Eurasian Journal of Educational Research 87 (2020) 101-118

Eurasian Journal of Educational Research
www.ejer.com.tr

Examining of Internal Consistency Coefficients in Mixed-Format Tests in Different Simulation Conditions*

Hatice GURDIL EGE[1], Ergul DEMIR[2]

ARTICLE INFO

Article History:
Received: 12 Mar. 2019
Received in revised form: 24 Feb. 2020
Accepted: 19 May. 2020
DOI: 10.14689/ejer.2020.87.5

Keywords
mixed format test, reliability, stratified α, angoff-feldt, feldt-raju

ABSTRACT

Purpose: The present study aims to evaluate how the reliabilities computed using α, Stratified α, Angoff-Feldt, and Feldt-Raju estimators may differ when sample size (500, 1000, and 2000) and item type ratio of dichotomous to polytomous items (2:1; 1:1, 1:2) included in the scale are varied.

Research Methods: In this study, Cronbach's α, Stratified α, Angoff-Feldt, and Feldt-Raju reliability coefficients were estimated on simulated datasets (sample sizes 500, 1000, 2000) and the number of dichotomous versus polytomous item ratios (2:1, 1:1, 1:2).

Findings: In the simulation conditions of this research, in all sample size conditions, estimated Angoff-Feldt, and Feldt-Raju reliability coefficients were higher when the number of dichotomous items in the item-type ratio was higher than that of polytomous items. This was also the case for the estimated α and Stratified α reliability coefficients when the item-type ratio was reversed. While all different reliability estimators gave similar results in the large samples (n≥1000), there were some differences in reliability estimates depending on the item-type ratio in the small samples (n=500).

Implications for Research and Practice: In the light of the findings and conclusions obtained in this study, it may be advisable to use α and Stratified α for mixed-type scales when the number of polytomously scored items in the scale is higher than that of the dichotomously scored items. On the other hand, the coefficients Angoff-Feldt and Feldt-Raju are recommended when the number of items scored dichotomously is higher.

* This article was derived from the first author's a master's dissertation conducted under the supervision of the second author.

[1] Corresponding Author, Atam Primary School, TURKEY, e-mail: haticegurdil1985@gmail.com, ORCID: https://orcid.org/0000-0002-0079-3202

[2] Ankara University, Educational Science Faculty, TURKEY, e-mail:erguldemir@ankara.edu.tr, ORCID: https://orcid.org/0000-0002-3708-8013

Introduction

Tests in which different item types are used in the same test are called mixed-format tests. Berger (1998) considers mixed-format test from different perspectives and describes it as a test that emerges when a combination of the item types that require different scoring forms, such as dichotomous and polytomous scoring. Because the mixed-format tests are composed of different scoring items, the total test score is defined as a composite score. In this context, estimating reliability depends on how to obtain these composites.

In the general framework, reliability, a feature that must be present in mixed-format tests, as well as in all tests, is defined as the reproducibility of measurements of a given characteristic applied to the same individuals in similar conditions (Crocker & Algina, 2008). One or more application-based methods are used to estimate reliability. The test-retest method and equivalent form methods are based on multiple applications, while Cronbach's α is based on a single application. Among these methods, Cronbach's α is easier to use because they need just one application. On the other hand, when tests contain heterogeneous substance types, the classical reliability coefficients (e.g., Cronbach' α coefficient) may yield misleading results in mixed-format tests (Zinbarg et al., 2005). If the parts differ in their standard deviations but are tau equivalent, Cronbach's α is appropriate. However, if the two parts comprise heterogeneous item types, a less well-known estimate, the Angoff-Feldt coefficient, is appropriate (Feldt & Charter, 2003).

There are several methods, including Stratified α, Raju, Angoff-Feldt, Feldt-Gilmer, Kristof coefficients, to estimate the reliability of the composite score of the mixed-format test (Osburn, 2000). All these coefficients are estimated by considering the strata or subtests of the test. Young and Yoon (1998) stated that the total score of a mixed-format test is a composite score, and this composite score is stratified with the different types of items or tasks. In this context, it is obvious that open-ended items and multiple-choice items can be defined as the different strata or subtest of a mixed-format test. Stratified α assumed that the components of a composite can be divided into subgroups based on content or difficulty. When the components of a composite can be grouped into subtests, Stratified alpha may provide a better estimation of the reliability than Cronbach's α coefficient computed on the same composites (He, 2009; Cronbach, Schonemanve McKie, 1965 as cited in Osburn, 2000). The following formula is given by Feldt and Brennan (1989) for the Stratified α coefficient:

Stratified $\alpha = \dfrac{1-\sum \sigma_i^2(1-\alpha_i)}{\sigma_x^2}$

σ_i^2 : subtest i variance

σ_x^2 : test total variance

α_i: α coefficient in subtest i

The Angoff-Feldt coefficient can be used when the length of the test parts, called as sub-tests or parts of different item types, are arbitrary. The Angoff-Feldt coefficient (r_{AF}), which is less known than the Cronbach's α coefficient, is used if the two parts contain heterogeneous item types or are not equal concerning functional length. The Angoff-Feldt coefficient assumed that test could be divided into only two parts of arbitrary length; the scores could only be congeneric equivalent and the sum of the error variances for the two parts is equal to the error variance of the total test. Feldt and Brennan (1989) have given the following formula for the Angoff-Feldt coefficient:

$$\rho = \frac{4\sigma_{12}}{\sigma_x^2 - \left(\frac{\sigma_1^2 - \sigma_2^2}{\sigma_x}\right)^2}$$

