ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2021; 12(2); 147-162

Using Generalizability Theory to Investigate the Reliability of Scores Assigned to Students in English Language Examination in Nigeria *

Olufunke Favour AKINDAHUNSI ** Eyitayo Rufus Ifedayo AFOLABI ***

Abstract
The study investigated the reliability of scores assigned to students in English language in National Examinations Council (NECO). The population consisted of all the students who sat for NECO Senior School Certificate Examination (SSCE) in 2017 in Nigeria. A sample of 311,138 was selected using the proportionate stratified sampling technique. The Optical Marks Record (OMR) sheet containing the responses of the examinees was the instrument for the study. The data was analyzed using lme4 package of R language and environment for statistical computing, factor analysis and Tucker index of factor congruence. The psychometric properties of the data were determined by estimating the generalizability (g) coefficient, phi (Φ) coefficient and construct validity. The results indicated the g-coefficient to be 0.90 and Φ coefficient as 0.87, which is an indication of high reliability of scores. The result also showed that a decrease in the number of the items resulted in a decrease in both g- and phi coefficients in D-study. The construct validity of 0.99 obtained from the result affirms the credibility of the items. Hence, it was concluded that the scores were dependable and generalizable.

Key Words: Reliability, validity, English language, score, Generalizability theory.

INTRODUCTION

Generalizability theory is a statistical method used to analyze the results of psychometric tests, such as performance tests like the objective structured clinical examination, written or computer-based knowledge tests, rating scales, or self-assessment and personality tests (Breithaupt, 2011). It involves separating various sources of error and recognizing that multiple sources of error such as error attributed to items, occasions, and forms may occur simultaneously in a single measurement process, thereby forming the basic approach underlying generalizability theory (g-theory) which is to decompose an observed score into a component for the universe score and one or more error components. Its main purpose is to generalize from an observation at hand to the appropriate universe of observations. It is also advantageous in that it can estimate the reliability of the mean rating for each examinee while simultaneously accounting for both interrater and intra-rater inconsistencies as well as discrepancies due to various possible interactions, which are impossible in Classical Test Theory (CTT) (Brennan, 2001). In generalizability theory, various sources of error contributing to the inaccuracy of measurement are explored. It is a valuable tool in judging the methodological quality of an assessment method and improving its precision. It gives the opportunity of disentangling the error components of measurement and is also interested in the reliability or dependability of behavioral measurement, that is, the certainty that the score is reliable to generalize.

All test scores, just like any other measurement, contain some errors. It is this error that affects the reliability or consistency of test scores. When there are variations in the measurement under the same conditions, then error comes in. Error in measurement can be defined as the difference between a person's observed score and his/her true score. Error is not a mistake in statistics; it is bound to occur.

* The paper was a part of thesis and the whole study is not published anywhere.
** PhD. Student, Obafemi Awolowo University, Department of Educational Foundations and Counselling, Ile – Ife-Nigeria, olufavour@ymail.com, ORCID ID: 0000-0002-5041-7088
*** Prof., Obafemi Awolowo University, Department of Educational Foundations and Counselling, Ile – Ife-Nigeria, eriafolabi@gmail.com, ORCID ID:0000-0002-0014-0711

To cite this article:
Akindahunsi, O. F., & Afolabi, E. R. I. (2021). Using generalizability theory to investigate the reliability of scores assigned to students in English language examination in Nigeria. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 147-162. doi: 10.21031/epod.820989

Received: 04.11.2020
Accepted: 30.05.2021

Breithaupt (2011) identified two types of measurement errors in the examination of items and test scores: random error and systematic error. The author expressed that random error is a source of bias in scores and an issue of validity while systematic error is a measurement error that can be estimated in reliability studies. Its estimates permit the test developer to determine the possible size and sources of construct irrelevant variation in test scores. Thus, it is assumed that the skill, trait, or ability measured is a relatively stable defined quantity during testing. Therefore, variation in obtained scores is usually attributed to sources of error and thus poses the challenge of determining the psychometric property of a test. The goal of the psychometric analysis is to estimate and minimize, if possible, the error variance so that the observed score (X) is a good measure of the true score (T). Understanding whether the test error is due to high variance is important in measurement. It is generally assumed that the exact or true value exists based on how what is being measured is defined. Though the true value exactly may not be known, attempts can be made to know the ideal value. In CTT any observed score is seen as the combination of a true component and a random error component, even though the error could be from various sources. However, only a single source of measurement error can be examined at any given time. CTT treats error as random and cannot be used to differentiate the systematic error from random error. Generalizability theory also focuses on the universe score, or the average score that would be expected across all possible variations in the measurement procedure (e.g., different raters, forms, or items). This universe score is believed to represent the value of a particular attribute for the object of measurement (Crocker & Algina, 2008). The universe is defined by all possible conditions of the facets of the study. It also gives the opportunity to judge whether the score differences observed between the subject could be generalized to all items and occasions (de Gruijter & van der Kamp, 2008). This means that g-theory helps to know whether the means observed over a sample of items and a sample of occasions could be generalized to the theoretical universe of items and occasions. Since g-theory focuses on the simultaneous influence of multiple sources of measurement error variance, it more closely fits the interest of researchers.

