate the test-retest reliability.

Data Analysis

While developing PMS-A and PMS-P forms in the study, Principal Component Analysis was used to present the factor design of the forms as the factorization method. In this study, it was investigated whether there is a similarity between the structure of the theory that helps the behavior to be understood with EFA and factors or not. Next, CFA was done to test the structural validity of the forms. In order to overcome the missing data problem to prepare the data sets for the analysis, the median replacement was preferred since it was suggested that all possible strategies for ….. missing data would have similar results (Tabachnick & Fidell, 2001) the data was collected by ordinal type of scale (Hastie, Tibshirani, Sherlock, Eisen, Brown & Botsein, 1999).

RESULTS

Principal Component Analysis was chosen for the factorization method for PMS-A and PMS-P and oblique method was chosen as direct oblimin method.

For both of the forms, the data set before the EFA and CFA were checked in terms of size of the sample, missing data, multivariate and univariate normality, linearity, multivariate and univariate outliers analysis, multicollinearity and singularity. For both of the forms, findings obtained from the hypothesis tests done before the EFA and CFA were given. After determining that the hypothesis was met, the EFA process started for the forms. For both of the forms, eigenvalue greater than 1, scree plot, the contribution of factors to the variance and the results of Horn`s Parallel Analysis were assessed altogether while determining the factor numbers. From the items that are cyclical or the factor load value of which are under .32 (Tabachnick & Fidell, 2001); first the cyclical ones and later those the factor load value of which is .32 were removed from the scale.

Testing Assumptions for EFA in PMS-A Data Set

While testing the structure validity in the study group -1A for EFA, 361 adolescents were contacted. Whether the sample size is adequate for EFA was tested by Kaiser-Meyer-Olkin statistics and the value was found .95 for PMS-A that was sufficient for the process. While checking missing data, there was no parameter of missing data rate over %5 and parameters under %5 was done median designation. Multivariate normality was assessed by Bartlett Globality Test and the multivariate normality for PMS-A was met ($\chi^2_{(231)}$ = 4937.986; p<.05). Univariate normality was examined by Levene Test and significance level seemed to be met the present assumption since it was bigger than .05 (LF$_{(2,358)}$ = .754, p>.05). Linearity was checked by the scatter diagram and the elliptical shape of the diagram also met that assumption. For multicollinearity VIF, case index (CI) and tolerance value were checked while for the singularity problem the correlation coefficient between pairs of items were checked. Accordingly, VIF value is smaller than 10, CI is smaller than 30 and tolerance value is bigger than .10 that indicates there is not any multicollinearity problems (Çokluk, Şekercioğlu & Büyüköztürk, 2014). Singularity is the state of correlation coefficient being rxy=1.00 between the item pairs (Şencan, 2005). Accordingly, in PMS-A (VIF=1.000, CI=1.000 and tolerance rate = 1.000; the correlation coefficient between the pairs of the item is .24-.71) there is no multicollinearity or singularity problem. In multivariate and univariate outlier analyses assessed by Mahalonobis Distance and Z points, there were no outliers.

EFA Process Steps for PMS-A

As a result of the analysis for the 22 items based on EFA in PMS-A, it is appropriate to assess the scale by the three-factor structure. Two items were excluded from the analysis in items examination. In consequence of EFA of 20 items, three subscales are labeled as \`control/restriction\`, \`monitoring\` and \`active mediation\`. The contribution of the factors on the total variance is 27.08% for \`control/restriction\`, 18.86% for \`active mediation\` and 15.80% for \`monitoring\`. The total contribution of those three factors on the variance is 61.74%. Factor loading for each factor are presented in Table 1.

Table 1. Factor Load and Common Factor Variance for PMS-A

Items	1. factor (Control/ restriction)	2. factor (active mediation)	3. factor (monitoring)
He/She monitors the games I play on the Internet.	.79	.15	.03
My family determines a rule about turning off a device that I can access the Internet (such as phone, computer) on a definite time.	.82	.02	.16
He/She checks what I do on the Internet.	.85	.02	.09
He/She checks my correspondence on social networking sites.	.66	.04	.19
He/She takes precautions to prevent my access to unsafe Internet websites.	.55	.20	.04
He/She checks the Internet websites I visited.	.66	.05	.17
He/She checks what I shared on social networking sites.	.59	.06	.18
He/She limits the time that I use on the Internet.	.65	.07	.05
He/She checks the people I texted on my mobile phone.	.62	.03	.24
He/She checks whether I made a video chat with strangers or not.	.57	.17	.08
He/She monitors whether I exceed the time that I suppose to spend on the Internet or not.	.53	.27	.09
He/She encourages me to use the Internet to get information.	.06	.84	.06
He/She encourages me to use the Internet to do my homework or to support my lessons.	.09	.86	.04
He/She encourages me to share the new information I learnt from the Internet with him/her.	.06	.73	.05
He/She listens to me when I share the new information I learnt from the Internet with him/her.	.02	.70	.14
He/She talks about the negativeness of writing people that I don\`t know.	.19	.60	.07
He/She asks me to tell or show my personal information to him/her before I share them on the Internet.	.19	.30	.47
He/She knows my passwords for social networking sites.	.05	.06	.89
He/She checks my e-mail correspondences.	.37	.06	.63
He/She knows my e-mail password.	.33	.01	.85

Testing Assumptions for CFA in PMS-A Data Set

Similar to EFA, in CFA process which is made to confirm the structures resulted from the EFA for PMS-A, first of all, assumptions were tested. The assumptions in the data set of 355 participants for PMS-A were tested. The results of the Kaiser-Meyer-Olkin statistics were .96 for PMS-A and that proved to be reached out the sufficient sample size. While checking missing data, there was no parameter of missing data rate over %5 and parameters under %5 was done median designation. The results of

Barlett Globality Test met the multivariate normality in PMS-A ($\chi^2_{(190)} = 4954.237$; p<.05). The results of Levene Test indicated that the univariate normality was met (LF $_{(2,352)}$ = .317, p>.05). The elliptical appearance of scatter diagram proved linearity. For multicollinearity, VIF, CI and tolerance values and for singularity problem, the correlation coefficient between pairs of items were checked. The assessments show that the assumptions for the multicollinearity and singularity were met (VIF=1.000, CI=1.000 and tolerance value=1.000 the correlation coefficient between the pairs of items is .25 - .72). In multivariate outliers analysis, 12 data were excluded from the data set since they are above the critical chi-square value. However, there were no outliers based on univariate outlier analyses. After determining that all the assumptions were met for CFA, the analysis procedure was initiated.

Steps of CFA Process for PMS-A

In CFA, t values for each item are between 11.81 – 11.76 (Figure 1) and standardized analysis values are between .59 and .84 (Figure 2). Calculated t values for all the items are significant at p<.01 level.

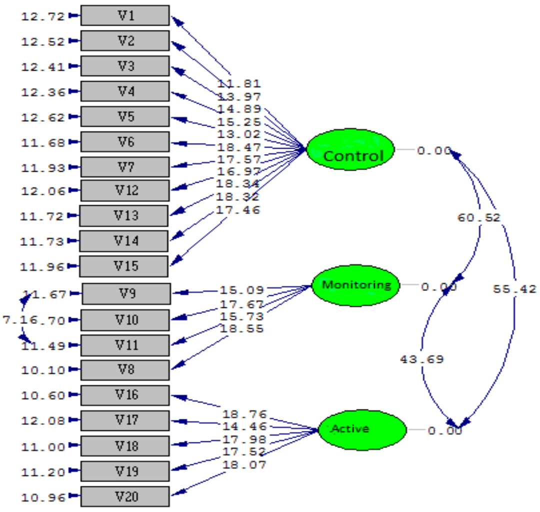


Figure 1. t values of the Items

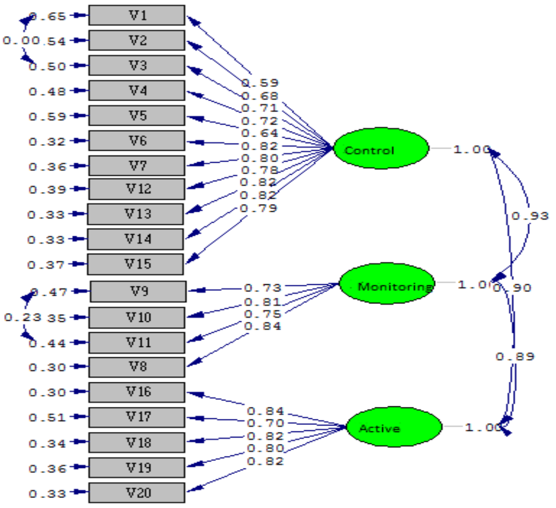


Figure 2. The standardized Factor Loadings of the Items

When the fit index resulting from CFA is examined, p-value of χ^2 value is significant (p<.05). This finding indicated that there was a significant difference between expected and monitored covariance matrix. Therefore, the χ^2/sd ratio (620.33/167) was calculated and the rate was 3.71. In larger samples, even the p-value is significant, χ^2/sd rate under 5 shows sufficient fit (Çokluk, Şekercioğlu & Büyüköztürk, 2014). Regarding the examinations with other fit indexes showing that the RMSEA value was not at the desired level (RMSEA= .11) and the χ^2/sd rate was close to 5, modification suggestions were examined. Accordingly, some modifications were done with the items 11(He/She knows my e-mail password) and 9 (He/She knows my social networking sites passwords) and with the items 1 (He/She monitors the games I play on the Internet) and 3(He/She checks what I do on the Internet) since they were under the same factor and have close meanings. In the modifications, when the corrections were added to the model for the errors between the items 11 and 9, the decreased in chi-square value was 77.9 and similarly, it was 41.7 between the items 1 and 3. When the goodness fit values were checked in the model after the modifications, the χ^2/sd rate was 2.94. This value showed a perfect fit in large samples (Sümer, 2000; Kline, 2005). However, finding the significance level p<.05 could be originated from the large sample (Çokluk, Şekercioğlu & Büyüköztürk, 2014). Thus, the rate was proof of model data fit. When the goodness fit values of the model were checked, RMSEA indicated (.07) good fit (Hooper, Coughlan & Mullen, 2008; Sümer, 2000). The GFI (.88) and AGFI (.84) values indicated acceptable fit, SRMR (.040) indicated perfect fit (Brown, 2006), NFI (.98) and NNFI (.98) indicated good fit (Tabachnick & Fidell, 2001); and CFI (.99) also indicated perfect fit (Hu & Bentler, 1999; Sümer, 2000; Thompson, 2004). According to the CFA results, the three-factor model was acceptable.

The Cronbach Alfa reliability coefficient obtained from study group -2A for 20 items of the scale was found to be .95. The Cronbach Alfa internal consistency coefficients of Control/Restrict, Active mediation and Monitoring were calculated as .91, .79, and .78 in order. This value indicated that the reliability of the scores obtained from the scale was high. In findings of study group 3A, the reliability coefficient of the scale regarding its test-retest reliability was .82 for the whole scale. The test-retest reliability coefficient of the control/restriction, active mediation and monitoring were .89, .81, and .78, respectively.

Testing the Assumptions for EFA in PMS-P Data Set

For the study group -1P, 432 parents were contacted. Whether the sampling size was adequate for EFA was tested by Kaiser-Meyer-Olkin statistics and the value was found to be .97 and showed sufficient sample size was met for EFA in the data set. While checking missing data, there was no parameter of missing data rate over %5 and parameters under %5 was done median resignation. Multivariate normality was assessed by Barlett Globality Test and the multivariate normality was met ($\chi^2_{(630)}$ = 11052.844; p<.05). Univariate normality was examined by the Levene Test and significance level seemed to be met the present assumption since it was bigger than .05 ($LF_{(2,429)}$ = 1.581, p>.05). Linearity was checked by scatter diagram and the elliptical shape of the diagram also met that assumption. For multicollinearity VIF, case index (CI) and tolerance value were checked while for the singularity problem the correlation coefficient between the pairs of the items were checked. Findings of the form showed that there was not any multicollinearity or singularity problem in PMS- P. (VIF=1.000, CI=1.000, and tolerance rate=1.000; the correlation coefficient between the pairs of the items is .31 - .84) In multivariate and univariate outlier analyses assessed by Mahalonobis Distance and Z points, there was no outliers.

EFA Process Steps for PMS-P

As a result of the assessments done to determine the factor numbers for the 23 items based on EFA in PMS-P, analysis continued with the two-factor structure of the scale. Five items were excluded from the scale after examining the items. In consequence of EFA of 18 items, factors were labeled as “control/restriction”, and “active mediation”. The contribution of the factors on the total variance was 32.46% for “control/restriction” and 25.28% for the “active mediation”. The total contribution of those two factors on the variance was 63.74%. The factor loadings for each factor are presented in Table 2.

Testing the Assumptions for CFA in PMS-P Data Set

Similar to EFA, in CFA process which was made to confirm the structures resulted from the EFA for PMS-P, first of all, assumptions were tested. The assumptions were tested in the data set of 296 participants for study group -2P. The result of the Kaiser-Meyer-Olkin statistics was .96 for PMS-P and that proved to be reached out the sufficient sample size. In missing data control, 25 parameters the missing data rate of which was above %5 were excluded from the data set. Parameters under %5 were done median resignation. The results of Barlett Globality Test met the multivariate normality ($\chi^2_{(153)}$ = 4293.491; p<.05). The results of the Levene test indicated that the univariate normality was met ($LF_{(2,293)}$ = .067, p>.05). The elliptical appearance of the scatter diagram proved linearity. For multiple connectedness problem VIF, case index (CI) and tolerance value and for singularity problem, the correlation coefficient between the pairs of items was checked. The assessment showed that the assumptions for the multicollinearity and singularity were met. (VIF=1.000, CI=1.000, and tolerance value=1.000 .The correlation coefficient between the pairs of items was .31 - .87. In multivariate outlier analyses, 12 data points were excluded from the data set since they were above the critical chi-square value. However, in univariate outlier analyses, there were no outliers. After determining that all the assumptions were met for CFA, analysis procedure was initiated.

Table 2. Factor Loads for PMS-P

Items	1. factor (Control/ Restriction)	2. factor (active mediation)
I check who he/she adds as a friend on social networking sites.	.78	.28
I check his/her immediate text messages..	.78	.28
I know his/her passwords for the social networking site.	.75	.13
I check his/ her e-mail correspondence.	.74	.26
I check the applications he/she downloads.	.74	.26
I check what he/she shares on social networking sites.	.71	.40
I ask him/her to tell or show me his/her personal information before he/she shares it on the Internet.	.69	.31
If I see any inappropriate correspondence with his/her friend, I make sure that he/ she will exclude that friend from his/her friend list.	.68	.29
I ask him/her to show me the photos or videos of our family, friends or his/her friends before he/she uploads them.	.68	.36
I check the websites that he/ she visits.	.68	.40
While my child is online, I go next to him/her and watch him/her.	.67	.44
I limit the time that he/she spends on the net.	.60	.39
I use a filtration method to prevent him/her to access inappropriate content.	.59	.32
I ask him/her to tell me anything that disturbs him/her in his/her Internet correspondence.	.34	.84
I talk to my child about the negative aspects of texting to someone that he/she doesn`t know.	.38	.81
I talk to my child about unsafe websites.	.37	.78
If my child asks for my help about the Internet, I do my best to help him/her.	.18	.78
I listen to my child when he/she shares the new information that he/she learnt from the Internet.	.32	.77

The standardized factor loadings of each item in PMS-P was between .64 - .91. t values that were assessed to determine whether the standardized analysis value was significant or not were between 12.02 and 19.71. Calculated t values were significant at p<.01 level for all of the items. t values are shown in Figure-3 and standardized loadings are shown in Figure-4.

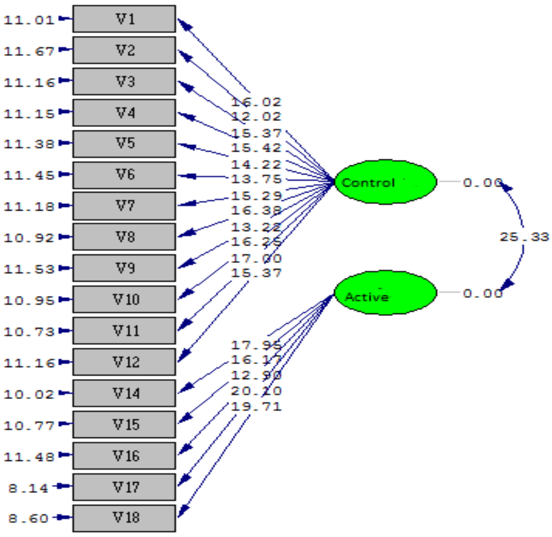


Figure 3. t values of the items

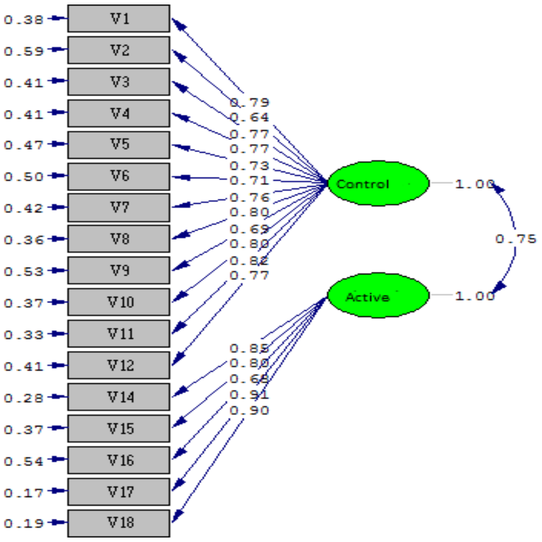


Figure 4. Standardized Factor Loadingsof the Items

When the fit indexes obtained from the results of the CFA were examined, χ^2/df rate indicated (2.08) perfect fit (Sümer, 2000; Kline, 2005), RMSEA value (.06) indicated good fit (Hu & Bentler, 1999; Thompson, 2004), GFI and AGFI values indicated (.91, .88) good fit (Hooper, Coughlan & Mullen, 2008; Sümer, 2000); SRMR value indicated (.03) perfect fit (Brown, 2006; Burne, 1994), NFI and NNFI values (98, .99) also indicated perfect fit (Tabachnick and Fidell, 2001); CFI value (.99) indicated perfect fit (Hu and Bentler, 1999, Sümer, 2000).

After finalizing the scale, the Cronbach Alfa reliability coefficient obtained from CFA study group of 18 items was calculated as .95. The same coefficient was found as .95 for control/restriction subscale and .93 for active mediation subscale. This value showed the high internal consistency of the scale. In the findings obtained from the 3P study group, the reliability coefficient of test-retest of the scale was calculated .87 for the whole scale. The reliability coefficients of the test-retest in the control/restriction and active mediation subscales were .89 and .86, respectively.

Finally, the structures obtained by EFA for PMS-A and PMS-P were confirmed by CFA. Both of the scales can be stated as the appropriate measurement tools for Turkish culture to evaluate the parental mediation strategies in Internet usage for adolescents between the ages of 10 and 14.

DISCUSSION and CONCLUSION

In this study, it was aimed to develop two forms regarding the statements of adolescents and parents to assess the parental mediation strategies in Internet usage of adolescents. First of all, item pools were formed for both of the forms. While forming the item pools, the measurement tools from international studies in this field and the theoretical base of the topic were considered and two items that were supposed to evaluate the same features were written. In these item pools, the statements of five adolescents for the adolescent form and the statement of five parents for the parent form were applied. The item which was stated to be more comprehensible than the two items written in accordance with the opinions received by individual interviews with both parents and adolescents was included in the measurement tool. Thus, there were 25 items each in adolescents and parent forms. Then, an expert opinion form in 3 points Likert type that aims to assess the items in terms of appropriateness and comprehensibility was given to a specialist clinical psychologist, a psychological consultant and guidance specialist, two academicians in the field of psychology of education, a Turkish language specialist and a measurement and evaluation specialist. Regarding the experts` suggestions, the adolescent form was formed of 22 items and the parent form was formed of 23 items. The items were finalized by the pre-application and then the main application started. The validity of the structure for both of the scales done by EFA for PMS-A and PMS-P was proved. After the items were examined in terms of cross-loadings and magnitude of the factor loading, PMS-A was formed of 20 items while PMS-P was formed of 18. In the three-dimensional structure of the PMS-A, the variance was found to be 61.74% and in the two-factor structure of the PMS-P, the variance was found to be 63.74%. For social sciences, the explained variance between 40% and %60% is sufficient (Scherer, Wiebe, Luther and Adams, 1988). The explained variance for both of the developed forms is at a good level. When the factor loading values obtained from the measurement tools were examined in terms of magnitude, it is possible to describe it from "good" to "perfect" (Comrey & Lee, 2013). In PMS-A form, the subdimensions were called as "control/restriction", "active mediation" and "monitoring". However, there was no monitoring subscale in PMS-P and subdimensions were called as "control/restriction" and "active mediation".

In CFA for adolescent form, the RMSEA value was not at the desired level and therefore the model fit indexes suggested by the package program were examined. After the parameter predictions and indexes are examined, the researchers could make modifications to the model to have a better fit or more complex model (Schreiber, Nora, Stage, Barlow & King, 2006) and those modifications should match up with the theoretical structure (Diamantopoulos & Siguaw, 2000). Thus, two modifications were made in adolescent form and in this way RMSEA values seemed in acceptance boundary. For parent form, no modifications were done and the first structure was supported by the CFA. Finally, In CFA both for adolescents and parents, model fit indexes were at acceptance boundary. In assessments for reliability, the internal consistency coefficient indicates that the form has high reliability and the results of test-retest show that forms have high stability.

Parental mediation strategies for Internet usage may be defined as the concept that expresses the attitudes and behaviors of parents about their children`s Internet usage. There are studies that show that both quality and quantity of mediation strategies used by parents in adolescence are different from those in childhood (Davies & Gentile, 2012; Gentile, Nathanson, Rasmussen, Reimer & Walsh, 2012). For instance, "restrictive mediation" is used less by parents in adolescence (Davies & Gentile, 2012). Because of the findings that show the quality and quantity of parental mediation strategies have changed, in this study it was aimed to develop a measurement tool especially for the individuals in adolescence.

When the parental mediation strategies are evaluated, it is important to consider the differences between the statements of adolescents and parents. Studies in recent years have expressed that parents tend to do higher mediation than adolescents (Rideout, Foehr & Roberts, 2010). This condition is related to the one in which parents declare higher mediation to obtain social appreciation so it is stated that the result of the adolescents` statements would be more realistic (Gentile, Nathanson, Rasmussen, Reimer & Walsh, 2012). However, adolescence is a period in which autonomy develops and independence from parents begins. When it is assessed through this point of view, adolescents would like to state a high autonomy level so it may cause a low level of parental mediation strategies results. Thus, the forms in this study developed according to the statements of both parents and adolescents.

The early studies regarding parental mediation strategies have been conducted on television. In the theoretical base of the concept, mediation strategies that were active mediation, control/restriction, monitoring and co-use were defined. There is a similar structure in studies of mediation strategies for internet usage with the mediation strategies used for watching television (Sonck, Nikken & de Haan, 2013). The basic differences between the dimensions of the mediation strategies used in television and used in Internet usage are detected in co-use and technical restrictions. In studies of Sonck, Nikken and de Haan (2013), co-use was not determined as a dimension and this condition was attributed to that Internet usage has a more individual activity unlike, television usage. Technical restrictor mediation, different from the television, is among the mediation strategies for Internet usage. For instance, in a study made by Livingstone (2017), technical control is determined as a dimension. The dimension that contains the items mostly about software regulation is considered sensible not to be discussed in researches of parental mediation strategies in television usage. One of the studies regarding parental mediation strategies in internet usage was conducted in Brazil and three dimensions that are "active mediation", "co-using" and "restrictive mediation" were determined (Cabello-Hutt, Cabello & Claro, 2017). The data obtained from eight European countries revealed the dimensions of "active mediation for Internet usage", "child-initiated support", "active mediation of Internet safety", "technical controls", "parental monitoring" and "parental restriction" (Livingstone et al., 2017). Unlike those cultures studied, parental mediation strategies were assessed with dimensions of "restrictive mediation", "active mediation", "co-using" and "no mediation" in a study made in Korea (Lee & Kim, 2017). In Turkey, however, no instruments were observed in terms of evaluating parental mediation strategies. In the two forms developed in this study, the dimensions of "active mediation", "monitoring" and "control/restriction" were determined since they were the most determined dimensions in studies made in different cultures. While forming the item tool, the scales that were developed before in abroad literature studies to measure the theoretical base and parental mediation strategies were grounded on. In those scales, the most discussed items were assessed and it was aimed to prepare items that would include all kinds of socio-economic levels of parents and adolescents regardless of digital skills. Finally, those three sub-dimensions are considered to support the structure of the two forms developed for the study.

A tool that was developed by Eijden (2007) to measure the parents' attitudes in Internet usage was adapted to Turkish by Ayas and Horzum (2013). The theoretical base of that tool is based on the parenting style model of Baumrind (1991). The attitudes and behaviours of parents are assessed in quartet structure formed according to the different levels of control and proximity dimensions. These structures are called permissive, laissez-failure, authoritarian and authoritative. The authoritative attitude that is expressed by high parents control and closeness contains desirable attitudes and behaviours. In assessments of items base, the items discussed in terms of closeness and control dimensions in Parental Attitude Scale are addressed in different sub-dimensions in PMS. The differences in the theoretical base affect the formation of dimensions in measurement tools. Internet Family Attitude Scale differs from the PMS in terms of its assessment criteria. In Internet Family Attitude Scale, an assessment can be done regarding the total points in closeness and control dimensions or the results can be examined according to the quartet structure. The parents' behaviours that get points lower than 3 in control items and higher than 5 in closeness items are evaluated as permissive. Authors declare that this evaluation type is the one that is used in the original form of the scale. However, the applications in all those items that are discussed under different dimensions in PMS indicate the increase in the level of parental mediation strategies and that is interpreted as a desirable condition. Finally, since there is not a measurement tool developed in the theoretical base of parental mediation strategies in Turkish Literature, it has revealed the necessity to develop a measurement tool that contains the self-report of both parents and adolescents for that purpose.

In this study, two forms were developed in terms of the self-reports of adolescents and parents. In PMS-A form, three-dimensional structures of "control/restrict", "active mediation" and "monitoring" appeared and in PMS-P form two-dimensional structures of "control/restrict" and "active mediation" were confirmed. Some of the items in adolescent form in monitoring sub-dimensions ("He/She knows the passwords of my social networking site", "He/She asks me to show him/her my personal information before I share them on the net.") take part in the control/restrict sub-dimension in parent form. This situation indicates that the behaviors of parents for Internet usage were perceived differently by parents and adolescents. Adolescence is a period in which egocentric thoughts dominate (Steinberg, 2007). "Imaginary audience" is one of the basic concepts for this thought. According to that, the adolescence thinks that everyone around him/her watches him/her and all the attention are on him/her all the time (Elkind, 1974). Therefore, the parents' knowledge of their passwords of the social networking sites may be perceived as they are being monitored and followed for adolescents while just having the password is a control/restrict method for parents. The items in control/restrict sub-dimension in parent form taking part in monitoring in adolescent form indicates a structure formed as a result of the egocentric way of thinking in adolescents.

In this study, two forms were developed for the adolescents between the ages 11-14 (6th-8th grade) and their parents which are developed according to the self-reports of parents (PMS-P) and adolescents (PMS-A) to assess the parental mediation strategies in Internet usage. Those forms have value of use, they would provide an archive of data collected from parents and adolescents and they will contribute to the practical education programs and researches in the future. The purpose of this study was narrowed down to develop the tools to measure parental mediation strategies. In further research, using the forms of adolescents and parents, the relation between autonomy and parental mediation strategies would be evaluated longitudinally considering the basic criticism that it doesn`t support the autonomy which is a basic variable of adolescence and which is evaluated in control/restrict subdimension of parental mediation strategies. Moreover, it is an important issue to assess whether the mediation strategies provide any change in the quality and quantity of Internet usage or not regarding Turkish culture. At this point, time-lagged panel designs that would discuss Internet usage features and parental mediation strategies together and that would reveal the cause and effect relationship between them might be suggested. This study is limited to a group of adolescents that are in the period of preadolescents and midadolescent and the parents of those adolescents. In future studies, developing the tools to assess the parental mediation strategies for both children and for individuals of pre-adolescent period will provide to assess the mediation strategies for different periods of life regarding that period`s features. What is more, the measurement tools are limited to the adolescents who have an e-mail and a social networking site accounts and their parents. This restriction is originated from the theoretical structure of parental mediation strategies in Internet usage. Nowadays, considering the position of the Internet just for accessing social media or for communication purposes, whether having an account or not will be a variable that will affect the mediation strategies. Therefore, having those accounts are determined as prerequisite for this study. It should also be discussed as a necessary feature to be asked in the personal information form for future studies.

REFERENCES

Barkin, S., Ip, E., Richardson, I., Klinepeter, S., Finch, S.& Krcmar, M. (2006). Parental media mediation styles for children aged 2 to 11 years. *Archieves of Pediatrics& Adolescent Medicine*, 160(4), 395-401.

Bayraktar, F. (2017). Çevrimiçi riskler ve ebeveyn aracılık stratejileri: Türkiye'de ve Avrupa'da yaşayan türk kökenli çocuk/ergenlerin karşılaştırılması. *Eğitim ve Bilim*, *42*(190), 25-37. DOI: 10.15390/EB.2017.6323

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Cabello-Hutt, T., Cabello, P. & Claro, M. (2017). Online opportunities and risks for children and adolescents: The role of digital skills, age, gender and parental mediation in Brazil. *New Media & Society*, 20(7)2411-2431. DOI:10.1177/14614444817724168.

Chang, F. C., Chiu, C. H., Miao, N. F., Chen, P. H., Lee, C. M., Chiang, J. T. & Pan, Y. C. (2015). The relationship between parental mediation and Internet addiction among adolescents, and the association with cyberbullying and depression. *Comprehensive psychiatry*, *57*, 21-28. DOI: 10.1016/j.comppsych.2014.11.013

Chen, V. H. H. & Chng, G. S. (2016). Active and restrictive parental mediation over time: Effects on youths' self-regulatory competencies and impulsivity. *Computers & Education*, *98*, 206-212. DOI: 10.1016/j.compedu.2016.03.012

Clark, L. S. (2011). Parental mediation theory for the digital age. *Communication theory*, *21*(4), 323-343. DOI:10.1111/j.1468-2885.2011.01391.x

Comrey, A. L., & Lee, H. B. (2013). A first course in factor analysis. NY: Psychology Press.

Çokluk Ö., Şekercioğlu G.& Büyüköztürk, Ş. (2014). *Sosyal Bilimler için çok değişkenli istatistik SPSS ve LISREL uygulamaları* (3. Baskı). Pegem Akademi, Ankara.

Davies, J. J. & Gentile, D. A. (2012). Responses to children's media use in families with and without siblings: A family development perspective. *Family Relations*, *61*(3),410-425. DOI:10.1111/j.1741-3729.2012.00703.x

Diamantopoulos A, Siguaw JA. Introducing LISREL: A Guide For The Uninitiated. London: SAGE; 2000. p.102-22.

Elkind, D. (1974). *Children and adolescents: Interpretive essays on Jean Piaget*. Oxford, Enfland: Oxford U. Press.

Festl, R. & Langmeyer, A. N. (2018). The Role of Internet Parenting for the Internet use of Children in Pre-, Primary and Secondary School. *Praxis der Kinderpsychologie und Kinderpsychiatrie*, *67*(2), 154-180.

Fikkers, M. K., Piotrowski, T. J.& Valkenburg, M. P. (2017). A matter of style? Exploring the effects of parental mediation styles on early adolescents media violence exposure and aggression. *Computers in Human Behavior*, 70, 407-415. DOI: 10.1016/j.chb.2017.01.029.

Fraenkel, R. J. & Wallen, E. N. (1993). *How to design and evaluate Research in Education*. New York: McGrow-Hill.

Gentile, D. A., Nathanson, A. I., Rasmussen, E. E., Reimer, R. A. & Walsh, D. A. (2012). Do you see what I see? Parent and child reports of parental monitoring of media. *Family Relations*, 61, 470–487. DOI:10.1111/j.1741-3729.2012.00709.x

Glatz, T., Crowe, E. & Buchanan, C. M. (2018). Internet-specific parental self-efficacy: Developmental differences and links to Internet-specific mediation. *Computers in Human Behavior*, *84*, 8-17. DOI: 10.1016/j.chb.2018.02.014

Gómez, P., Harris, S. K., Barreiro, C., Isorna, M. & Rial, A. (2017). Profiles of Internet use and parental involvement, and rates of online risks and problematic Internet use among Spanish adolescents. *Computers in Human Behavior*, *75*(2017), 826-833. DOI: 10.1016/j.chb.2017.06.027

Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., & Botsein, D. (1999). *Imputing missing values for gene expression arrays*. Technical report, Divison of Biostatistics, Stanford University.

Hooper, D., Coughlan, J. ve Mullen, M. R. (2008). Structural equation modelling: guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53-60.

Hu, L. & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisiplinary Journal*, *6*(1), 1-55. DOI: 10.1080/10705519909540118.

Kılınç, F. (2017). *Çocukların çevrimiçi ortamda karşılaştıkları risk türleri ile aracılık uygulamaları arasındaki ilişki* (Master's thesis). Mersin Üniversitesi, Mersin. Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Kirwil, L. (2009). Parental mediation of children's internet use in different European countries. *Journal of Children and Media*, *3*(4), 394-409. DOI: 10.1080/17482790903233440.