σ_{12} : 1. and 2. item covariance

σ_x^2 : Total variance

σ_1^2 : First subtest variance

σ_2^2 : Second subtest variance

σ_x: Total covariance

The Feldt-Raju coefficient is a coefficient obtained by combining the Raju coefficient and the Feldt coefficient. It has been developed to estimate the reliability when a different number of items are placed in sub-tests (expressed as part of a test) or in sections comprised of different item types. The Feldt-Raju coefficient assumed that the parts of a test are most congeneric. Osbourn (2000) has given the following formula for the Feldt-Raju coefficient:

$$F - R\rho tt' = \frac{\left(\sigma_i^2 - \sum \sigma_i^2\right)}{\left(1 - \sum \lambda_i^2\right)\sigma_i^2}$$

σ_i^2 : Item i variance

σ_t^2 : Total point variance

σ_{it} : Covariance between item i and total

$\lambda_i = \sigma_{it} / \sigma_t^2$: The functional additive of the first component

As seen in the formula above, for mixed-format tests, the accuracy of estimating the reliability depends on the strata or subtests and their characteristics. Defining the

strata or subtests, the number of items each subtest, item types, item, and total test scoring mechanism are crucial for accuracy.

When the studies on mixed-format tests are examined in general, it has been seen that vast majority of these studies are based on mixed-format test equating (Bastari, 2000; Cao, 2008; Gubes, 2014; He, 2011; Hu, 2018; Kim & Lee, 2018; Kirkpatrick, 2005; Lee & Lee, 2016; Li, Chen, & Li, 2018; Uysal & Kilmen, 2016) and the weighting of the items in mixed-format tests by different methods (Gultekin, 2011; Saen-amnuaiphon, Tuksino & Nichanong, 2012). There is also research on the comparison of different weighting methods (Ercikan et al.,1998; Gultekin, 2011), comparison of different calibration methods, item analysis methods and scale transformation methods in mixed-format tests (Kim & Lee, 2006; Kinsey, 2003), scoring of mixed-format tests (Donoghue, 1993; Skyes et al., 2001), the classification accuracy of mixed-format tests (Kim and Lee 2019; Wang, Drasgow & Liu, 2016), and opinions of teachers and students on using the mixed-format tests in classrooms (Eren, 2015).

As mentioned above, there are many research studies on the mixed-format tests, yet only a few of them focus on the reliability of mixed-format tests in particular (Falk & Savalei, 2011; Osburn, 2000). These studies investigate how different item types used in different ratios influence reliability estimates for given conditions. As mentioned before, there are some options. However, as far as we know, there are no studies comparing various methods by examining both the effects of item-type ratios and sample sizes on the resulting reliability estimations. It is of great importance that researchers should be able to choose the reliability estimator best suited to their particular study conditions involving mixed item-type tests with different item type ratios given the sample sizes at hand. In this context, for mixed-format tests, it should be clarified which reliability coefficients would be more accurate and how they change under some specific conditions.

The present study aims to define how the Cronbach's α, Stratified α, Angoff-Feldt, and Feldt-Raju change and the descriptive relationship between the estimations of these coefficients; when the sample size (500, 1000 and 2000) and the proportion (2:1; 1:1 and 1:2) of the item types used in the mixed-formed tests vary. As mentioned before, α is not appropriate for the mixed-format test (DeVellis, 2003; Lucke, 2005a; Zinbarg et al., 2005). However, it is also one of the very common techniques. In this study, Cronbach's α was used as a criterion for assessing the other coefficients.

Method

Research Design

In this study, several different reliability estimates of test scores were examined under the conditions that were determined on the basis of mixed-format simulation design. This research was conducted as basic research.

Simulation Design

In this study, response patterns were produced for a 30-item test, including both dichotomously and polytomously scored items in different ratios. The factors that

were held constant in the conditions were the number of item types (2: dichotomous and polytomous) used generating the model (unidimensional 2PL for dichotomous items and unidimensional PCM for polytomous items), the total number of items (k=30), the number of response categories (2 for dichotomous and 4 for polytomous), the total scoring method, the item discriminations and the item difficulties.

The Partial Credit Model was chosen for the items that were categorized in the generating data since this model was developed by Master (1982) for the analysis of multistep test items. While determining the total number of items, it was considered that it was preferred often 30-item test length for the studies in the related literature of previous years (Baker, 1998; Kinsey, 2003) and item numbers of the subtests in the large-scale tests applied in Turkey generally ranges between 20 and 30 (KPSS and LGS). While determining the number of response categories, it was considered that in applications where the item and ability parameters are predicted, TIMSS and PISA use quadripartite response according to the IRT scale for the success variable. For the total scoring method, it has been considered that Wainer (1976) recommended the use of equal weighting in mixed-format tests. As of the item discrimination, it has been considered that Hambleton, Swaminathan and Rogers (1991) stated that the item discrimination index (a) in the IRT model is expressed as the defined normal range (0.00-2.00). For the item difficulty, it is produced in a uniform distribution (-3.00; 3.00), considering that it is close to real values. When the literature related to sample size is examined, it is taken into account the use of 500 (Baker, 1998; Odabas, 2016), 1000 (Odabas, 2016) and 2000 (Gao & Chen, 2005; Spray, 1990) individuals were taken into consideration.