The reliability coefficients under CTT are usually focused on the consistency of the test results. For instance, test-retest reliability considers only the time/occasions of testing, parallel-forms reliability considers only the forms of the test and internal consistency considers the items as the only source of error. Some authors (Mushquash and O'Connor, 2006; Webb, Shavelson, & Haertel, 2006) noted that the effects of various sources of variance can be tested using CTT models within which it is only possible to examine a single source of measurement error at a given time, but that it is impossible to examine the interaction effects that occur among these different sources of error. Generalizability theory is particularly useful in this regard; each feat of the measurement situation is a source of error in test scores and its termed facet. Therefore, the inadequacy of explanation of numerous sources of error as pointed out by several authors (Brennan, 2001; Johnson & Johnson, 2009) and the researchers' dissatisfaction with CTT's inability to identify possible sources of error and simultaneously examining them led to the development of g-theory which was an extension of CTT. It offers a broader framework than the CTT for estimating reliability and errors of measurement. Generalizability theory involves two types of study: generalizability study (G-study) and Decision study (D-study). The main purpose of a G-study is to estimate components of score variance that are associated with various sources, while a D-study takes these estimated variance components to evaluate and optimize among alternatives for subsequent measurement. Two types of decision and error variance, relative and absolute, are made in G-study, but only relative decisions are made in CTT (Brennan, 2001; Yin & Shavelson, 2008).

Alkharusi (2012) explained that an observed score for any student obtained through some measurement procedure could be decomposed into the true score and a single error. Since the performances of students in National Examinations Council (NECO) Senior School Certificate Examination (SSCE) are based on the sum of their total scores, that is, CTT, there is a need to consider the psychometric properties (difficulty, discrimination, reliability, validity) of the test in taking decisions on the observable performance of candidates in order to improve upon test construction, administration and analysis. Reliability and validity are two technical properties that indicate the quality and usefulness of tests as well as major factors to be considered in the construction of test items for examinations. Junker (2012) described reliability as the extent to which the test would produce

consistent results if it is administered again under the same conditions. It also reflects how dependably a test measures a specific characteristic. This consistency is of three types: over time (test-retest reliability), across items (internal consistency), and across different researchers (inter-rater reliability). Many reasons can be adduced for an individual not getting exactly the same test score every time he or she takes the test. These include the test taker's temporary psychological or physical state, multiple raters and test forms. These factors are sources of chance or random measurement error in the assessment process. If there are no random errors of measurement, the individual will get the same test score, that is, the individual's true score each time. The degree to which test scores are unaffected by measurement errors is an indication of the reliability of the test.