Kline, R. B. (2005).Principles and practice of structural equation modelling. N. Y: Guilford Press.

Lee, S. J. (2013). Parental restrictive mediation of children's internet use: Effective for what and for whom?. *New Media & Society*, *15*(4), 466-481. DOI: 10.1177/1461444812452412.

Lee, C. & Kim, O. (2017). Predictors of online game addiction among Korean adolescents. *Addiction Research & Theory*, *25*(1), 58-66. DOI: 10.1080/16066359.2016.1198474

Lenhart, A., Simon, M. & Graziano, M. (2001). *The Internet and education: Findings of the pew Internet & American life Project, Washington.*

Livingstone, S., Ólafsson, K., Helsper, E. J., Lupiáñez-Villanueva, F., Veltri, G. A. & Folkvord, F. (2017). Maximizing opportunities and minimizing risks for children online: The role of digital skills in emerging strategies of parental mediation. *Journal of Communication*, *67*(1), 82-105. DOI: 10.1111/jcom.12277

Mares, M. L., Stephenson, L., Martins, N. ve Nathanson, A. I. (2018). A house divided: Parental disparity and conflict over media rules predict children's outcomes. *Computers in Human Behavior*, 81, 177-188. DOI: 10.1016/j.chb.2017.12.009

Nathanson, A.I. (1999). Identifying and explaining the relationship between parental mediation and children's aggression. *Communication Research*, *26*(2), 124-143.

Nikken, P. & Schols, M. (2015). How and why parents guide the media use of young children. *Journal of Child and Family Studies*, *24*(11), 3423–3435.

Padilla-Walker, L. M., Coyne, S. M. & Fraser, A. M. (2012). Getting a high-speed family connection: associations between family media use and family connection. *Family Relations*, *61*(3), 426–440. DOI:10.1111/j.1741-3729.2012.00710.x

Pasquier, D., Simões, J. A. & Kredens, E. (2012). Agents ofmediation and sources of safety awareness: Acomparative overview. (Eds. S.Livingstone,L.Haddon ve A.Görzig), Children, risk and safety on the Internet. Bristol, England: Policy Press.

Rideout, V. J., Foehr, U. G. & Roberts, D. F. (2005). *Generation M: Media in the lives of 8-18 year-olds*. Henry J. Kaiser Family Foundation.

Sabina, C., Wolak, J. & Finkelhor, D. (2008). The nature and dynamics of internet pornography exposure for youth. *Cyberpsychology and Behavior*, *11*(6), 691–693. DOI: 10.1089/cpb.2007.0179

Scherer, R. F., Wiebe F. A., Luther, D. C.,& Adams J. S. (1988). Dimensionality of Coping: Facor Stability Using the Ways of Coping Questionnaire, *Psychological Reports*, *62*(3), 763-770. DOI:10.2466/pr0.1988.62.3.763

Schreiber, J.B., Nora, A., Stage, F.K., Barlow, E.A. & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *The Journal of Educational Research*, *99*(6), 323-38. DOI: 10.3200/JOER.99.6.323-338

Shin, W.& Huh, J. (2011). Parental mediation of teenagers’ video game playing: Antecedents and consequences. *New Media & Society*, *13*(6), 945-962. DOI: 10.1177/1461444810388025

Shin, W. & Kang, H. (2016). Adolescents' privacy concerns and information disclosure online: the role of parents and the Internet. *Computers in Human Behavior*, *54*, 114-123. DOI: 10.1016/j.chb.2015.07.062

Sonck, N., Nikken, P. & de Haan, J. (2013). Determinants of internet mediation. *Journal of Children and Media*, 7(1), 96-113. DOI: 10.1080/17482798.2012.739806

Spada, M. M. (2014). An overview of problematic Internet use. *Addictive Behaviors*, 39, 1, 3-6. DOI: 10.1016/j.addbeh.2013.09.007

Steinberg, L. (2007) Ergenlik (Ed. Figen Çok). Ankara: İmge Kitabevi Yayınları.

Sümer, N. (2000). Yapısal eşitlik modelleri: Temel kavramlar ve örnek uygulamalar. *Türk Psikoloji Yazıları*, *3*(6), 49-74.

Sütçü, S. S. (2017). The reactions of the children towards restrictions on their use of ınformatıon technologies. *Journal of Theory and Practice in Education*, 13(2), 301-315.

Şencan, H. (2005). *Sosyal ve davranışsal ölçümlerde güvenilirlik ve geçerlilik*, Hüner Şencan, Ankara.

Tabachnick, B. G., & Fidel, L. S. (2001). *Using multivarite statistics*, MA: Allyn ve Bacon.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC, US: American Psychological Association.

Treuer, T., Fabian, Z. & Füredi, J. (2001). Internet addiction associated with features of impulse control disorder: is it a real psychiatric disorder? *Journal of Affective Disorders*, *66*(2-3), 283.

TÜİK (Türkiye İstatistik Kurumu) (2017). *Hane Halkı Bilişim Teknolojileri Kullanım Araştırması*. http://www.tuik.gov.tr/PreTablo.do?alt_id=1028 İnternet adresinden 28.08.2017 tarihinde edinilmiştir.

Valkenburg, P.M., Krcmar, M., Peeters, A.L. & Marseille, N. M. (1999). Developing a scale to assess three styles of television mediation: “Instructive mediation,” “restrictive mediation,” and “social coviewing. *Journal of Broadcasting & Electronic Media*, *43*(1),52-66.

Valkenburg, P. M., Piotrowski, J., Hermanns, J., & de Leeuw, R. (2013). Developing and validating the Perceived Parental Media Mediation Scale: a self- determination perspective. *Human Communication Research*, *39*(4), 445-469. DOI: 10.1111/hcre.12010

Vaterlaus, J. M., Beckert, T. E., Tulane, S. & Bird, C. V. (2014). “They always ask what I'm doing and who I'm talking to”: Parental mediation of adolescent interactive technology use. *Marriage & Family Review*, *50*(8), 691-713. DOI: 10.1080/01494929.2014.938795

Wang, R., Bianchi, S. M. & Raley, S. B. (2005). Teenagers’ Internet use and family rules: A research note. *Journal of Marriage and Family*, *67*(5), 1249-1258. DOI: 10.1111/j.1741-3737.2005.00214.x

Warren, R. (2001). In words and deeds: Parental involvement and mediation of children's television viewing**.** *Journal of Family Communication*, *1*(4), 211-231. DOI: 10.1207/S15327698JFC0104_01

Widyanto, L. & McMurran, M. (2004). The psychometric properties of the internet addiction test. Cyberpsychology & Behavior, *7*(4), 443-450. DOI: 10.1089/cpb.2004.7.443

Williams, A. L. & Merten, M. J. (2011). iFamily: Internet and social media technology in the family context. *Family and Consumer Sciences Research Journal*, *40*(2), 150–170. DOI: 10.1111/j.1552-3934.2011.02101.x

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(2);133-148

Performances Based on Ability Estimation of the Methods of Detecting Differential Item Functioning: A Simulation Study*

İbrahim UYSAL** Levent ERTUNA*** F. Güneş ERTAŞ****
Hülya KELECİOĞLU*****

Abstract
The aim of the study is to examine differential item functioning (DIF) detection methods—the simultaneous item bias test (SIBTEST), Item Response Theory likelihood ratio (IRT-LR), Lord chi square ($\chi 2$), and Raju area measures—based on ability estimates when purifying items with DIF from the test, considering conditions of ratio of the items with DIF, effect size of DIF, and type of DIF. This study is a simulation study and 50 replications were conducted for each condition. In order to compare DIF detection methods, error (RMSD) and coefficient of concordance (Pearson's correlation coefficient) were calculated according to estimated and initial abilities for the reference group. As a result of the study, the lowest error and the highest concordance were seen in the case of 10% uniform DIF in the test and the method of IRT-LR, considering all other conditions. Moreover, for the method of SIBTEST and IRT-LR in all conditions, it was found that the error obtained by purifying items with C level DIF is lower than the error obtained by purifying items with both B and C level DIF. Similarly, for the method of SIBTEST and IRT-LR in all conditions, it was seen that the concordance coefficient found by purifying C level DIF is higher than the coefficient by purifying items with both B and C level DIF.

Key Words: Differential item functioning, simulation, ratio of the items with DIF, type of DIF

INTRODUCTION

Tests which are used in education and psychology for various purposes should meet specific standards, such as validity, reliability, and practicality. According to Messick (1995) these characteristics are not only the fundamental principles of measurement, but also the social values used by decision-makers in addition to measurement. In this regard, items in the test should not provide advantages or disadvantages for any subgroup at the same ability level. Otherwise, the test will be biased for specific groups. Bias can be defined as a systematic error in test scores depending on a group of individuals (Camilli & Shepard, 1994). When viewed from this aspect, bias is a major threat for validity and objectivity of a test (Clauser & Mazor, 1998; Kristanjansonn, Aylesworth, McDowell, & Zumbo, 2005).

The process of investigating item bias starts with examining differential item functioning (DIF), which is based on more objective results and may be a measurement of item bias. DIF is defined as differentiation of the probability of correctly responding to an item if individuals are at the same ability level but from different groups (Hambleton, Swaminathan, & Rogers, 1991). It is mentioned in the literature that group differences can be caused by two reasons. One of these is real ability

* This study presented as an oral presentation at the International Meeting of the Psychometric Society (IMPS) in Asheville/USA in 2016.

** Dr., Abant İzzet Baysal University, Faculty of Education, Bolu-Turkey, e-mail: ibrahimuysal@ibu.edu.tr, ORCID ID: 0000-0002-6767-0362
*** Res. Assist., Sakarya University, Faculty of Education, Sakarya-Turkey, e-mail: leventertuna@sakarya.edu.tr, ORCID ID: 0000-0001-7810-1168
**** Res. Assist., Boğaziçi University, Faculty of Education, İstanbul-Turkey, e-mail: gunes.ertas@boun.edu.tr, ORCID ID: 0000-0001-8785-7768
***** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, e-mail: hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:
Uysal, İ., Ertuna, L., Ertaş, F., G. & Kelecioğlu, H. (2019). Performances based on ability estimation of the methods of detecting differential item functioning: A simulation study. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 133-148. DOI: 10.21031/epod.534312

Received: 01.03.2019
Accepted: 26.05.2019

difference between subgroups, which is also called item impact. Item impact refers to the fact that different level subgroups perform differently on items, and this difference does not mean that the item is biased. The other reason is item bias. Different performances can be observed in subgroups due to the item. This means that the item causes one or more of the parameters to be too high or too low, depending on the group (Camilli & Shepard, 1994; Zumbo, 1999).

DIF is classified as uniform and non-uniform functions in terms of its occurrence (Mellenbergh, 1982). The basis of this differentiation is that the ability level and group membership together influence the probability of correct response to an item. Accordingly, uniform DIF occurs when the probabilities of correct response to an item for two groups at the same ability level is constant across all ability levels. On the other hand, non-uniform DIF occurs when the probabilities of correct response to an item for two groups at the same ability level is incoherent at different ability levels (Camilli & Shepard, 1994; Penfield & Lam, 2000; Zumbo, 1999).

Methods of detecting DIF are basically classified according to Classical Test Theory (CTT) and Item Response Theory (IRT). According to CTT, methods of detecting DIF are analysis of variance, chi-square, converted item index, logistic regression, Mantel-Haenszel (MH), and the simultaneous item bias test (SIBTEST). IRT methods are Lord's chi square ($\chi 2$), Raju's area measure, and IRT-likelihood ratio (IRT-LR) (Camilli & Shepard, 1994; Oshima & Morris, 2008). In this study, SIBTEST, IRT-LR, Lord's $\chi 2$, and Raju's area measure are examined; the below provides a brief introduction to these tests.

SIBTEST: DIF in the SIBTEST method is based on the comparison of the response rate of the tested item in the focal group and reference group according to true score. This method tests the null hypothesis that the expected value of differences between specified ratios is equal to zero. In this regard, it can be decided whether or not DIF is present and the level of DIF (Roussos & Stout, 1996). Moreover, on a theoretical basis, this method uses regression-based corrections in order to reduce Type I error (Cheng, 2005).

IRT-LR: In this method, proposed by Thissen, Steinberg, and Wainer (1993), item parameters are estimated for the focal and reference groups. For the item parameters, constrained and extended models are generated. While in the constrained model it is assumed that item parameters are equal for both groups, in the extended model it is assumed that item parameters for each tested item are different for focal and reference groups and the same for all other items. The likelihood ratio is calculated for the constrained and extended models for each item, and the null hypotheses are tested for these values (Thissen, 2001).

Lord's $\chi 2$: In the Lord's $\chi 2$ method, variance and covariance of items are calculated for the focal and reference groups in order to detect DIF. These values calculated for the two groups are scaled for the purpose of comparison. These scaled values are calculated by using Lord's $\chi 2$. Then, the null hypothesis of no DIF is tested by comparing with critical values and it is decided whether DIF exists or not (Cromwell, 2002).

Raju's Area Measure: In this method, proposed by Raju (1990), item characteristic curves are considered while detecting DIF. In the calculation stages, item characteristic curves are drawn based on the probability of correct response to the item for focal and reference groups. If the probabilities of responding to the item are different for two groups, a specific area occurs between the curves, and this area is defined as the area index.

In a test, it is important not only to detect DIF, but also to decide what will be done after detecting items with DIF. It may be required to purify DIF items in order to provide unbiasedness. However, if the item is compulsory or essential for a latent trait or construct, it may not be appropriate to remove the item. Sometimes, editing a relevant item may result in removing DIF, although sometimes this solution may not be enough (Golia, 2015). When items with DIF exist in the test, it is known that these items will affect test statistics, results, and individual scores; however, it is not known what the effect will be (Li & Zumbo, 2009). If it is decided to purify the item from the test, the validity of the test may decrease, depending on the decreasing number of items of test. Moreover, the level at which purifying items with DIF will affect the ability estimation cannot be predicted. In this study, this is the question to answer. Also, the effects of purifying items with middle level (B) DIF from the test are examined.

In the literature, studies exist about how test statistics change when items are discarded from the test in the case of dichotomous scoring (Lee & Zhang, 2017; Li & Zumbo, 2009; Roznowski & Reith, 1999; Rupp & Zumbo, 2003, 2006; Wells, Subkoviak & Serlin, 2002) and polytomous scoring (Golia, 2010, 2015; Tennant & Pallant, 2007). Some of these studies examined cases within the context of item parameter invariance (Roznowski & Reith, 1999; Rupp & Zumbo, 2003, 2006; Well, Subkoviak & Serlin, 2002), and some of these regard the cases as parameter invariances within the context of DIF as is the case in this current study (Golia, 2010, 2015; Lee & Zhang, 2017; Li & Zumbo, 2009; Tennant & Pallant, 2007). It can be stated that the studies in this direction are limited. Tennant and Pallant (2007) examined the effects of discarding items with uniform DIF from the test. The results of this study, which was conducted on five categorical items, found that discarding items in significant levels causes differences in individual and group levels. Li and Zumbo (2009) focused on the number of items with DIF and the size of DIF conditions in their study, which aimed to investigate the impacts of keeping and discarding items with uniform DIF. In the study, it was pointed out that when there are few items with DIF and a low size of DIF, even if the items in the test show DIF, the error and the effect size do not change significantly; when the size of DIF increases, discarding items with DIF from the test increases the error. Golia (2010) considered the effects of keeping and discarding three items with uniform DIF in different sizes and found that if there are few items with DIF, keeping them in the test does not affect ability estimations negatively; on the contrary, discarding them from the test has a negative impact on ability estimations. Golia (2015) also studied the effects of having items with DIF in a 15-item test and indicates that when there are three items with DIF or the size of DIF is large, the ability estimation is affected by these conditions. Lee and Zhang (2017) studied uniform DIF and investigated the conditions of the ratio of items with DIF and the existence of items with B and C levels. They also determined items with DIF by using MH methods in their study and they found that when the ratio of items with DIF increased, the ability estimations differed in individual and group levels. Moreover, the study shows that if the items with DIF are in C level, then the ability differences between reference and focal groups will be larger. Similar to this current study, several studies have compared DIF detection methods in the literature. Finch (2005) has compared the methods of MH, SIBTEST, IRT-LR, and MIMIC by considering the ratio of items with DIF. This study indicated that the method of IRT-LR was affected more than other methods when the ratio of items with DIF increased. Finch and French (2007) studied non-uniform DIF and compared the methods of logistic regression, SIBTEST, IRT-LR, and confirmatory factor analysis with the variables of DIF size, sample size, ability distribution, and IRT model. The study, which was conducted on 30 dichotomous items, showed that SIBTEST was the best in terms of Type 1 error and power, but factors that were manipulated did not have significant impact on the methods in terms of Type 1 error. Atalay Kabasakal, Arsan, Gök, and Kelecioğlu (2014) compared the methods of MH, SIBTEST, and IRT-LR in a simulation study conducted on uniform DIF. In this study, the ratio of items with DIF was studied and effect size of DIF was fixed at B level. The results of the study, conducted on dichotomously scored items, indicated that the largest Type 1 error was in SIBTEST method and the smallest Type 1 error was in the IRT-LR method. It also showed that when the ratio of items with DIF was increased, the error increased in IRT-LR and SIBTEST methods, with a larger increase in the SIBTEST method.

This study is different from the other simulation studies (Golia, 2015; Lee & Zhang, 2017; Li & Zumbo, 2009) in terms of the method used to detect DIF, number of items in the test, and number of response categories; from this point of view, it aims to evaluate the conditions. This has not been previously covered in the literature. This research also differs from other studies in the literature in terms of purifying the DIF items identified in the methods.

Purpose of the Study

In this study, the aim is to investigate how the errors will change depending on the ability estimates for the DIF detection methods -SIBTEST, IRT-LR, Lord's χ2 and Raju's area measures- when the items with DIF are purified from the test under the ratio of the number of items with DIF, effect size of DIF, and type of DIF.

METHOD

Research Design

Because the performances of different DIF detection methods are examined under specific conditions and based on the ability estimation obtained by purifying items with DIF from the test, this study was conducted as a Monte Carlo simulation study.

Simulation Conditions

The study investigates DIF detection methods—SIBTEST, IRT-LR, Lord's χ2 and Raju's area measures—through purifying items with DIF according to ratio of items with DIF, effect size of DIF (for SIBTEST and IRT-LR), and the type of DIF. The reason for choosing these four methods in the research is that they are frequently preferred in DIF researches and they are curious about the performance of these methods when item purifying applied. Atalay Kabasakal et al. (2014), Finch (2005), Finch and French (2007), and Lopez's (2012) studies investigated DIF according to IRT and even though SIBTEST is a CTT-based and a non-parametric method they have used SIBTEST method in their studies. For this reason SIBTEST was included in the current study. Hence, Finch (2005) compared the IRT-based IRT-LR method and the SIBTEST method in his study and pointed out that the SIBTEST provided effective results for the short tests. Also, researchers have included the SIBTEST method in a DIF study based on IRT and CAT (Lei, Chen, & Yu, 2006).

In the current study, sample size, test length, ability distribution, item type, and type of IRT model are constant. In the first place, Item type, test length, and IRT model are determined as simulation conditions. Thirty dichotomous items (1-0) were generated according to 3PLM (the three parameter logistic model), which considers the case of responding correctly by chance. Thirty-item tests were selected because the number of items is close to the number of items in high stakes tests in Turkey. Moreover, Downing and Haladyna (2004) indicate that usually a minimum of 30 items are used in achievement tests in order to be representative for the investigating area. Glas and Meijer (2003) used 30 items for the short test form in their simulation study conducted with item response theory. Suh (2016) also created a 30-item test form in their study about multidimensional IRT and DIF.

Secondly ability distribution and sample size are decided as simulation conditions. Ability parameters consisting of 1000 people were generated using normal distribution. Shepard, Camilli, and Averill (1981) stated that it is required to use at least 1000 people in order to obtain stable results.

In this study, the first condition tested for impact was the ratio of the items with DIF. The ratio of the items with DIF was determined to be 10% and 20%. Narayanan and Swaminathan (1994) stated that a 20% DIF item ratio is the worst scenario. In their research, Jodoin and Gierl (2001) studied the 10% and 20% items with DIF ratios. Thus, in 30-item tests, three and six items were made with DIF. The second condition tested for impact was the effect size of DIF. The effect sizes were examined in two ways as C level and B & C level for the methods of IRT-LR and SIBTEST. B & C and C levels were included in the study in order to evaluate the effect of items with middle level (B level) DIF on the ability estimation. The types of DIF were examined through the determination of uniform DIF, non-uniform DIF, and both uniform and non-uniform DIF. The simulation conditions are summarized in Table 1.

Table 1. Simulation Conditions

	Rates of items with DIF	SIBTEST		IRT-LR		Lord χ^2	Raju Area Measure
		B Level	B & C Level	B Level	B & C Level		
Non-uniform	10%	√	√	√	√	√	√
	20%	√	√	√	√	√	√
Uniform	10%	√	√	√	√	√	√
	20%	√	√	√	√	√	√
Non-uniform and uniform	10%	√	√	√	√	√	√
	20%	√	√	√	√	√	√

Data Generation

Firstly, item parameters were generated. In accordance with 3PLM, item parameters were obtained through the software WINGEN 3 (Han, 2007). While generating parameters, the item parameters that are usually encountered in real test applications were used. From the item parameters, a discrimination parameter was generated using lognormal distribution with a mean of 0 and a standard deviation of 0.2; the difficulty parameter was generated by normal distribution with a mean of 0 and standard deviation of 1; the guessing parameter was generated by beta distribution with an a-value of 8 and a b-value of 32. Kim and Lee (2004) also used similar distributions and values while obtaining test forms in their simulation study. The generated test form is shown in Table 2.

Table 2. Item Parameters in the Test Form

Item No	Model	Number of Cathogory	a	b	c	Item No	Model	Number of Cathogory	a	b	c
1	3PLM	2	1.130	-.727	.216	16	3PLM	2	1.114	-1.353	.322
2	3PLM	2	.791	-1.606	.241	17	3PLM	2	1.384	-1.817	.125
3	3PLM	2	1.491	0.928	.197	18	3PLM	2	1.118	.361	.222
4	3PLM	2	1.252	.348	.173	19	3PLM	2	.911	.276	.273
5	3PLM	2	1.236	1.488	.177	20	3PLM	2	1.723	-.044	.208
6	3PLM	2	.913	-2.291	.151	21	3PLM	2	.993	.525	.336
7	3PLM	2	.824	-.840	.122	22	3PLM	2	1.045	.207	.239
8	3PLM	2	.680	-1.333	.178	23	3PLM	2	.785	.591	.159
9	3PLM	2	1.008	-.669	.088	24	3PLM	2	.963	.064	.213
10	3PLM	2	1.128	-.253	.201	25	3PLM	2	1.259	.047	.116
11	3PLM	2	.781	1.036	.145	26	3PLM	2	.933	-1.285	.267
12	3PLM	2	.994	1.524	.162	27	3PLM	2	1.109	.984	.148
13	3PLM	2	.822	.464	.261	28	3PLM	2	1.077	-.296	.171
14	3PLM	2	.957	1.879	.146	29	3PLM	2	.952	-.462	.164
15	3PLM	2	1.106	-.267	.195	30	3PLM	2	.949	.947	.219

After generating item parameters, ability parameters were generated by normal distribution with a mean of 0 and standard deviation of 1. For the tests consisting of uniform and non-uniform or both types of DIF items, the ability parameters were obtained similarly. Mazor, Clauser, and Hambleton (1993) examined non-uniform DIF and generated abilities for a reference group with a similar distribution and values. In order to make sure that the results are stable, this was repeated 50 times in the study. Harwell, Stone, Hsu, and Kirisci (1996) reported that this should be repeated at least 25 times in Monte Carlo simulation studies. Finally, 1-0 data were created by applying the items to the individuals.

The obtained 1-0 data were rescaled using the software PARSCALE 4.1 (Muraki & Bock, 2003). This process was done to obtain 50 ability parameters by using items without DIF and to fix abilities for each condition. The a-parameter was increased by .75 for displaying some items in the test to display non-uniform DIF. A similar rate was used in the study of Mazor, Clauser and Hambleton (1993). They stated that by considering the b-parameter, the difference in a-parameter over a value of .50 increased the rate of detection. Furthermore, the b-parameter was increased by .60 for displaying items in the test uniform DIF. Because the rate of DIF item conditions were being examined, in the first case, this process was applied to three items (Items 7, 12, and 26) and in the second case it was applied to six items (Items 6, 9, 12, 17, 21, and 29).

For displaying both uniform and non-uniform DIF items in the test, in the case of three items, DIF b-parameters of two items were increased by .60 and the a-parameter of one item was increased by .75; in the case of six items, DIF b-parameters of four items were increased by .60 and a-parameters of two items were increased by .75. DIF was randomly assigned to the items. Items with DIF were applied to an individual by using WINGEN; thus, 1-0 data were obtained for focal and reference groups. Simulation conditions were checked by comparing the parameters obtained from focal and reference groups.

Data Analysis

Binary data of focal and reference groups were analyzed using SIBTEST (Li & Stout 1994), IRTLRDIF (Thissen, 2001), and the difR package in R software (Magis, Beland, Tuerlinckx, & De Boeck, 2010; Magis, Beland, & Raiche 2013). For each condition in the SIBTEST and IRTLRDIF software, items with C level DIF and then items with B & C level DIF were removed from the response matrix and estimated using PARSCALE 4.1 software. Using the difR package, items that demonstrated significant DIF according to Lord χ2 and Raju's area measures were removed from the response matrix and estimated similarly with PARSCALE 4.1 software. In order to compare the methods, root mean squared difference (RMSD) and the coefficient of concordance (Pearson correlation coefficient) were calculated from estimated and initial abilities. Below, the criteria used are explained in detail.

RMSD (root mean squared difference)

To calculate RMSD, first the square of the difference between estimated and real ability values were found and summed. After that, this value was divided by the frequency of ability level and the square root of the result was calculated. The following is the equation of the RMSD:

θ: Real ability level
θ*: Estimated ability level
f: Frequency of ability level

$$\text{RMSD} = \sqrt{\frac{\sum_i f_i(\theta^* - \theta)^2}{\sum_i f_i}} \quad (1)$$

Coefficient of concordance

The coefficient of concordance was calculated depending on the mean of Pearson correlation coefficients between estimated and real abilities of an individual.

In order to determine the effectiveness of DIF detecting methods, all RMSD values and coefficients of concordance that were obtained as a result of repetition according to simulation conditions were examined with the significance tests. For this, firstly the normality of data according to DIF detecting methods were examined and, if the normality conditions were not met, the methods were compared using a Kruskal-Wallis H test. Group comparisons were made by nonparametric multiple comparison test. The η2 value was calculated to determine the effect of DIF detecting methods on RMSD and coefficient of concordance coefficients. The size of the eta square of .01, .06 and .14 respectively shows small, medium and large effect size (Green & Salkind, 2005). The following is the equation of the η2:

χ2: Chi square value

N: Sample size

$$\eta 2 = \chi 2 / (N-1)$$

RESULTS

The research results were examined within the framework of the research question and the DIF detecting methods were compared using the error (RMSD) and coefficient of concordance.

The results, obtained from detecting items with DIF and removing them with the different methods according to 10% and 20% item rates and uniform, non-uniform, and both uniform and non-uniform DIF types, are shown in Table 3.

Table 3. The Coefficients of Error and Concordance for DIF Conditions

	DIF	SIBTEST		IRT-LR		Lord χ^2		Raju Area Measure	
	Rates	RMSD	Pearson	RMSD	Pearson	RMSD	Pearson	RMSD	Pearson
Non-Uniform	%10	.581435	.751599	.584374	.748612	.586027	.746705	.610559	.714193
	%20	.585285	.747123	.598210	.734050	.598814	.733239	.599150	.732580
Uniform	%10	.579508	.753530	.511010	.781589	.583243	.749699	.586162	.746380
	%20	.590214	.742381	.565683	.753397	.589310	.742946	.587441	.744388
Non-Uniform	%10	.578935	.753578	.521621	.777584	.578815	.753490	.579444	.752800
&Uniform	%20	.587103	.745318	.602092	.726336	.592590	.739431	.593482	.738539

Table 3 illustrates that when the rate of DIF items increases, removing DIF items increases the error. Only when using Raju's area measures for the non-uniform DIF type, removing DIF items decreased the error when the rate of DIF items increased. As a result of removing items with DIF in all conditions, the method of IRT-LR showed the minimum error in the 10% rate of DIF and uniform DIF type. If the coefficients of concordance were examined, after removing DIF items, the method of IRT-LR showed the maximum correlation in the 10% rate of DIF and uniform DIF type. Furthermore, it is possible to state that, generally, for all types of DIF, correlation coefficients calculated by removing DIF items decrease when the rate of DIF increases. Only in the condition of non-uniform DIF does the coefficient of concordance calculated as a result of removing DIF items increase according to the rate of DIF for the Raju method. Table 4 shows whether the RMSD and the coefficients of concordance have a significant difference according to the DIF detection method.

Table 4. The Results of Kruskal-Wallis H Test of RMSD and Coefficients of Concordance According to DIF Detecting Methods

	DIF detection method	N	Mean Rank	df	χ^2	p	Difference
RMSD	SIBTEST	300	549.53	3	10.584	.014	SIBTEST - Raju Area Measure
	IRT-LR	300	595.53				
	Lord χ^2	300	623.47				
	Raju Area Measure	300	633.47				
Pearson	SIBTEST	300	653.77	3	11.684	.009	SIBTEST - Lord χ^2 SIBTEST - Raju Area Measure
	IRT-LR	300	606.14				
	Lord χ^2	300	577.12				
	Raju Area Measure	300	564.98				

Table 4 shows that there is a significant difference between coefficients of RMSD obtained from the simulation conditions according to DIF detecting methods [$\chi 2$=10.584, *p*=.014]. The nonparametric multiple comparisons which were conducted to investigate which groups this difference occurs between indicate that the difference in RMSD coefficients are between the methods of SIBTEST and Raju's area measures. Therefore, it can be stated that the mean rank of SIBTEST (549.53) is lower than the mean rank of Raju area measure (633.47). In addition, the median of SIBTEST (.585) is lower than the median of Raju area measure (.588). This means that the error value (RMSD) of SIBTEST is lower than Raju area measure. The η2 value was calculated to determine the effect of DIF detecting methods on RMSD coefficients. Consequently, the effect size (η2=.01) was found to be low (Green & Salkind, 2005). Similarly, it can be seen that there is a significant difference between coefficients of concordance obtained from the simulation conditions according to DIF detecting methods [χ^2=11.684, p=.009]. The nonparametric multiple comparisons, which were conducted to investigate which groups this difference occurs between, indicate that the difference in concordance coefficients are between the methods of SIBTEST and Lord χ2, as well as SIBTEST and Raju's area measures. Therefore, it can be stated that the mean rank of SIBTEST (653.77) is higher than the mean ranks of Raju area measure (564.98) and Lord χ2 (577.12). In addition, the median of SIBTEST (.749) is higher than the medians of Raju area measure (.745) and Lord χ2 (.744). This means that the coefficient of concordance of SIBTEST is higher than Raju area measure. The η2 value was calculated to determine the effect of DIF detecting methods on concordance coefficients; thus, the effect size (η2=.01) was found to be low level (Green & Salkind, 2005).

In order to assess the effect of purifying items with B level DIF from the test on ability estimation, firstly items with C level DIF and then items with B & C level DIF in the methods of SIBTEST and IRT-LR were extracted from test; the abilities were estimated later. The error and coefficient of concordance values calculated from the ability levels which were obtained in both cases are shown in Table 5.

Table 5. The Effect of Extracting B-Level DIF Items on the Error and Concordance Coefficients

			SIBTEST		IRT-LR	
		DIF Effect Level	RMSD	PEARSON r	RMSD	PEARSON r
Non-uniform	10 %	C	.576762	.756118	.380603	.839445
		B & C	.581435	.751599	.584374	.748612
	20 %	C	.576233	.756162	.583750	.744341
		B & C	.585285	.747123	.598210	.734050
Uniform	10 %	C	.574934	.757978	.000000	1.00000
		B & C	.579508	.753530	.511010	.781589
	20 %	C	.570526	.761617	.000000	1.00000
		B & C	.590214	.742381	.565683	.753397
Non-uniform and uniform	10 %	C	.572760	.759623	.046230	.980451
		B & C	.578935	.753578	.521621	.777584
	20 %	C	.569988	.762370	.081300	.966065
		B & C	.587103	.745318	.602092	.726336

Table 5 shows that in the methods of SIBTEST and IRT-LR the error values obtained from purifying C level DIF items are lower than the errors obtained from purifying B & C level DIF items when the rate of DIF items are 10% and 20% and when the type of DIF changes. Both methods at the rate of 10% and 20% DIF showed that the correlation coefficients calculated by purifying C level DIF items in all DIF type conditions were higher than the correlation coefficients calculated by purifying B & C level DIF items.

DISCUSSION and CONCLUSION

This study aims to investigate the effect of purifying DIF items from a test by using different DIF detection methods on individuals' ability estimates. For this purpose, a simulation study was conducted and firstly item parameters and depending on this the ability parameters were generated. In the fifty-replication study, the data set were generated according to 1000 participants' responses to 30 items and the ability estimates were rescaled after purifying items with DIF.