Data Production

The data for each sample size were produced in the size of the sample concerned, with the ability estimates of the individuals fixed and the normal distribution of individuals with an actual score average of 0.00 and standard deviations of 1.00. 20 items scored dichotomously were produced by using a Two-Parameter Logistic Model. Then, 10 items scored polytomously were produced with a response category number of four. Twenty-five replications were performed to obtain the corresponding number of response patterns for each run. The number of samples (500, 1000 and 2000) and item rates (2:1, 1:1 and 1:2) were changed and the first five steps were repeated for each of these conditions. It is expected that the simulations considered should present data that are reasonably close to real-world conditions. Real-world includes the conditions that poltyhomously scored items are less than dichotomously scored items. However, this study includes the conditions that poltyhomously scored items are equal or more than dichotomously scored items. The reason for that is to observe and define the effects of the item rates in a more clear way.

In this study, response data were generated to be unidimensional. Lord (1980) states that unidimensionality is also a sign of local independence. Also, because there are no missing and time limitations in generating process, data do not show the speed test structure. Harwell, Stone, Hsu and Kirisci (1996) state that errors decrease and the effect approaches 1.0 after 25 replications. Also, there is some research using this

criterion. For example, Gul (2015) also used 25 replications. Considering this situation, 25 replications were made in generating data. Thus, 9x25= 225 different datasets were produced for 3x3 = 9 different experimental conditions. The WinGen program was used to produce data suitable for the conditions determined for this research. All simulation conditions used in the study are presented in Table 1.

Table 1

Simulation Conditions

Sample Size	Number of Items		Item Type Ratio
	Polytomous Items	Dichotomous Items	
500	20	10	2:1
	15	15	1:1
	10	20	1:2
1000	20	10	2:1
	15	15	1:1
	10	20	1:2
2000	20	10	2:1
	15	15	1:1
	10	20	1:2

According to Table 1, sample sizes are 500, 1000 and 2000; item type ratios are 1: 2, 1: 1 and 2:1 and the first part of the test consists of polytomous items.

Data Analysis

For Cronbach's α and standard error value for each condition and replication, SPSS was used for each dichotomous and polytomous items. Mean values of Cronbach's α coefficients obtained from 25 replication and standard error values were taken and tabulated. Table values are evaluated and interpreted at a descriptive level by considering the average and standard errors. Stratified α, Angoff-Feldt, and Feldt-Raju reliability coefficients formulas were written to Excel; standard error values were calculated separately for each dataset using SPSS. For 25 replication, the obtained Stratified α, Angoff-Feldt, and Feldt-Raju reliability coefficients and standard error values were averaged and tabulated. The table values were evaluated and interpreted at a descriptive level, taking into consideration the averages and standard errors. A mixed ANOVA was run in which the reliability coefficient estimation technique is the within-subject factor. Sample size and ratio are the between-subject factor.

Results

The values of Cronbach's α reliability coefficient calculated according to changing sample size and item rates are given in Table 2.

Table 2

A Reliability Estimates for Different Sample Sizes and Item Type Ratio in 25 Replication

Sample Size	Number of Items	Item Type*	Mean (concerning 25 replications)	Standard Error
500	20	D	.634	.0039
	10	P	.812	.0015
	30	Total	.796	.0019
	15	D	.516	.0049
	15	P	.881	.0010
	30	Total	.835	.0011
	10	D	.338	.0098
	20	P	.908	.0087
	30	Total	.859	.0058
1000	20	D	.743	.0017
	10	P	.817	.0009
	30	Total	.852	.0007
	15	D	.639	.0026
	15	P	.873	.0005
	30	Total	.857	.0007
	10	D	.484	.0031
	20	P	.901	.0004
	30	Total	.873	.0006
2000	20	D	.754	.0010
	10	P	.810	.0008
	30	Total	.858	.0004
	15	D	.639	.0015
	15	P	.874	.0004
	30	Total	.868	.0003
	10	D	.516	.0027
	20	P	.908	.0003
	30	Total	.892	.0003

**D: Dichotomous, P: Polytomous*

According to Table 2, when the sample size is considered, as the sample size increases, the Cronbach's α coefficient increases or tends to increase in all item ratios. When the item rates are taken into account, it is observed that the number of items that scored dichotomously in all the sample sizes, and accordingly, the Cronbach's α values decrease as the ratio of the dichotomous items increase. Table 3 shows the comparison of the reliability coefficient values calculated on the basis of the changing sample size and item rates.

Table 3

Reliability Estimates for Different Sample Sizes and Item Type Ratio in 25Replication

		Reliability Coefficient- Standard Errors								
Sample Size	**Item Ratio**	**True Reliability**	**Cronbach's α**	**SE**	**Stratified α**	**SE**	**Angof-Feldt**	**SE**	**Feld-Raju**	**SE**
	2:1	.802	.796	.0019	.797	.0018	.852	.0027	.849	.0026
500	1:1	.839	.835	.0011	.835	.0014	.810	.0023	.808	.0023
	1:2	.876	.859	.0058	.867	.0010	.751	.0041	.748	.0041
	2:1	.858	.852	.0007	.852	.0007	.892	.0009	.891	.0008
1000	1:1	.862	.857	.0007	.857	.0007	.857	.0017	.855	.0018
	1:2	.877	.873	.0006	.873	.0006	.818	.0019	.817	.0019
	2:1	.860	.858	.0004	.857	.0005	.892	.0017	.892	.0016
2000	1:1	.870	.868	.0003	.869	.0003	.861	.0011	.860	.0011
	1:2	.894	.892	.0003	.892	.0003	.840	.0013	.839	.0013

*Item Ratio= Dichotomous/ Polytomous (Total Item Number= 30)

When Table 3 is examined, considering the sample size, it is observed that as the sample size increases, the Stratified α values increase or tend to increase at all item ratios. When the item rates are considered, it is observed that as the number of items dichotomously scored in all samples increases, the value of Stratified α decreases. This situation could also be due to increased scale sensitivity provided by the categorical item scoring since polytomous item scoring can, ideally, be more precise, and consequently, a higher reliability coefficient estimates that of dichotomous item scoring.