Reliability is threatened when errors occur in measurement. When a measure is consistent over time and across items, one can conclude that the scores represent what they intend to; meanwhile, there is more to it because a measure can be reliable but not valid. Reliability and validity are therefore needed to assure adequate measurement of the construct of interest. Validity refers to what characteristic the test measures and how well the test measures that characteristic. In other words, it determines the extent to which a measure adequately represents the underlying construct that it is supposed to measure. Valid conclusions cannot be drawn from a test score unless one is sure that the test is reliable. Even when a test is reliable, it may not be valid. Therefore, care should be taken to ensure that any test selected is both reliable and valid for the situation. The accuracy and validity of the interpretation of test results are determined by the inferences made from test scores. Validity of inferences is concerned with the negative consequences of test score interpretation that is traceable to construct under-representation or construct-irrelevance variance. The focus should be on the theoretical dimensions of the construct a test is intending to measure in order to prevent inappropriate consequences from test score interpretation. Generally, in testing, it is necessary to consider how test-takers' abilities can be inferred based on their test scores. Student marks are affected by various types of errors of measurement which always exist in them, and these reduce the accuracy of measurement. The magnitude of measurement error is incorporated in the concept of reliability of test scores, where reliability itself quantifies the consistency of scores over replications of a measurement procedure. Also, it is often expected that test score variation should only be due to an artifact of test-takers' differing abilities and task demands. But in reality, it is being proven that test-takers' scores are most of the time affected by other factors, including test procedures, personal attributes other than abilities, and other random factors. A single score obtained on one occasion on a particular form of a test with a single administration as done by NECO is not fully dependable because it is unlikely to match that person's average score over all acceptable occasions, test forms, and administrations. A person's score would usually be different on other occasions, on other test forms, or with different administrators. Which are the most serious sources of inconsistency or error? Where feasible, it is expected that error variances that arise from each identified source be estimated. Regardless of the strengths of g-theory, it has not been widely applied specifically to estimate the dependability of scores of students in secondary school examinations in Nigeria.

In Nigeria, at the end of secondary school education, students are expected to write certification examinations such as the SSCE conducted by the West Africa Examination Council (WAEC) and the NECO, or the National Business and Technical Certificate Education (NBTCE) conducted by the National Business and Technical Examination Board (NABTEB). The NECO conducts the SSCE in June/July and November/December every year. It was established in 1999 to reduce the workload of WAEC, especially to mitigate the burden of testing a large number of candidates. It was also to democratize external examination by providing candidates with a credible alternative. While some Nigerians saw NECO's arrival as an opportunity for choice of examination body for candidates to patronize, others doubted its capacity to conduct reliable examinations that could command widespread national and international respect and acceptability.

English language education is a colonial legacy that has deeply entrenched in Nigerian heritage and apparently become indispensable. It is widely recognized as an instrument par excellence for socio-cultural and political integration as well as economic development. Its use as a second language as well as the language of education provided a speedy access to modern development in science and

technology (Olusoji, 2012). It is for the above reasons that much importance is attached to English Language education nationwide and at all levels of the nation's educational system. To date, the English language remains the major medium of instruction at all levels of education in Nigeria, and no student can proceed to the tertiary level without a minimum of pass in the English language. In addition, considering the importance of the English language as an international language and its influence on Nigerian secondary school students' performance, it is imperative that generalizability theory be used to examine the credibility of secondary school examinations, hence this study.

Purpose of the Study

The objectives of the study are to:

1. Determine the generalizability coefficient of the English Language items;
2. Estimate the phi (dependability) coefficient of the English Language items; and
3. Determine the validity of the English Language items.
4. Conduct a D-study to determine the generalizability and phi coefficients based on the results of G- study.

METHOD

The study adopted the ex post facto research design. This type of design examines the cause and effect through selection and observation of existing variables without any manipulation of existing relations.

Sample

The total population of students who sat for NECO SSCE English Language examination in the year 2017 in Nigeria was 1,037,129, out of which 311,138 candidates constituted the study sample. The sample was selected using a proportionate stratified sampling technique. Thirty percent of the candidates were randomly selected from each state. The detail is presented in Table 1.

Data Collection Techniques

The data used in the study were responses of the candidates (to the 100-item multiple-choice test) who wrote the NECO June/July 2017 English language SSCE in Nigeria as indicated on the Optical Marks Record (OMR) sheets obtained from NECO office.

Instrument

The instrument used for the study was the OMR sheets for the NECO June/July 2017 English language objective items. The OMR sheets contained the responses of examinees to the NECO June/July 2017 English Language objective items paper III. The English Language examination is a dichotomously scored multiple-choice examination consisting of 100 items with five options length. The responses of the examinees were scored 1 and 0 for correct and incorrect responses. The minimum score for an examinee from computation was zero while the maximum score was 100.