The abilities determined and scaled through items without DIF are accepted as real abilities. The cases of 10% and 20% DIF items rates in the uniform, non-uniform and both uniform and non-uniform DIF types were examined. Different methods to detect DIF (SIBTEST, IRT-LR, Lord's χ2, Raju’s area measures) were used and discussed the effects of these methods on ability estimations. For two conditions in three items with DIF and six items with DIF, the abilities were estimated again after purifying DIF items determined by the methods and then the concordance and error coefficients were calculated according to each method. For the methods of SIBTEST and IRT-LR, purifying only C-level DIF items the ability estimates were calculated and then purifying B & C level DIF items the abilities were estimated. Since there is no such distinction for the methods of Lord’s $\chi 2$ and Raju’s area measures, the values were compared by purifying DIF items at one time.

DIF is caused by the fact that the probability to respond an item correctly of a group is more or less relative to other group depends on not the ability level but the group (Osterlind, 1983; Zumbo 1999). Therefore, the existence of DIF items in the test can cause bias and error in individuals’ ability estimations (Camilli, 1993). In other words, DIF is an indicator of systematic error of measurement (Camilli & Shepard, 1994; Kelecioğlu, Karabay & Karabay, 2014). Although DIF items are threats for the validity, since DIF items will cause a bias in ability estimation (Golia, 2015) purifying items may be seen as an appropriate solution to estimate abilities accurately. Lee and Zhang (2017) have found differences in estimations of ability when the ratio of items with DIF increased. Golia (2015) examined how the ability estimations would change in instrument that belonged the polytomously scored items with DIF. If the test belonged more than one items with DIF, there was a significant bias in estimations of ability. Golia (2010) investigated the effects of keeping and purifying three items with uniform DIF in 15 items tests and found that the goodness of ability estimations was not influenced by this condition when the test belonged a few number of items with DIF. Li and Zumbo (2009) studied on the number of items with DIF and the size of DIF by conducting a simulation study. They pointed out that if there was quite a little number of items with DIF or there was a small number of items with DIF and the size of DIF was small, then there was no bias in ability estimations. They also observed that when the number of items with DIF and the size of DIF increased then the errors changed. The studies indicated that if the size of DIF and the ratio of DIF increase, this increase causes the bias in ability estimations. Therefore, in the current conducted study the effects of purifying items with DIF which are determined by the DIF detecting methods were examined when the ratio of items with DIF 10% and 20%. In this way, not only the effects of purifying items with DIF from the test were observed but also the DIF detecting methods were compared. Concordance and error between ability estimation after purifying item which is detected as with DIF through methods, and true abilities in the case of no items with DIF. Thus, the results state that the error which shows the ability estimation differences, increases when the ratio of items with DIF even if these items are discarded. Tennant and Pallant (2007) indicated that there may be differences in individual ability estimations after purifying items with DIF. Similarly, Golia (2010) studied on polytomously (6) scored 15 items and pointed out that purifying 3 items with DIF from the test negatively affected ability estimations.

According to findings, purifying items with DIF determined by the method of IRT-LR yielded the most concordant and the least inaccurate results with the real abilities. The highest error and the lowest concordance were obtained in the estimation through excluding items with DIF determined by the method of Raju’s area measure. When the number of items with DIF increases, errors generally increase but in the method of Raju’s area measure the error may decrease. Atalay Kabasakal, Arsan, Gök and Kelecioğlu (2014) compared DIF detecting methods (MH, SIBTEST, IRT-LR) in a simulation study and found that IRT-LR method had the smallest error. In this study which compares methods according to ability estimations, the similar relationship was found in RMSD and Pearson Correlation concordance index. On the other hand, Finch (2005) compared the methods of MH, SIBTEST, IRT-LR and MIMIC and stated that the increase in the number of items with DIF was more effective on IRT-LR method. However, in some different studies under the different conditions different results were obtained according to methods. Therefore, it will be more appropriate to discuss which method under which conditions gave results with the highest concordance and the lowest error. Considering the error and concordance in the nonparametric comparisons based on ability estimations under the conditions of this study, SIBTEST & Lord’s $\chi 2$ and SIBTEST & Raju’s area measure produced different results. Finch and French (2007) conducted a study on nonuniform DIF and compared the methods LR, SIBTEST, IRT-LR and confirmatory factor analysis. They indicated that DIF size, sample size, ability distributions and IRT model had no significant impact on methods when the error was considered. In the current study, it was found that the manipulated factors did not cause a significant difference for the methods of IRT-LR and SIBTEST.

The methods of Lord's $\chi 2$ and Raju's area measures are based on the parameter estimations. Therefore, while determining DIF these methods may be affected by the algorithms used in item parameter estimations (Cohen & Kim, 1993). As a result of this, it is thought that the concordance coefficients of these methods may be lower than the others. Furthermore, in the method of Raju's area measure the situation of when the number of items with DIF increases the error decreases may be caused by the characteristics that the methods are based on.

In this study, for only the methods of SIBTEST and IRT-LR, both the cases of excluding C-level DIF items and the case of excluding B & C level DIF items were examined and compared. In the methods of SIBTEST and IRT-LR under the conditions of 10% and 20% DIF items ratio, when only C-level DIF items were extracted, the error ratio was found to be lower and the concordance index were found to be higher. Lee and Zhang (2017) remark that when the items with DIF is in C level instead of B level, the difference in ability estimations will be larger. The results support this finding. Since items in B level do not affect ability estimations negatively as in C level, keeping B level items in test may decrease the error of ability estimations. Furthermore, purifying items in B and C level decreases the number of items in test. This situation may cause finding the larger error after purifying items in B and C level. In this situation, for SIBTEST and IRT-LR under this condition, it can be said that the error of ability estimation increases when items with B-level DIF are extracted from the test. Therefore, for the conditions in this study it may be suggested that items with B-level DIF should not be excluded from the test in the methods of SIBTEST and IRT-LR.

In the scope of this study, for the investigation of the effect of purifying DIF items from the test on the ability estimations, different methods were compared according to uniform, non-uniform, both uniform and non-uniform DIF types under the 10% and 20% DIF item ratios. There were differences between the methods in terms of the error and concordance coefficients. Further studies may repeat this under similar conditions by using different IRT estimation methods. Moreover, when the conditions and methods change the obtained results will be different. Therefore, the effect of purifying items with DIF on ability estimations may be examined under different conditions and using different methods.

REFERENCES

Atalay Kabasakal, K., Arsan, N., Gök, B., & Kelecioğlu, H. (2014). Comparing performances (type I error and power) of IRT Likelihood Ratio SIBTEST and Mantel-Haenszel methods in the determination of differential item functioning. *Educational Sciences: Theory & Practice, 14*(6), 2175–2193. doi: 10.12738/estp.2014.6.2165

Camilli, G. (1993). The case against item bias detection techniques based on internal criteria: Do item bias procedures obscure test fairness issues? In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 397–418). New York: Routledge.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. California: Sage.

Cheng, C. M. (2005). *A study on Differential Item Functioning of the basic mathematical competence test for junior high schools in Taiwan* (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database (UMI No. 3189625).

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31-47. doi: 10.1111/j.1745-3992.1998.tb00619.x

Cohen, A. S., & Kim, S. (1993). A comparison of Lord's Chi Square and Raju's Area Measures in detection of DIF. *Applied Psychological Measurement*, *17*(1), 39–52. doi: 10.1177/014662169301700109

Cromwell, S. (2002, February). *A primer on ways to explore item bias*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327-333. https://doi.org/10.1046/j.1365-2923.2004.01777.x

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT Likelihood Ratio. *Applied Psychological Measurement, 29*(4), 278-295. doi: 10.1177/0146621605275728

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning a comparison of four methods. *Educational and Psychological Measurement, 67*(4), 565-582. doi: 10.1177/0013164406296975

Glas, C. A. W., & Meijer, R. R. (2003). A Bayesian approach to person fit analysis in Item Response Theory models. *Applied Psychological Measurement, 27*(3), 217-233. doi: 10.1177/0146621603252216

Golia S. (2010). The assessment of DIF on Rasch measures with an application to job satisfaction. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation, 1*(1) 16–25. doi: 10.1285/i2037-3627v1n1p16

Golia S. (2015). Assessing the impact of uniform and nonuniform differential item functioning items on Rasch measure: The polytomous case. *Computational Statistics, 30*, 441–461. doi: 10.1007/s00180-014-0542-x

Green, S., & Salkind, N. (2005). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (4th Ed). Upper Saddle River, NJ: Pearson.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. California: Sage.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459. doi: 10.1177/0146621607299271

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in Item Response Theory. *Applied Psychological Measurement, 20*(2), 101-125. doi: 10.1177/014662169602000201

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

Kelecioğlu, H., Karabay, B., & Karabay, E. (2014). Investigation of placement test in terms of item biasness. *Elementary Education Online, 13*(3), 934–953.

Kim, S., & Lee, W. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report 2004-5). Iowa City, IA: Act, Inc.

Kristanjansonn, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response model. *Educational and Psychological Measurement, 65*(6), 935-953. doi: 10.1177/0013164405275668

Lee, Y. H., & Zhang, J. (2017). Effects of differential item functioning on examinees' test performance and reliability of test. *International Journal of Testing, 17*(1), 23-54. https://doi.org/10.1080/15305058.2016.1224888

Lei, P-W., Chen, S-Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement, 43*(3), 245-264. https://doi.org/10.1111/j.1745-3984.2006.00015.x

Li, H. H., & Stout, W. (1994). SIBTEST: A fortran V program for computing the simultaneous item bias DIF statistics. Department of Statistics, University of Illinois, Urbana Champaign.

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicológica, 30*, 343-370.

Lopez, G. E. (2012). *Detection and classification of DIF types using parametric and nonparametric methods: A comparison of the IRT-Likelihood Ratio test, Crossing-SIBTEST, and Logistic Regression procedures* (Doctoral dissertation). Retrieved from http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=5327&context=etd

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous Differential Item Functioning. *Behavior Research Methods, 42*(3), 847–862. doi: 10.3758/BRM.42.3.847

Magis, D., Beland, S., & Raiche, G. (2013). difR: Collection of methods to detect dichotomous Differential Item Functioning (DIF) in psychometrics. R package version 5.0. http: //www.CRAN.R-project.org/package=difR

Mazor, K. M., Clauser, R. E., & Hambleton, R. K. (1993, March). *Identification of nonuniform Differential Item Functioning using a variation of the Mantel-Haenszel procedure.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105–118. doi: 10.3102/10769986007002105

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi: 10.1037/0003-066X.50.9.741

Muraki, E., & Bock, R. D. (2003). PARSCALE 4 for Windows: IRT based test scoring and item analysis for graded items and rating scales [Computer software]. Skokie, IL: Scientific Software International, Inc.

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential. *Applied Psychological Measurement, 18*(4), 315-328. doi: 10.1177/014662169401800403

Oshima, T. C., & Morris, S. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice, 27*(3), 43-50. doi: 10.1111/j.1745-3992.2008.00127.x

Osterlind, S. J. (1983). *Test item bias.* Newbury Park, California: Sage.

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

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

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*(2), 215-230. Retrieved from http://www.jstor.org/stable/1435184

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59*(2), 248-269. doi: 10.1177/00131649921969839

Rupp, A. A., & Zumbo, B. D. (2003). Which model is best? Robustness properties to justify model choice among unidimensional IRT models under item parameter drift. *Alberta Journal of Educational Research, 49*, 264-276.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84. doi: 10.1177/0013164404273942

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*(4), 317–375. doi: 10.2307/1164616

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

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if Differential Item Functioning makes a difference. *Rasch Measurement Transactions, 20*(4), 1082-1084.

Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio tests for Differential Item Functioning. L.L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, NC.

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

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*(1), 77–87. doi: 10.1177/0146621602261005

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

Değişen Madde Fonksiyonu Belirlemede Yöntemlerin Yetenek Kestirimine Dayalı Performansları: Bir Benzetim Çalışması

Giriş

Madde yanlılığın incelenme süreci daha nesnel sonuçlara dayanan ve madde yanlılığının bir ölçüsü olabilecek değişen madde fonksiyonunun (DMF) incelenmesi ile başlar. DMF aynı yetenek düzeyinde fakat farklı gruplardaki kişilerin bir maddeyi doğru yanıtlama olasılıklarının birbirinden farklılaşması olarak tanımlanmaktadır (Hambleton, Swaminathan ve Rogers, 1991). DMF ortaya çıkışı açısından tek biçimli (uniform) ve tek biçimli olmayan (non-uniform) fonksiyonlar şekilde sınıflandırılır (Mellenbergh, 1982). Bu farklılaşmanın temelinde yatan gerçek ise yetenek düzeyi ile grup üyeliğinin birlikte maddeyi doğru yanıtlama olasılığını etkilemesidir. Buna göre tek biçimli DMF, aynı yetenek düzeyindeki iki grubun bir maddeye doğru yanıt verme olasılıklarının tüm yetenek düzeyleri için sabit bir değer olması durumunda meydana gelir. Buna karşın tek biçimli olmayan DMF ise aynı yetenekteki iki grubun maddeye doğru yanıt verme olasılıkları farklı yetenek düzeylerinde tutarsız olduğu durumda oluşur (Camilli ve Shepard, 1994; Penfield ve Lam, 2000; Zumbo, 1999).

DMF belirleme teknikleri temelde Klasik Test Kuramı (KTK) ve Madde Tepki Kuramına (MTK) göre sınıflandırılmaktadır. KTK’ya göre DMF belirleme yöntemleri varyans analizi, ki-kare, dönüştürülmüş madde indeksi, lojistik regresyon, Mantel-Haenszel (MH) ve SIBTEST’tir. MTK yöntemleri ise Lord’un χ2’si, Raju’nun alan ölçüsü ve MTK-olabilirlik oranı (MTK-OO)’dır (Camilli ve Shepard, 1994; Oshima ve Morris, 2008).

Bir testte DMF’nin belirlenmesin yanında DMF gösteren madde bulunduğunda ona ne yapılacağına karar verilmesi önemlidir. Yansızlığı sağlamak adına ilgili maddenin testten çıkarılması gerekebilir. Buna karşın ilgili madde ölçülen örtük özellik ya da yapının önemli ya da zorunlu maddesiyse maddenin atılması uygun olmayabilir. Bazen ilgili maddenin yeniden ifade edilmesi DMF’nin ortadan kalkmasını sağlayabilirken bazen bu çözüm yeterli olmayabilir (Golia, 2015). Testte DMF’li maddeler bulunduğunda bu maddelerin test istatistiklerini, sonuçları, bireylere ait puanları etkileyeceği bilinmekte fakat bu etkinin nasıl olacağı bilinmemektedir (Li ve Zumbo, 2009). Eğer maddenin testten çıkarılmasına karar verilirse, testteki madde sayısının azalmasına bağlı olarak testin geçerliliği düşürebilir. Bununla birlikte DMF’li maddelerin testten çıkarılmasının yetenek kestirimini hangi düzeyde etkileyeceği kestirilememektedir. Bu çalışmada bu soruya yanıt aramaktadır. Bununla birlikte orta (B) düzeydeki DMF’li maddelerin testten çıkarılmasının etkileri de incelenmektedir.

Alanyazında maddelerin ikili puanlandığı (Lee ve Zhang, 2017; Li ve Zumbo, 2009; Roznowski ve Reith, 1999; Rupp ve Zumbo, 2003, 2006; Wells, Subkoviak ve Serlin, 2002) ve çoklu puanlandığı (Golia, 2010, 2015; Tennant ve Pallant, 2007) durumlarda testten madde çıkarılmasının teste ilişkin istatistikleri nasıl değiştiğine dair çalışmalar bulunmaktadır. Bu çalışmaların bir kısmı madde parametreleri değişmezliği kapsamında bu durumu incelerken (Roznowski ve Reith 1999; Rupp ve Zumbo, 2003, 2006; Well, Subkoviak ve Serlin, 2002), bazıları ise ilgili durumu bu çalışmada olduğu gibi DMF kapsamında parametre değişmezliği olarak ele almıştır (Golia, 2010, 2015; Lee ve Zhang, 2017; Li ve Zumbo, 2009; Tennant ve Pallant, 2007).

Bu araştırmada DMF belirleme yöntemlerinden SIBTEST, MTK-OO, Lord’un χ2’si ve Raju’nun alan ölçüsünün DMF’li madde oranı ve DMF etki büyüklüğü altında DMF’li maddelerin testten çıkarılması durumunda yetenek kestirimine dayalı olarak hataların nasıl değiştiğinin incelenmesi amaçlanmaktadır.

Yöntem

Araştırmada farklı DMF belirleme yöntemlerinin performansları, belirli koşullar altında DMF’li maddelerin testten çıkarılmasıyla elde edilen yetenek kestirimine dayalı olarak incelendiğinden bir Monte Carlo benzetim çalışması yürütülmüştür.

Araştırma SIBTEST, MTK-OO, Lord χ2, Raju’nun alan ölçüleri DMF belirleme yöntemlerini DMF’li madde oranları, DMF etki büyüklüğü (SIBTEST ve MTK-OO için) ve DMF türüne göre, tespit edilen DMF’li maddelerin testten çıkarılmasıyla incelemektedir. Bu araştırmada sıklıkla kullanılan DMF yöntemleri seçilmiştir. Bunun sebebi sıklıkla kullanılan bu yöntemlerin maddelerin testten çıkarılması durumundaki performanslarını belirlemektir. SIBTEST KTK’ya dayalı olması ve parametrik olmayan bir yöntem olmasına rağmen araştırmaya dahil edilmiştir. Bunun sebebi SIBTEST yönteminin Atalay Kabasakal vd. (2014), Finch (2005), Finch ve French (2007), Lopez

(2012) gibi araştırmacılar tarafından madde tepki kuramında gerçekleştirilen DMF çalışmalarına dahil edilmesidir. Nitekim Finch (2005) araştırmasında bir MTK yöntemi olan IRTLR ile SIBTEST yöntemini karşılaştırmış ve kısa testlerde SIBTEST'in etkili sonuçlar verdiğini belirlemiştir. CAT temelinde ve MTK'ya dayalı olarak gerçekleştirilen bir DMF araştırmasında (Lei, Chen ve Yu, 2006) da SIBTEST'e yer verildiği görülmektedir.

Örneklem büyüklüğü, test uzunluğu, yetenek dağılımı, madde türü, MTK model türü koşulları araştırmada sabit tutulmuştur. Araştırmada belirlenen koşullardan ilki madde türü, test uzunluğu ve MTK modelidir. Araştırmada ikili puanlanan (1-0) 30 madde şansla doğru cevaplama olasılığını da dikkate alan (Baker, 2001) 3PLM'ye göre oluşturulmuştur. 30 maddelik testler Türkiye'de geniş ölçekli testlerde karşılaşılan madde sayısına yakın olduğu için seçilmiştir. İkinci koşul yetenek dağılımı ve örneklem büyüklüğüdür. 1000 kişiden oluşan yetenek parametreleri normal dağılım kullanılarak oluşturulmuştur. Shepard, Camilli ve Averill (1981) kararlı sonuçlar elde edebilmek için en az 1000 bireyden oluşan örneklemler kullanılması gerektiğini belirtmiştir.

Araştırmada etkisi test edilen koşullardan ilki DMF'li madde oranıdır. DMF'li madde oranı %10 ve %20 olarak belirlenmiştir. Narayanan ve Swaminathan (1994) %20 DMF madde oranının testlerdeki en kötü senaryo olduğunu belirtmiştir. Böylece 30 maddelik testlerde 3 ve 6 madde DMF'li hale getirilmiştir. Etkisi test edilen ikinci koşul DMF etki büyüklüğüdür. MTK-OO ve SIBTEST yöntemleri için etki büyüklükleri C düzeyinde, B ve C düzeyinde olmak üzere iki durum altında incelenmiştir. C, B ve C düzeyleri orta düzeydeki (B düzeyi) DMF'li maddelerin yetenek kestiriminde bulunmasının etkisini değerlendirmek amacıyla araştırmaya dâhil edilmiştir. DMF türü tek biçimli, tek biçimli olmayan, hem tek biçimli hem tek biçimli olmayan DMF'nin tespiti üzerinden incelenmiştir.

Verilerin türetilmesi aşamasında öncelikle madde parametreleri 3PLM'e uygun olarak WINGEN 3 (Han, 2007) programıyla elde edilmiştir. Parametreler elde edilirken gerçek test uygulamalarında genellikle karşılaşılan madde parametreleri kullanılmıştır. Madde parametrelerinden ayırıcılık parametresi ortalaması 0, standart sapması ,2 olan lognormal dağılımla, güçlük parametresi ortalaması 0 standart sapması 1 olan normal dağılımla, şans parametresi ise a değeri 8, b değeri 32 olan beta dağılımıyla oluşturulmuştur.

Madde parametrelerinin türetilmesinin ardından ortalaması 0 standart sapması 1 olan normal dağılımla yetenek parametreleri türetilmiştir. Tek biçimli, tek biçimli olmayan ya da her iki DMF türündeki maddelerin bir arada yer aldığı testler için yetenek parametreleri benzer dağılımlarla elde edilmiştir. Sonuçların kararlılığından emin olmak amacıyla araştırmada 50 tekrar yapılmıştır. Harwell, Stone, Hsu ve Kirisci (1996) Monte Carlo benzetim çalışmalarında en az 25 tekrar kullanılması gerektiğini belirtmiştir. Son olarak bireylere maddeler uygulanarak 1-0 verilerinin elde edilmesi sağlanmıştır.

Elde edilen 1-0 verileri PARSCALE 4.1 (Muraki ve Bock, 2001) programıyla tekrar ölçeklenmiştir. Bu işlem 50 yetenek parametresinin DMF'siz maddeler üzerinden elde edilmesi ve her bir koşul için yeteneklerin sabitlenmesi için gerçekleştirilmiştir. Bazı maddelerin tek biçimli olmayan DMF göstermesi için a parametresi ,75 arttırılmıştır. Benzer oran Mazor, Clauser ve Hambleton (1993)'ın çalışmasında kullanılmıştır. Mazor, Clauser ve Hambleton (1993) b parametresi de dikkate alınarak a parametresinin ,50 üzerindeki farkının tespit oranını yükselttiği belirtilmiştir. Bunun yanında testteki maddelerin tek biçimli DMF göstermesi için b parametresine ,60 oranında arttırım uygulanmıştır. Bu işlem; DMF'li madde oranı koşulları incelendiği için ilk durumda 3 maddeye (7, 12 ve 26. maddeler), ikinci durumda ise 6 maddeye (6, 9, 12, 17, 21 ve 29. maddeler) uygulanmıştır. Testteki maddelerin hem tek biçimli hem de tek biçimli olmayan DMF göstermesi için ise 3 maddenin DMF'li olduğu durumda 2 maddenin b parametresine ,60 oranında, 1 maddenin a parametresine ,75 oranında; 6 maddenin DMF'li olduğu durumda 4 maddenin b parametresine ,60 oranında, 2 maddenin a parametresine ,75 oranında arttırım uygulanmıştır. DMF, maddelere seçkisiz olarak atanmıştır. DMF'li maddeler WINGEN programıyla bireylere uygulanmış ve böylece odak ve referans grupları için 1-0 verileri elde edilmiştir.

Odak ve referans gruplarına ait ikili puanlan veriler SIBTEST (Li ve Stout 1994), IRTLRDIF (Thissen, 2001) ve R programında yer alan difR (Magis, Beland, Tuerlinckx ve De Boeck, 2010; Magis, Beland ve Raiche 2013) paketi kullanılarak analiz edilmiştir. SIBTEST ve IRTLRDIF programlarında her koşul için öncelikle C ve sonrasında B ve C düzeyinde DMF'li bulunan maddeler cevap matrisinden çıkarılarak PARSCALE 4.1 programıyla kestirim yapılmıştır. difR paketi ile Lord χ2, Raju'nun alan ölçülerine göre anlamlı DMF gösteren maddeler cevap matrisinden çıkarılarak PARSCALE 4.1 programıyla benzer şekilde kestirim yapılmıştır. Yöntemleri karşılaştırabilmek için referans gruplar için kestirilen yetenekler ve ilk yetenekler üzerinden hata (RMSD) ve uyum katsayısı (Pearson korelasyon katsayısı) hesaplanmıştır.

DMF belirleme yöntemlerinin etkililiğini belirlemek amacıyla benzetim koşullarına göre yapılan tekrarlar sonucunda elde edilen tüm RMSD ve uyum katsayıları anlamlılık testleriyle incelenmiştir. Bunun için öncelikle verilerin DMF belirleme yöntemlerine göre normalliği incelenmiş ve normallik koşulları sağlanmadığından Kruskal-Wallis H testi ile yöntemler karşılaştırılmıştır. Yöntemler arasında ortaya çıkan farklılığın hangi yöntemlerden kaynaklandığını belirlemek üzere nonparametric çoklu karşılaştırma testi kullanılmıştır. η2 değeri aracılığıyla ortaya çıkan farka ilişkin etki büyüklükleri hesaplanmıştır.

Sonuç ve Tartışma

Bu araştırma, farklı DMF belirleme yöntemleri kullanılarak bir testte DMF'li maddelerin çıkarılma durumlarının bireylerin yetenek kestirimine olan etkisini incelemeyi amaçlamaktadır. Araştırmanın sonuçlarına göre MTK-OO yöntemiyle belirlenen DMF'li maddelerin testten çıkarılması gerçek yeteneklerle en uyumlu ve en az hatalı sonucu vermiştir. En yüksek hata ve en düşük uyum ise Raju'nun alan ölçeği yöntemi ile belirlenen DMF'li maddelerin testten çıkarılmasıyla yapılan kestirimde görülmüştür. DMF'li madde sayısı arttığında hatalar genel olarak artarken Raju'nun alan ölçüleri yönteminde hata miktarı azalabilmektedir. Atalay Kabasakal, Arsan, Gök ve Kelecioğlu (2014) DMF belirleme yöntemlerini karşılaştırdıkları benzetim çalışmasında MTK-OO yönteminin Tip 1 hata dikkate alındığında en düşük hatayı verdiğini bulmuştur. Aynı çalışmada SIBTEST yöntemi güç açısından MTK-OO yönteminden daha üstün bulunmuştur. Yetenek kestirimleri üzerinden yöntemlerin karşılaştırıldığı bu çalışmada da benzer bir ilişki RMSD hata ve Pearson korelasyonu uyum indeksi açısından bulunmuştur. Diğer bir yandan Finch (2005), MH, SIBTEST, MTK-OO ve MIMIC yöntemlerini karşılaştırmış ve DMF'li madde sayısının arttığında MTK-OO'nun daha etkili olduğunu belirtmiştir. Ancak birçok farklı çalışmada farklı koşullar altında yöntemlere ilişkin farklı sonuçlar elde edilmektedir. Bu yüzden hangi yöntemin hangi koşullar altında en uyumlu ve en az hatalı sonuçlar verdiğini tartışmak daha doğru olacaktır. Bu çalışmanın koşulları altında yetenek kestirimleri üzerinden yapılan nonparametrik karşılaştırmalarda hata ve uyum dikkate alındığında SIBTEST ve Lord'un χ2'si ile SIBTEST ve Raju alan ölçüleri yöntemlerinin birbirlerinden farklı sonuçlar verdiği görülmektedir. Finch ve French (2007) çalışmalarında tek biçimli olmayan DMF'li maddeler üzerinde lojistik regresyon, SIBTEST, MTK-OO ve doğrulayıcı faktör analizi yöntemlerini karşılatırmış ve DMF büyüklüğü, örneklem büyüklüğü, yetenek dağılımı ve MTK modelinin hata açısından anlamlı bir etkisinin olmadığını belirtmiştir. Bu çalışmada da, manipüle edilen faktörlerin MTK-OO ve SIBTEST yöntemlerinde anlamlı bir farklılığa sebep olmadığı bulunmuştur.

SIBTEST ve MTK-OO yöntemleri için sadece C düzeyinde belirlenmiş maddeler atıldığı, B ve C düzeyinde belirlenmiş maddelerin birlikte atıldığı durumlar araştırmada incelemiş ve karşılaştırılmıştır. SIBTEST ve MTK-OO yönteminde hem %10 hem de %20 DMF'li madde oranı koşullarında sadece C düzeyinde madde atıldığı durumda hata oranı daha düşük ve uyum indeksi daha yüksek bulunmuştur. Bu durumda SIBTEST ve MTK-OO için bu çalışma koşulları altında B düzeyinde belirlenen DMF'lerin testten çıkarılması durumunda yetenek kestirimdeki hataların arttığı söylenebilir. Bu nedenle araştırmada yer alan koşullarda SIBTEST ve MTK-OO yöntemlerinde B düzeyindeki maddelerin testten çıkarılmaması önerilebilir. Lee ve Zhang (2017) araştırmasında

DMF’li maddelerin C düzeyinin altında olmasının testlerde daha düşük etki yaratacağını belirtmektedir.

DMF’li maddelerin çıkarıldığı testlerin bireylerin yetenek kestirimine olan etkilerinin araştırılmasında bu çalışma kapsamında tek biçimli, tek biçimli olmayan, hem tek biçimli hem tek biçimli olmayan DMF türünde %10 ve %20 DMF’li madde barındıran koşullarda farklı yöntemler karşılaştırılmıştır. Hata ve uyum katsayıları açısından yöntemler arasında farklılıklar bulunmuştur. Bundan sonraki çalışmalar benzer koşullarda farklı MTK kestirim yöntemleri kullanılarak tekrarlanabilir. Ayrıca koşullar ve yöntemler değiştikçe elde edilen sonuçlar farklılaşmaktadır. Bu yönde farklı koşullar ve yöntemler kullanılarak DMF’li maddelerin testten çıkarılmasının yetenek kestirimine etkisi incelenebilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(2);149-164

The Effect of Item Weighting on Reliability and Validity*

Abdullah Faruk KILIÇ** Nuri DOĞAN ***

Abstract

The purpose of this study is to examine the effect of the item weighting method developed by researchers on the construct validity of the test. For this purpose, a Monte Carlo simulation study was carried out. Test length, average factor loadings, and sample size were considered as simulation conditions. Item weighting method was defined as follows: If average score of the individuals (calculated as individual's test score/the number of items) plus item difficulty index is 1 and over then item reliability index added to individual's item score (1 or 0); if not, then the item score of the individual (1 if 1, 0 if 0) is preserved. As a result of the research, it was observed that the weighting method contributes to the construct validity. According to the results of confirmatory factor analysis, the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) values were improved. According to the research findings, the weighting method used in this research can be recommended.

Key Words: item weighting, validity, reliability, EFA, CFA

INTRODUCTION

The validity of the scores obtained from tests used in the psychological field is among the most important subjects of the psychological measurement field. Validity is considered as a feature of the scores obtained from the applied tests (American Educational Research Association [AERA], American Educational Research Association [APA], & National Council on Measurement In Education [NCME], 2014) and can be collected under the construct validity as an umbrella term (Messick, 1995). In the process of collecting evidence for construct validity, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are frequently used (Nunnally & Bernstein, 1994). EFA is based on covariance structures and is a technique for obtaining fewer latent variables (factors) than the covariance matrix between observed variables (Daniel, 1989). In EFA, the aim is to reveal the factor structure of the clusters formed by the variables. In CFA, the theoretical construct is tested, and the structural properties of the variables measured before the analysis are known. For this reason, the purpose of CFA is to try to verify the predicted factor structure based on the measurements obtained from the measurement instrument (Stevens, 2009). In the process of collecting evidence for construct validity, both analyses have high importance. Nunnally (1978) emphasizes the importance of factor analysis by saying that it is at the heart of the measurement of psychological constructs.

Different measures may be taken to increase the validity of the scores obtained from the test. Following the scale development procedure during the development phase of the measurement instrument, knowing the theoretical subset well and reflecting it on the measurement instrument, and taking some measures in the implementation phase of the tests are examples. However, it has been thought that some weighting operations on the scores obtained after the test application can increase the validity of the scores obtained from the test, and studies on item weighting have been conducted accordingly (Erkuş, 2014; Ghiselli, 1964; Gulliksen, 1950; Rotou, Headrick, & Elmore, 2002).

* This study was conducted as an oral presentation during the Second International Social Sciences Symposium held in Antalya on May 18–20, 2017.

** Research Assistant, Hacettepe University, Graduate School of Educational Science, Ankara-Turkey, e-mail: abdullahfarukkilic@gmail.com, ORCID ID: 0000-0003-3129-1763
*** Professor Dr., Hacettepe University, Faculty of Educational Science, Ankara-Turkey, e-mail: nurid@hacettepe.edu.tr, ORCID ID: 0000-0001-6274-2016

To cite this article:
Kılıç, A. F., & Doğan, N. (2019). The effect of item weighting on reliability and validity. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 149-164. DOI: 10.21031/epod.516057

Geliş Tarihi: 22.01.2019
Kabul Tarihi: 26.05.2019

When studies on item weighting are examined, it is seen that they were conducted mostly in the second half of the 20th century (Burt, 1950; Dick, 1965; Ghiselli, 1964; Guilford, 1954). Among the recommended methods related to item weighting in addition to the use of methods such as assigning values by multiple regression, assigning piecewise regression coefficients (Guilford, 1954), weighting with item discrimination indices (Birnbaum, 1968), and using factor analysis (Burt, 1950), some authors suggest methods like item weighting using test variance or item variance (Dick, 1965), weighting the items related to more important topics, taking into account the context in which the items are linked (Ghiselli, 1964), and weighting item clusters instead of individual items (Gulliksen, 1950). In a more recent study, Rotou, Headrick, and Elmore (2002) proposed weighting items by using multidimensional item response theory parameters and a hybrid weighting method based on the use of total score calculation in the classical test theory to calculate individual scores. When the results of item weighting studies are examined in general, it is observed that different weighting of the items for shorter tests is more efficient, item weighting has little effect when the number of items is between 10 and 20 (Ghiselli, 1964), the best weighting method for long tests is to make weights 1 of all items. When the average of item correlations is low, item weighting gives better results (Guilford, 1954), and when the number of components in the test decreases, scoring items differently is more effective in ranking individuals compared to the total points obtained by equally weighting (Ghiselli, 1964) (e.g., scoring for the correct answer, giving 0 for the wrong answer, and collecting the items that are correctly answered).