As shown in Table 3, it is generally observed that as the sample size increases, the Angoff-Feldt reliability coefficient values increase or tend to increase. For example, in Table 3, for a 1:2 item ratio, if the sample size is 500, then r_{AF}=.751, if sample size is 1000, then r_{AF}=.818, and if the sample size is 2000, then r_{AF}=.840. When the item rates are considered, it is observed that the Angoff-Feldt reliability coefficient values increase or tend to increase as the number of items dichotomously scored in all samples increases. Accordingly, higher Angoff-Feldt estimates are obtained if the number and ratio of the polytomously scored items in a mixed-format test are higher. According to Table 3, when the sample size is considered, the Feldt-Raju reliability coefficient values increase at all item ratios as the sample size increases. When the item rates are taken into consideration, it is observed that the Feldt-Raju values increase as the number of items dichotomously scored in all samples increases. In addition, when

Table 3 is examined, it is seen that Angoff-Feldt and Feldt-Raju values have almost the same values.

When Table 3 is examined, and the sample size is considered, it is observed that as the sample size increases, the reliability coefficient values obtained with four reliability coefficients in all item ratios are also increased. That is, in large samples, all four reliability coefficients tend to give higher estimates. In addition, as the sample size increases, the standard errors for the averages of the estimates of the four reliability coefficients approach zero by decreasing. This is an expected result since the standard error is inversely proportional to the sample size.

When item ratios are taken into consideration, it is observed that as the number of items polytomously scored in all samples increases, Cronbach's α and Stratified α values increase, while Angoff-Feldt and Feldt-Raju values decrease. In other words, Cronbach's α and Stratified α values give higher predictions for the tests with a high polytomous item ratio, while Angoff-Feldt and Feldt-Raju give higher estimates for the tests with high dichotomous item ratio. From a different view point, it can be conceivable that in the case of number and rate of dichotomous items were higher than polytomous items,α and Stratified α give lower limit in these four reliability coefficients. And also, it can be conceivable that α and Stratified α give upper limit in these four reliability coefficients when the number and rate of dichotomous items are lower than polytomous items. It is also possible to express this finding by taking into account the Angoff-Feldt and Feldt-Raju coefficients. Thus, if the rate of dichotomous items is higher, it can be conceivable that Angoff-Feldt and Feldt-Raju give the upper limit in these four reliability coefficients. Also, when the standard error values are examined, it is seen that α and Stratified α have values lower than Angoff-Feldt and Feldt-Raju in almost all conditions. Besides, when Table 3 is examined, it is seen that α and Stratified α are closer to true reliability coefficients. Angoff-Feldt and Feldt-Raju give upper limits among these four coefficients when the ratio of dichotomous items is high, but the standard error values appear to give high values in these four coefficients. However, when Table 3 is examined, the values calculated with Angoff-Feldt and Feldt-Raju have given higher values than the true reliability value. A two way ANOVA was conducted to investigate the impacts of sample size and item ratio on the reliability coefficient. There was a significant main effect of sample size and item ratio, ($F [2, 221] = 136{,}924$ $p < .001$, $\eta^2 = .55$). There is a great effect ($\eta^2 > .14$). Paired comparisons were made after the analysis with the Tukey method for the significant F values and the results were found to be significant.

These increases and decreases in the reliability coefficient values depending on the item rate are lower in the sampling for 1000 and 2000 individuals than the sampling of 500 individuals. As the sample size increases for all the four reliability coefficients, the effect of the item ratio on the reliability estimate appears to decrease or tend to decrease. It is also seen that the difference between the standard error values for the average of the estimates for the four reliability coefficients and the increase in the rate of the polytomous items is more evident for the samples with 500 samples than for the samples with 1000 and 2000 individuals. When Table 3 was examined, it was seen that all reliability coefficients were close to each other due to the increase in sample size in

the 1: 1 item-type ratio. It is also seen that these values are close to the real reliability coefficient values. Therefore, as the sample size increases, it can be considered that these coefficients can be used interchangeably in 1:1 item type ratio.

As another result, Cronbach's α and Stratified α coefficient values are observed to be similar in all sample sizes and item ratios. Similarly, the Angoff-Feldt and Feldt-Raju reliability coefficients are close to each other or give the same estimates. In other words, Cronbach's α and Stratified α, Angoff-Feldt, and Feldt-Raju coefficients can be considered to work similarly. Also, when the standard error values are examined, it is seen that Cronbach's α with Stratified α and Angoff-Feldt with Feldt-Raju have similar values. While in all sample size and item type ratio Angoff-Feldt and Feldt-Raju give similar standard error values; Cronbach's α and Stratified α give more similar standard error values in large samples. Accordingly, it can be said that Angoff-Feldt and Feldt-Raju tend to work more similarly. This also indicates that the similar study trend of Angoff-Feldt and Feldt-Raju was not affected by sample size; however, the similar operating tendency of α with Stratified α was affected by sample size, indicating that this trend increased in large samples.