Table 1. Population and Sample Size of English Language Candidates Who Sat for NECO Senior School Certificate Examination in 2017

States	Population	Sample size
Abia	10405	3121
Adamawa	37320	11196
Akwa Ibom	23059	6917
Anambra	20509	6152
Bauchi	41413	12424
Bayelsa	4346	1304
Benue	40196	12059
Borno	27439	8232
Cross Rivers	17583	5275
Delta	16647	4994
Ebonyi	10540	3162
Edo	21659	6498
Ekiti	11429	3429
Enugu	26231	7869
FCT	18517	5555
Gombe	25526	7658
Imo	23587	7076
Jigawa	21387	6416
Kaduna	51860	15558
Kano	88227	26468
Katsina	34613	10384
Kebbi	26567	7970
Kogi	28157	8447
Kwara	22079	6624
Lagos	52392	15718
Nasarawa	35950	10785
Niger	33414	10024
Ogun	25212	7564
Ondo	26558	7967
Osun	26126	7838
Oyo	54828	16448
Plateau	34391	10317
Rivers	11484	3445
Sokoto	25379	7614
Taraba	19874	5962
Yobe	17063	5119
Zamfara	25162	7549
Total	**1037129**	**311138**

Data Analysis

The data were analyzed using "lme4" package of R language and environment for statistical computing, factor analysis and Tucker index of factor congruence. The generalizability study was conducted with fitting linear mixed-effect models using lme4 package of R language and environment for statistical computing to find the g-coefficient and phi coefficient. Factor analysis was conducted to identify one dimension underlying the English language test for male and female samples. Thereafter the extracted factor loadings for the test under male and female samples were compared. The comparison of the extracted factor loadings in two samples was made using Tucker index of factor congruence.

RESULTS

One-facet (pxi) design of generalizability theory was adopted to determine the generalizability coefficient. This is because there is a single facet; the items (i) and the persons (p) are the objects of measurement. However, to conduct the analysis under generalizability theory, two levels of analysis were conducted as recommended by Shavelson and Webb (1991). The analysis includes the generalizability (G) study and the decision (D) study. First, the G-study was conducted, and thereafter

the D-study was conducted based on the result of the G-study for the extraction of the generalizability coefficient. The analysis was conducted with fitting linear mixed-effect models using lme4 package (Bates, Mächler, Bolker and Walker, 2015) of R language and environment for statistical computing.

Table 2 presents the estimated variances from the G study. The table shows the magnitude of error in generalizing from a candidate's scores on 2017 NECO English language test to the universe score. A useful exploratory approach for interpreting the variances that are estimated in a G study is to calculate the percentage of the total variance that each variance component represents. These percentages are presented in the last column of Table 2.

Table 2. Parameters of G-Study for 2017 NECO English Language Test

Source	Variance Component	Estimated Variance	Percent of Variability
Person	σ_p^2	0.0142	6.0
Item	σ_i^2	0.0747	31.60
Residual	$\sigma_{pi,e}^2$	0.1472	62.30

The table shows that the variance component for candidates (i.e., the universe score variance) accounts for only 0.0142 or 6.0% of all the variance, and this is rather low. Furthermore, the variance component for the items (0.0747, or 31.6% of the total variance) is large relative to the universe score variance but smaller than the residual variance (0.1472 or 62.3% of the total variance).

Figure 1 presents the histogram that calculates the percentage of items that each candidate got correct. The Figure shows that none of the participants got all the items correct or incorrect and that the overwhelming majority of participants got 60% or 70% of the items correct on the test (i.e., 60 to 70 correct answers). This tight clustering accounted for the observed low universe score variance.

Table 3 shows the proportion of correct items obtained by the candidates for the 100 items 2017 NECO English language test. The table shows that the proportion of item correct ranges from .02 to .91, which reflects a lot of variation and corroborates the high percent of variation accounted for by the items. The large residual variance captures both the person by item interaction and the random error (which we are unable to disentangle). Maybe some items were more easily answered by some participants or maybe there was systematic variation such as the physical environment where the test was administered, or possibly other random variation like fatigue during the assessment. Whatever the cases, these sources could not be disentangled from one another in this variance component.