Research on weighting in Turkey has been carried out, but the emphasis was usually on the option weighting for multiple choice items (Akkuş & Baykul, 2001; Erdem, Ertuna, & Doğan, 2016; Gözen-Çıtak, 2010; Özdemir, 2004), and it has been seen that the research conducted on item weighting is limited. In the research carried out by Yurdugül (2010), the evaluation was based on the total scores of the individuals. The limitation in this research was that it focused on the total scores and rankings of individuals. In the current study, research was carried out on the construct validity. In addition, in contrast to other studies, a new weighting method was developed in current research.

When the methods have generally been evaluated, it has been asserted that the effort for item weighting will not be worth it (Guilford, 1954; Phillips, 1943). However, besides increasing the validity and reliability of the results obtained from the item weighting test, as it should maximize the difference between individuals (Horst, 1936), item weighting will help to better discriminate the individuals. Since today's highly developed computer technology also reduces the labor required for item weighting, even if it does not make an excessive contribution to validity and reliability, item weighting may be recommended due to piecewise contributions.

Although item weighting improves the validity and reliability of the results obtained from the test, and for this reason it is clearly important, it is surprisingly used in a small number of studies (Burt, 1950). Nowadays, due to the limited research conducted on the effects of item weighting, the item weighting method developed in this research was examined under different conditions. In the study, the following sub-problems were asked in order to search for answers to the question, "What is the effect of item weighting on the validity and reliability of the test?"

1. How does the explained variance ratio change that is described as a result of EFA, which is based on the matrix of converted item scores obtained by the proposed item weighting method?

2. How do the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and chi-square values change, which are described as a result of CFA, which is based on the matrix of converted item scores obtained by the proposed item weighting method?

3. How do the Cronbach’s alpha reliability coefficient values change, which are described as a result of reliability analysis, which is based on the matrix of converted item scores obtained by the proposed item weighting method?

METHOD

This research is a Monte Carlo simulation study conducted to examine the effect of the weighting method on the validity and reliability of the test, which is proposed by researchers. Research data are limited by 1–0 scoring because of the multiplicity of categorical data types and with the idea that they should be studied separately. Another limitation is the generation of data in unidimensional construct in the case of multidimensional data, as it is difficult to deal with many conditions such as data type, number of dimensions, inter-dimensional relations, and number of items in dimensions.

As simulation conditions sample size (250, 1000, and 3000), the number of items (20, 30, and 40), and an average factor loading (0.5 and 0.7) are examined.

Regarding to sample size, small (250), medium (1000), and large (3000) samples were formed. Rhemtulla, Brosseau-Liard, & Savalei (2012) stated that 200 sample sizes are common in psychology literature. But in this study, exploratory and confirmatory factor analysis was conducted. 250, 1000 and 3000 sample sizes were chosen to avoid of the sample size requirement of the factor analysis (Comrey, 1988; Floyd & Widaman, 1995; Gorsuch, 1974; Guadagnoli & Velicer, 1988).

Because of unidimensional constructs were examined in this research, tests consist of 20, 30 and 40 items were formed as simulation condition. Tests consist of 20 items was used commonly in practice (MEB, 2013, 2019). So, the condition of 20 items was added simulation. 30 and 40 items were added to examine the effect of the number of items on the weighting method.

Factor loadings were between 0.30 and 0.50 could be considered low and above 0.70 could be considered high (Trierweiler, 2009). Thus, in this research, 0.50 (low) and 0.70 (high) was used for average factor loadings condition.

When all conditions were considered, a total of 18 simulated conditions were researched, and 1000 replications were made for each condition. The simulation conditions are presented in Table 1.

Table 1. Simulation Factors and Conditions

Fixed Conditions		Simulation Conditions		
Data Type	Number of Factors	Sample Size	Test Length	Average Factor Loading
		250	20	0.50
1-0	Unidimensional	1000	30	0.70
		3000	40	

When the conditions in Table 1 are considered together, 1 (data type) x 1 (number of factors) x 3 (sample size) x 3 (test length) x 2 (average factor loading) = 18 conditions are obtained. Since 1000 replications were made for each condition, the research was carried out on 18000 data files. EFA, CFA, and Cronbach’s alpha reliability coefficient calculations were performed separately on 1000 replicated data files produced for each condition and the averages of the obtained values were calculated, and these values were compared with each other.

The averages of the descriptive statistics of replicated data generated according to the simulation conditions are presented in Table 2.

When Table 2 is examined, average values of data sets obtained from 1000 replications are seen. When the data sets are examined according to the simulation conditions, the mean skewness for the items are 0 and the mean kurtosis values are around 1. It can be stated that the average discrimination values change according to the average factor loading condition. The skewness values of the total score are around 0, and the kurtosis values are about 2 in the case where the average factor loading is 0.5 and 1.8 in the case of 0.7. When the psych package (Revelle, 2016) is used in two categorical data production, a cut-off score is entered for the skewness value. According to this cut-off point, the skewness of the data is negative, positive, or around zero. However, according to the cut-off point entered, the kurtosis is automatically adjusted according to the skewness. For example, in skewed data, the kurtosis values may be even higher.

Table 2. Descriptive Statistics of Simulation Data

Simulation Conditions			Item Statistics		Total Score Statistics After Conducting Item Weighting Procedure								
Sample Size	Number of Items	Factor Loading	Mean of Kurtosis Values of Items	Mean of Skewness Values of Items	Mode	Median	Arithmetic Mean	Maximum	Minimum	Skewness	Kurtosis	Mean Difficulty	Mean Discrimination
250	20	0.5	1.016	0.002	9.958	9.992	9.992	19.771	0.190	0.001	2.257	0.500	0.505
		0.7	1.017	0.001	10.172	9.986	9.993	20.000	0.000	0.001	1.822	0.500	0.690
	30	0.5	1.016	0.002	15.010	14.969	14.983	29.313	0.670	0.004	2.260	0.499	0.488
		0.7	1.017	0.002	15.076	14.950	14.985	29.999	0.000	0.004	1.823	0.500	0.680
	40	0.5	1.016	0.001	20.130	20.028	19.993	38.775	1.166	-0.003	2.261	0.500	0.479
		0.7	1.017	0.004	19.431	19.951	19.960	39.995	0.004	0.002	1.825	0.499	0.673
1000	20	0.5	1.004	-0.001	10.032	10.003	10.005	19.998	0.002	0.002	2.252	0.500	0.506
		0.7	1.004	-0.001	9.963	10.008	10.004	20.000	0.000	0.000	1.811	0.500	0.692
	30	0.5	1.004	0.001	14.843	14.993	14.994	29.910	0.099	0.000	2.260	0.500	0.488
		0.7	1.004	-0.002	14.892	15.020	15.012	30.000	0.000	-0.002	1.815	0.500	0.681
	40	0.5	1.004	0.000	20.063	19.995	20.003	39.695	0.305	0.002	2.260	0.500	0.480
		0.7	1.004	-0.003	20.439	20.043	20.028	40.000	0.000	-0.004	1.817	0.501	0.675
3000	20	0.5	1.001	0.000	9.977	10.000	9.998	20.000	0.000	0.001	2.250	0.500	0.506
		0.7	1.001	0.001	9.914	9.998	9.996	20.000	0.000	0.000	1.811	0.500	0.692
	30	0.5	1.001	0.000	14.965	14.997	14.998	29.999	0.001	0.001	2.258	0.500	0.489
		0.7	1.001	-0.001	15.212	15.000	15.004	30.000	0.000	-0.001	1.814	0.500	0.681
	40	0.5	1.001	0.000	20.102	20.001	20.003	39.974	0.037	0.000	2.258	0.500	0.480
		0.7	1.001	0.000	19.985	19.995	19.998	40.000	0.000	0.000	1.815	0.500	0.675

Weighting Method Used

Test scores are open to having random errors during the development, implementation, and scoring of the tests. Since the amount and direction of random errors are not known, it is not possible to eliminate them from the measurement results. Random errors can be estimated only on a group basis using statistical methods, not individual-base. Reliability values obtained for statistical tests or items are the values that help in the estimation.

With this in mind, the item difficulty index (pj) of each item and the average score (Ii) of the individual from the test were calculated. It was checked whether the sum of these two variables was greater than 1. If this sum was greater than 1, the item reliability index was added to the answer of the individual. If less than 1, the item score of the individual was unchanged. In this way, a new matrix of item scores was established.

This weighting method was developed by researchers based on the following explanations. In the study, item scores were first produced, and then a weighting process was carried out through an item scores matrix. For this purpose,

$$f(x_{ij}) = \begin{cases} x_{ij} + item\ reliability\ index, & p_i + I_j \geq 1 \\ x_{ij}, & p_i + I_j < 1 \end{cases} \qquad (1)$$

function is used. Where, p_i represents the difficulty of item, and I_j represents the average score of the individual j, which can be expressed as follows:

$$I_j = \sum_{i=1}^{n} \frac{x_i}{n} \qquad (2)$$

Where, xi refers to the score of the individual taken from the item i (0 or 1) and n refers to the total number of items. Thus, the average score is calculated for each individual. Accordingly, the average score of the individual will be between 0 and 1.

The weighting function is defined as a piecewise function. When the function is examined, xij expresses the answer of the individual j to the item i. According to this, the answer to the item i given by the individual j can take a value of 1 or 0. Regarding the piecewise function, if the sum of the average score of the individual and the item difficulty indices is 1 or more; the item reliability index is added to the item score of the individual (0 or 1). If this sum is less than 1, the item score of the individual is kept the same (0 or 1).

The purpose of defining the item weighting function as specified in Equation 1 is to try to correct the random error involved in the measurement results. When the function is examined, it is based on the principle of correcting the answer given by a successful individual to an easy question carelessly or due to different random error sources. Likewise, a minimally successful individual can also receive correction scores for the item difficulty that he or she can answer. This situation is shown schematically in Figure 1.

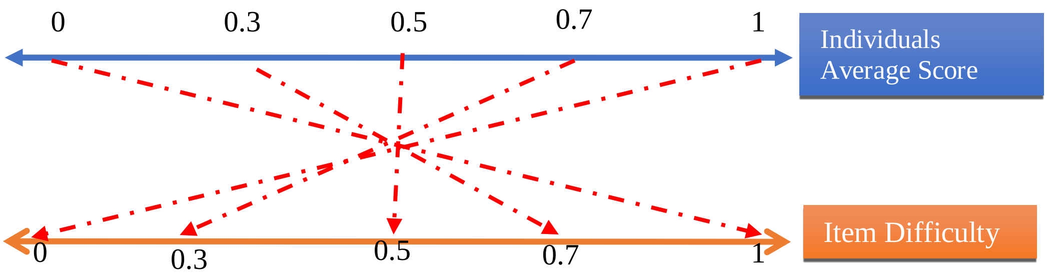


Figure 1. Item Weighting Function Chart

Figure 1 shows that when the average score of the individual is 0.3, for the individual to get correction points (1 + item reliability if the item is answered correctly, and 0 + item reliability if answered wrong), the item difficulty must be 0.7 or above, so the item must be easy. As the average score of the individual increases, the success of the individual increases, and the difficulty index of the item is also decreases, as it is getting more difficult. In this case, the weighting function can also work for an individual with low success. The important point in the function is that the individual has an average that he or she can respond correctly to that item. In Figure 1, a match between item difficulty and individual average score is presented for clarification of item weighting procedure. For example, if the average of the individual is 0.8, then the item difficulty is 0.20, and for the items above (0.20 and easier items), the weighting function will work.

Here is an example to explain this function: Assume that the average test score (total score / number of items) of a student is 0.62. In this case, an item's item difficulty index must be 0.38 (1-0.62) or above for the weighting of this student's score. The students' average scores for the items will not be weighted unless the item difficulty index 0.37 or lower. However, the students' average scores for the items will be weighted for the items with difficulty index 0.38 or higher. For more clarification, additional examples were given in Table 3.

When Table 3 was examined, it could be seen that if a student's average score + item difficulty index ≥ 1, the item's score will be weighted. But if it is smaller than 1, the item's score will not be weighted.

Table 3. Item Weighting Function Examples

Student Average Score	Item Difficulty Index	Whether Item Score Will Weight
0.20	0.80	Yes
0.20	0.85	Yes
0.20	0.79	No
0.20	0.50	No
0.50	0.50	Yes
0.50	0.60	Yes
0.50	0.30	No
0.50	0.20	No
0.70	0.30	Yes
0.70	0.40	Yes
0.70	0.20	No

To understand rationale of the method, let us suppose that the average of the individual is 0.9 and the difficulty of the item is 0.5. Then it is natural to expect that an individual who answers correctly to 90% of the items can also answer correctly to item with an average difficult. Similarly, an individual with an average of 0.20 may be expected to answer correctly to an item with an item difficulty index of 0.95. For these cases, weighting is performed by adding item reliability to the item score of the individual. If the individual gave the wrong answer to this item, an item reliability index is added to the item score to prevent it from getting 0 from that item. If the individual has already answered correctly that item, then the item score of the individual rises to the item reliability index. The reason for using the item reliability index here is that both item discrimination and the standard deviation of the item can be achieved at the same time. Thus, with the defined function, item scores of the individual are corrected by combining the item difficulty, item discrimination, and standard deviation of the item.

Process

In the study, data sets for each simulation condition were first generated using the psych package (Revelle, 2016) found in the R program (R Core Team, 2018). The aim was that the skewness value in data production is close to 0. For this reason, the cut-off score for the data produced in the dichotomously was taken as 0 (Revelle, 2016). The kurtosis values were based on the cut-off point.

After the data sets (1000 data sets for each condition) were generated according to the simulation conditions (18 conditions), the weighting process was applied to the data sets. The function presented in Equation 1 was used for weighting. The code was written by the researchers in the R program with the purpose of applying weighting to all data sets. Thus, a matrix of weighted item scores was generated from the generated data sets (1–0 form).

EFA, CFA, and reliability analyses were performed on all data sets before and after weighting (1000 replicative data sets of 18 conditions scored 1–0, and 1000 replicative data sets of 18 conditions). The psych package was used for EFA (Revelle, 2016). Since the weighted item scores matrix for EFA consisted of continuous data, Pearson correlation matrix was used for both non-weighted and weighted data sets for comparison. The Mplus software was used for CFA, but the Mplus Automation package in the R program was also utilized (Hallquist & Wiley, 2017). To conduct CFA, maximum likelihood (ML) estimation method was used for both non-weighted and weighted data sets for comparison.

The reported explained variance ratio calculated for the EFA in the study is the average explained variance ratio disclosed, which was obtained from 1000 replications for each condition. The average factor loading was obtained from 1000 replications for each condition, and then the average factor loading of the test was calculated according to the number of the items. CFI, RMSEA, and chi-square values calculated for CFA were obtained as an average as a result of the analysis of the data sets obtained from 1000 replications. However, since the research was a simulation study, the average expression was not used. The average factor loading expression was used in the tables because the factors were calculated by taking the factor loading average (since the test shows the average factor loading).

It was examined that whether the explained variance ratio was normally distributed via Shapiro-Wilk test. Because they were normally distributed, t-test was used for comparison of explained variance ratio for original (1-0 data set) and weighted data set. To compare average factor loadings for original (1-0 data set) and weighted data set, Fisher's z-test was used. While Cohen's d was used for effect size to compare average of explained variance ratio, Cohen's q was used for effect size to compare average explained variance as two correlation coefficients. While Cohen's d interpreted as 0.2 small, 0.5 medium and 0.8 large, Cohen's q interpreted as 0.1 small, 0.3 medium and 0.5 large (Cohen, 1988, 1992).

RESULTS

Results are presented in the order of sub-problems.

Ratio of Variance Obtained as EFA Result and Findings for Average Factor Loadings

The explained variance ratios obtained by the simulation conditions as a result of the research and the average factor loadings are presented in Table 4.

The explained variance ratios are between 16.1% and 32.8% in the non-weighted data sets for all simulation conditions, and they range from 40.0% to 57.1% in weighted data sets. When the differences between the explained variance ratios that correspond to deviance before and after weighting are taken into account (explained variance ratio from after weighting minus explained variance ratio for binary scores), it is observed that they changed from 10.2% to 17.7% and the average was 13.8%. It is additional to note that these ratio values are derived by subtracting binary scores from the explained variance ratio for weighted scores.

For the simulation condition in which the average factor loading is 0.5, the increase in the explained variance ratio was less, and for simulation conditions with an average factor loading of 0.7, the explained variance ratio increased more. For simulation conditions with average factor loadings of 0.5 and 0.7, the increase in the sample size also increased the difference in the explained variance ratio. When the sample size was 3000, the increase of the number of items increased the difference in the explained variance ratio between before and after weighting. However, as the number of items decreased in the 250- and 1000-person samples, the explained variance ratio increased. As a result of the t-test performed to compare the explained variance ratios, it was observed that all differences were statistically significant ($\alpha<0.05$). When the effect size values were examined, it was observed that differences of the explained variance ratio had large effect size.

When the differences of the average factor loadings between before and after weighting were examined, it was observed that the differences varied between 0.096 and 0.138 and increased by an average of 0.119. In the case where the sample sizes were 250 and 1000, it was observed that with the reduction of the number of items the average factor loading difference before and after weighting was increased. On the other hand, when the average factor loading used in the simulation condition was 0.7 for a sample of 3000 individuals, the average factor loading difference between before and after weighting was increased. In addition, the difference of EVR increases with increasing the AFL in the samples of 250 and 1000 people. However, the difference in EVR decreases as the number of items increases. In the other hand, the difference for EVR was same as the number of items increased for 3000 sample size. According to these results, it could be said that the increase in the sample size, the effect of the increase in the number of items on the EVR decreases. As a result of the Fisher's z-test performed to compare the average factor loadings, it was observed that all differences were statistically significant (α<0.05). When the effect size values were examined, it was observed that differences of the average factor loadings had a small and medium effect size.

Table 4. Explained Variance Ratios and Average Factor Loadings

Simulation Conditions			Results Scoring 1–0		Results After Weighting		Difference		Cohen d for EVR	Cohen q for AFL
Sample Size	Number of Items	AFL	EVR	AFL	EVR	AFL	EVR	AFL		
250	20	0.5	0.164	0.401	0.277	0.522	0.113*	0.122†	5.86 (L)	0.15 (S)
	20	0.7	0.328	0.571	0.498	0.704	0.170*	0.133†	7.07 (L)	0.23 (M)
	30	0.5	0.164	0.400	0.269	0.506	0.105*	0.105†	3.85 (L)	0.13(S)
	30	0.7	0.328	0.571	0.483	0.682	0.154*	0.110†	3.61 (L)	0.18(S)
	40	0.5	0.163	0.400	0.266	0.496	0.102*	0.096†	3.23 (L)	0.12(S)
	40	0.7	0.327	0.570	0.474	0.667	0.147*	0.097†	2.73 (L)	0.16(S)
1000	20	0.5	0.162	0.401	0.278	0.526	0.116*	0.125†	11.60 (L)	0.16(S)
	20	0.7	0.327	0.571	0.500	0.707	0.173*	0.136†	14.60 (L)	0.23 (M)
	30	0.5	0.161	0.401	0.270	0.510	0.109**	0.109†	4.83 (L)	0.14(S)
	30	0.7	0.327	0.571	0.485	0.684	0.158*	0.113†	4.09 (L)	0.19(S)
	40	0.5	0.162	0.401	0.267	0.502	0.105*	0.101†	3.52 (L)	0.13(S)
	40	0.7	0.326	0.571	0.477	0.668	0.151*	0.097†	2.87 (L)	0.16(S)
3000	20	0.5	0.161	0.401	0.278	0.527	0.116**	0.125†	20.44 (L)	0.16(S)
	20	0.7	0.326	0.571	0.500	0.707	0.174*	0.136†	25.03 (L)	0.23 (M)
	30	0.5	0.161	0.401	0.280	0.529	0.119*	0.128†	22.15 (L)	0.16(S)
	30	0.7	0.326	0.571	0.502	0.709	0.176*	0.137†	27.29 (L)	0.24 (M)
	40	0.5	0.161	0.401	0.281	0.530	0.120*	0.129†	24.67 (L)	0.17 (S)
	40	0.7	0.326	0.571	0.503	0.709	0.177*	0.138†	27.28 (L)	0.24 (M)
Mean							0.138	0.119		

EVR: explained variance ratio; AFL: average factor loading,
*in terms of t-test result it is statistically significant (α<0.05)
†in terms of Fisher's z-test it is statistically significant (α<0.05)
(S): Small, (M): Medium, (L): Large

The impact of weighting on CFA results was examined before and after weighing in CFA as well as EFA.

Findings for Chi-Square Values and Results Obtained from CFA Fit Indexes

The CFI, RMSEA, and chi-square values obtained as a result of the CFA performed before (1–0 item score matrix) and after the weighting process applied to the item scores matrix are presented in Table 5.

When Table 5 is examined, it is seen that CFI values generally improve after weighting, and RMSEA and chi-square values tend to decrease. When the differences between CFI values between before and after weighting are examined, a change is observed between -0.003 (12th row) and 0.311 (1st row). The CFI values increased by an average of 0.112 for all simulation conditions after weighting. The CFI value was decreased when there were only 1000 subjects, 40 items, and an average factor loading of 0.7. There seems to be some improvement for all other conditions. When the average factor loading was 0.5, the improvement in the CFI was higher than the conditions when it was 0.7. When the sample size and number of item conditions are examined, when the sample size decreases and the number of items increases it can be said that the weighting tends to increase the CFI index.

Table 5. CFA, RMSEA, and Chi-Square Values Obtained from CFA

	Simulation Conditions			1–0 Scored Results			Weighting after Results			Difference		
Row Number	Sample Size	Number of Items	Average Factor Loading	CFI	RMSEA	Chi-Square	CFI	RMSEA	Chi-Square	CFI	RMSEA	Chi-Square
1	250	20	0.5	0.498	0.076	416.092	0.809	0.069	376.754	0.311	-0.006	-39.338
2		20	0.7	0.824	0.076	419.620	0.932	0.065	354.892	0.108	-0.011	-64.728
3		30	0.5	0.685	0.052	679.066	0.843	0.051	679.724	0.158	-0.001	0.658
4		30	0.7	0.869	0.055	718.176	0.915	0.056	763.311	0.046	0.001	45.135
5		40	0.5	0.751	0.042	1063.197	0.846	0.043	1124.488	0.095	0.002	61.291
6		40	0.7	0.877	0.047	1156.833	0.896	0.052	1351.126	0.019	0.005	194.293
7	1000	20	0.5	0.518	0.074	1102.929	0.822	0.067	941.487	0.304	-0.007	-161.442
8		20	0.7	0.844	0.071	1039.637	0.944	0.059	775.492	0.100	-0.012	-264.145
9		30	0.5	0.715	0.049	1362.145	0.862	0.047	1363.485	0.147	-0.001	1.340
10		30	0.7	0.900	0.048	1325.547	0.933	0.047	1513.180	0.033	0.000	187.633
11		40	0.5	0.799	0.036	1723.846	0.871	0.038	1989.171	0.071	0.002	265.325
12		40	0.7	0.923	0.037	1735.951	0.920	0.041	2550.180	-0.003	0.004	814.229
13	3000	20	0.5	0.522	0.074	2946.655	0.824	0.067	2462.226	0.302	-0.007	-484.428
14		20	0.7	0.847	0.070	2710.133	0.946	0.058	1917.261	0.099	-0.012	-792.872
15		30	0.5	0.721	0.048	3216.713	0.895	0.043	2709.393	0.174	-0.005	-507.320
16		30	0.7	0.905	0.046	2996.236	0.966	0.038	2192.891	0.061	-0.008	-803.345
17		40	0.5	0.806	0.036	3583.580	0.926	0.032	3064.073	0.120	-0.003	-519.507
18		40	0.7	0.931	0.035	3401.912	0.975	0.029	2596.216	0.044	-0.006	-805.696
	Mean									0.112	-0.004	-159.606

When the differences of the RMSEA index between before and after weighting are examined, these differences are seen to have changed between -0.012 (14th row) and 0.005 (6th row), and the average value is 0.004. Most of the simulation conditions result in a decrease in the RMSEA value. The biggest decrease was in the 20-item test with 3000 individuals and with an average 0.7 factor loading. When the number of samples decreases, and the number of items increases, the difference in RMSEA values also decreases. While the increase in the number of items for 250 and 1000 sample sizes resulted in a decrease in RMSEA, RMSEA values was improved for all conditions for 3000 sample size.

When the differences between before and after weighting of the chi-square value are examined, it is seen that these differences changed between -805.696 (18th row) and 814.229 (12th row). Chi-square values decreased after weighting in all conditions for the sample sizes of 3000. When the sample sizes were 250 and 1000, the weighting process caused a decrease in the chi-square values in 20-item tests. However, when the number of items were 30 and 40, chi-square values increased.

Findings for Coefficient of Reliability

The findings for the Cronbach's alpha reliability coefficient obtained as a result of the reliability analysis conducted before the weighting was applied to item matrix scores (1–0 item matrix scores) and after applying the weighting process are presented in Table 6.

Table 6. Cronbach's Alpha Values Obtained as a Result of Reliability Analysis

Simulation Conditions			1–0 Scored Results	Results after Weighting	Difference	Cohen's q for Cronbach's Alpha
Sample Size	Number of Items	Average Factor Loading	Cronbach's Alpha	Cronbach's Alpha	Cronbach's Alpha	
250	20	0.5	0.792	0.882	0.091†	0.31 (M)
	20	0.7	0.906	0.952	0.046†	0.35 (M)
	30	0.5	0.851	0.952	0.101†	0.59 (L)
	30	0.7	0.935	0.962	0.027†	0.28 (M)
	40	0.5	0.884	0.930	0.047†	0.26 (M)
	40	0.7	0.950	0.970	0.020†	0.26 (M)
1000	20	0.5	0.793	0.884	0.091†	0.31 (M)
	20	0.7	0.906	0.952	0.046†	0.35 (M)
	30	0.5	0.851	0.912	0.061†	0.28 (M)
	30	0.7	0.935	0.963	0.027†	0.29 (M)
	40	0.5	0.885	0.932	0.047†	0.28 (M)
	40	0.7	0.951	0.971	0.020†	0.27 (M)
3000	20	0.5	0.793	0.885	0.091†	0.32 (M)
	20	0.7	0.906	0.952	0.046†	0.35 (M)
	30	0.5	0.852	0.921	0.069†	0.33 (M)
	30	0.7	0.936	0.968	0.032†	0.35 (M)
	40	0.5	0.885	0.940	0.055†	0.34 (M)
	40	0.7	0.951	0.976	0.025†	0.36 (M)
Mean					0.052	

†in terms of Fisher's z-test it is statistically significant (α<0.05)
(S): Small, (M): Medium, (L): Large

When Table 6 is examined, it is seen that the Cronbach's alpha coefficient changes between 0.792 and 0.951 for the non-weighted data sets and between 0.882 and 0.976 for the weighted data sets. When the differences between before and after weighing are examined, it is observed that they show changes between 0.020 (6th row) and 0.101 (3rd row). The Cronbach's alpha coefficients increased by 0.052 on average for all simulation conditions after weighting. For simulation conditions with an average factor loading of 0.7, the increase in the Cronbach's alpha coefficient was observed to be lower than the increase for the simulation conditions with an average factor loading of 0.5. As a result of the increase of the number of items in the simulation condition, the effect of the weighting process on the Cronbach's alpha coefficient is decreased. It has been observed that the increase in sample size generally results in a further increase in the confidence coefficient obtained after the weighting process. When the sample size and item number were evaluated together, it was observed that in the cases where the sample size increased, and the number of items decreased, the weighting process increased the Cronbach's alpha coefficient more. As a result of the Fisher's z-test performed to compare the average factor loadings, it was observed that all differences were statistically significant (α<0.05). When the effect size values were examined, it was observed that differences of the average factor loadings had a small and medium effect size.

DISCUSSION and CONCLUSION

As a result of the research, it was observed that the proposed weighting method increased the explained variance ratio by 13.8%. As the average factor loading increases, the effect of the weighting process on the explained variance increases. Accordingly, the effect of the weighting process increases when the relationship between the items increases. This differs from the result obtained by the nominal weighting method used by Ghiselli (1964). The nominal weighting method decreases the average correlation between the components, and it has been reported that weighted scores are more effective in ranking individuals (Ghiselli, 1964). In the current study, as the average factor loading of the items increased, the explained variance ratio also increased. Accordingly, it is recommended to use the weighting method used in the current research in tests with high relation between items.

When the effect of the weighting process on the explained variance ratio is examined, it can be said that the explained variance ratio also increases when the factor loading, number of items, and sample increase. When the averages obtained are examined, we do not agree with Guilford (1954) and Phillips (1943), who argued that item weighting would not be worth the effort. Increasing the ratio of explained variance in a psychological construct by around 13% is an important gain, and with the help of computer programs to operate weighting it is not difficult.

When the results of CFA are examined, it can be said that in general the weighting process improves CFI and RMSEA values. When chi-square values are examined, although there is no improvement for some models, it is observed that there was a decrease in chi-square values when evaluated as average. CFA estimation method can cause that result. To provide comparison of weighted and non-weighted analysis result, ML estimation method was used for CFA. On the other hand, when the change in the chi-square values which is close to zero, it can be said that the change in CFI values are quite high. So, it can be stated that there is a better fit in terms of CFI values.

When the reliability analysis results are examined, it is observed that the reliability coefficient on average increased to the 0.05 level. This result is similar to the research findings of Guilford, Lovell, and Williams (1942). However, when both EFA and CFA results are evaluated together, it is believed to be sufficient when the reliability coefficient is not reduced, because the increase in the number of items also increases the reliability coefficient. All calculated reliability coefficients show that the weighting results can be used. It is estimated that it may be sufficient for the weighting process not to have a lowering effect.

According to the results of the research, the weighting method recommended by researchers can be used by both researchers and policy practitioners. This weighting method contributes to the construct, but it should not be overlooked that it is being investigated for one-dimensional constructs. It may be advisable to researchers to investigate how the proposed weighting method produces results for two-dimensional tests or greater.

REFERENCES

American Educational Research Association (AERA), American Educational Research Association (APA), & National Council on Measurement In Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Akkuş, O., & Baykul, Y. (2001). Çoktan seçmeli test maddelerinin puanlamada, seçenekleri farklı biçimde ağırlıklandırmanın madde ve test istatistiklerine olan etkisinin incelenmesi. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *20*, 9–15.

Birnbaum, A. (1968). Some latent 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). Addison-Wesley: Reading, MA.

Burt, C. (1950). The influence of differential weighting. *British Journal of Statistical Psychology*, *3*(2), 105–125. https://doi.org/10.1111/j.2044-8317.1950.tb00288.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Comrey, A. L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology*, *56*(5), 754–761. https://doi.org/10.1037/0022-006X.56.5.754

Daniel, L. G. (1989). Comparisons of exploratory and confirmatory factor analysis. In *Mid-South Educational Research Association*. Little Rock, AR: (ERIC Document Reproduction Service No. ED314447).

Dick, W. (1965). *Item weighting: Test parameter effects and comparison of the efficiency of various weighting methods*. (Doctoral Dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3564813).

Erdem, B., Ertuna, L., & Doğan, N. (2016, September). Çoktan seçmeli testlerde seçenek ağırlıklandırma yöntemlerinin testin faktör yapısına etkisinin incelenmesi (pp. 148–149). Paper presented at the Fifth International Congress on Measurement And Evaluation in Education And Psychology, Antalya, TR. Abstract retrieved from http://epod2016.akdeniz.edu.tr/_dinamik/333/53.pdf.

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

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*(3), 286–299. https://doi.org/10.1037/1040-3590.7.3.286

Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York: McGraw-Hill.

Gorsuch, R. L. (1974). *Factor analysis*. Toronto: W. B. Saunders.

Gözen-Çıtak, G. (2010). Klasik Test ve Madde Tepki Kuramlarına göre çoktan seçmeli testlerde farklı puanlama yöntemlerinin karşılaştırılması. *İlköğretim Online*, *9*(1), 170–187.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265–275.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Guilford, J. P., Lovell, C., & Williams, R. M. (1942). Completely weighted versus unweighted scoring in an achievement examination. *Educational and Psychological Measurement*, *2*, 15–21.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Hallquist, M., & Wiley, J. (2017). MplusAutomation: Automating Mplus Model Estimation and Interpretation. Retrieved from https://cran.r-project.org/package=MplusAutomation

Horst, P. (1936). Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika*, *1*(1), 53–60. https://doi.org/10.1007/BF02287924

MEB. (2013). Temel Eğitimden Ortaöğretime Geçiş. Retrieved March 24, 2015, from http://oges.meb.gov.tr/docs2104/sunum.pdf

MEB. (2019). Sınavla öğrenci alacak ortaöğretim kurumlarına ilişkin merkezî sınav başvuru ve uygulama kılavuzu. Retrieved May 15, 2019, from https://www.meb.gov.tr/meb_iys_dosyalar/2019_04/03134315_Kilavuz2019.pdf

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037//0003-066X.50.9.741

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd. ed.). New York, NY: McGraw-Hill.

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.

Phillips, A. J. (1943). Further evidence regarding weighted versus unweighted scoring of examinations. *Educational and Psychological Measurement*, *3*(1), 151–155. https://doi.org/10.1177/001316444300300114

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/.

Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research. Evanston, Illinois. Retrieved from https://cran.r-project.org/package=psych

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.

https://doi.org/10.1037/a0029315

Rotou, O., Headrick, T. C., & Elmore, P. B. (2002). An investigation of difference scores for a grade-level testing program. *International Journal of Testing, 2*(2), 83–105. https://doi.org/10.1207/S15327574IJT0202

Stevens, J. P. (2009). *Applied multivariate statistics for the social science* (5th ed.). London: Routledge.

Trierweiler, T. (2009). *An evaluation of estimation methods in confirmatory factor analytic models with ordered categorical data in LISREL*. (Doctoral dissertation). Available from ProOuest Dissertations and Theses database. (UMI No. 3416004).

Yurdugül, H. (2010). Farklı madde puanlama yöntemlerinin ve farklı test puanlama yöntemlerinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 1*(1), 1–8.

Madde Ağırlıklandırmanın Güvenirlik ve Geçerliğe Etkisi

Giriş

Psikolojik alanda kullanılan testlerden elde edilen puanların geçerliği, psikolojik ölçme alanının en önemli konuları arasında yer almaktadır. Geçerlik, uygulanan testlerden elde edilen puanların bir özelliği olarak düşünülmekte ve şemsiye bir kavram olarak yapı geçerliği altında toplanabilmektedir.

Yapı geçerliğine yönelik kanıt toplama sürecinde genellikle açımlayıcı ve doğrulayıcı faktör analizinden yararlanılmaktadır. Açımlayıcı faktör analizi (AFA) kovaryans yapıları üzerine kurulmuş olup gözlenen değişkenler arasındaki kovaryans matrisinden daha az sayıda gizil değişkenler (faktörler) elde etmeye yarayan bir tekniktir. AFA'da amaç, değişkenlerin oluşturduğu kümenin faktör yapısını ortaya çıkarmaktır. DFA'da ise teorik olarak ortaya konulan kuramsal yapının test edilmekte ve analiz öncesinde ölçülen değişkenin yapısal özellikeri bilinmektedir. Bu nedenle DFA'da amaç ölçme aracından elde edilen ölçümlere dayanarak öngörülen faktör yapısının doğrulanmaya çalışılmasıdır. Yapı geçerliğine yönelik kanıt toplama sürecinde her iki analizin de önemi yüksektir. Nunnally (1978), faktör analizi psikolojik yapıların ölçümünün kalbinde yer almaktadır diyerek faktör analizinin önemi vurgulamaktadır.

Testten elde edilen puanların geçerliğini artırmak amacıyla farklı önlemler alınabilir. Buna örnek olarak ölçme aracının geliştirilme aşamasında ölçek geliştirme prosedürünü takip etmek, kuramsal alt yapıyı iyi bir şekilde bilmek ve ölçme aracına yansıtabilmek, testlerin uygulanma aşamasında bazı önlemlerin alınması gösterilebilir. Ancak test uygulandıktan sonra elde edilen puanlar üzerinde bazı ağırlıklandırma işlemleri ile de testten elde edilen puanların geçerliğinin artırılabileceği düşünülmüş ve madde ağırlıklıklandırmasına yönelik araştırmalar yürütülmüştür.

Madde ağırlıklandırmaya yönelik araştırmalar incelendiğinde çoğunlukla 1900'lü yılların ilk yarısında yer aldığı görülmektedir. Madde ağırlıklandırmayla ilgili olarak önerilen yöntemler arasında çoklu regresyon yoluyla değer atama, kısmı regresyon katsayılarını atama Guilford (1954), madde ayırıcılık indeksleri ile ağırlıklandırma Birnbaum (1968), faktör analizini kullanma (Burt, 1950) gibi yöntemlerin kullanılmasının yanında test varyansını ya da madde varyansını kullanarak madde ağırlıklandırma (Dick, 1965), maddelerin bağlantılı olduğu içerik dikkate alınarak daha önemli konularla ilişkili olan maddeleri ağırlıklandırma (Ghiselli, 1964) tek tek maddeler yerine madde kümelerinin ağırlıklandırma şeklinde yöntemler öneren yazarlarda bulunmaktadır (Gulliksen, 1950). Günümüze daha yakın bir çalışmada ise Rotou, Headrick ve Elmore (2002) çok boyutlu madde tepki kuramı parametreleri kullanarak maddeleri ağırlıklandırmayı ve bireylerin puanlarını hesaplamak için klasik test kuramındaki toplam puan hesaplamasını kullanmaya dayanan bir hibrit ağırlıklandırma yöntemi önermiştir.

Türkiye'de ağırlıklandırmaya yönelik araştırmalar yürütülmüş ancak genellikle çoktan seçmeli maddeler için seçenek ağırlıklandırma üzerinde durulmuş (Akkuş & Baykul, 2001; Erdem, Ertuna, & Doğan, 2016; Gözen-Çıtak, 2010; Özdemir, 2004) madde ağırlıklandırma üzerinde yürütülen araştırmaların sınırlı olduğu görülmüştür (Yurdugül, 2010). Yurdugül (2010) tarafından yürütülen araştırmada bireylerin toplam puanları üzerinden değerlendirme yapılmıştır. Mevcut araştırmada ise yapı geçerliğine yönelik araştırmada bulunulmuştur.

Yöntemler genel olarak değerlendirildiğinde, Guilford (1954) ve Phillips (1943) madde ağırlıklandırma için harcanan emeğe değmeyeceğini belirtmiştir. Ancak madde ağırlıklandırmayla testten elde edilen sonuçların geçerliği ve güvenirliğinin artırılmasının yanında bireyler arasındaki farkı da maksimize etmesi gerektiğinden (Horst, 1936) bireyleri daha iyi ayırmayı sağlayacaktır. Günümüzde bilgisayar teknolojisinin oldukça gelişmiş olması madde ağırlıklandırma işlemine harcanacak emeği de azaltacağından geçerlik ve güvenirliğe aşırı katkı yapmasa bile kısmi katkıları sebebiyle kullanılması önerilebilir.

Madde ağırlıklandırmanın sonuçları genel olarak incelendiğinde, kısa testler için maddelerin farklı ağırlıklandırmanın daha verimli olduğu, madde sayısının 10 ila 20 arasında olduğunda madde ağırlıklandırmanın çok az etkili olduğu (Ghiselli, 1964), uzun testler içinse en iyi ağırlıklandırmanın tüm maddeler için 1 olarak seçilmesi olduğu belirtilmektedir. Maddeler arası korelasyonların ortalaması düşük olduğunda madde ağırlıklandırmanın daha iyi sonuçlar verdiği (Guilford, 1954) ve testteki bileşen sayısı azaldıkça maddeleri farklı puanlamanın eşit ağırlıklandırma yoluyla elde edilen toplam puana (Örneğin, doğru cevap için 1, yanlış cevap için 0 puan vermek ve doğru cevap verilen maddeleri toplamak gibi.) göre bireyleri sıralama üzerinde daha etkili olduğu ifade edilmiştir (Ghiselli, 1964).

Madde ağırlıklandırmanın testten elde edilen sonuçların geçerliğini ve güvenirliğini artırıcı etki yapması ve bu nedenle de açıkça önem arz etmesine rağmen çok az sayıda çalışmada kullanılması Burt (1950) tarafından da şaşırtıcı olarak ifade edilmiştir. Günümüzde madde ağırlıklandırmanın etkilerine yönelik olarak yürütülen araştırmaların sınırlı olması nedeniyle bu araştırmada araştırmacılar tarafından geliştirilen madde ağırlıklandırma yöntemi farklı koşullar altında incelenmiştir. Araştırmada “Madde ağırlıklandırmanın testin geçerlik ve güvenirliğine etkisi nasıldır?” sorusuna yanıt aramak amacıyla i) önerilen madde ağırlıklandırma yöntemiyle elde edilen dönüştürülmüş madde puanları matrisi üzerinden yürütülen AFA sonucunda açıklanan varyans oranı nasıl değişmektedir?, ii) önerilen madde ağırlıklandırma yöntemiyle elde edilen dönüştürülmüş madde puanları matrisi üzerinden yürütülen DFA sonucunda CFI, RMSEA ve ki-kare değerleri nasıl değişmektedir?, iii) önerilen madde ağırlıklandırma yöntemiyle elde edilen dönüştürülmüş madde puanları matrisi üzerinden yürütülen güvenirlik analizi sonucunda Cronbach Alfa güvenirlik katsayısı değerleri nasıl değişmektedir? Sorularına yanıt aranmıştır.

Yöntem

Araştırmacılar tarafından önerilen ağırlıklandırma yönteminin testin geçerlik ve güvenirliği üzerindeki etkisinin incelenmesi amacıyla Monte Carlo simülasyon çalışması yürütülmüştür. Araştırmanın verileri, kategorik veri türlerinin çokluğu ve ayrı çalışılması gerektiği düşüncesiyle 1-0 puanlamayla sınırlandırılmıştır. Diğer bir sınırlılık ise veri setlerinin tek boyutlu üretilmesidir. Bunun nedeni ise çok boyutlu verilerde veri türü, boyut sayısı, boyutlar arası ilişkiler, boyutlardaki madde sayıları vb. gibi birçok koşulu bir arada ele almanın çalışmayı amacından uzaklaştırabileceği düşüncesidir.

Simülasyon koşulu olarak örneklem büyüklüğü (250, 1000 ve 3000), madde sayısı (20, 30 ve 40) ve ortalama faktör yükü (0.5 ve 0.7) ele alınmıştır. Örneklem büyüklüğü olarak küçük, orta ve büyük olacak şekilde örneklemler oluşturulmuştur. Ağırlıklandırma sonrasında madde sayısının, faktör analizi ve puanların güvenirliğine etkisini incelemek için madde sayısı da simülasyon çalışmasına koşul olarak eklenmiştir. Faktör yükleri ortalaması da tek boyutluluğun güçlü ya da zayıf olması durumunda ağırlandırmanın etkisini görmeyi sağlayacağı düşüncesiyle ele alınmıştır. Bütün koşullar ele alındığında toplamda 18 simülasyon koşulu araştırılmış ve her bir koşul için 1000 replikasyon yapılmıştır.

Araştırmada kullanılan ağırlıklandırma yönteminde,

$$f(x_{ij}) = \begin{cases} x_{ij} + madde\ güvenirlik\ indeksi, & p_i + I_j \geq 1 \\ x_{ij}, & p_i + I_j < 1 \end{cases} \tag{1}$$

fonksiyonu kullanılmıştır. Burada p_i, i maddesinin madde güçlüğünü, I_j ise j. bireyin ortalama puanını ifade etmektedir. Yani;

$$I_j = \sum_{i=1}^{n} \frac{x_i}{n} \tag{2}$$

şeklinde ifade edilebilir. Burada x_i, bireyin i. maddeden aldığı puanı (0 ya da 1), n ise toplam madde sayısını ifade etmektedir. Böylece her bireyin ortalama puanı hesaplanmaktadır. Buna göre bireyin ortalama puanının 0 ile 1 arasında değer alacağı söylenebilir.

Ağırlıklandırma fonksiyonu parçalı fonksiyon olarak tanımlanmıştır. Fonksiyon incelendiğinde x_{ij}, j bireyinin i maddesine verdiği yanıtı ifade etmektedir. Buna göre j bireyinin i maddesine verdiği yanıt 1 ya da 0 değerinin alabilir. Parçalı fonksiyonun kuralları incelendiğinde eğer bireyin testten aldığı ortalama puan yani I_j ile i maddesinin madde güçlük indeksinin toplamı 1 ve daha büyükse o zaman bireyin madde puanına (0 ya da 1) i maddesinin madde güvenirlik indeksi eklenmektedir. Eğer bu toplam 1'den küçükse bu durumda bireyin madde puanı aynen korunmaktadır.

Madde ağırlıklandırma fonksiyonunun Denklem 1'de belirtilen şekilde tanımlanmasının amacı, ölçme sonuçlarına karışan tesadüfi hatayı düzeltmeye çalışmaktır. Fonksiyon incelendiğinde başarılı bir bireyin kolay bir soruya dikkatsizlikle veya farklı tesadüfi hata kaynakları nedeniyle verdiği cevabın düzeltilmesi esasına dayanmaktadır. Aynı şekilde düşük başarılı bir birey de kendi cevaplayabileceği madde güçlüğü için düzeltme puanı alabilmektedir.

Sonuç ve Tartışma

Araştırma sonucunda önerilen madde ağırlıklandırma yönteminin açıklanan varyans oranını ortalama %13.8 artırdığı gözlenmiştir. Ortalama faktör yükü arttıkça ağırlıklandırma işleminin açıklanan varyans üzerindeki etkisi artmıştır. Buna göre maddeler arasındaki ilişki arttıkça ağırlıklandırma işleminin etkisinin arttığı söylenebilir. Bu sonuç Ghiselli (1964) tarafından belirtilen nominal ağırlıklandırma yönteminden elde edilen sonuçla farklılaşmaktadır. Nominal ağırlıklandırma yönteminde bileşenlere farklı ağırlıklıklar atanmaktadır. Ghiselli (1964) tarafından belirtilen nominal ağırlıklandırma yöntemiyle bileşenler arası ortalama korelasyon azaldıkça ağırlıklandırılmış puanların bireylerin sıralanmasında daha fazla etkili olduğu raporlanmıştır. Mevcut araştırmada ise maddelerin ortalama faktör yükü arttıkça açıklanan varyans oranının da arttığı gözlenmiştir. Buna göre maddeleri arasındaki ilişkileri yüksek olan testlerde mevcut araştırmada kullanılan ağırlıklandırma yönteminin kullanılması önerilebilir.

Ağırlıklandırma işleminin açıklanan varyans oranına etkisi incelendiğinde faktör yükü, madde sayısı ve örneklem arttıkça açıklanan varyans oranın da arttığı söylenebilir. Guilford (1954) ve Phillips (1943) harcanan emeğe nazaran elde edilen iyileşmenin önemsiz olduğunu vurgulamıştır. Ancak mevcut araştırmadan elde edilen ortalamalar incelendiğinde madde ağırlıklandırma yönteminin kullanılmasının harcanan efora değecek sonuçlar ürettiği düşünülmektedir. Diğer bir deyişle kullanışlık açısından araştırmada önerilen ağırlıklandırma yönteminin önerilebileceği söylenebilir. Bir psikolojik özellikteki açıklanan varyans oranını %13 civarında arttırmak önemli bir kazançtır ve artık bilgisayar programlarının da yardımıyla ağırlıklandırma yapmak çok da zor olmamaktadır.

Doğrulayıcı faktör analizi sonuçları incelendiğinde ise genel olarak ağırlıklandırma işleminin CFI ve RMSEA değerlerinde iyileşme sağladığı söylenebilir. Ki-Kare değerleri incelendiğinde bazı modeller için iyileşme olmadığı gözlense de ortalama olarak değerlendirildiğinde ki-kare değerlerinde de bir düşme olduğu gözlenmiştir.

Güvenirlik analizi sonuçları incelendiğinde ise ortalama olarak 0.05 düzeyinde güvenirlik katsayısının yükseldiği gözlenmiştir. Bu sonuç Guilford, Lovell, & Williams (1942) araştırma bulgularıyla da benzerdir. Ancak AFA ve DFA sonuçları birlikte değerlendirildiğinde güvenirlik katsayısının azalmamasının yeterli olabileceği düşünülmektedir. Çünkü madde sayısının artması güvenirlik katsayısını da arttırmaktadır. Hesaplanan tüm güvenirlik katsayıları ağırlıklandırma sonuçlarının kullanılabileceğini göstermektedir. Ağırlıklandırma işleminin güvenirliği düşürücü bir etki yapmamasının yeterli olabileceği değerlendirilmektedir.

Araştırma sonuçlarına göre araştırmacılar tarafından önerilen ağırlıklandırma yönteminin kullanılması hem araştırmacılara hem de politika uygulayıcılarına önerilebilir. Yapı geçerliğine katkı sunan bu ağırlıklandırma yönteminin tek boyutlu yapılar için araştırıldığı gözden kaçırılmamalıdır. Araştırmacılara iki yada daha çok boyutlu testler için önerilen ağırlıklandırma yönteminin nasıl sonuçlar ürettiğini araştırılması önerilebilir.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(2);165-178

JMETRIK: Classical Test Theory and Item Response Theory Data Analysis Software

Gökhan AKSU*   Cem Oktay GÜZELLER**   Mehmet Taha ESER***

Abstract
The aim of this study is to introduce the jMetric program which is one of the open source programs that can be used in the context of Item Response Theory and Classical Test Theory. In this context, the interface of the program, importing data to the program, a sample analysis, installing the jmetrik and support for the program are discussed. In sample analysis, the answers given by a total of 500 students from state and private schools, to a 10-item math test were analyzed to see whether they shows differentiating item functioning according to the type of school they attend. As a result of the analysis, it was found that two items were showing medium-level Differential Item Functioning (DIF). As a result of the study, it was found that the jMetric program, which is capable of performing Item Response Theory (IRT) analysis for two-category and multi-category items, is open to innovations, especially because it is open-source, and that researchers can easily add the suggested codes to the program and thus the program can be improved. In addition, an advantage of the program is producing visual results related to the analysis through the item characteristic curves.

Keywords: jMetrik, item response theory, classical test theory, differential item functioning.

INTRODUCTION

For researchers nowadays, technology has almost the same meaning as the software that they use every day. Software products offer solutions for many challenges faced by the users. Technology extended the usage of software analysis by accessing to a wider audience and by this means researchers at each specialization level may develop themselves and experience different software products relevant to their field. The use of software, which has a great importance in all fields in scientific terms, is of great importance in terms of calculation, evaluation and development of statistics on measurement results in the field of measurement and evaluation and psychometry. The statistics calculated and used in the context of classical test theory (CTT) and item response theory (IRT), which have a very important role in psychometrics, are complex, difficult in terms of manual calculation and time consuming, which encourages researchers to use software. At this point, the needs of the researchers may change over time and the need towards a software that is easy to use, cheaper or free, fast, open-to-development increases day-by-day.

Classical Test Theory, which is also called as true score model occasionally, covers the mathematical computations laying in the background of measurement tool development process. CTT is almost 100 years old and it is still widely used. The statistics, such as correlation among items, covariance, difficulty index, discrimination power index, reliability coefficient, variance/standard deviation of the sample, measurement error, etc. are calculated by CTT, which is mainly used for the purpose of developing and improving the reliability and validity of measurement tools (Crocker and Algina, 1986; Mcdonald, 1999). Most of the statistics covered in CTT are based on mean, ratio and correlation. The

* Dr., Adnan Menderes University, Aydın Vocational School, Aydın-Turkey, e-mail: gokhanaksu1983@hotmail.com ORCID ID: 0000-0003-2563-6112
** Prof. Dr., Akdeniz University, Tourism Faculty, e-mail: cguzeller@gmail.com ORCID ID: 0000-0002-2700-3565
*** Öğr. Gör., Akdeniz University, Statistical Research Center, e-mail: tahaeser@gmail.com ORCID ID: 0000-0001-7031-1953

To cite this article:
Last name, A., & Last name, B. (201?). JMETRIK: Classical test theory and item response theory data analysis software. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 165-178. DOI: 10.21031/epod.483396
Received: 15.11.2018
Accepted: 12.03.2019

theory has a constant perspective to deal with important problems related to measurement. The need of seeking for another test theory emerged due to several weaknesses of CTT, including: item and test statistics depend on the test and on the group, which the test was applied; a single error estimation is obtained for all ranges of skill level; and the weaknesses on test linking/equating. These weaknesses led to the development of IRT that is seen as a significant innovation in the field of Psychometry (Hambleton and Swaminathan, 1985; Embretson and Reise, 2000; Meyer, 2010). Individuals often get low scores from difficult tests and higher scores from easy ones whereas their skill level stays constant. This caused the development of another test theory, which is IRT, originally presented in the manuscript of Lord and Novick (1968). Compared to CTT, IRT is stronger regarding the applications of linking/equating, differential item functioning (DIF) and individualized computer test. Since the statistics of IRT have more complex structure in terms of both computation and interpretation, compared to the statistics of CTT, various software products were developed to facilitate the tasks of the researchers in every sense. There are software products performing the computations of both CTT's and IRT's statistics, as well as software products solely performing the calculations according to CTT or IRT. CITAS, ITEMAN, Lertap, and TAP are the packages that are widely used by researchers, using which only the analysis of CTT applications can be performed; BILOG-MG, flexMIRT, ICL, MULTILOG, PARSCALE, PARAM-3PL, Winsteps and Xcalibre, IRT PRO, NOHARM, TESTFACT, flexMIRT are the packages using which the analysis of IRT applications can be performed; whereas jMetrik, R and Mplus are the popular software products can compute statistics for CTT and IRT. Regarding the software packages, which considerably facilitate the computations of IRT, researchers may only use the complete edition of jMetrik, PARAM-3PL, NOHARM and R free of charge.

This study aims to provide information about the functionality of jMetrik software, which has been developed to help statistical and psychometric procedures related to both CTT and IRT, and to indicate the differences between jMetrik and the other software products. For this purpose, the readers are informed about the functionality, installation, interface, strength and support of the software, and the outputs of an analysis performed by the software were illustrated as an example.

jMetrik is a free, open source psychometric software. It can be run on any Windows, Mac, OSX or Linux-based platform with a current Java version. The first version of the software has been released in 2009, then the second version with two major revisions was released, followed by the third release to which some statistical methods and interface changes were added. The current version of jMetrik is jMetrik 4.1.1. Dr. Meyer, the developer and copy right owner of jMetrik, continues his work at the University of Virginia.

jMetrik is a user-friendly software, it is designed to facilitate working in a production environment and to enable each researcher to use advanced psychometric procedure . Compared to similar software products, it provides a more integrated system in terms of carrying out psychometric analysis for research and operational purposes free of cost, unlike some other psychometric software. jMetrik provides comprehensive statistical and psychometric procedures such as descriptive statistics, IRT parameter estimation, linking scales and score equalization. Moreover, jMetrik helps to create various graphs and tables for the visualization of the data. The structure of software's graphical user interface is intuitive and easy to learn. In addition, it scales according to the experience of the user. New users can execute psychometric procedures via pop-up menus with marks, whereas experienced users can use jMetrik commands to automate the analysis. Another significant feature of jMetrik is being an integrated database that allows users to easily organize and manage data. Results obtained from an analysis can be saved in the database and they can be used as input for another analysis. There is no need to manipulate or reshape the data between each psychometric procedure, which significantly reduces the time required for a complete and comprehensive psychometric analysis as well as the efforts made for analysis. jMetrik can perform many statistical and psychometric methods. The most important of these are undoubtedly the analytic and psychometric methods that are related to IRT.

Although the frequency of use of jMetrik in international studies increased day by day, its national use has not reached the desired level yet. jMetrik software is not known sufficiently, which may explain the reason of this fact. In the literature, there is only one national study concerning the introduction of jMetrik (Aksu, Reyhanlıoğlu and Eser, 2017) which provides information about the functionality and strengths of jMetrik, thus this study is considered to be important. It is believed that the introduction of this software, which is easy to use and free of charge, to the researchers who want to perform analysis within the scope of both IRT and CTT, will make significant contributions to future studies

FUNCTIONALITY of the SOFTWARE, STRENGTHS of the SOFTWARE, SOFTWARE INTERFACE, SAMPLE ANALYSIS and SOFTWARE SUPPORT

Functionality of the Software

jMetrik is used for the calculation of statistics, reliability estimation, test scaling, DIF, nonparametric IRT applications, Rasch measurement models, IRT models (3PLM, 4PLM, GPCM vb.), IRT linking and equalization. jMetrik 4 has a great importance in the use of parametric and non-parametric IRT applications. Ramsay (1991, 2000) has used Kernel Regression to directly estimate the item characteristic curves for two-category and multi-category items. Kernel regression is a method used not only to predict characteristic curves of the items but also to estimate curves for both groups (DIF). In jMetrik, non-parametric IRT procedures can be easily saved in color, as .jpg or .png files. Nonparametric characteristic curves provide an easy and fast tool to examine the data and analyze the relationship between latent traits and correct responses. The only limitation of nonparametric characteristic curves is the actual difficulty of each item and the subjective interpretation of discrimination. Parametric IRT makes it easier to quantify these properties, compare the items, or compare two different groups of the same item.

jMetrik offers two estimation options in terms of parametric IRT. Software uses maximum likelihood estimation for Rasch, partial credit and rating scale models (Wright and Masters, 1982). Partial credit model is formulated by the item difficulty parameter and two or more threshold parameters. Regarding rating scale model, it can be said that it is a special case of partial credit model with threshold parameters. jMetrik uses Proportional Curve Fitting Algorithm (Meyer and Hailey, 2012) instead of Newton-Raphson Method for individual, item and threshold parameters. The software computes goodness of fit statistics for the items and the individuals in addition to parameter estimations. In addition, scale quality statistics such as separation and reliability can be calculated within the scope of the software.

jMetrik, uses marginal maximum likelihood estimation (MMLE) for two-category and multi-category IRT models, including 3-Parameter Logistic Model (3PLM), 4-Parameter Logistic Model (4PLM), and generalized partial credit model (GPCM). In addition to MMLE, the software offers Bayes Model Estimation and normal, lognormal, four-parameter a priori beta distribution options for each item parameter. Generalized S-X2 Statistics are used in terms of item fit of these models (Kang and Chen, 2007; Orlando and Thissen, 2000). jMetrik has three options for scoring the individual characteristics, which are maximum likelihood, maximum a posteriori (MAP) and expected a posteriori (EAP). Software options allow the creation of output tables with analysis results, which can be used as inputs in the procedures such as linking the scales, etc. (Meyer, 2018).

jMetrik offers two options for depicting the analysis results of IRT. The first method provides item characteristic curves, information functions and standard error functions for all items separately and for the whole test. The software uses the information contained in the output tables to automatically select the appropriate IRT model and produce these graphics quickly. The second method provide item maps in the analysis results. Item mapping method is quite common within the scope of the Rasch measurement model and it illustrates the distribution of individual's skill estimates and the distribution of item parameter's estimates in the form of two histograms with a common axis. The method is useful in terms of assessing the quality of match between individuals and items, and to determine whether more (or less) items are needed to obtain a more precise (or more effective) estimate of individual's skill. In other words, the method is a tool that guides the selection of items for the test development process and the test.

Psychometry experts usually need to link the items in different test formats in a common scale or equalize the scores . jMetrik offers several options for linking under IRT and the equalization of non-equivalent group design in the data collected according to a common item. These options are simultaneous calibration, constant common item parameter and conversion coefficient methods. Within the scope of jMetrik, simultaneous calibration and constant common item parameter methods are limited to Rasch model family at the moment. The conversion coefficient methods covered in jMetrik include a wider range of IRT models, including 3PLM, GPCM and graded response model (PRM) (2PLM, Rasch and partial credit model options, which are special cases of these models are included within the scope of the software). Linking the scales can be executed in a combination with one of these models or with any model (mixed test model). Conversion coefficient methods covered in jMetrik includes mean/average (Loyd and Hoover, 1980), mean/sigma (Marco, 1977), Haebara (Haebara, 1980) and Stocking-Lord (Stocking and Lord, 1983) procedures. For the characteristic curve methods, the type of distribution that minimizes the criterion function can be selected. Evenly spaced points can be used from a normal or uniform distribution, four points and weights, or a histogram of estimated individual skills' values. jMetric has the option of minimizing these distributions by forward, backward or symmetrical moves of criterion function (Kim and Kolen, 2007).

Linking scales places the parameters on a common scale. Linking scales is sufficient for achieving comparable scores at the point where the conversion of the participant's skill occurs in the metric to be reported. On the other hand, an additional score equalization step is needed if the reporting metric is a conversion of the observed score (Cook and Eignor, 1991; Meyer, 2018). In such cases, jMetrik allows users to perform IRT real-time score equalization procedure through single-format or mixed-format tests

Regarding DIF, Mantel-Haenszel, Joint Probability Effect Size, Standardized p-DIF effect size and ETS DIF classification levels can be obtained within the scope of jMetrik. These statistics help to evaluate the statistical and practical importance of DIF.

Advantages of the Software

jMetrik is a Java application and it works on Windows, Max OSX, or Linux operating systems with Java 7 or a higher version. jMetrik does not require more than 512 megabytes of available memory. This memory allocation is sufficient for large samples up to 1,000,000, but it can be increased when needed.

Another advantage of jMetrik is that it uses a single frame to combine psychometric methods that require multiple software, which allows a researcher to quickly switch from one analysis method to another (For example, the output of parameter estimation of jMetrik is input for linking scales). This tight integration contradicts other software. A researcher who has not used jMetrik may need a maximum of three software to re-shape and manage the data, to estimate item parameters, and to establish a scale linking. Even with a software like R, it is important to be able to operate functions of a package efficiently with the functions of another package. jMetrik is designed to avoid this hassle by integrating the workflow for various psychometric procedures.

jMetrik has a user-friendly interface that is easy to use. Analysis can be performed from point-and-click menus and dialog boxes. This feature allows new users to learn the software quickly and also makes teaching a lot easier. With conventional software, the time devoted to the course is consumed by the time required to debug old archaic syntax and Fortran format expressions. The point-and-click interface of jMetrik prevents these struggles and allows trainers to regain their time for teaching the theory.

The point-and-click interface is the most obvious way to perform an analysis in jMetrik, but this is not the only way. Each analysis can also be executed and automated through syntax. The task of analysis windows is to generate code in the background. All codes executed by the software are saved in the log. A user can save the log and scripts for later use and edit them. In jMetrik, the command log is separated from the error log. The command log keeps a record of all methods executed by a user. It can be saved and used to run the analysis again with a few changes (for example, changing the names of the data tables).

With jMetrik, a more transparent approach was adopted to psychometric calculation. The use of personal software that are closed to improvement is seen as a statistical norm in the development of large-scale tests but this limitation makes it difficult for stakeholders to check and verify the integrity of the software. There are two publicly available Java libraries for jMetrik: jMetrik library contains all interface and database codes; whereas psychometry library contains all measurement and psychometric methods. All codes of the software are available online at www.github.com, a storage space integrated with git software. Anyone can browse and install source codes. Programmers who know Java programming language can make changes in the codes, they can add patches and new features (Code changes are added to the library after review and approval). Psychometry library provides royalty-free use in special software without licensing and any conditions, allowing the institutions to use the psychometry library to create registered systems at institutional level using public tools.

www.ItemAnalysis.com is the official website of the software. The website also includes sample data files, quick procedures for the software, and answers to frequently asked questions. Questions about the software are answered very quickly by the software developer himself. jMetrik is an open source application distributed under General Public License version 3 or higher. Source code of psychometric procedures covered by jMetrik is also open source and it is distributed under Apache License version 2 and it can be installed from https://github.com/meyerjp3/psychometrics address.

At the same time, factor analysis with various rotation method options, polychoric and polyserial correlation analysis can be performed within the scope of jMetrik.

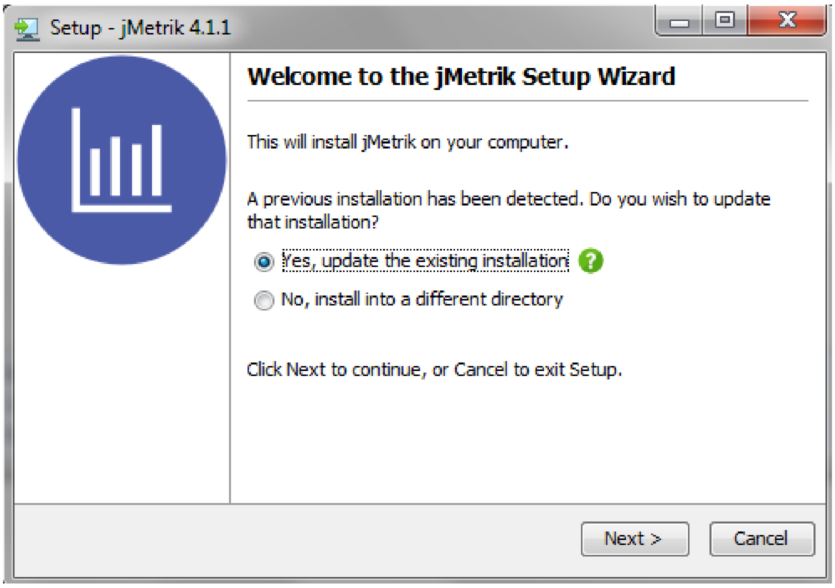
CTT analysis within the scope of JMetrik includes item and test analysis and test scaling. Classical item analysis includes the options such as item statistics, reliability analysis and conditional standard error of measurement. The analysis output of CTT, which includes item statistics, test statistics, and reliability analysis, can be saved as a text file. The output contains item difficulty, standard deviation, and two different item correlations (biserial correlation and point-biserial correlation) for each of the multiple choice and structured open-ended items. In addition, five different reliability calculation methods are available, namely Guttman’s Lambda, Cronbach’s Alfa, Feldt-Gilmer, Feldt-Brennan and Raju’s Beta. Decision consistency and accuracy estimates are provided for item analysis: Huynh’s Raw Agreement, Huynh’s Kappa, KR-21,Beta-binamial alpha and Beta-binomial beta. The classical item analyzes offered by jMetrik are very comprehensive for all user levels in both research and practical environments.

Test scaling options of CTT are very easy to use in jMetrik and many options are available. Users can quickly convert the data to overall, classifying percentage, Kelley’s true score and normalized score. At the same time, users can determine the constraints for minimum, maximum and precision points, in addition to these they can also perform an optional linear conversion. CTT analysis provided by the software offers a point-and-click type interface similar to SPSS and Excel.