When Table 3 is examined, it can be seen that as the sample size increases, the difference between the highest and the lowest value obtained from the four reliability coefficients in all item ratios decreases. The decrease in the difference between the reliability values depending on the sample size can be interpreted as the difference will decrease gradually when larger samples are used. Besides, it is seen that the difference between the standard error values decreases as the sample size increases for the four reliability coefficients. This situation has been evaluated that differences from true reliability values will decrease depending on the sample size increase in the 2:1 and 1:1 item type ratio.

Discussion, Conclusion and Recommendations

As expected, in this study, it is seen that there is a relationship between sample size and reliability estimates. Similar results have been obtained with Gay (1987). Cronbach's α, Angoff-Feldt, and Feldt-Raju reliability coefficients tend to give high estimates as sample size increases. Also, as the sample size increases, the standard error decreases. This indicates that the possibility of approaching the true reliability value increases. Lord and Novick (1968) also found that the reliability coefficient and error variance values approach real values with the increase in the sample size.

Another conclusion of this study is that there is a relationship between the ratios of different item types in mixed-format tests and reliability estimates. A mixed ANOVA results showed that this effect was significant. Nunnally (1964) and Mehren and Lehmann (1973) also found that the polytomous and dichotomous item rates affect the reliability. Saen-AmnuAiphon, Tuksino, and Nichanong (2012) indicated that increasing the number of items that are dichotomously scored reduced the value of reliability Cronbach's α. As a result of this study, Cronbach's α and Stratified α values increased with the increase in the number of items scored polytomously in all samples; Angoff-Feldt and Feldt-Raju values increased or tended to increase as the number of items scored dichotomously increased. In other words, in a mixed-format test with

dichotomous items, α and Stratified α values have the upper limit in these four coefficients when the number and ratio of the items scored polytomously are high; if there are a large number of items scored dichotomously, it gives the lower limit. This is also true for the Angoff-Feldt and Feldt-Raju coefficients.

Another conclusion reached in this study is that the changes in the observed reliability estimates are affected by the sample size, depending on the number and ratio of the items. A mixed ANOVA results showed that this effect was significant. The differences between the coefficients are more evident in the 500-person samples than in the 1000- and 2000-person samples. That is, as the sample size increases, the effect of the item ratio on the reliability estimates seems to decrease or tend to decrease. Charter (1999) also found that as the sample size increased, the difference between the different reliability coefficients decreased.

In this study, it was seen that some of the reliability coefficients tend to run similarly and their estimates are also similar to each other. In all sample sizes and item ratios, Cronbach's α and Stratified α run similarly. This is also true for the relationship between Angoff-Feldt and Feldt-Raju coefficients. The conclusion that Angoff-Feldt and Feldt-Raju coefficients tend to run in a similar way is consistent with the findings of Warrens (2016). Warrens (2016) found that different coefficients may tend to work similarly, and also, if the larger standard deviation is less than 30% and if the difference between the lengths is at most 0.20, than the differences between the values is less than 0.07. Osburn (2000) and Feldt and Charter (2003) were also showed that the different coefficient produces very similar values in a variety of situations using simulated data.

There are studies on how Cronbach's α and Stratified α tend to give the same results (Cronbach, Schöneman & McKie, 1965; Cronbach & Shavelson, 2004; Osburn, 2000). These studies show that the use of Stratified α is more appropriate in a mixed-format test involving different item types. Because the coefficient α estimate reliability is lower than as it is. Similarly, there are other studies that show that the Cronbach's α coefficient generally gives lower estimates than the other reliability coefficients (Feldt, 2002; Guttman, 1945; Raykov & Shrout, 2002; Sijtsma, 2009). As a result, the related literature shows that Cronbach's α tends to estimate lower than Stratified α.

Like the above-mentioned sources, in this study, a similar result with the literature was observed in the sample of 500 individuals. For samples larger than 500, Cronbach's α and Stratified α coefficient values either give the same value or differentiate after the three digits. It has been assessed that this may be related to sample size. Accordingly, it is predicted that when the sample size reaches a certain value, Cronbach's α and Stratified α values will be fixed by giving the same value.

According to the results of this study, as the sample size increases, the difference between the highest value and the lowest value obtained from four reliability values in all item ratios decreases. Depending on the sample size, the gradual decrease in the differences between the reliability values indicates that the difference will gradually decrease when larger samples are used and that four reliability values will give similar results after a certain sample size. This is also supported by the tendency to decrease

in standard error values and its approach to zero. This can be interpreted as the reliability that can be predicted more accurately with the increase in the sample size.

In the light of the findings and conclusions obtained in this study, to obtain more reliable results, it may be conceivable to use Cronbach's α and Stratified α, if the number of items polytomously scored is higher. However, when the literature is examined, it is seen that there are some studies indicating that choosing an internal consistency estimation technique should not decide which procedure provides the highest coefficient (Qualls, 1995). On the other hand, the coefficients Angoff-Feldt and Feldt-Raju can be used when the number of items scored dichotomously is higher. In smaller samples (500 individuals or less), it can be considered to increase the number of items scored polytomously. The Cronbach's α, Stratified α, Angoff-Feldt, and Feldt-Raju reliability coefficients can be used in larger samples (2000 individuals and over) since they give the same values for calculating reliability values for a mixed-format test with a 1:1 item ratio. And also, it was seen that all coefficient values were near to the true reliability coefficient.