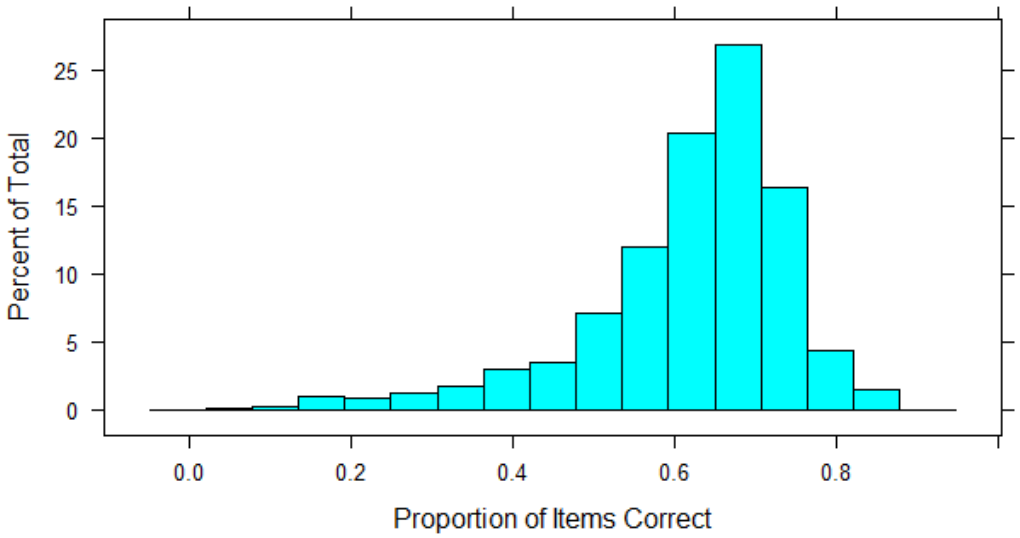

Figure 1. Distribution of Candidates' Proportion of Item Correct

Table 3. Means of 2017 NECO English Language Test Source

Item	Mean	Item	Mean	Item	Mean	Item	Mean
1	.79	26	.21	51	.70	76	.16
2	.32	27	.48	52	.82	77	.81
3	.87	28	.78	53	.89	78	.81
4	.74	29	.84	54	.36	79	.76
5	.79	30	.86	55	.91	80	.34
6	.69	31	.70	56	.81	81	.05
7	.84	32	.83	57	.87	82	.22
8	.81	33	.83	58	.74	83	.28
9	.85	34	.79	59	.29	84	.60
10	.66	35	.61	60	.83	85	.74
11	.33	36	.83	61	.33	86	.79
12	.75	37	.81	62	.83	87	.80
13	.73	38	.84	63	.44	88	.14
14	.86	39	.75	64	.82	89	.13
15	.83	40	.78	65	.84	90	.70
16	.44	41	.27	66	.76	91	.08
17	.88	42	.86	67	.81	92	.72
18	.80	43	.24	68	.40	93	.04
19	.84	44	.83	69	.71	94	.08
20	.71	45	.86	70	.51	95	.09
21	.86	46	.37	71	.36	96	.06
22	.70	47	.84	72	.02	97	.02
23	.74	48	.83	73	.77	98	.11
24	.84	49	.83	74	.85	99	.53

Generalizability Coefficient of 2017 NECO English Language Test

The generalizability coefficient is similar to the reliability coefficient in CTT. It is the ratio of the universe score to the expected observed score variance. For relative decisions and a pxi random-effects design, the generalizability coefficient is calculated as:

$$Ep^2_{X_p} i.u^p = Ep^2 = \frac{E_p(\mu_p - \mu)^2}{E_p E_i(X_{pi} - \mu_i)^2} = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} \quad (1)$$

$$\frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} = \frac{0.0142}{0.0142 + 0.0015} = 0.9046$$

where σ_p^2 is the variation of students' test scores (the universe-score variance), σ_δ^2 is the relative error variance (Desjardins & Bulut, 2018) Table 4 presents the result.

Table 4. Generalizability Coefficient

Source	Estimate
Variance of person	0.0142
Relative error variance	0.0015
Generalizability coefficient	0.9046

Table 4 shows the parameter used for the estimation of the generalizability coefficient of the 100-item 2017 NECO English language test. The table shows that the generalizability coefficient of the NECO test was .90. The generalizability coefficient of the test was high, suggesting that the test was highly reliable.

To determine the dependability coefficient, D-study was conducted based on the G-study conducted in objective 1. Thereafter, the dependability of the NECO test was extracted from the D-study. As in the case of the generalizability coefficient, lme4 package was used for the analysis. The dependability coefficient is calculated with:

$$Dependability\ coefficient = \Phi = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{abs}^2} \quad (2)$$

$$\Phi = \frac{0.0142}{0.0142 + 0.0022} = 0.8659$$

where σ_s^2 is the variation of students' test scores (the universe-score variance), and σ_{abs}^2 is the absolute error variance (Desjardins & Bulut, 2018). Table 5 presents the result.