SPSS files (.sav) can be directly imported to jMetrik and the software can convert the data set to a file with .sav extension and export it.

Installing the Software

Researchers may install 4.1.1 version of jMetrik software, released in February 2018 from the address https://itemanalysis.com/jmetrik-download/after determining the appropriate operating system for their computers. During the installation of the software, if the java application on the computer is an older version, the software will direct you to the latest java application. The minimum java version required for jMetrik software is 1.8, otherwise the software will not function properly. To install the software, install it here first link under the heading “install jMetrik version 4.1.1” should be clicked. After installing the setup file to your computer, the install process is continued by clicking the Run button. Then, the installation is continued by clicking Next button on the screen shown in Figure 1.



Welcome to the jMetrik Setup Wizard

This will install jMetrik on your computer.

A previous installation has been detected. Do you wish to update that installation?

● Yes, update the existing installation

○ No, install into a different directory

Click Next to continue, or Cancel to exit Setup.

Figure 1. jMetrik Software Setup Screen

After this operation, click the option to accept the terms of the license agreement as in the installation of other software and click Next button. The shortcut of jMetrik software will be added on your desktop after completing the install process as described above.

Software Interface

The main interface of jMetrik software is shown in Figure 2. This interface consists of a) the main menu and toolbar area, b) a list of database tables, c) a tabbed pane showing a view of the data and analysis output, and d) a status bar providing feedback to the user. The main menu allows you to access software procedures. For example, data management features such as import and export can be accessed through the manage menu; whereas psychometric procedures can be accessed through transform, analyze or graph menus.

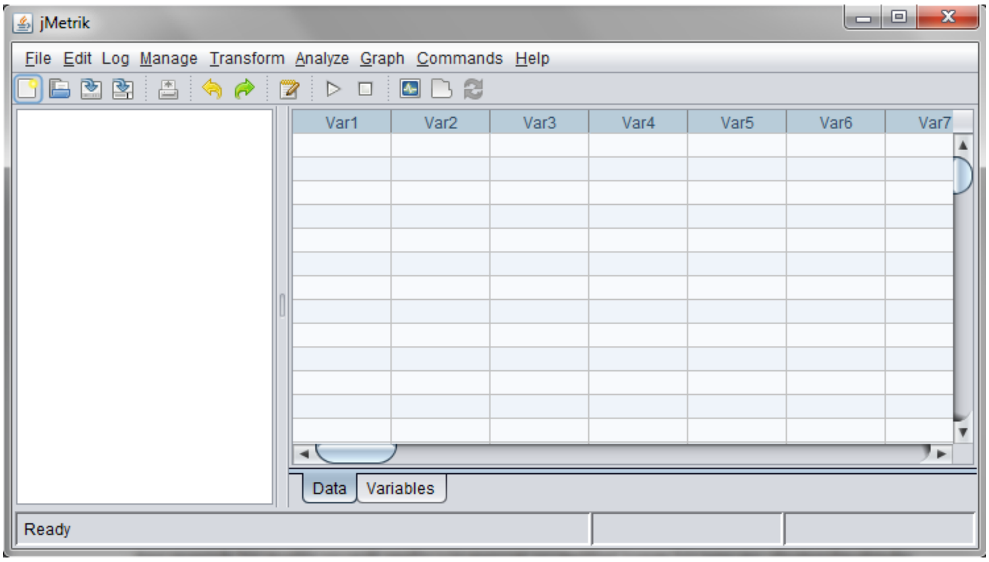

jMetrik
File Edit Log Manage Transform Analyze Graph Commands Help
Var1 Var2 Var3 Var4 Var5 Var6 Var7
Data Variables
Ready

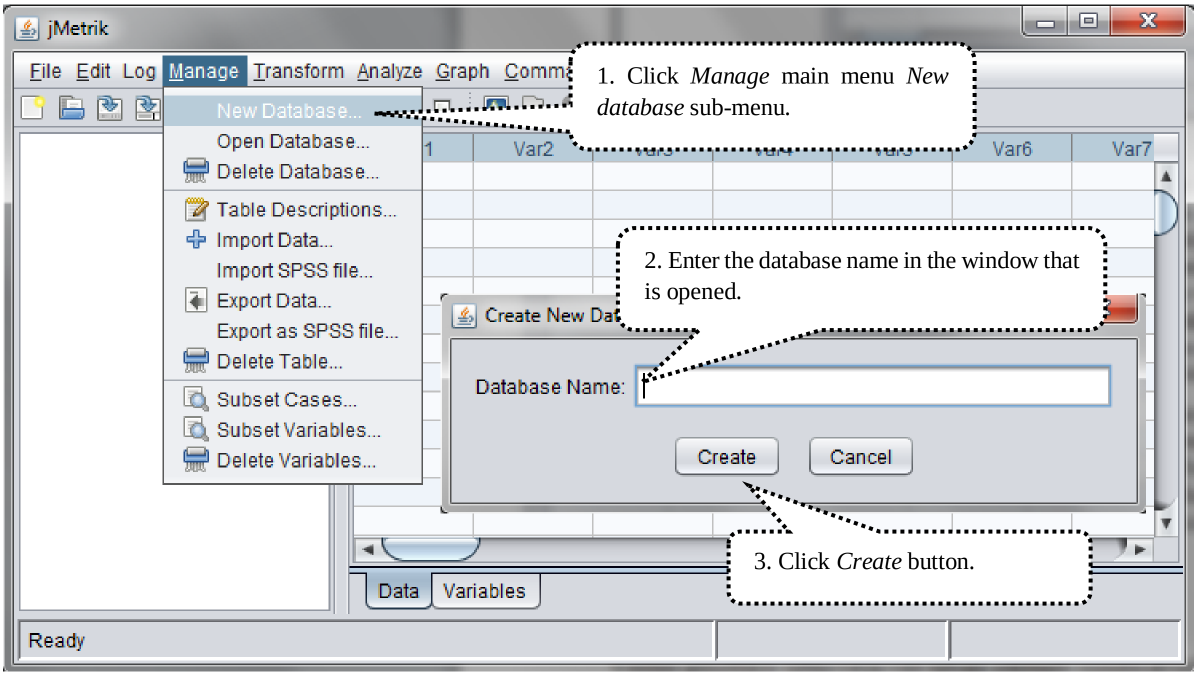
Figure 2. jMetrik Software Main Screen

Selecting an item in the main menu creates a window with options available for analysis. The components in the windows (radio buttons, checkboxes, etc.) directly correspond to various jMetrik commands. The user has to select the options she/he needs for her/his analysis and then press the run button. When the run button is pressed, the instruction for the analysis is automatically recorded into the log file of the software and the analysis is performed. To view the command (and all other commands that run during a session), log> script log path in the main menu is followed. Errors and other problems are recorded in a separate area that can be displayed by following log> view log path.

Any analysis in jMetrik can be performed by typing the commands or by copying and pasting them into command index. A new text file is opened by following File> New path and the command is entered into the new text file. The file directory can be saved by following File> Save As path. To perform the analysis, Commands> Run Commands path is followed in the main menu after entering the command (or multiple commands).

Preparing Data for Analysis

Data is usually presented in the form of tables or databases. However, the data storage method of jMetrik software is the input of data that you previously obtained from different databases. Therefore, when performing analysis with jMetrik software, the database must be defined first. Figure 3 shows the procedures to be followed to create a database.



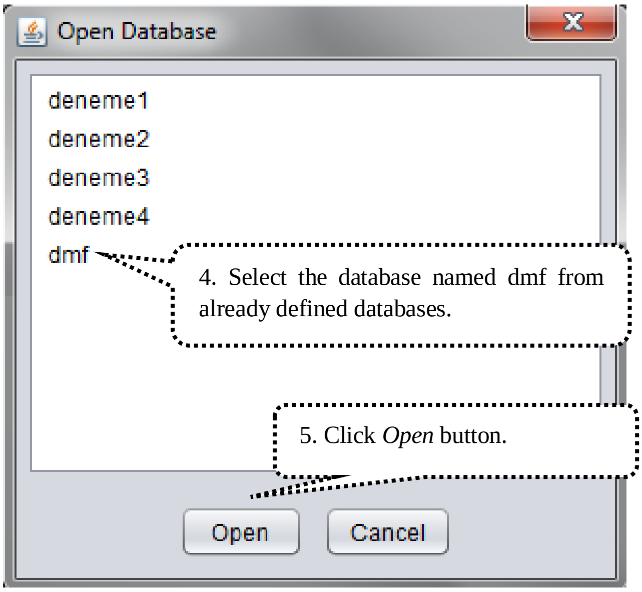
1. Click *Manage* main menu *New database* sub-menu.

2. Enter the database name in the window that is opened.

3. Click *Create* button.

Figure 3. JMetrik Software Database Creation Window

After this process, the creation of the database defined as dmf in the example application is completed by clicking *Create* button. The appearance of *ready* sign at the bottom left of the screen means that the database is created. The next step is to open this database by following Manage >> Open Database steps. Figure 4 shows how to open already defined database.

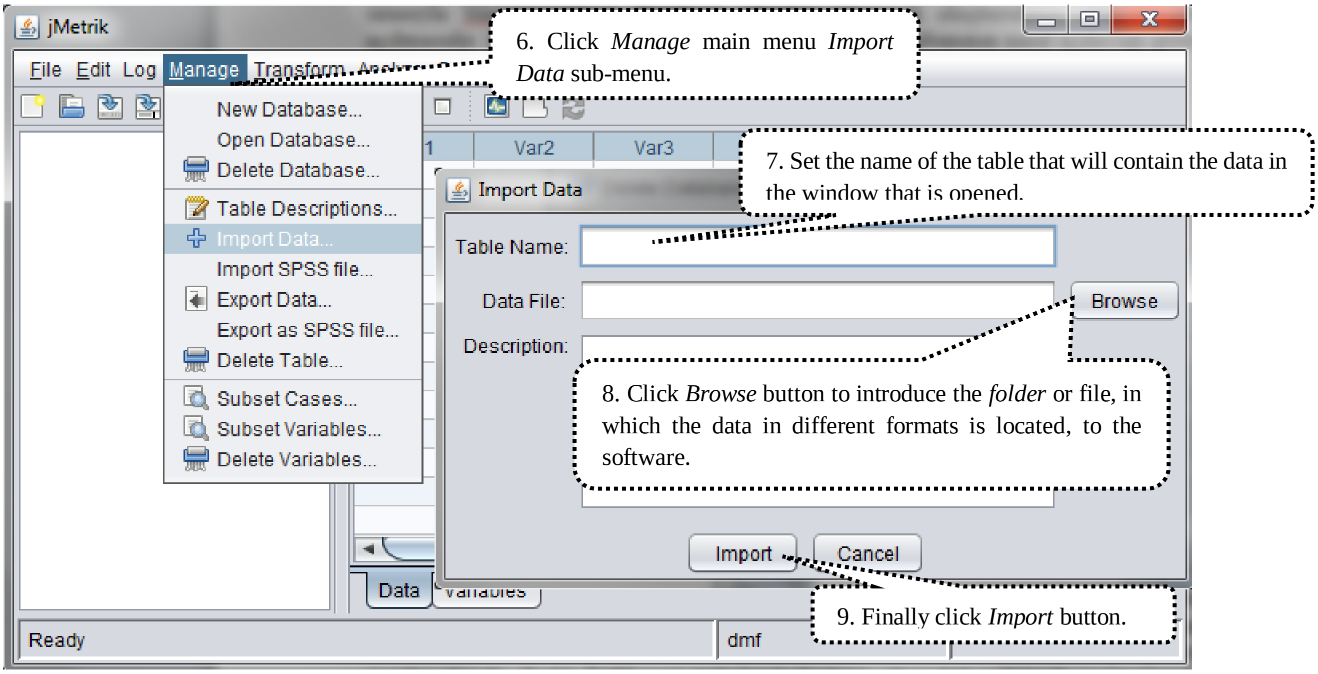


4. Select the database named dmf from already defined databases.

5. Click *Open* button.

Figure 4. jMetrik Software Database Opening Window

After this process, the name of the created database will appear in the open database window as in the Figure 4. After opening the database, the data file that will be used for the analysis should be imported into the software. jMetrik software can process different data types easily, in order to transfer data to jMetric, select Manage >> Import data. After this operation, the main screen shown in Figure 5 will appear.



6. Click *Manage* main menu *Import Data* sub-menu.

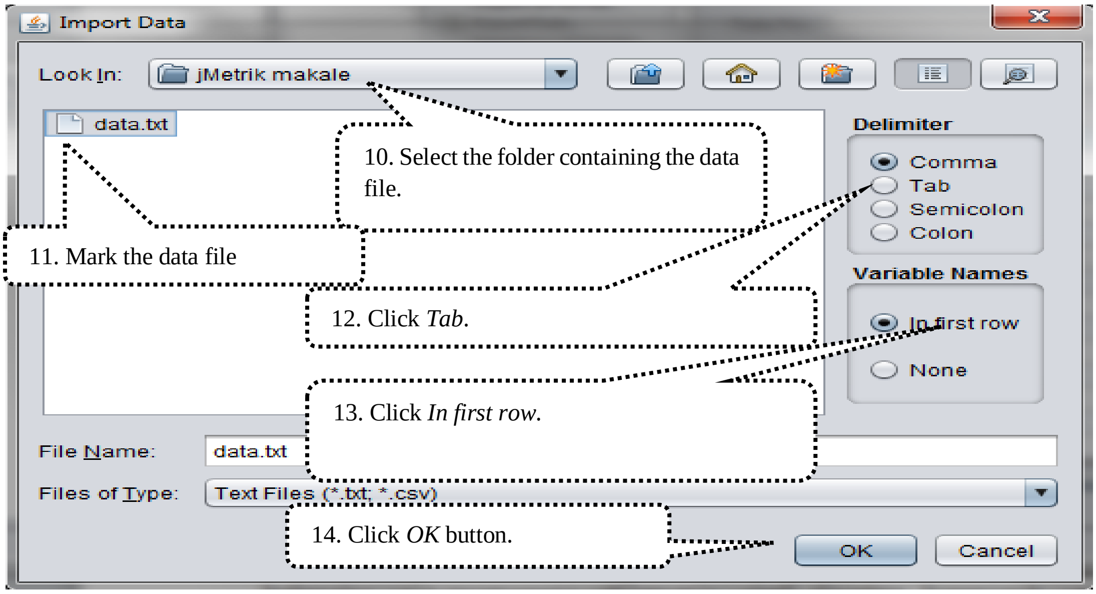
7. Set the name of the table that will contain the data in the window that is opened.

8. Click *Browse* button to introduce the *folder* or file, in which the data in different formats is located, to the software.

9. Finally click *Import* button.

Figure 5. jMetrik Software Data Transfer Window

In order to transfer already prepared data file to the software, click the Browse button shown in Figure 4 and make the definitions in the window that is opened. In the data definition window shown in Figure 6, enter the information such as type of data file and how to define the data and if first row contains data names.



10. Select the folder containing the data file.

11. Mark the data file

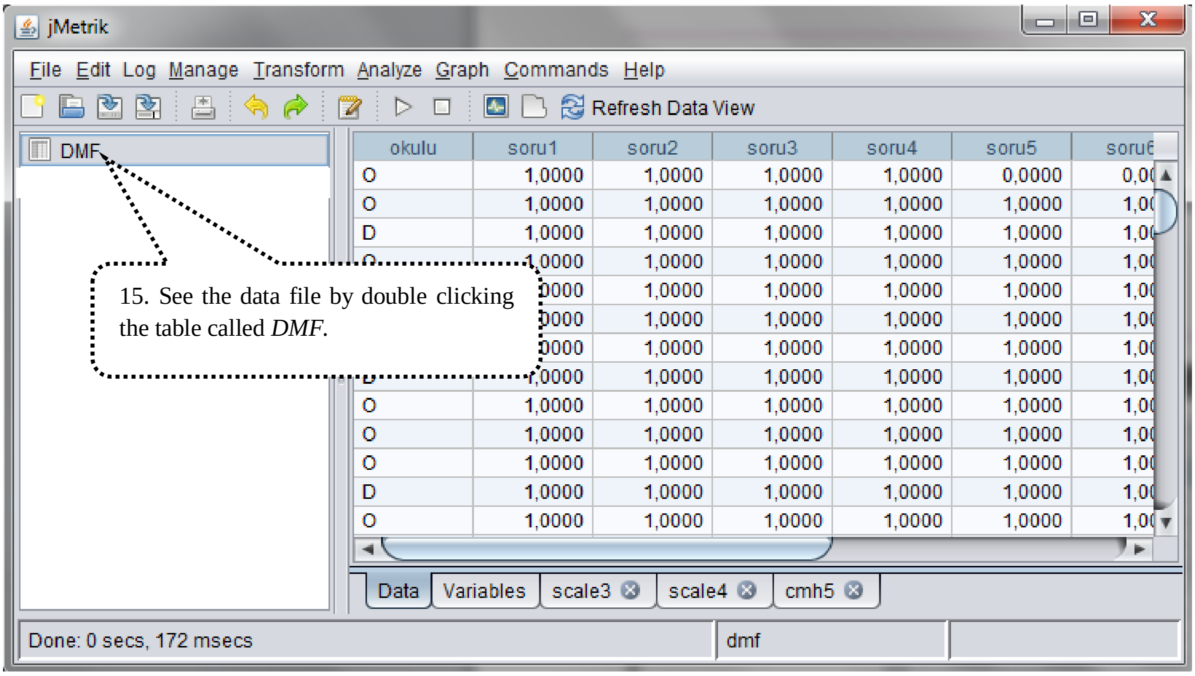
12. Click *Tab*.

13. Click *In first row*.

14. Click *OK* button.

Figure 6. jMetrik Software Data Definition Window – I

After the necessary definitions, click the OK button and the table containing the data will be imported to the software. If the import operation is carried out correctly, the table named "dmf" will be shown in the window on the left side of the jMetrik main screen. Sample data file contains the answers given to a 10-question math test by the students from two different schools, defined as state and private school. The data file can be seen by double-clicking the dmf table in the window shown in Figure 7.

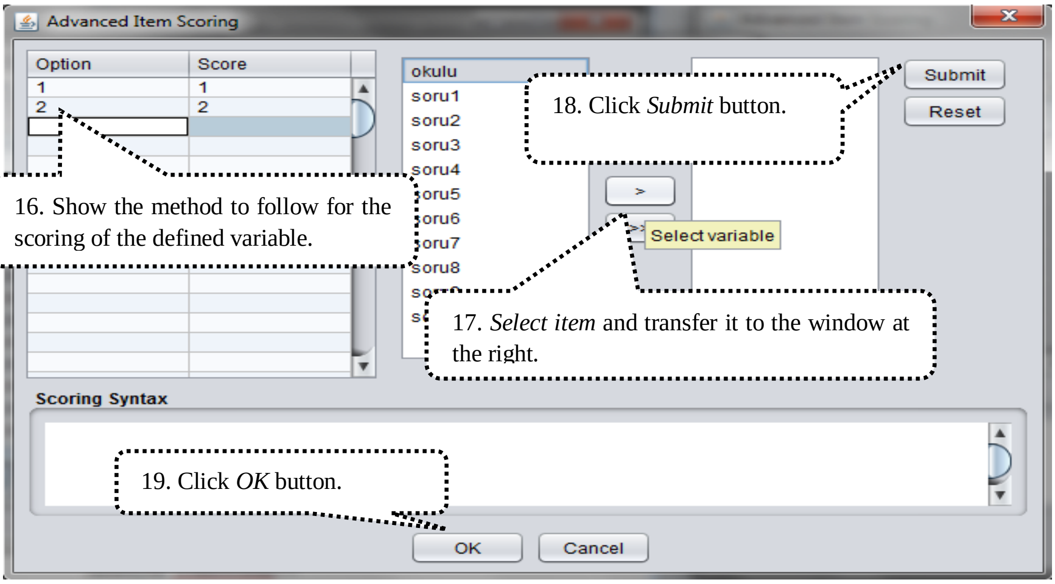


15. See the data file by double clicking the table called *DMF*.

okulu	soru1	soru2	soru3	soru4	soru5	soru6
O	1,0000	1,0000	1,0000	1,0000	0,0000	0,0
O	1,0000	1,0000	1,0000	1,0000	1,0000	1,0
D	1,0000	1,0000	1,0000	1,0000	1,0000	1,0
O	1,0000	1,0000	1,0000	1,0000	1,0000	1,0
[illegible]	0000	1,0000	1,0000	1,0000	1,0000	1,0
[illegible]	0000	1,0000	1,0000	1,0000	1,0000	1,0
[illegible]	0000	1,0000	1,0000	1,0000	1,0000	1,0
D	,0000	1,0000	1,0000	1,0000	1,0000	1,0
O	1,0000	1,0000	1,0000	1,0000	1,0000	1,0
O	1,0000	1,0000	1,0000	1,0000	1,0000	1,0
O	1,0000	1,0000	1,0000	1,0000	1,0000	1,0
D	1,0000	1,0000	1,0000	1,0000	1,0000	1,0
O	1,0000	1,0000	1,0000	1,0000	1,0000	1,0

Figure 7. jMetrik Software Data Definition Window – II

Researchers will encounter a window with which they are familiar from different data analysis software. In this window there are two tabs; the one called Data contains the data and the other called Variables contains the variables. If required, researchers may easily correct missing or incorrect data from this window. The most important process to be done after the introduction of the data to jMetrik software is defining how to score this data. Otherwise, jMetrik software will not perform any analysis. For this reason, Transform >> Advance Item Scoring steps should be followed and the way of scoring the data should be defined. At this stage, meaning of the numbers in the options must be defined in this step for each variable, as shown in the main screen illustrated in Figure 8.



16. Show the method to follow for the scoring of the defined variable.

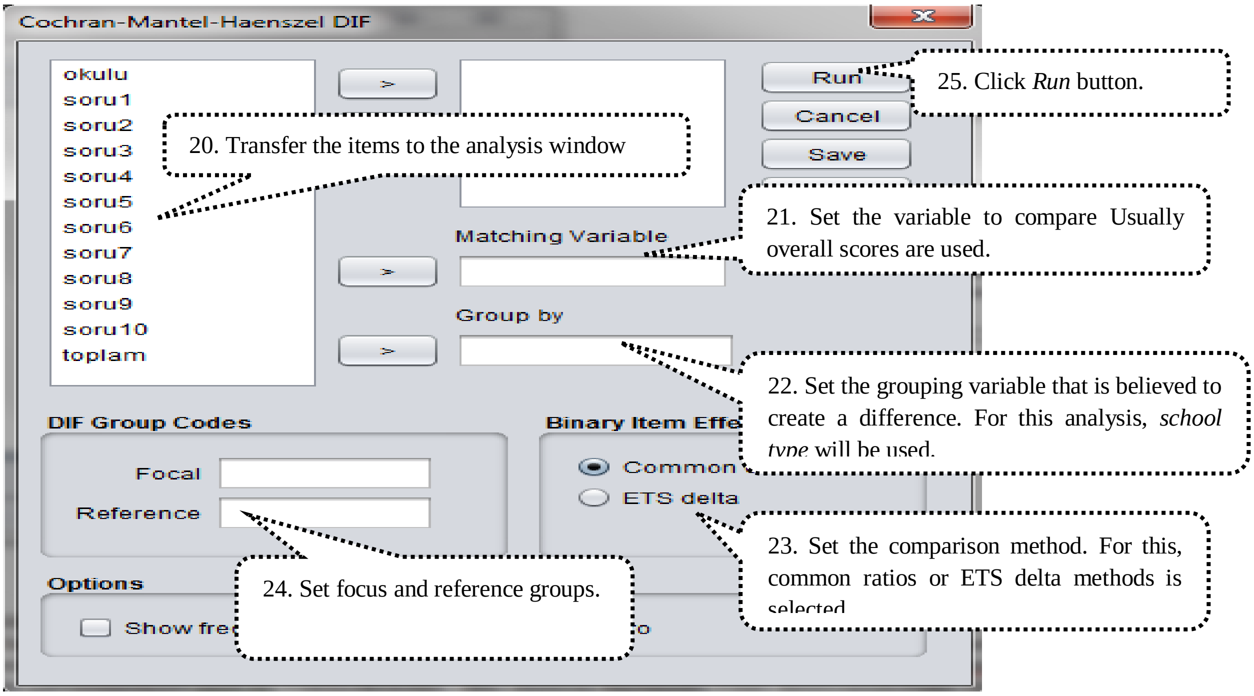
17. *Select item* and transfer it to the window at the right.

18. Click *Submit* button.

19. Click *OK* button.

Figure 8. jMetrik Software Item Scoring Window

Similarly, the items included in the test are scored as 1-0 in terms of being right and wrong, and OK button is clicked. Differential Item Functioning (DIF) analysis performed as an example in the study addressing whether the responses given to 10 items differ significantly according to students' school type (state or private) or not. The steps to be followed for DIF analysis referred in the context of the bias study, which is considered to be one of the evidences to be presented regarding the item validity, are Analyze>>DIF: Mantel-Haenszel steps. You will then see the main screen shown in Figure 9.



20. Transfer the items to the analysis window

21. Set the variable to compare Usually overall scores are used.

22. Set the grouping variable that is believed to create a difference. For this analysis, *school type* will be used.

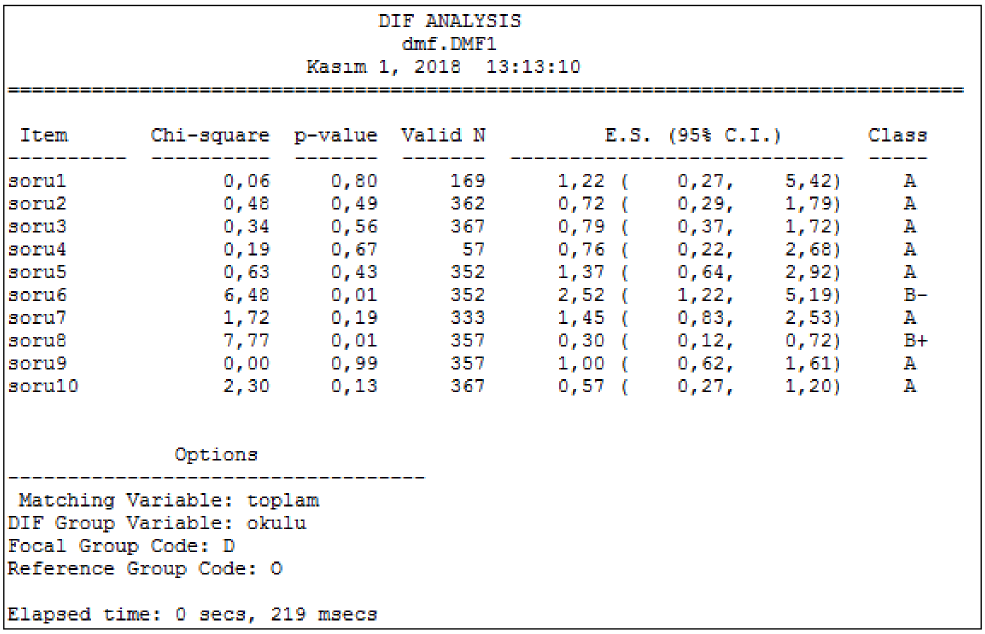
23. Set the comparison method. For this, common ratios or ETS delta methods is selected

24. Set focus and reference groups.

25. Click *Run* button.

Figure 9. jMetrik Software DIF Analysis Window

Test scaling menu is used to create the total scores in jMetrik software. In DIF analysis, the total scores obtained from all items will be used as the comparison variable; after the necessary descriptions, RUN button is clicked to complete the analysis process. The result of the analysis is the Output file shown in Figure 10.



```
                               DIF ANALYSIS
                                 dmf.DMF1
                          Kasım 1, 2018  13:13:10
================================================================================

 Item       Chi-square  p-value  Valid N        E.S. (95% C.I.)         Class
----------  ----------  -------  -------   ----------------------------  -----
soru1             0,06     0,80      169      1,22 (    0,27,     5,42)    A
soru2             0,48     0,49      362      0,72 (    0,29,     1,79)    A
soru3             0,34     0,56      367      0,79 (    0,37,     1,72)    A
soru4             0,19     0,67       57      0,76 (    0,22,     2,68)    A
soru5             0,63     0,43      352      1,37 (    0,64,     2,92)    A
soru6             6,48     0,01      352      2,52 (    1,22,     5,19)    B-
soru7             1,72     0,19      333      1,45 (    0,83,     2,53)    A
soru8             7,77     0,01      357      0,30 (    0,12,     0,72)    B+
soru9             0,00     0,99      357      1,00 (    0,62,     1,61)    A
soru10            2,30     0,13      367      0,57 (    0,27,     1,20)    A


              Options
-----------------------------------
 Matching Variable: toplam
DIF Group Variable: okulu
Focal Group Code: D
Reference Group Code: O

Elapsed time: 0 secs, 219 msecs
```

Figure 10. jMetrik Software DIF Analysis Results Window

As a result of DIF analysis, it was found that question 6 has a moderate differential item functioning in favor of the focus group, whereas question 8 has a moderate differential item functioning in favor of the reference group. Other questions in the test were found to have negligible DIF (A).

Support for the Software

To get an insight about the software, new users can read quick start guide that can be found in https://itemanalysis.com/jMetrik-quick-start/ address. For reaching frequently asked questions about the software and the answers https://itemanalysis.com/jMetrik-faq/ can be visited, whereas https://groups.google.com/forum/#!forum/jMetrik-user-group can be visited for more detailed question. *Applied Measurement with jMetrik*, which was written by Dr. Meyer, the developer of the software, is the source book containing the theoretical information about CTT and IRT and the sample analysis of the software. The book can be used as a guide containing general information about CTT and IRT and the use of jMetrik. The book consists of 10 chapters, namely *Data Management*, *Item Scoring*, *Test Scaling*, *Item Analysis*, *Reliability*, *Differential Item Functioning*, *Rasch Model*, *Multi-Category Rasch Models*, *Graphical Representation of Item and Test Properties*, *IRT-Based Scale Linking and Score Equalization.*

RESULTS and DISCUSSION

jMetrik is a software by which CTT and IRT based data analysis can be performed under a single roof without requiring any other software. IRT analysis of two-category and multi-category items can be carried out by jMetrik. Differential item function analysis can be performed based on IRT. Analyzes that can be performed using pop-up windows help researchers to perform analyzes very easily. In addition, each analysis can be carried out through commands. Being an open source project, jMetrik is a software that can be installed and used at no cost. Owing of being open source, either the codes created by people contributing to jMetrik can be used or the researchers can create their own codes. In this way, the researchers can contribute both to themselves and to other researchers who use jMetrik across the world. It is thought that being free of charge puts the software one step ahead of the other software packages that can perform the same analyzes but can only be used by paying a fee. *Applied Measurement With jMetrik*, the book written by the developer of the software can be used as a reference guide for the software.

In addition to these advantages, some aspects of jMetrik should be developed. Currently, the calculations of Multidimensional Item Response Theory (MIRT) cannot be performed within scope of the software. But given that the software is open source and therefore the researchers have the opportunity to contribute to the development of the software, this limitation is thought to be overcome easily in the future. Although the reference book *Applied Measurement With jMetrik,* which has been written by the developer of the software, contains information about how to use the software, the version used while writing the article is different from the current version, therefore, a very comprehensive resource, including information on additional analyzes and software features in the current version, will be helpful for users. The web address created for reaching frequently asked questions and answers about the software and the web address created for answering more detailed questions about the software need further improvements. Consequently, considering the advantages and disadvantages of jMetrik it is thought that the software will help researchers in answering many problems related to CTT and IRT and the execution of the application; it will reduce the workload of the researchers considering that it is free and open source and the analysis can be performed easily and quickly; having knowledge about the formulas working in background algorithms in account of being open source and the quality of the outputs in terms of readability will also contribute to ease researchers' workload. Psychometry and measurement software review studies in the future, can be carried out considering the strength and weaknesses of the parameters related to the analysis that can be performed by the software to be compared.

In addition, the jMetrik program reports the results of the analysis on its own interface compared to the other IRT analysis programs. As with other software, researchers can easily move these outputs to their work areas. In addition, the parameters obtained in jMetrik are very similar to the programs such as IRTPRO, BILOG and PARSCALE. Since the theoretical foundations of the program are inspired by the most commonly used IRT programs in the literature, it provides a great advantage to the users in interpreting and reporting the results of the analysis. In relation to that Aksu, Reyhanlıoğlu and Eser (2017) found that the results obtained from BILOG, IRT PRO and JMETRİK programs have correlation values of .99 and above in terms of both item parameters and ability estimations and it is an indication of how the results of the program are consistent with other programs used in the field. jMetrik program can perform analyzes in a very short time compared to other programs. The only negative feature related to the program is that the analysis cannot be performed without performing the database creation process which is not defined in other IRT programs. As a matter of fact, since jMetrik is Java based, researchers should first create a database where they can perform analyzes and then transfer their data to this database.

REFERENCES

Aksu, G., Reyhanlıoğlu, Ç., Eser M. T. (2017). Examining the two categorical datas by jMetrik, Bilog-MG and IRTPRO with application of mathematics exam. *European Scientific Journal*, 13 (33), 20-43. doi: dx.doi.org/10.19044/esj.2017.v13n33p20

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart & Winston.

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

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, 22(3), 144–149.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory principles and applications. Boston-USA: Kluwer-Nijhoff Publishing.

Kim, S. & Kolen, M. J. (2007). Effects of scale linking on different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*(4), 371–397.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Loyd, B. H. & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179–193. doi: http://dx.doi.org/10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139–160. doi: http://dx.doi.org/10.1111/j.1745-3984.1977.tb00033.x

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

Meyer, J. P. (2010). Understanding measurement: Reliability. New York: Oxford University Press.