When the Cronbach's alpha coefficient is low from the other reliability coefficients, the Angoff-Feldt coefficient gives or tends to give a high value from the other reliability coefficients. Feldt and Charter (2003) suggested that when the use of the Cronbach's α coefficient is not appropriate, the use of the Angoff-Feldt coefficient is a higher and more accurate estimation. It can be understood in Feldt and Brennan (1989) that the Angoff-Feldt coefficient would be more accurate than Cronbach's α coefficient in a mixed-format test, which consists of different types of items. However, when findings evaluated with standard error values, standard error values are higher in cases where Angoff-Feldt values higher than Cronbach's α. However, when the true reliability coefficient was examined, it was seen that coeffient alpha is near the true reliability coefficient.

References

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153-169.

Bastari, B. (2000). *Linking MCQ and CR Itemsto a common proficiency scale (*Unpublished doctoral dissertation). University of Massachusetts Amherst, USA.

Berger, M. P. (1998). Optimal design of tests with dichotomous and polytomous items. *Applied Psychological Measurement*, *22*(3), 248-258.

Cao, Y. (2008). *Mixed-format test equating: Effects of test dimensionality and common-item sets*(Doctoral dissertation). Retrived from https://drum.lib.umd.edu/handle/1903/8843

Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, *21*(4), 559-566.

Crocker, L., &Algina J. (2008). *Introductiont a classical and modern test theory*. N.Y.: Nelson Education.

Cronbach, L. J., Schönemann, P., &McKie, D. (1965). α coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, *25*(2), 291-312.

Cronbach, L. J., &Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement*, *64*(3), 391-418.

DeVellis, R. F. (2003). *Scale development: Theory and application*. Sage Publications: California.

Donoghue, J. R. (1993). An empirical examination of the IRT information in polytomously scored reading items. *ETS Research Report Series*, *1993*(1).

Ercikan, K., Schwarz, R., Julian, M.W., Burket, G.R., Weber, M.W., &Link, V. (1998). Calibration and scoring of tests with multiplie-choice and constructed response test item type. *Journal of Educational Measurement*, *35*(2), 137-154.

Eren, B. (2015). *The comparison of student achievements, students' and teachers' views for multiple choice and mixed format test applications (*Unpublished master's dissertation). Ankara üniversitesi, Ankara.

Falk, C. F., & Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *Journal of personality assessment*, *93*(5), 445-453.

Feldt, L. S., &Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement*(3rd ed.,pp.105-146). New York: Macmillan.

Feldt, L. S. (2002). Estimating the internal consistency reliability of tests composed of test lets varying in length. *Applied Measurement in Education*, *15*(1), 33-48.

Feldt, L. S., & Charter, R. A. (2003). Estimation of internal consistency reliability when test parts vary in effective length. *Measurement and Evaluation in Counseling and Development*, *36*(1), 23-27.

Gao, F., & Chen, L. (2005). Bayesia nor non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, *18*(4), 351-380.

Gay, L. R. (1996). *Educational research: competencies for analysis and application* (5th ed). By Prentice-HallInc.: USA.

Gubes, N. Ö. (2014). *The effects of test dimensionality, common item format, ability distribution and scale transformation methods on mixed - format test equating*(Doctoral dissertation). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/. (Accession Number: 399465)

Gul, E. (2015). *Examining multidimensional structure in view of unidimensional and multidimensional item response theory* (Doctoral dissertation). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/. (Accession Number: 419288)

Gultekin, S. (2011). *The evaluation based on Item Response Theory of the psychometric characteristics in multiple choice, constructed response and mixed format tests* (Doctoral dissertation). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/. (Accession Number: 302033)

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255-282.

Hambleton, R. K., Swaminathan, H., Rogers, H. (1991), *Fundamentals of Item Response Theory*. Newbury Park CA: Sage Publications.

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.

He, Q. (2009). Estimating the reliability of composite scores. Retrieved from https://pdfs.semanticscholar.org/0f54/d8c356f82fbca4fd2326239c1d21fbc9b778.pdf

He, Y. (2011). *Evaluating equating properties formixed-format tests (*Unpublished doctoral dissertation). University of Iowa, Iowa City.

Hu, B. (2018). Equating Errors and Scale Drift in Linked-Chain IRT Equating with Mixed-Format Tests. *Journal of applied measurement*, *19*(1), 41-58.

Kim, S. H., & Lee, W.-C. (2006). An extension of four IRT linking methods for mixedformat tests. *Journal of Educational Measurement*, 43, 53-76.

Kim, S. Y., & Lee, W. C. (2018). Simple-Structure MIRT True-Score Equating for Mixed-Format Tests. *Mixed-Format Tests: Psychometric Properties with a Primary Focus on Equating (Volume 5)*, 127.

Kim, S. Y., & Lee, W. C. (2019). Classification consistency and accuracy for mixed-format tests. *Applied Measurement in Education*, *32*(2), 97-115.

Kinsey, T. L. (2003). *A comparison of IRT and rasch procedures in a mixed-item format test* (Doctoral dissertation).Retrieved from ProQuest Digital Dissertations.

Kirkpatrick, R. K. (2005). *The effects of item format in common item equating(*Unpublished doctoral dissertation). University of Iowa, Iowa City.