Table 5. Dependability Coefficient

Source	Estimate
Variance for person	0.0142
Absolute error variance	0.0022
Dependability coefficient	0.8659

Table 5 shows the parameter used for the estimation of the dependability coefficient of the 100-item 2017 NECO English Language test. It shows that the dependability coefficient of the NECO test was .87. The result showed that the 2017 NECO English test scores were highly dependable. This implies that candidates' scores obtained on the 2017 NECO English language test were highly dependable in terms of reflecting the ability of the candidates.

Table 6. Decision Study

Number of items	Relative error var.	Absolute error var.	G coefficients	Phi coefficients
90	0.0017	0.0024	.90	.86
80	0.0019	0.0028	.88	.84
70	0.0021	0.0031	.87	.82
60	0.0025	0.0037	.85	.79
50	0.003	0.0044	.81	.76

As can be seen from Tables 4 and 5, the G and phi coefficients for 100-items fully crossed random designs were estimated as .90 and .87 respectively. Table 6 shows the D-study results obtained by reducing the number of items. When the number of items was reduced from 90 to 80, the relative error variance increased from 0.0017 to 0.0019; the absolute error variance also increased from 0.0024 to 0.0028; the g-coefficient decreased from .90 to .88 and phi coefficient also decreased from .86 to .84. The D-study is particularly useful in determining which combination of various measurement methods can be employed to obtain reliable coefficients.

Two levels of analysis were conducted to determine the extent to which the test was able to measure the same trait among male and female students. Factor analysis was conducted to identify one dimension underlying the English language test for male and female samples. Thereafter the extracted factor loadings for the test under male and female samples were compared. The comparison of the extracted factor loadings in two samples was made using Tucker index of factor congruence. The congruence coefficient is the cosine of the angle between two vectors and can be interpreted as a standardized measure of the proportionality of elements in both vectors. It is evaluated as:

$$\phi(x, y) = \frac{\sum_{n=i}^{N} x_i \, y_i}{\sqrt{\sum_{n=i}^{N} x_i^2 \; \sum_{n=i}^{N} y_i^2}} \tag{3}$$

where x_i and y_i are loadings of variable i on factor x and y, respectively, i = 1, 2, 3, …, n (in this case n = 100). Usually, the two vectors are columns of a pattern matrix. Therefore, how large should the coefficient be before two factors from two samples can be considered highly similar? Lorenzo-Seva and Ten Berge (2006) suggested that a value in the range of .85-.94 corresponds to a fair similarity, while a value higher than .95 implies that the two factors or components compared can be considered equal. The estimated factor loadings and other parameters for the estimation of the congruence index are presented in Appendix.

The table shows the parameters of the Tuckers index for congruence estimation. These parameters were substituted for in Equation 3. The result is presented as follows.

$\sum_{n=i}^{N} x_i\, y_i = 34.06, \quad \sum_{n=i}^{N} x_i^2 = 32.31, \quad \sum_{n=i}^{N} y_i^2 = 35.98$. Therefore,

$\phi(x,y) = \frac{34.06}{(32.31)(35.98)} = \frac{34.06}{\sqrt{1162.514}} = \frac{34.06}{34.10} = 0.9988$

The result showed that Tucker congruence index of similarity of the factors estimated under male and female candidates' samples was .99. This indicates that the factor underlying the performance of male candidates was almost identical with the factor underlying the female candidates' performance. The implication of the result is that the construct validity of the 2017 NECO English language test was very high and the test measured to a great extent the proficiency of students in the English language, and there was no other nuisance factor(s).

DISCUSSION and CONCLUSION

The findings of this study also showed the magnitude of error in generalizing from a candidate's score on 2017 NECO English language test to a universe score, as shown in Table 2. All 100 dichotomously scored items were analyzed using generalizability theory (G- theory) in a single-facet crossed study of persons (*p*) crossed with items (*i*). The variance component for candidates (i.e., the universe score variance) accounts for a smaller percentage of all the variance, corresponding to the largely similar scores obtained by the examinees. In order to reach more reliable results, it is generally desired that the number of moderate difficult items in the test is higher and the number of easy and difficult items relatively less; most of these items are of moderate difficulty. Therefore, none of the examinees scored all the items correct or incorrect; the majority of them scored between 60% and 70% of the items correct in the test. The tight clustering accounted for the observed low universe score variance. Furthermore, the variance component for the items is large relative to the universe score variance but smaller than the residual variance. The proportion of items that is correct reflects a lot of variations which corroborate the high percentage of variation accounted for by the items. The large residual variance captures both the person by item interaction and the random error, which cannot be disentangled. The high estimated variance component for persons crossed with items and the error is an indicator that almost 2/3 of the variability (random error) lies within this relationship and provides an estimate in the changes in the relative standing of a person from item to item (see Table 2). The result is in agreement with the findings of de Vries (2012) that the majority of error variance for the examination could be due to the interaction of persons with items, and lowering this variance would lead to an increase in dependability.