Meyer, J. P. & Hailey, E. (2012). A study of Rasch partial credit, and rating scale model parameter recovery in WINSTEPS and jMetrik. *Journal of Applied Measurement*, *13*(3), 248–258.

Meyer, J. P. (2014). *Applied Measurement with jMetrik*. New York: Routledge.

Meyer, J. P. (2018). jMetrik. In W. van der Linden (Ed.). *Handbook of Item Response Theory* (pp.557-567). Boca Raton, FL: Taylor & Francis.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. doi: http://dx.doi.org/10.1177/014662168300700208

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(2);179-201

Estimation and Standardization of Variance Parameters for Planning Cluster-Randomized Trials: A Short Guide for Researchers

Metin BULUŞ*   Sakine GÖÇER ŞAHİN**

Abstract
A review of literature covering the past decade indicates a shortage of cluster-randomized trials (CRTs) in education and psychology in Turkey, the gold standard that is capable of producing high-quality evidence for high-stake decision making when individual randomization is not feasible. Scarcity of CRTs is not only detrimental to collective knowledge on the effectiveness of interventions but also hinders efficient design of such studies as prior information is at best incomplete or unavailable. In this illustration, we demonstrate how to estimate variance parameters from existing data and transform them into standardized forms so that they can be used in planning sufficiently powered CRTs. The illustration uses publicly available software and guides researchers step by step via introducing statistical models, defining parameters, relating them to notations in statistical models and power formulas, and estimating variance parameters. Finally, we provide example statistical power and minimum required sample size calculations.

Key Words: cluster-randomized trials, variance estimation, statistical power analysis, minimum required sample size.

INTRODUCTION

Cluster randomized trials (CRTs) are experimental designs where subjects are not assigned to treatment conditions independently but rather as a group. There has been an increasing interest in CRTs over the past decade in educational research (Spybrook, Shi, & Kelcey, 2016). Merely using “CRT” as a searching keyword, more than 1000 articles related to CRTs are found in educational research area in the academic journals on the Web of Science database. Although CRTs are not as efficient as individual-randomized trials, the nature of an intervention may warrant assignment of clusters (groups of individuals) to treatment conditions. There are a couple of reasons for this. First, it may be more viable to implement an intervention at the cluster level. Second, using existing clusters can be highly beneficial in terms of cost reduction and implementation convenience. Third, it may not be ethical to deprive some subjects from the intervention within the same organization. For example, providing some students with a promising intervention while excluding others from the study could be considered an unfair practice in education. Furthermore, CRTs can reduce the risk of treatment contamination that might occur if individuals in the same organization were to be randomized to treatment conditions.

However, compared to individual-randomized trials, CRTs are more complicated to design, need more participants to obtain similar statistical power, and anticipated statistical analyses are more complicated (Hayes & Moulton, 2017). Statistical methods that ignore clustering might produce misleading results, because they assume that all subjects, regardless of which clusters they come from, provide independent observations. In education settings, the assumption of independent observations is often violated as a result of contextual effects. For example, observations may not be independent

[*] Research Associate / Lecturer, Adıyaman University, Adıyaman, TURKEY, e-mail: bulusmetin@gmail.com, ORCID ID: orcid.org/0000-0003-4348-6322
[**] Postdoctoral Researcher, University of Wisconsin-Madison, WI, USA, e-mail: sgocersahin@gmail.com, ORCID ID: orcid.org/0000-0002-6914-354X

To cite this article:
Buluş, M., & Göçer Şahin, S. (2019). Estimation and standardization of variance parameters for planning cluster-randomized trials: a short guide for researchers. *Journal of Measurement and Evaluation in Education and Psychology, 10*(2), 179-201. DOI: 10.21031/epod.530642

Received: 22.02.2019
Accepted: 13.06.2019

from each other because students in the same classroom have an experienced teacher or collaboration among them is encouraged. Similarly, students and teachers within the same school share resources such as library or laboratory that differ from other schools, which may have similar contextual effect. Applying methods that ignore clustering (e.g. ordinary least squares) in such cases can prompt confidence intervals that are excessively narrow and yield p-values that are very small (Bland, 2004). In the case of experimental designs, narrower confidence intervals and smaller p-values can misguide researchers as they may indicate significant differences when, in fact, there is actually none.

There are different ways of addressing clustering depending on statistical methodology and sampling scheme. One solution is to make inferences based on cluster-robust standard errors (e.g. Cameron & Miller, 2015). If results pertain to a specific subpopulation consisting of a few clusters and not to be generalized, another alternative is to include cluster membership as fixed effects in the statistical model along with the treatment indicator. Nonetheless, applying Hierarchical Linear Models (HLM, Raudenbush & Bryk, 2002) is more common in education. Even if researchers can use cluster-robust standard errors, or depending on the sampling scheme, use fixed effects estimation method, it is not straightforward to decompose variance to within and between clusters, a strategy we will use throughout this guide to estimate and standardize variance parameters. Therefore, in parallel with studies in education effectiveness research we adopt HLM formulation.

By the same token, when planning studies that have similar nesting structure (student within classroom within schools) contextual effects should be taken into consideration, as power analysis procedures rely on the standard error of the estimate. There are various studies that have derived approximate standard error formulas with which a researcher can estimate power rate ahead of an experimental study (a priori power analysis) given sample size and other characteristics (e.g., Bloom, 1995; Bloom, 2006; Bloom, Bos, & Lee, 1999; Dong & Maynard, 2013; Hedges & Rhoads, 2010; Konstantopoulos, 2009a, 2009b).

Despite the increasing trend in the use of CRTs across many education systems and countries around the world, our review of literature in the past decade indicates a shortage in educational and psychological research in Turkey. Also, statistical power analysis in existing studies are either absent or have not considered nesting structure of the sample. We examined 174 experimental studies in education field published in Turkish journals on the Ulakbim Tubitak Journal Park Database to see whether they report power analysis procedure to determine effective sample size. Although none of the experiments utilized CRT, none of the authors reported power analysis procedure either. As a result, in these papers, results mostly suffer from small sample size where the experiment possibly could not detect a significant treatment effect when in fact there was.

One particular issue with a priori power analysis is that variance parameters used in the approximate formulas are not known. Other parameters needed for power calculations either have commonly accepted standards or does not need estimation or require extensive methodological expertise. For example, standard practice in educational research is to keep power rate at 80%, have type I error rate of 5%, and to conduct two-tailed hypothesis testing of the treatment effect (Dong & Maynard, 2013). Moreover, sample size information (e.g., the average number of students per school) can be obtained from administrative records or calculated via descriptive statistics.

While there is an emerging body of literature reporting standardized variance parameters from existing data (e.g., Hedberg & Hedges, 2014; Hedberg, 2016; Hedges & Hedberg, 2013; Spybrook, Westine, & Taylor, 2016; Westine, 2016; Westine, Spybrook, & Taylor, 2014; Zopluoglu, 2012), the majority of which focuses on K-12 academic outcomes within the United States, results may not apply to other subjects, grades, or geographical areas. Variance parameters are often sample and subject specific and should be obtained either from prior research in the literature or empirical data, preferably as close as possible to the geographical area of interest, and as similar as possible to the subject under scrutiny. Thus, estimation and standardization of variance parameters from earlier research of the same kind become an indispensable tool to researchers, especially where there is little or no prior information.

Purpose of the Study

The purpose of this study is to guide researchers in education and psychology toward planning efficient CRTs in light of little or no prior knowledge. Specifically, the study aims to provide readers with a short tutorial on estimating variance parameters from existing data using HLM, standardizing them in terms of well-known variance parameters such intra-class correlation coefficients and R-squared values, and using standardized parameters in statistical power and minimum required sample size calculations for planning CRTs.

METHOD

We provide models for two- and three-level CRTs in HLM and mixed-model forms and define parameters as in Dong and Maynard (2013). We also illustrate how to estimate treatment effect and obtain variance parameters via `lme4` (Bates, Maechler, Bolker, & Walker, 2015) library in the R environment (R Core Team, 2019). Finally, we use estimated variance parameters (unstandardized) to calculate some of the standardized design parameters (i.e., intra-class correlation coefficients and R-squared values) and use them in statistical power analysis via `PowerUpR` (Bulus, Dong, Kelcey, & Spybrook, 2019). In most instances using two libraries in the R environment will be sufficient to analyze and plan CRTs, however, depending on the complexity of the task, researchers can use any other preferred software or platform.

Ideally, results from a CRT should be informative with respect to variation in the outcome, explanatory power of covariates, and the treatment effect, which can be obtained via several statistical models. Minimally sufficient models that can inform researchers in both planning and analysis of CRTs are null and full models. Null model (also known as unconditional model) can be used to get a sense of unconditional variation in the outcome (i.e., dependent variable), whereas full model can be used to estimate both the treatment effect and conditional variation in the outcome. Null and full models for two- and three-level CRTs are described below.

Two-level CRTs

Null Model to Estimate Unconditional Variation

The following unconditional model can be used to obtain variance parameters σ^2 and τ^2 as defined below, which will be used to calculate standardized variance parameters along with parameters from full model.

HLM formulation:

Level 1: $Y_{ij} = \beta_{0j} + r_{ij}$

Level 2: $\beta_{0j} = \gamma_{00} + \mu_{0j}$

Mixed model formulation:

$$Y_{ij} = \gamma_{00} + \mu_{0j} + r_{ij}$$

where r_{ij} and μ_{0j} are level 1 and level 2 residuals, following normal distributions as $r_{ij} \sim N(0, \sigma^2)$ and $\mu_{0j} \sim N(0, \tau^2)$, respectively. Thus, σ^2 and τ^2 are variances in the outcome between level 1 and level 2 units, respectively. Y_{ij} is level 1 outcome of interest for subject i in cluster j, β_{0j} is level 1 intercept (in this case, the mean of subjects in cluster j), γ_{00} is level 2 intercept (in this case, the mean of all subjects in all clusters - grand mean).

Full Model to Estimate Treatment Effect and Conditional Variation

The following model can be used to obtain variance parameters $\sigma^2_{|X}$ and $\tau^2_{|W}$ as defined below, which are used to calculate standardized variance parameters along with parameters from unconditional model.

HLM formulation:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \delta T_j + \gamma_{01}W_j + \mu_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

Mixed model formulation:

$$Y_{ij} = \gamma_{00} + \delta T_j + \gamma_{01}W_j + \gamma_{10}X_{ij} + \mu_{0j} + r_{ij}$$

where r_{ij} and μ_{0j} are level 1 and level 2 conditional residuals, following normal distributions as $r_{ij} \sim N(0, \sigma^2_{|X})$ and $\mu_{0j} \sim N(0, \tau^2_{|W})$, respectively. Thus, $\sigma^2_{|X}$ and $\tau^2_{|W}$ are conditional variances in the outcome between level 1 and level 2 units, respectively. Y_{ij} is level 1 outcome of interest for subject i in cluster j, X_{ij} is level 1 covariate for subject i in cluster j, T_j is treatment condition (1 if cluster j assigned to the treatment, 0 if not) and W_j is level 2 covariate for cluster j, β_{0j} is level 1 intercept, γ_{00} is level 2 intercept, δ is the treatment effect, β_{1j} or γ_{10} is regression weight for level 1 covariate X_{ij}, γ_{01} is regression weight for level 2 covariate W_j.

We can calculate standardized variance parameters based on unstandardized variance parameters from unconditional and full models. $\rho = \tau^2/(\tau^2 + \sigma^2)$ represents proportion of variance in the outcome between level 2 units (also referred to as intra-class correlation coefficient in the literature), $R^2_1 = 1 - \sigma^2_{|X}/\sigma^2$ is proportion of variance in the outcome explained by level 1 covariates, $R^2_2 = 1 - \tau^2_{|W}/\tau^2$ is proportion of variance in the outcome explained by level 2 covariates. The treatment effect can also be standardized in the form of Cohen's d as $\delta^* = \delta/\sqrt{\tau^2 + \sigma^2}$, hereafter often referred to as effect size.

In the full model, we can get an estimate for the treatment effect and the associated t statistics. The hypothesis of "there is a treatment effect" is tested against the null hypothesis of "there is no treatment effect". By comparing the t statistics from the full model to the critical t value given Type I error rate (α, probability of detecting treatment effect when in fact there is none in the underlying population), we can inspect whether results can be explained beyond chance factor. Similarly, knowing t statistics, we can have an idea about Type II error rate (β, probability of detecting no treatment effect when in fact there is an effect in the underlying population). In practice we are interested in the probability of detecting a treatment effect when in fact there is an effect in the underlying population, and that is statistical power $(1 - \beta)$. To calculate statistical power, we can use t statistics after an experiment, although it may not be useful, as the experiment has already been completed. However, we can plan for an experiment such that sample size will likely produce adequate statistical power had it been repeated many times. To calculate statistical power prior to an experiment, we need some information from earlier studies; an estimate of what would be a meaningful treatment effect (often set as 0.20 or 0.25 in education, but may be increased if there is sufficient evidence that earlier interventions produced large treatment effects) and its standard error.

Standard Error Formula under Balanced Sample Size and Homogenous Variance

Assuming that level 1 variances are equal across J number of level 2 units, and level 1 sample sizes are balanced (e.g., n number of level 1 units per level 2 unit), standardized variance takes the form

$$Var(\delta^*) = \frac{\rho(1-R_2^2)}{p(1-p)J} + \frac{(1-\rho)(1-R_1^2)}{p(1-p)nJ}$$

Standard error of the treatment effect is $SE(\delta^*) = \sqrt{Var(\delta^*)}$, and if we know δ^* and $SE(\delta^*)$, we can calculate t statistics with which statistical power can be calculated. $\delta^*/SE(\delta^*)$ follows t distribution with $J-g-2$ degrees of freedom where g is number of covariates added at level 2 (Bloom, 2006, p. 17; Dong & Maynard, 2013, p. 51). Statistical power $(1-\beta)$ for two-tailed hypothesis testing can be calculated as

$$1-\beta = P\left(t_{df}(\lambda) > t_{df,1-\alpha/2}(0)\right) + P\left(t_{df}(\lambda) < t_{df,\alpha/2}(0)\right)$$

where $df = J-g-2$ for the two-level CRT, $t_{df,\alpha/2}(0)$ is the statistic associated with central t distribution with degrees of freedom df and probability $\alpha/2$, $t_{df}(\lambda)$ is the statistic associated with non-central t distribution with non-centrality parameter $\lambda = \delta^*/SE(\delta^*)$, degrees of freedom df, and α and β are Type I and Type II error rates (see, Hedges & Rhoads, 2010; Moerbeek & Safarkhani, 2018). In what follows we will demonstrate how to estimate variance parameters and how to calculate parameters needed in $Var(\delta^*)$ formula.

Estimation and Standardization of Treatment Effect and Variance Components

If not pre-installed, `lme4` and `PowerUpR` libraries should be installed in the R environment using `install.packages(c("lme4", "PowerUpR"))` command. They can be loaded into the current R session using `library(lme4)` and `library(PowerUpR)` commands.

In order to demonstrate variance estimation procedure in R, considering education settings, we simulate a simple two-level CRT data named `CRT2` which has 2,000 students across 100 schools (20 students per school). The data include five variables; school identification numbers (`schid`), a level 1 outcome variable (`outcome`), a level 2 treatment variable (`treatment`), a level 1 covariate (`covx`), and a level 2 covariate (`covw`). Number of level 1 or level 2 covariates will not change analysis strategy very much. Outcome is continuous and can be considered as any of the achievement indicator for a particular subject – such as mathematics, science, or reading scores. The treatment can be any intervention that aims at increasing student achievement scores such as a science, technology, engineering, and mathematics (STEM) program. Level 1 and level 2 covariates can be student pretest score and average school-level pretest score. First a few lines of the simulated data is printed below. Each school has a unique identification number (`schid`). Since schools are assigned to treatment conditions, the same school identification numbers will have the same values for treatment variable (`treatment`). Level 1 (students) and level 2 (schools) covariates (`covx` and `covw`) follows standard normal distributions, and outcome (`outcome`) is a linear function of these covariates with some level 1 and level 2 noise added (See data generation mechanism in Appendix A). From this point forward, R scripts are within shaded boxes. Along with code chunks, comments begin with `## --` and outputs begin with `##`.

```
head(CRT2)

##   schid treatment    outcome        covx      covw
## 1     1         0 -0.7145407 -0.37560287 0.2533185
## 2     1         0  0.2411899 -0.56187636 0.2533185
## 3     1         0 -0.8423327 -0.34391723 0.2533185
## 4     1         0 -0.9780591  0.09049665 0.2533185
## 5     1         0  3.2965023  1.59850877 0.2533185
## 6     1         0  1.7267023 -0.08856511 0.2533185
```

First, we estimate variance parameters for unconditional model to calculate the intra-class correlation coefficient. The output includes variance for two random effects indicating variation in the outcome that is between school means (`tau2`) and that is between students (`sigma2`). Sum of the two is roughly same as variance of the outcome. Thus, proportion of variance in the outcome that is between schools, also known as intra-class correlation (`rho2`), can be calculated.

```
## -- install.packages(c("lme4", "PowerUpR"))
library(lme4) # for estimation
library(PowerUpR) # for power analysis

## -- null model (unconditional model)
null.model <- lmer(outcome ~ (1 | schid), data = CRT2)
print(VarCorr(null.model), comp = "Variance")

##  Groups   Name        Variance
##  schid    (Intercept) 1.2253
##  Residual             1.9601

## -- variance parameters
tau2 <- 1.2253
sigma2 <- 1.9601


## -- intra-class correlation coefficient
rho2 <- tau2 / (tau2 + sigma2)
round(rho2, 2)

## [1] 0.38
```

Next, we estimate variance parameters for the full model to calculate R-squared values along with variance parameters from unconditional model. The output, again, includes variances for two random effects indicating conditional variation in the outcome that is between schools (`tau2w`) and students (`sigma2x`) beyond what is explained by level 2 and level 1 predictors. As some of the variation between schools and students are explained by level 2 and level 1 predictors respectively, note that variance components are reduced compared to the null model. Using proportion of reduction in the variance for level 2 and level 1, we can calculate R-squared values for each (`r21` and `r22`).

```
## -- full model
full.model <- lmer(outcome ~ treatment + covx + covw + (1 | schid),
data = CRT2)
print(VarCorr(full.model), comp = "Variance")

##  Groups   Name        Variance
##  schid    (Intercept) 0.85332
##  Residual             0.98335

## -- variance parameters
tau2w <- 0.8533
sigma2x <- 0.9834

## -- R-squared values for level 1 and level 2
r21 <- 1 - (sigma2x / sigma2)
r22 <- 1 - (tau2w / tau2)

round(r21, 2)

## [1] 0.5

round(r22, 2)

## [1] 0.3
```

We can also extract and standardize the treatment effect (`delta`) by the variance of the outcome in the form of Cohen's d (`es`). In this way, the effect is comparable to previous literature, can be compared to in future studies, and also be used in statistical power analysis procedures, if needed.

```
## -- treatment effect
coef(summary(full.model))["treatment",]

##   Estimate Std. Error    t value
##  0.9849094  0.1930537  5.1017374

delta <- 0.9849
es <- delta / sqrt(sigma2 + tau2)
round(es, 2)

## [1] 0.55
```

Statistical Power and Minimum Required Sample Size Calculations

Before we find statistical power and minimum required sample size, there are a few things to clarify. Earlier, we estimated and standardized variance parameters so that we can use them in power analysis procedures, however, there are other parameters needed, most of which are either have commonly accepted standards or known (or can be obtained via simple procedures that does not require methodological expertise). In education research, it is common to find power for an effect size (`es`) of 0.20 or 0.25, have a Type I error rate (`alpha`) of .05, and assume a two-tailed (`two.tailed`) hypothesis testing. Other way around, when the interest is in finding minimum required sample size, additionally, the power rate is assumed to be 80%. Furthermore, assigning half of the schools to treatment group (`p`) produces optimal power rate or optimal minimum required sample size (note that $p(1-p)$ in the denominator of standard error formula is maximum when $p = .50$). In our case, we know there are 20 students per school (`n`), and 100 schools (`J`) in total. Now we can calculate statistical power as

```
## -- power analysis
design <- power.cra2r2(es = .20, alpha = .05, two.tailed = TRUE,
                       rho2 = .38, r21 = .50, g2 = 1, r22 = .30,
                       p = .50, n = 20, J = 100)

##
## Statistical power:
## ---------------------------------------
##  0.463
## ---------------------------------------
## Degrees of freedom: 97
## Standardized standard error: 0.106
## Type I error rate: 0.05
## Type II error rate: 0.537
## Two-tailed test: TRUE
```

where, in addition to parameters defined earlier, `g2` is the number of covariates added at level 2. Parameters obtained from the data produce a power rate of 46.3%, which means if we repeat this experiment for a large number of times we will detect a statistically significant treatment effect 46.3% of the time, if in fact there is an effect in the underlying population. This is under recommended benchmark power rate of 80% in power analysis literature. In other words, this is worse than flipping a coin in order to decide whether or not an intervention would be effective. Figure 1 demonstrates how far we are from the benchmark power rate. By visual inspection, it seems a sample consisting of somewhere between 200 to 250 schools is capable of producing results with 80% power rate.

```
plot(design, ypar = "power", locate = TRUE, xlim = c(50, 250))
```

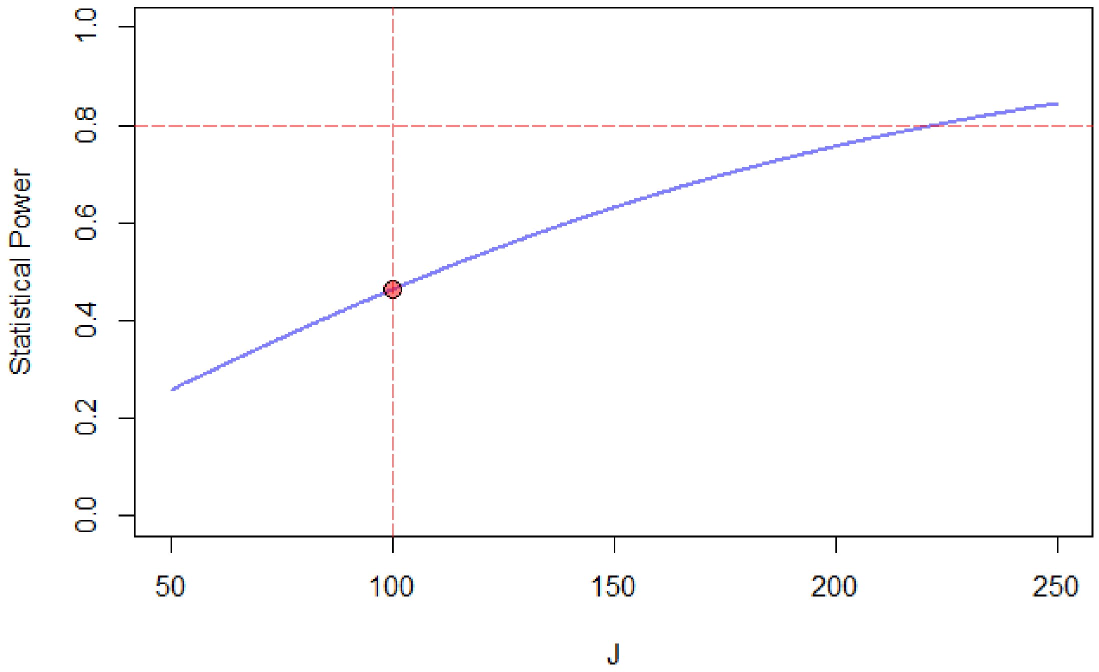
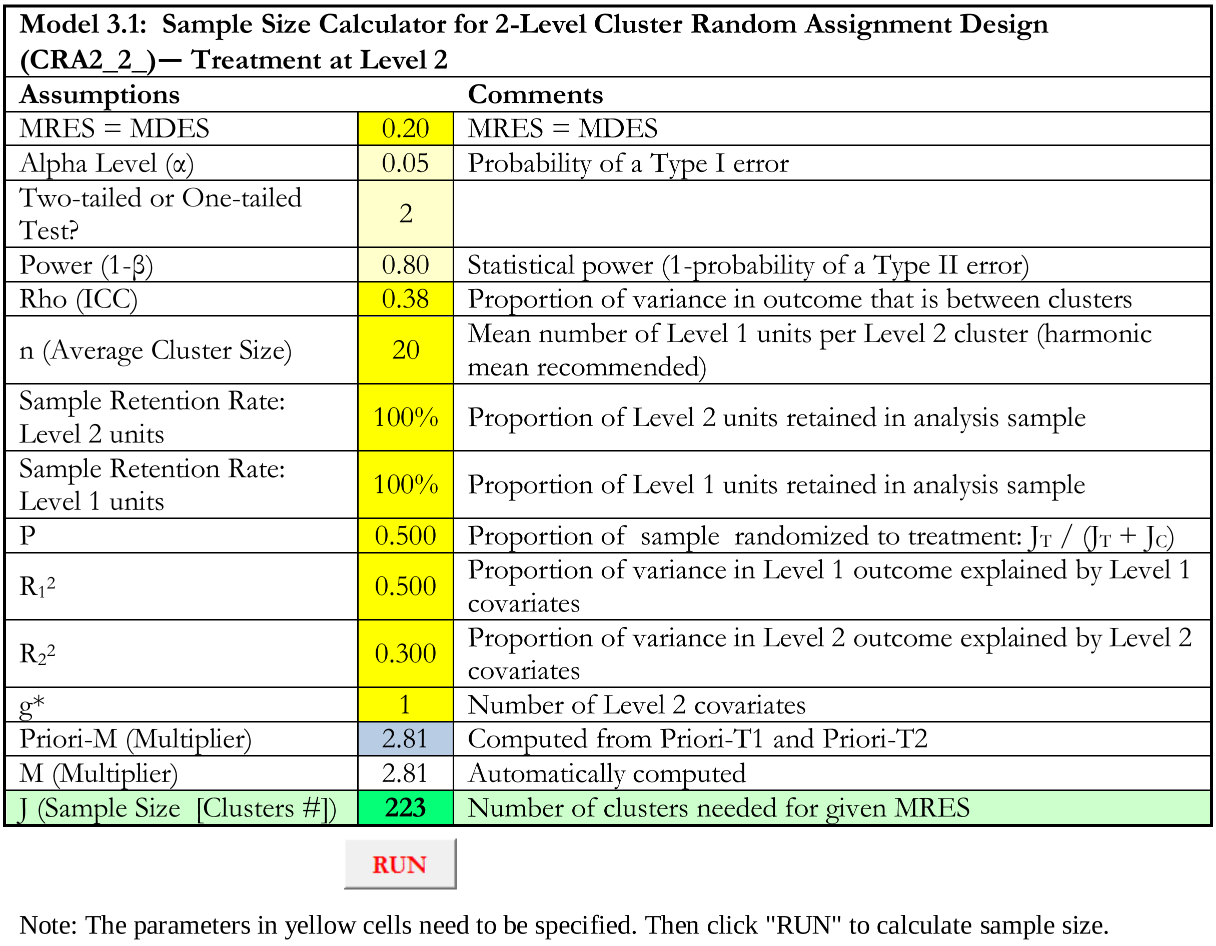


Figure 1. Statistical Power as a Function of Number of Schools for Two-level CRT Example

Precise number of schools to detect an effect size of 0.20 with 80% power rate can be found via calculating minimum required number of schools in PowerUpR (script below) or PowerUp! (Figure 2) as

```
# -- minimum required sample size
mrss.cra2r2(power = .80, es = .20, alpha = .05, two.tailed = TRUE,
            rho2 = .38, r21 = .50, g2 = 1, r22 = .30,
            p = .50, n = 20)

## J = 223
```



Model 3.1: Sample Size Calculator for 2-Level Cluster Random Assignment Design (CRA2_2_)— Treatment at Level 2		
Assumptions		**Comments**
MRES = MDES	0.20	MRES = MDES
Alpha Level (α)	0.05	Probability of a Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1-probability of a Type II error)
Rho (ICC)	0.38	Proportion of variance in outcome that is between clusters
n (Average Cluster Size)	20	Mean number of Level 1 units per Level 2 cluster (harmonic mean recommended)
Sample Retention Rate: Level 2 units	100%	Proportion of Level 2 units retained in analysis sample
Sample Retention Rate: Level 1 units	100%	Proportion of Level 1 units retained in analysis sample
P	0.500	Proportion of sample randomized to treatment: J_T / (J_T + J_C)
R_1^2	0.500	Proportion of variance in Level 1 outcome explained by Level 1 covariates
R_2^2	0.300	Proportion of variance in Level 2 outcome explained by Level 2 covariates
g*	1	Number of Level 2 covariates
Priori-M (Multiplier)	2.81	Computed from Priori-T1 and Priori-T2
M (Multiplier)	2.81	Automatically computed
J (Sample Size [Clusters #])	**223**	Number of clusters needed for given MRES

Note: The parameters in yellow cells need to be specified. Then click "RUN" to calculate sample size.

Figure 2. Minimum Required Number of Schools for Two-level CRT Example

With a sample similar to what we have in terms of average of number of students per school ($n = 20$), intra-class correlation coefficient ($\rho = .38$), explanatory power of covariates at level 1 ($R_1^2 = .50$), and at level 2 ($R_2^2 = .30$), we need at least 223 schools to detect an effect size of 0.20 with a power rate of 80% and type I error rate of 5% for a two-tailed hypothesis testing of the treatment effect.

Explanatory Power of Covariates

Researchers often have control over sample size to increase power rate prior to implementing a two-level CRT. However, in some cases, sampling more units is not feasible or induces prohibitive cost. In this case, explanatory power of covariates for a level can be increased via collecting more information, which in turn improves the power rate. The question naturally comes to mind is whether to collect more information on level 1 or level 2 units. To address this question, we demonstrate to what extent changes in R_1^2 or R_2^2 lead to changes in variance for treatment effect via taking first derivative of $Var(\delta^*)$ with respect to R_1^2 or R_2^2. What becomes apparent is that changes in $Var(\delta^*)$ occur in the opposite direction with changes in R_1^2 or R_2^2 (note negative signs). This means if we increase R_1^2 or R_2^2 this will reduce $Var(\delta^*)$, which in turn improves power rate.

$$\frac{\partial Var(\delta^*)}{\partial R_2^2} = -\frac{\rho}{p(1-p)J}$$

$$\frac{\partial Var(\delta^*)}{\partial R_1^2} = -\frac{(1-\rho)}{p(1-p)nJ}$$

Due to limited resources, researchers may favor collecting information on a level that reduces $Var(\delta^*)$ comparably more. In this case, increasing R_2^2 reduces the variance $(\rho n)/(1-\rho)$ times more compared to the reduction induced by increasing R_1^2 by the same amount (obtained from the ratio of the two derivatives). Therefore, focusing on increasing explanatory power of covariates at level 2 is a more efficient strategy.

For example, for the two-level CRT example, increasing R_2^2 from .40 to .50 (.10 increment) reduces variance from 0.01126 to 0.00974 (a reduction of 0.00152), which, in turn, increases power rate from 46.3% to 51.9%. However, increasing R_1^2 from .30 to .40 (.10 increment) marginally reduces variance from 0.01126 to 0.011136 (a reduction of 0.000124), which, in turn, increases power rate marginally from 46.3% to 46.7%. The ratio of variance reductions is precisely what one would obtain if they use $(\rho n)/(1-\rho)$ formula, which is 12.26. This means increasing R_2^2 by .10 reduces variance 12.26 times more compared to the variance reduction induced by increasing R_1^2 by the same amount.

Three-level CRTs

Null Model to Estimate Unconditional Variation

The following unconditional model can be used to obtain variance parameters σ^2, τ_2^2 and τ_3^2 as defined below, which will be used to calculate standardized variance parameters along with parameters from the full model.

HLM formulation:

Level 1: $Y_{ijk} = \beta_{0jk} + r_{ijk}$

Level 2: $\beta_{0jk} = \gamma_{00k} + \mu_{0jk}$
Level 3: $\gamma_{00k} = \xi_{000} + \varsigma_{00\mathrm{k}}$,

Mixed model formulation:

$$Y_{ijk} = \xi_{000} + \varsigma_{00\mathrm{k}} + \mu_{0jk} + r_{ijk}$$

where r_{ijk}, μ_{0jk}, and $\varsigma_{00\mathrm{k}}$ are level 1, level 2, and level 3 residuals, following normal distributions as $r_{ijk} \sim N(0, \sigma^2)$, $\mu_{0jk} \sim N(0, \tau_2^2)$, and $\varsigma_{00\mathrm{k}} \sim N(0, \tau_3^2)$, respectively. Thus, σ^2, τ_2^2 and τ_3^2 are variances in the outcome between level 1, level 2 and level 3 units, respectively. β_{0jk} is level 1 intercept (in this case, mean of subjects in sub-cluster j and cluster k), γ_{00k} is level 2 intercept (in this case, mean of subjects in all sub-clusters in cluster k), ξ_{000} is level 3 intercept (in this case, mean of all subjects in all sub-clusters in all clusters - grand mean).