Lee, G., & Lee, W. C. (2016). Bi-factor MIRT observed-score equating for mixed-format tests. *Applied Measurement in Education*, *29*(3), 224-241.

Li, Z., Chen, H., & Li, T. (2018). Exploring the Accuracy of MIRT Scale Linking Procedures for Mixed-format Tests. *arXiv preprint arXiv:1805.00189*.

Lord, F. M. (1980). *Applications of item response theory topractical testing problems*. London: Routledge.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading MA: Addison-Welsley Publishing Company.

Lucke, J. F. (2005a). The α and ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurements*. 29(1), 65-81.

Masters, G. N. (1982). A Rasch model forpartial credit scoring. *Psychometrika*, *47*(2), 149-174.

Mehren, W.A. & Lehmann I.J. (1973). *A measurement and evaluation in educationand psychology*. New York: Holt. Rinehartand Winston.

Nunnally, J.C. (1964). *Educational measurement and evaluation* (6th ed.). New York: McGraw- Hill Book Company.

Odabas, M. (2016). *The comparison of DINA model signed difference index, standardization and logistic regression techniques for detecting differential item functioning (*Unpublished doctoral dissertation). Hacettepe Üniversitesi, Ankara.

Osbourn, H.G. (2000). Coefficient α and related internal consistency reliability coefficients. *Psychological Methods*, 5, 343-355.

Qualls, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education*, *8*(2), 111-120.

Raykov, T., &Shrout, P. E. (2002). Reliability of scaleswith general structure: Point andinterval estimation using a structural equation modeling approach. *Structural Equation Modeling*, *9*(2), 195-212.

Saen-amnuaiphon, R., Tuksino, P.,& Nichanong, C. (2012). The Effect of Proportion of Mixed-Format Scoring: Mixed-Format Achievement Tests. *Procedia-Social and Behavioral Sciences, 69*, 1522-1528.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's α. *Psychometrika*, *74*(1), 107.

Spray, J. A. (1990). Comparison of Two Logistic Multidimensional Item Response Theory Models. (Research Report ONR90-8). ACT, Inc., Iowa City, IA.

Sykes, R. C., Truskosky, D.,&White, H. (11-12 April 2001), *Determining The Representation of Constructed Response Items in Mixed-Item-Format Exams.* Paperpresented at Annual Meeting of the National Council on Measurement in Education, ABD: Seattle.

Tekin, H. (1991). *Measurement and evaluation in education.* Ankara: Yargı yayınevi.

Uysal, İ., & Kilmen, S. (2016). Comparison of Item Response Theory Test Equating Methods for Mixed Format Tests. *International Online Journal of Educational Sciences*, *8*(2), 1-11.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make nonever mind. *Psychological Bulletin*, *83*(2), 213.

Wang, W., Drasgow, F., & Liu, L. (2016). Classification accuracy of mixed format tests: A bi-factor item response theory approach. *Frontiers in psychology*, *7*, 270.

Warrens, M. J. (2016). A comparison of reliability coefficients for psychometric tests that consist of two parts. *Advances in Data Analysis and Classification*, *10*(1), 71-84.

Young, M. J., &Yoon, B. (1998). *Estimating the consistency and accuracy of classification in a standards-referenced assessment*.Retrieved from https://cresst.org/wp-content/uploads/TECH475.pdf.

Zinbarg, R. E., Revelle, W., Yovel, I. & Li, W. (2005). Cronbach's α, Revelle's, β and McDonalds ω: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70(1), 1-11.

Karma Testlerde İç Tutarlılık Kestirimlerinin Farklı Benzetim Koşullarında İncelenmesi

Atıf:

Gurdil Ege, H., & Demir, E. (2020). Examining of internal consistency coefficients in mixed-format tests in different simulation conditions. *Eurasian Journal of Educational Research, 87,* 101-118. DOI: 10.14689/ejer.2020.87.5

Özet

Problem Durumu: Alanyazın incelendiğinde karma testlerde güvenirlik üzerine yapılan az sayıda araştırma olduğu görülmüştür. Bu araştırmalarda farklı madde tipleri farklı oranlarda kullanılmıştır. Ancak madde tipi oranlarının ve örneklem büyüklüğünün birlikte güvenirlik üzerindeki etkisini inceleyerek bu yöntemlerin karşılaştırıldığı bir araştırmaya rastlanmamıştır. Karma testlerde kullanılacak madde tipleri ve bunların sayısı, ayrıca güvenirlik kestirimleri için gerekli örneklem büyüklüğü, önemli tartışma ve sorun alanları arasındadır. Bu bağlamda; karma testlerde iç tutarlılık anlamında güvenirlik hesaplamalarında kullanılan katsayıların, karma testlerin yapısını belirleyen önemli değişkenler olarak farklı madde tipi oranı ve örneklem büyüklüğü dikkate alındığında, bu katsayıların nasıl değiştiği, ideal/ daha uygun katsayıların hangileri olduğu bu araştırmanın problem durumunu oluşturmaktadır.

Araştımanın Amacı: Bu araştırmanın araştırmada, karma testlerde örneklem büyüklüğü (500, 1000 ve 2000) ve kullanılan madde tiplerinin oranı (2:1; 1:1 ve 1:2) değişimlendiğinde; α, Tabakalı α, Angoff-Feldt ve Feldt-Raju güvenirlik katsayılarının nasıl değiştiğinin incelenmesi ve bu güvenirlik katsayıları arasındaki betimsel ilişkinin ortaya konması amaçlanmıştır.