For relative decisions and a random-effects design, the generalizability coefficient is highly reliable. The dependability coefficient, Φ, an index that reflects the contribution of the measurement procedure to the dependability of the examination was also very dependable. As claimed by Brennan (2003) and Strube (2002), values approaching one (1) indicate that the scores of interest can be differentiated with a high degree of accuracy despite the random fluctuations of the measurement conditions. An important advantage of Φ is that it can be used to determine the sources of error that reduce classification accuracy and the methods to best improve such classifications, although most authors examined variability across facets to determine which one will be of greater benefit to generalizability. These results are consistent with the findings of Gugiu, Gugiu and Baldus (2012), Fosnacht and Gonyea (2018), Tasdelen-Teker, Sahin and Baytemir (2016), Nalbantoglu-Yilmaz (2017), Kamis and Dogan (2018) and Rentz (1987) who reported that the acceptable standards for dependability should be ≥ .70.

The study is also in contrast to the findings of Uzun Aktas, Asiret and Yorulmaz (2018), de Vries (2012) and Solano-Flores and Li (2006), who argued that each test item poses a unique set of linguistic challenges and each student has a unique set of linguistic strengths and weaknesses. Therefore, a certain number of items would be needed to obtain dependable scores. Uzun et al. (2018) and de Vries (2012) also pointed out that increasing the number of raters or occasions would increase the score dependability when rater and occasion are considered as facets. Li, Shavelson, Yin and Wiley (2015)

confirmed that increasing the number of items reduces error variance and increases both G and phi coefficients.

Based on the outcome of Tucker congruence index of similarity of the factors estimated under male and female candidates' samples (.99), the factor underlying the performance of male candidates was almost identical with the factor underlying the female candidates' performance. This implies that the examination measures to a great extent proficiency of students in the English Language. The result is in agreement with Zainudin (2012), who reported that the factor loading for an instrument must be higher or equal to .50. Also, Lorenzo-Seva and Ten Berge (2006) suggested that a value in the range of .85-.94 corresponds to a fair similarity, while a value higher than .95 implies that the two factors or components compared can be considered equal.

Conclusion

The study reflected that the reliability was high, which established that the scores assigned to candidates were dependable and generalizable. Also, the item validity was high because it measured the underlying construct, which underscores the good credibility of the items.

Recommendation

Prospective users of a measurement procedure are therefore advised to consider explicitly various sources of variation. They have to state whether they are interested in making absolute or relative decisions and whether they wish to generalize overall or only certain facets of a measurement procedure. However, there is a need to apply this concept to all school subjects to ensure the generalizability of the certification examinations.

REFERENCES

Alkharusi, H. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education, 2*(1), 184-196. doi: 10.5296/jse.v2i1.1227

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

Breithaupt, K. (2011). *Medical licensure testing: White paper for the assessment review task force of the medical council of Canada*. Retrieved from https://www.mcc.ca/wp-content/uploads/Technical-Reports-Breithaupt-2011.pdf

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brennan, R. L. (2003). *Coefficients and indices in generalizability theory* (CASMA Research Report Number 1). Iowa: Centre for Advanced Studies in Measurement and Assessment.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. U.S.A: Cengage Learning.

de Gruijter, D. N., & van der Kamp, L. J. Th. (2008). *Statistical test theory for the behavioural sciences*. New York: Chapman & Hall/CRC.

de Vries, I. M. (2012). *An analysis of test construction procedures and score dependability of a paramedic recertification exam* (Master's thesis). Queen's University Kingston, Ontario, Canada.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R* (1st Ed). Parkway, Boca Raton: Chapman and Hall/CRC Press. doi: 10.1201/b20498

Fosnacht, K., & Gonyea, R. M. (2018). The dependability of the updated NSSE: A generalizability study. *Research and Practice in Assessment 13*, 62-73. Retrieved from https://eric.ed.gov/?id=EJ1203503

Gugiu, M. R., Gugiu, P. C., & Baldus, R. (2012). Utilizing generalizability theory to investigate the reliability of the grades assigned to undergraduate research papers. *Journal of Multidisciplinary Evaluation, 8*(19), 26-40. Retrieved from https://journals.sfu.ca/jmde/index.php/jmde_1/article/view/362

Johnson, S., & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Coventy: Ofqual

Junker, B. W. (2012). *Some aspects of classical reliability theory and classical test theory*. Department of Statistics, Carnegie Mellon University, Pittsburgh.