Full Model to Estimate Treatment Effect and Conditional Variation

The following model can be used to obtain variance parameters $\sigma^2_{|X}$, $\tau^2_{2|W}$ and $\tau^2_{3|V}$ as defined below, which are used to calculate standardized variance parameters along with parameters from the unconditional model.
HLM formulation:

Level 1: $Y_{ijk} = \beta_{0jk} + \beta_{1jk}X_{ijk} + r_{ijk}$

Level 2: $\beta_{0jk} = \gamma_{00k} + \gamma_{01k}W_{jk} + \mu_{0jk}$
$\beta_{1jk} = \gamma_{10k}$
Level 3: $\gamma_{00k} = \xi_{000} + \delta T_k + \xi_{001}V_k + \varsigma_{00\mathrm{k}}$
$\gamma_{01k} = \xi_{010}$
$\gamma_{10k} = \xi_{100}$

Mixed model formulation:

$$Y_{ijk} = \xi_{000} + \delta T_k + \xi_{001}\mathrm{V_k} + \xi_{010}W_{jk} + \xi_{100}X_{ijk} + \varsigma_{00\mathrm{k}} + \mu_{0jk} + r_{ijk}$$

where r_{ijk}, μ_{0jk}, and $\varsigma_{00\mathrm{k}}$ are conditional residuals following normal distributions as $r_{ijk} \sim N(0, \sigma^2_{|X})$, $\mu_{0jk} \sim N(0, \tau^2_{2|W})$, and $\varsigma_{00\mathrm{k}} \sim N(0, \tau^2_{3|W})$, respectively. Thus, $\sigma^2_{|X}$, $\tau^2_{2|W}$ and $\tau^2_{3|V}$ are residual variances at level 1, level 2 and level 3, respectively, which are not accounted for by the full model. Y_{ijk} is level 1 outcome of interest for subject i in sub-cluster j which is in cluster k, X_{ijk} is level 1 covariate for individual i in sub-cluster j which is in cluster k, W_{jk} is level 2 covariate for sub-cluster j in cluster j, T_k is treatment condition (1 if cluster k assigned to treatment, 0 if not), and V_k is level 3 covariate. β_{0jk}, γ_{00k}, and ξ_{000} are level 1, level 2 and level 3 intercepts, respectively. δ is the treatment effect, β_{1jk} or γ_{10k} or ξ_{100} is regression weight for level 1 covariate X_{ijk}, γ_{01k} or ξ_{010} is regression weight for level 2 covariate W_{jk}, and $\xi_{001}$is regression weight for level 3 covariate V_k.

Similar to two-level CRT case, we can calculate standardized variance parameters based on unstandardized variance parameters from unconditional and full models. $\rho_2 = \tau^2_2/(\tau^2_3 + \tau^2_2 + \sigma^2)$ and represents proportion of variance in the outcome between level 2 units, $\rho_3 = \tau^2_3/(\tau^2_3 + \tau^2_2 + \sigma^2)$ and represents proportion of variance in the outcome between level 3 units, $R^2_1 = 1 - \sigma^2_{|X}/\sigma^2$ and is proportion of variance in the outcome explained by level 1 covariates, $R^2_2 = 1 - \tau^2_{2|W}/\tau^2_2$ and is proportion of variance in the outcome explained by level 2 covariates, and $R^2_3 = 1 - \tau^2_{3|V}/\tau^2_3$ and is proportion of variance in the outcome explained by level 3 covariates. The treatment effect can be standardized in the form of Cohen's d as $\delta^* = \delta/\sqrt{\tau^2_3 + \tau^2_2 + \sigma^2}$.

Standard Error Formula under Balanced Sample Size and Homogenous Variance

Assuming balanced sample sizes, that is, n number of level 1 units per level 2 unit, J number of level 2 units per level 3 unit, and also assuming variance within each level 2 and level 3 unit is same across JK number of level 2 units and K number of level 3 units, standardized standard error takes the form

$$Var(\delta^*) = \frac{\rho_3(1 - R^2_3)}{p(1-p)K} + \frac{\rho_2(1 - R^2_2)}{p(1-p)JK} + \frac{(1 - \rho_2 - \rho_3)(1 - R^2_1)}{p(1-p)nJK}$$

Similar to two-level CRT, standard error of the treatment effect is $SE(\delta^*) = \sqrt{Var(\delta^*)}$. If we know δ^* and $SE(\delta^*)$ we can calculate t statistics, and therefore statistical power can be calculated. $\delta^*/SE(\delta^*)$ follows t distribution with $K - g_3 - 2$ degrees of freedom where g_3 is number of covariates added at level 3 (Dong & Maynard, 2013, p. 52). Statistical power can be calculated as in the two-level CRT case.

Estimation and Standardization of Treatment Effect and Variance Components

Similar to two-level CRT case, considering education settings, we simulated a simple three-level CRT data named `CRT3` which has 6000 students across 300 classrooms in 100 schools (20 students per classroom and 3 classrooms per school). The data includes seven variables; school identification numbers (`schid`), classroom identification numbers (`clsid`), a level 1 outcome variable (`outcome`), a level 3 treatment variable (`treatment`), a level 1 covariate (`covx`), a level 2 covariate (`covw`), and a level 3 covariate (`covv`). First few lines of the simulated data are printed below. Each school and classroom have unique identification numbers (`schid` and `clsid`). Since schools are assigned to treatment conditions, the same school and classrooms therein will have the same values for treatment variable (`treatment`). Level 1 (students) and level 2 (classrooms), and level 3 (schools) covariates (`covx`, `covw`, and `covv`) follow standard normal distributions, and outcome (`outcome`) is a linear function of these covariates with some level 1, level 2, and level 3 noise added (See data generation mechanism in Appendix A).

```
head(CRT3)

##   schid clsid treatment    outcome       covx       covw      covv
## 1     1     1         0  3.0263592  0.5622673 -0.3756029 0.2533185
## 2     1     1         0  1.7124732 -0.0974125 -0.3756029 0.2533185
## 3     1     1         0  1.0353372  1.0164552 -0.3756029 0.2533185
## 4     1     1         0 -0.8436311 -1.1561674 -0.3756029 0.2533185
## 5     1     1         0  1.7452900  2.3208602 -0.3756029 0.2533185
## 6     1     1         0  0.6092003 -0.6035312 -0.3756029 0.2533185
```

As in two-level CRT case, first we estimate variance parameters for unconditional model to calculate intra-class correlation coefficients. The output includes variance for three random effects indicating variation in the outcome that is between school means (`tau23`), between classroom means (`tau22`) and that is between students (`sigma2`). Sum of the three is roughly same as variance of the outcome. Thus, proportion of variance in the outcome that is between schools and classrooms can be calculated (`rho3` and `rho2`).

```
## -- null model (unconditional model)
null.model <- lmer(outcome ~ (1 | schid) + (1 | clsid), data = CRT3)
print(VarCorr(null.model), comp = "Variance")

##  Groups   Name        Variance
##  clsid    (Intercept) 1.2593
##  schid    (Intercept) 0.9969
##  Residual             1.6160

## -- variance parameters
tau23 <- 0.9969
tau22 <- 1.2593
sigma2 <- 1.6160

## -- intra-class correlation coefficients for level 2 and level 3
rho2 <- tau22 / (tau23 + tau22 + sigma2)
rho3 <- tau23 / (tau23 + tau22 + sigma2)
round(rho2, 2)

## [1] 0.33

round(rho3, 2)

## [1] 0.26
```

The output for the full model, again, includes variance for three random effects indicating conditional variation in the outcome that is between schools (`tau23v`), classrooms (`tau22w`) and students (`sigma2x`) beyond what is explained by level 3, level 2 and level 1 predictors, respectively. As some of the variation between schools, between classrooms and between students are explained by level 3, level 2 and level 1 variables, using proportion of reduction in the variance for level 3, level 2 and level 1 we can calculate R-squared values for each (`r23`, `r22` and `r21`).

```
## --  full model
full.model <- lmer(outcome~ treatment + covx + covw + covv +
               (1 | schid) + (1 | clsid), data = CRT3)
print(VarCorr(full.model), comp = "Variance")

##  Groups   Name        Variance
##  clsid    (Intercept) 1.06824
##  schid    (Intercept) 0.71853
##  Residual             1.00901

## -- variance parameters
tau23v <- 0.7185
tau22w <- 1.0682
sigma2x <- 1.0090

## -- R-squared values for level 1, level 2 and level 3
r21 <- 1 - (sigma2x / sigma2)
r22 <- 1 - (tau22w / tau22)
r23 <- 1 - (tau23v / tau23)
round(r21, 2)

## [1] 0.38

round(r22, 2)

## [1] 0.15

round(r23, 2)

## [1] 0.28

## -- treatment effect
coef(summary(full.model))["treatment",]

##   Estimate Std. Error    t value
##  0.9323254  0.2124156  4.3891572

delta <- 0.9323
es <- delta / sqrt(sigma2 + tau22 + tau22)
round(es, 2)

## [1] 0.46
```

Statistical Power and Minimum Required Sample Size Calculations

Default parameters for power analysis are same as two-level CRT case. Different from two-level CRT case, there are 20 students per classroom (`n`), 3 classrooms per school (`J`), and 100 schools (`K`) in total. Now we can calculate statistical power as

```
## -- power analysis
design <- power.cra3r3(es = .20, alpha = .05, two.tailed = TRUE,
                       rho2 = .33, rho3 = .26,
                       r21 = .38, r22 = .15, g3 = 1, r23 = .28,
                       p = .50, n = 20, J = 3, K = 100)

##
## Statistical power:
## ---------------------------------------
##  0.458
## ---------------------------------------
## Degrees of freedom: 97
## Standardized standard error: 0.107
## Type I error rate: 0.05
## Type II error rate: 0.542
## Two-tailed test: TRUE
```

where, in addition to calculated parameters above, `g3` is number of covariates added at level 3. Parameters obtained from the data produce a power rate of 45.8%, which means if we repeat this experiment for a large number of times, we will detect a statistically significant treatment effect 45.8% of the time, if in fact there is a treatment effect in the underlying population. Figure 3 demonstrates how far we are from the benchmark power rate. By visual inspection, it seems a sample consisting of somewhere between 200 to 250 schools is capable of producing results with 80% power rate.

```
plot(design, ypar = "power", locate = TRUE, xlim = c(50, 250))
```

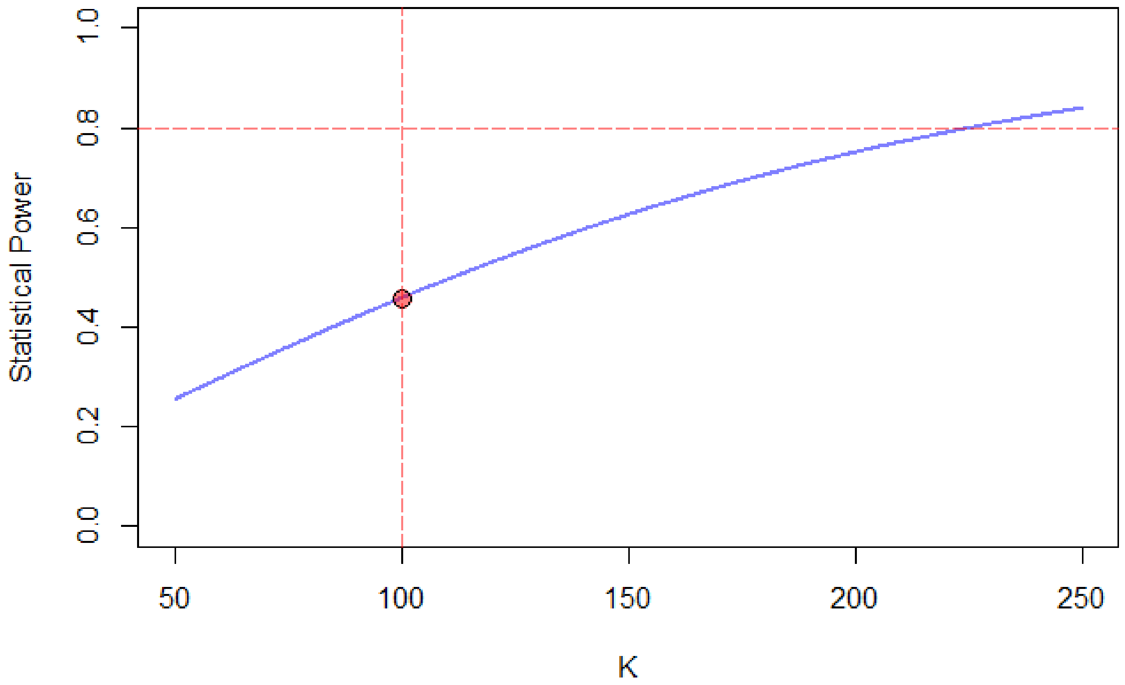
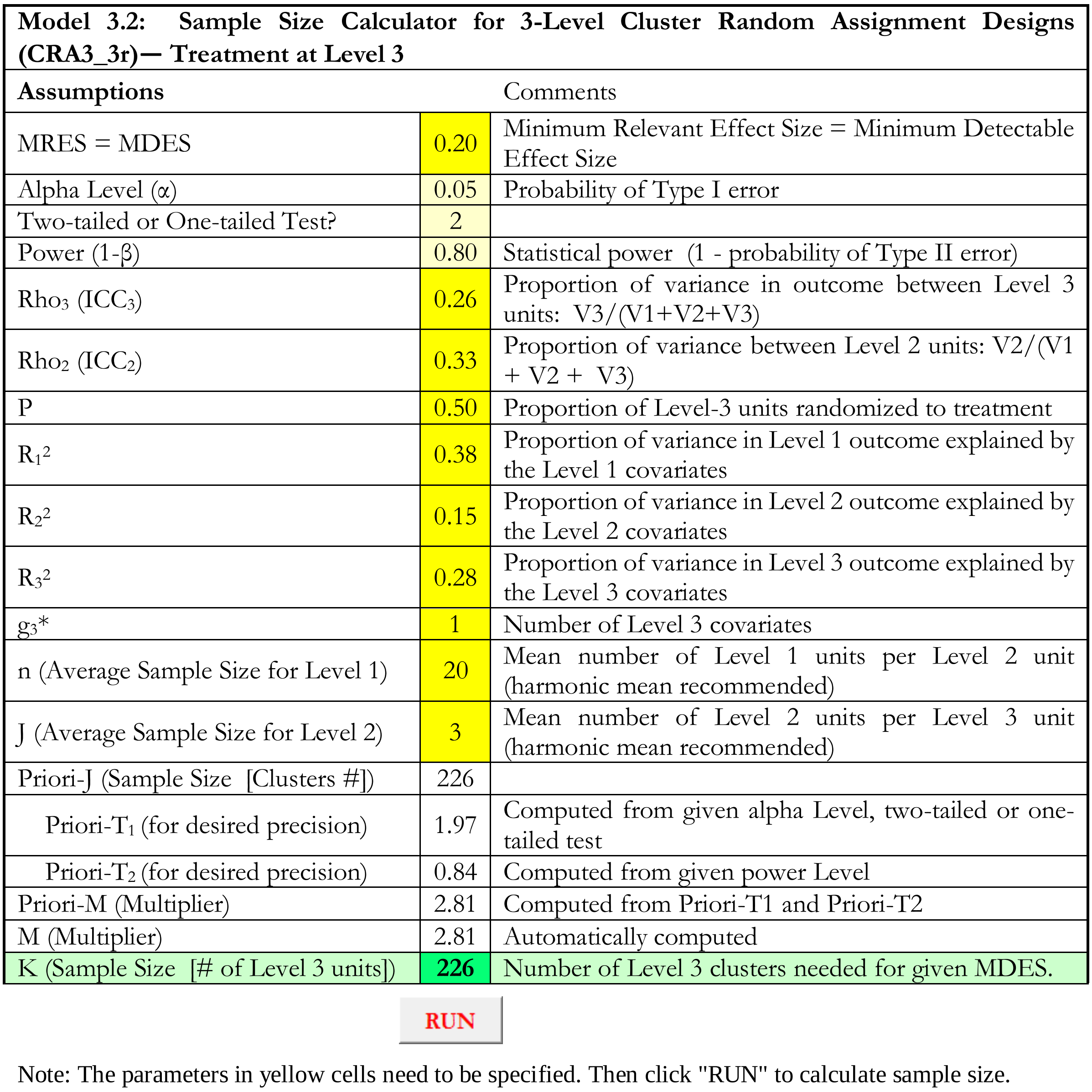


Figure 3. Statistical Power as a Function of Number of Schools for Three-level CRT Example

To find minimum required number of schools needed to detect an effect size of 0.20 with a power rate of 80% we can use PowerUpR (script below) or PowerUp! (Figure 4) as

```
# -- minimum required sample size
mrss.cra3r3(power = .80, es = .20, alpha = .05, two.tailed = TRUE,
            rho2 = .33, rho3 = .26,
            r21 = .38, r22 = .15, g3 = 1, r23 = .28,
            p = .50, n = 20, J = 3)

## K = 226
```



Model 3.2: Sample Size Calculator for 3-Level Cluster Random Assignment Designs (CRA3_3r)— Treatment at Level 3

Assumptions		Comments
MRES = MDES	0.20	Minimum Relevant Effect Size = Minimum Detectable Effect Size
Alpha Level (α)	0.05	Probability of Type I error
Two-tailed or One-tailed Test?	2	
Power (1-β)	0.80	Statistical power (1 - probability of Type II error)
Rho$_3$ (ICC$_3$)	0.26	Proportion of variance in outcome between Level 3 units: V3/(V1+V2+V3)
Rho$_2$ (ICC$_2$)	0.33	Proportion of variance between Level 2 units: V2/(V1 + V2 + V3)
P	0.50	Proportion of Level-3 units randomized to treatment
R_1^2	0.38	Proportion of variance in Level 1 outcome explained by the Level 1 covariates
R_2^2	0.15	Proportion of variance in Level 2 outcome explained by the Level 2 covariates
R_3^2	0.28	Proportion of variance in Level 3 outcome explained by the Level 3 covariates
g_3*	1	Number of Level 3 covariates
n (Average Sample Size for Level 1)	20	Mean number of Level 1 units per Level 2 unit (harmonic mean recommended)
J (Average Sample Size for Level 2)	3	Mean number of Level 2 units per Level 3 unit (harmonic mean recommended)
Priori-J (Sample Size [Clusters #])	226	
Priori-T$_1$ (for desired precision)	1.97	Computed from given alpha Level, two-tailed or one-tailed test
Priori-T$_2$ (for desired precision)	0.84	Computed from given power Level
Priori-M (Multiplier)	2.81	Computed from Priori-T1 and Priori-T2
M (Multiplier)	2.81	Automatically computed
K (Sample Size [# of Level 3 units])	**226**	Number of Level 3 clusters needed for given MDES.

Note: The parameters in yellow cells need to be specified. Then click "RUN" to calculate sample size.

Figure 4. Minimum Required Number of Schools for Three-level CRT Example

With a sample similar to what we have in terms of average number of students per classroom ($n = 20$), average number of classrooms per school ($J = 3$), intra-class correlation coefficients ($\rho_2 = .33$

and $\rho_3 = .26$), explanatory power of covariates at level 1 ($R_1^2 = .38$), level 2 ($R_2^2 = .15$), and at level 3 ($R_3^2 = .28$), power analysis result suggest that we need at least 226 schools to detect an effect size of 0.20 with a power rate of 80% and type I error rate of 5% for a two-tailed hypothesis testing of treatment effect.

Explanatory Power of Covariates

Due to the same reasons and similar to two-level CRT case, one should keep in mind that it is more efficient to increase explanatory power of covariates via including additional covariates at the third level. If we take first derivative of $Var(\delta^*)$ with respect to R_1^2, R_2^2, or R_3^2, what becomes apparent is that changes in $Var(\delta^*)$ occur in the opposite direction with changes in R_1^2, R_2^2, or R_3^2. This means increase in explanatory power for any of the R_1^2, R_2^2, or R_3^2 will reduce $Var(\delta^*)$, which improves the power rate.

$$\frac{\partial Var(\delta^*)}{\partial R_3^2} = -\frac{\rho_3}{p(1-p)K}$$

$$\frac{\partial Var(\delta^*)}{\partial R_2^2} = -\frac{\rho_2}{p(1-p)JK}$$

$$\frac{\partial Var(\delta^*)}{\partial R_1^2} = -\frac{(1-\rho_2-\rho_3)}{p(1-p)nJK}$$

Comparably, increasing R_3^2 reduces the variance $(\rho_3 J)/\rho_2$ times more compared to the reduction induced by increasing R_2^2 by the same amount, and $(\rho_3 nJ)/(1-\rho_2-\rho_3)$ times more compared to the reduction induced by increasing R_1^2. Therefore, focusing on increasing explanatory power of covariates at level 3 is a more efficient strategy.

For example, for the three-level CRT example, increasing R_3^2 from .28 to .38 (.10 increment) reduces variance from 0.011398 to 0.010357 (a reduction of 0.00104), which, in turn, increases power rate from 45.8% to 49.4%. Similarly, increasing R_2^2 from .15 to .25 (.10 increment) reduces variance from 0.011398 to 0.010957, which, in turn, increases power rate from 45.8% to 47.3%. The ratio of variance reductions is precisely what one would obtain if they use $(\rho_3 J)/\rho_2$ formula, which is 2.36. This means increasing R_3^2 by .10 reduces variance 2.36 times more compared to the variance reduction induced by increasing R_2^2 by the same amount. However, increasing R_1^2 from .48 to .58 (.10 increment) reduces variance marginally from 0.011398 to 0.011370, which, in turn, increases power rate marginally from 45.8% to 45.9%. Ratio of variance reductions is precisely what one would obtain if they use $(\rho_3 nJ)/(1-\rho_2-\rho_3)$ formula, which is 38. This means increasing R_3^2 by .10 reduces variance 38 times more compared to the variance reduction induced by increasing R_1^2 by the same amount.

DISCUSSION and CONCLUSION

In this tutorial, we demonstrated how to analyze and plan two- and three-level CRTs. We provided statistical models and estimated variance parameters to further use them in statistical power analysis procedures. Most of the power analysis programs require specification of standardized variance parameters. We also demonstrated how to standardize variance parameters into intra-class correlation coefficients and R-squared values. This guide will potentially assist researchers in their endeavors to plan two- and three-level CRTs with greater precision, thus, provide reliable results to evaluators, stakeholders and policy makers.

Statistical power calculations for two- and three-level CRTs can be conducted in any software program that allows standardized parameters as input (e.g., Optimal Design Plus, PowerUpR and PowerUp!).

Results from minimum required sample size (MRSS) calculations in PowerUp! and PowerUpR are compared to each other in nine slightly different designs (D1-D9 in Table B1) for two-level CRT, changing one parameter at a time. The same procedure is repeated for three-level CRT (D1-D12 in Table B2). Results indicate that MRSS calculations in both software programs are very much the same, rarely differ by one unit as a result of rounding difference in two different platforms.

We elaborated on the explanatory power of covariates and their relation to statistical power, demonstrated that collecting more information on higher level units and including them in statistical models as covariates improve power rate substantially. In contrast, covariates added at the individual level improve power rate only marginally. Thus, if there are financial and practical challenges to sampling more clusters, an alternative strategy would be focusing on improving explanatory power of covariates.

From the beginning of an intervention to the end, some clusters and individuals therein may refuse or discontinue participating, resulting in non-participation or attrition which deteriorates the power rate. Non-participation and attrition rates can also be obtained from prior research, for which minimum required sample size calculations can be adjusted accordingly. Thus, when analyzing existing data or reporting results, documenting non-participation and attrition rates will also help researchers to design CRTs with greater precision. One thing to keep in mind, in education context for example, is the fact that those students within schools cannot be oversampled while we can sample additional schools to adjust the sample size for non-participation or attrition.

There are some limitations to this guide. Although we demonstrated how to estimate variance parameters for CRTs, there might be other practical issues a researcher needs to deal with. For example, there might be missing data, outliers, or assumption of linearity may not hold. Researchers may also need to use weights, if they would like to plan for generalizable large-scale CRTs, and they have access to similar large-scale data sets. Such topics require an extensive treatment and are beyond the scope of this guide.

REFERENCES

Bland J. M. (2004). Cluster randomized trials in the medical literature: two bibliometric surveys. *BMC Medical Research Methodology*, *4*(21). DOI: https://doi.org/10.1186/1471-2288-4-21

Bloom, H. S. (1995). Minimum detectable effects a simple way to report the statistical power of experimental designs. *Evaluation Review,* *19*(5), 547-556. DOI: https://doi.org/10.1177/0193841X9501900504

Bloom, H. S. (2006). The core analytics of randomized experiments for social research. MDRC Working Papers on Research Methodology. New York, NY: MDRC. Retrieved from DOI: https://www.mdrc.org/sites/default/files/full_533.pdf.

Bloom, H. S., Bos, J. M., & Lee, S. W. (1999). Using cluster random assignment to measure program impacts statistical implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445-469. DOI: https://doi.org/10.1177%2F0193841X9902300405

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2019). PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments. R package version 1.0.4. DOI: https://CRAN.R-project.org/package=PowerUpR

Cameron, A. C., & Miller, d. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, *50*, 317-372. DOI: https://doi.org/10.3368/jhr.50.2.317

Dong, N., & Maynard, R. (2013). *PowerUp!*: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-experimental Design Studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24-67. DOI: https://doi.org/10.1080/19345747.2012.673143

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. DOI: https://doi.org/10.18637/jss.v067.i01

Hayes, R. J. & Moulton, L. H. (2017). *Cluster Randomized Trials* (2nd ed.). New York, NY: Chapman and Hall/CRC Press. DOI: https://doi.org/10.4324/9781315370286

Hedberg, E. C. (2016). Academic and behavioral design parameters for cluster randomized trials in kindergarten: an analysis of the Early Childhood Longitudinal Study 2011 Kindergarten Cohort (ECLS-K 2011). *Evaluation Review*, *40*(4), 279-313. DOI: https://doi.org/10.1177/0193841X16655657

Hedberg, E. C., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics: Results from a meta-analysis of district-specific values. *Evaluation Review*, *38*(6), 546-582. DOI: https://doi.org/10.1177/0193841X14554212

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445-489. DOI: https://doi.org/10.1177/0193841X14529126

Hedges, L. V., & Rhoads, C. (2010). *Statistical Power Analysis in Education Research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://files.eric.ed.gov/fulltext/ED509387.pdf

Konstantopoulos, S. (2009a). Using power tables to compute statistical power in multilevel experimental designs. *Practical Assessment, Research & Evaluation*, *14*(10).

Konstantopoulos, S. (2009b). Incorporating Cost in Power Analysis for Three-Level Cluster-Randomized Designs. *Evaluation Review*, *33*(4), 335-357. DOI: https://doi.org/10.1177/0193841X09337991

Moerbeek, M., & Safarkhani, M. (2018). The design of cluster randomized trials with random cross-classifications. *Journal of Educational and Behavioral Statistics*, *43*(2), 159-181. DOI: https://doi.org/10.3102/1076998617730303

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing [Computer software]. Vienna, Austria. Retrieved from https://www.R-project.org.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: an examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. International *Journal of Research & Method in Education*, *39*(3), 255-267. DOI: https://doi.org/10.1080/1743727X.2016.1150454

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: A multistate analysis. *AERA Open*, *2*(1). DOI: https://doi.org/10.1177/2332858415625975

Westine, C. D. (2016). Finding Efficiency in the Design of Large Multisite Evaluations: Estimating Variances for Science Achievement Studies. *American Journal of Evaluation*, *37*(3), 311-325. DOI: https://doi.org/10.1177/1098214015624014

Westine, C. D., Spybrook, J., & Taylor, J. A. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490-519. DOI: https://doi.org/10.1177/0193841X14531584

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 242-278.

Appendix A
Data Generation Process

Data Generating Model for Two-level CRT

The statistical model to generate data for two-level CRT is same as the statistical model described in the main text. Here we provide only the mixed model formulation, which is

$$Y_{ij} = \gamma_{00} + \delta T_j + \gamma_{01} W_j + \gamma_{10} X_{ij} + \mu_{0j} + r_{ij}$$

where parameters are explained elsewhere in the main text. The following parameter values are used in the simulation, while considering 20 students per school (n) and 100 schools in total (J).

$$\gamma_{00} = 0$$
$$\delta = 1$$
$$T_j \sim BERN(0.50)$$
$$\gamma_{01} = 0.50$$
$$W_j \sim N(0,1)$$
$$\gamma_{10} = 1$$
$$X_{ij} \sim N(0,1)$$
$$\mu_{0j} \sim N(0,1)$$
$$r_{ij} \sim N(0,1)$$

```
set.seed(123) # for replication
delta <- 1
js <- 100
ns <- rep(20, js)
id <- as.factor(rep(1:js, ns))
tj <- rep(rbinom(js, 1, .50), ns)
wj <- rep(rnorm(js), ns)
uj <- rep(rnorm(js), ns)
xij <- rnorm(sum(ns))
rij <- rnorm(sum(ns))
yij <- delta * tj + 0.50 * wj + xij + uj + rij

CRT2 <- data.frame("schid" = id,
                   "treatment" = tj,
                   "outcome" = yij,
                   "covx" = xij,
                   "covw" = wj)
```

Data Generating Model for Three-level CRT

The mixed model formulation for three-level CRT is

$$Y_{ijk} = \xi_{000} + \delta T_k + \xi_{001} \mathrm{V_k} + \xi_{010} W_{jk} + \xi_{100} X_{ijk} + \varsigma_{00\mathrm{k}} + \mu_{0jk} + r_{ijk}$$

where parameters are explained elsewhere in the main text. The following parameter values are used in the simulation, while considering 20 students per classroom (n), 3 classrooms per school (J), and 100 schools in total (K).

$$\xi_{000} = 0$$
$$\delta = 1$$
$$T_k \sim BERN(0.50)$$
$$\xi_{001} = 0.25$$
$$\mathrm{V_k} = N(0,1)$$
$$\xi_{010} = 0.50$$
$$W_{jk} \sim N(0,1)$$
$$\xi_{100} = 0.75$$

$$X_{ijk} \sim N(0,1)$$
$$\varsigma_{00k} \sim N(0,1)$$
$$\mu_{0jk} \sim N(0,1)$$
$$r_{ijk} \sim N(0,1)$$

```
set.seed(123) # for replication
delta <- 1
ks <- 100
js <- rep(3, ks)
ns <- rep(20, sum(js))

id3 <- as.factor(rep(rep(1:ks, js), ns))
id2 <- as.factor(rep(rep(1:sum(js), ns)))

tk <- rep(rep(rbinom(ks, 1, .50), js), ns)
vk <- rep(rep(rnorm(ks), js), ns)
sk <- rep(rep(rnorm(ks), js), ns)
wjk <- rep(rep(rnorm(sum(js)), ns))
ujk <- rep(rep(rnorm(sum(js)), ns))
xijk <- rnorm(sum(ns))
rijk <- rnorm(sum(ns))
yijk <- delta * tk + 0.25 * vk + 0.50 * wjk +  0.75 * xijk + sk + ujk +
rijk

CRT3 <- data.frame("schid" = id3,
                   "clsid" = id2,
                   "treatment" = tk,
                   "outcome" = yijk,
                   "covx" = xijk,
                   "covw" = wjk,
                   "covv" = vk)
```

Appendix B
PowerUpR and PowerUp! Comparisons

Table B1

Comparison for Two-level CRTs

Assumptions	**Base**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**
MRES = MDES	0.20	**0.40**	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Alpha Level (α)	0.05	0.05	**0.01**	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Two-tailed or One-tailed Test?	2	2	2	**1**	2	2	2	2	2	2
Power (1-β)	0.80	0.80	0.80	0.80	**0.20**	0.80	0.80	0.80	0.80	0.80
Rho (ICC)	0.40	0.40	0.40	0.40	0.40	**0.20**	0.40	0.40	0.40	0.40
n (Average Cluster Size)	20	20	20	20	20	20	**10**	20	20	20
P	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.30**	0.50	0.50
R_1^2	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.20**	0.50
R_2^2	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	**0.50**
J (Sample Size [# of Level 2 units]) in ***PowerUp!***	**234**	**60**	**348**	**184**	**41**	**128**	**246**	**239**	**241**	**171**
J (Sample Size [# of Level 2 units]) in **PowerUpR**	233	60	348	184	41	128	245	238	241	171

Note. g (number of covariates added at level 2) is fixed at 1 for all nine designs.

Table B2

Comparison for Three-level CRTs

Assumptions	Base	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
MRES = MDES	0.20	**0.40**	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Alpha Level (α)	0.05	0.05	**0.01**	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Two-tailed or One-tailed Test?	2	2	2	**1**	2	2	2	2	2	2	2	2	2
Power (1-β)	0.80	0.80	0.80	0.80	**0.20**	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Rho$_3$ (ICC$_3$)	0.30	0.30	0.30	0.30	0.30	**0.15**	0.30	0.30	0.30	0.30	0.30	0.30	0.30
Rho$_2$ (ICC$_2$)	0.30	0.30	0.30	0.30	0.30	0.30	**0.10**	0.30	0.30	0.30	0.30	0.30	0.30
P	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.30**	0.50	0.50	0.50	0.50	0.50
R$_1^2$	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.30**	0.50	0.50	0.50	0.50
R$_2^2$	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.40**	0.50	0.50	0.50
R$_3^2$	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.70**	0.50	0.50
n (Average Sample Size for Level 1)	20	20	20	20	20	20	20	20	20	20	20	**10**	20
J (Average Sample Size for Level 2)	2	2	2	2	2	2	2	2	2	2	2	2	**3**
K (Sample Size [# of Level 3 units]) in ***PowerUp!***	**183**	**47**	**272**	**144**	**33**	**125**	**145**	**217**	**184**	**194**	**136**	**187**	**162**
K (Sample Size [# of Level 3 units]) in **PowerUpR**	**183**	**47**	**272**	**144**	**33**	**125**	**145**	**217**	**184**	**194**	**135**	**186**	**162**

Note. g_3 (number of covariates added at level 3) is fixed at 1 for all 12 designs.