Araştırmanın Yöntemi: Araştırma için belirlenen koşullara uygun veri üretmek için; WinGen program kullanmıştır. Araştırma kapsamında oluşturulan koşullarda madde tipi sayısı, very türetmede kullanılan model, toplam madde sayısı, yanıt kategori sayısı, toplam puan alma yöntemi, madde ayırıcılığı ve madde güçlüğü sabit tutulurken; örneklem büyüklüğü ve madde tipi oranı için değişimlemeler yapılmıştır.

Sabit tutulan ve üzerinde değişimleme yapılan değişkenler için ilgili alanyazın dikkate alınmıştır. Theta, her bir örneklem büyüklüğü için; ortalaması 0.00 ve standart sapmaları 1.00 olan normal dağılıma uygun olacak şekilde üretilmiştir. İki kategorili puanlanan maddeler İki Parametreli Lojistik Model'le, çok kategorili puanlanan maddeler Kısmi Puan Modeli ile üretilmiştir. Örneklem sayısı (500, 1000 ve 2000) ve madde oranları (2:1, 1:1 ve 1:2) olacak şekilde değişimlenmiş ve ilk beş adım bu koşulların her biri için tekrarlanmıştır. Veri üretiminde 25 tekrar(replikasyon) yapılmıştır. Böylelikle, 3x3=9 farklı deneysel koşul için 9x25=225 farklı veri seti üretilmiştir. Elde edilen very setlerine ait her bir koşul ve tekrar için α, Tabakalı α, Angoff-Feldt ve Feldt-Raju değerleri hesaplanmış ve tablolaştırılmıştır. Bu tablo değerleri, ortalama ve standart hatalar dikkate alınarak betimsel düzeyde değerlendirilmiş ve yorumlanmıştır.

Araştırmanın Bulguları: Tüm madde oranlarında örneklem büyüklüğü arttıkça tüm güvenirlik kestirim değerlerinin de arttığı belirlenmiştir. Tüm örneklemlerde iki kategorili puanlanan madde sayısı arttıkça α ve Tabakalı α güvenirlik katsayı değerleri; çok kategorili puanlanan madde sayısı arttıkça ise Angoff-Feldt ve Feldt-Raju güvenirlik katsayı değerleri azalmaktadır. α ve Tabakalı α güvenirlik katsayı değerleri tüm örneklem büyüklükleri ve tüm madde oranlarında hemen hemen aynı değeri vermekteyken, 500 kişilik örneklemde madde oranları değişimine gore güvenirlik katsayı değerleri arasındaki farkın daha belirgin olduğu görülmüştür. Diğer bir bulgu olarak, iki madde tipi içeren bir karma testte iki kategorili puanlanan madde sayısı çok kategorili puanlanan madde sayısından daha fazla olduğunda α ve Tabakalı α güvenirlik katsayı değerlerinin bu dört güvenirlik katsayısı içersinde alt sınırı, aksi durumda üst sınırı verdiği görülmüştür.

Araştırmanın Sonuçları ve Önerileri: Örneklem büyüklüğü 500'ün üzerinde olduğunda α ve Tabakalı α güvenirlik katsayı değerlerinin benzer sonuçlar verdiği, diğer katsayıların kısmen farklılaştığı görülmüştür. Örneklem büyüklüğü 1000'in üzerinde olduğunda ise α, Tabakalı α, Angoff-Feldt ve Feldt-Raju değerleri arasındaki farkın azaldığı görülmüştür.Görece küçük örneklemlerde (n=500) kısmi farklılıklar görülmekle birlikte büyük örneklemlerde (n≥1000), farklı güvenirlik katsayılarının benzer değerler verdiği, örneklem büyüklüğü arttıkça madde oranının güvenirlik kestirimleri üzerindeki etkisinin de düştüğü ya da düşme eğiliminde olduğu sonucuna ulaşılmıştır. Bu araştırmada elde edilen bulgular ve sonuçlar doğrultusunda bu alanda uygulama yapacaklara daha güvenilir sonuçlar elde edebilmek için çok kategorili puanlanan madde sayısı daha fazlaysa α ve Tabakalı α; iki kategorili puanlanan madde sayısı fazlaysa Angoff-Feldt ve Feldt-Raju güvenirlik katsayıları kullanmaları önerilebilir. Küçük örneklemlerde (n≤500) yapılacak test uygulamalarında α ve Tabakalı α güvenirlik katsayısı kullanılacaksa, testin güvenirlik düzeyini artırmak için çok kategorili puanlanan maddelerin sayısı arttırılabilir. Büyük örneklemlerde (n≥1000) ise (2:1, 1:1, 1:2) madde oranlarından elde edilen güvenirlik değerleri birbirine çok yakın oldukları için bu madde oranlarından herhangi biri kullanılabilir. Büyük örneklemlerde (n≥1000), özellikle 1:1 madde oranından oluşan bir karma test için güvenirlik değeri hesaplanmasında aynı değerleri verdiğinden α, Tabakalı α, Angoff-Feldt ve Feldt-Raju güvenirlik katsayıları kullanılabilir. 500 kişiden

büyük örneklemlerde birbirleriyle hemen hemen aynı değerleri verdikleri için α yerine Tabakalı α; Angoff-Feldt yerine Feldt-Raju güvenirlik katsayısı kullanılabilir.

Anahtar Sözcükler: Karma test, güvenirlik, tabakalı α, angof-feldt, feldt-raju