Kamis, O., & Dogan, C. D. (2018). An investigation of reliability coefficients estimated for decision studies in generalizability theory. *Journal of Education and Learning*, 7(4), 103-113. doi: 10.5539/jel.v7n4p103

Li, M., Shavelson, R. J., Yin, Y., & Wiley, W. (2015). Generalizability theory. In *The* encyclopedia of clinical psychology (pp. 1322-1340). doi: 10.1002/9781118625392.wbecp352

Lorenzo-Seva, U., & Ten Berge, J. U. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2*(2), 57-64. doi: 10.1027/1614-2241.2.2.57

Mushquash, C., & O'Connor, B. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavioral Research Methods, 38*, 542-547. doi: 10.3758/BF03192810

Nalbantoglu-Yilmaz, F. (2017). Reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials prepared by teacher candidates. *Educational Sciences: Theory & Practice*, *17*(2), 395-409. doi: 10.12738/estp.2017.2.0098

Olusoji, O. A. (2012). Effects of English language on national development. *Greener Journal of Social Sciences, 2*(4), 134-139. doi: 10.15580/GJSS.2012.4.08291255

Rentz, J. O. (1987). Generalizability theory: A comprehensive method for assessing and improving the dependability of marketing measures. *Journal of Marketing Research, 24*(1), 19-28. doi: 10.1177/002224378702400102

Shavelson, R. J., & Webb, N. M.1(991). *Generalizability theory: A Primer.* Newbury Park CA: Sage.

Solano-Flores, G., & Li, M. (2006). The use of generalizability theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice, 25*(1), 13-22. doi: 10.1111/j.1745-3992.2006.00048.x

Strube, M. J. (2002). Reliability and generalizability theory. In L.G. Grimm and P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Tasdelen-Teker, G., Sahin, M. G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment. *Journal of Human Sciences, 13*(3), 5574-5586. Retrieved from https://j-humansciences.com/ojs/index.php/IJHS/article/view/4155

Uzun, N. B., Aktas, M. Asiret, S., & Yorumalz, S. (2018). Using generalizability theory to assess the score reliability of communication skills of dentistry students. *Asian Journal of Education and Training, 4*(2), 85-90. doi: 10.20448/journal.522.2018.42.85.90

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of Statistics*, *26*, 81-124. doi: 10.1016/S0169-7161(06)26004-8

Yin, Y., & Shavelson, R. J. (2008). Application of generalizability theory to concept map assessment research. *Applied Measurement in Education, 21*(3), 273-291. doi: 10.1080/08957340802161840

Zainudin, A. (2012). *Research methodology and data analysis* (5th Ed). Shah Alam: Universiti Teknologi MARA Publication Centre (UiTM Press).

Appendix. Factor Loading of English Test in Male and Female Examinees Groups

Item	Female (X)	Male (Y)	XY	X^2	Y^2
1	0.62	0.65	0.40	0.38	0.42
2	-0.46	-0.47	0.21	0.21	0.22
3	0.63	0.63	0.39	0.39	0.39
4	0.63	0.63	0.40	0.40	0.40
5	0.61	0.63	0.38	0.37	0.39
6	0.61	0.63	0.38	0.37	0.39
7	0.51	0.54	0.28	0.26	0.29
8	0.55	0.57	0.31	0.30	0.33
9	0.58	0.60	0.35	0.33	0.36
10	0.67	0.70	0.47	0.45	0.49
11	-0.48	-0.49	0.23	0.23	0.24
12	0.62	0.67	0.41	0.38	0.44
13	0.70	0.69	0.48	0.49	0.48
14	0.48	0.54	0.26	0.23	0.29
15	0.45	0.48	0.22	0.20	0.23
16	-0.55	-0.57	0.31	0.30	0.32
17	0.50	0.54	0.27	0.25	0.29
18	0.51	0.53	0.27	0.26	0.29
19	0.61	0.60	0.37	0.37	0.36
20	0.72	0.73	0.53	0.52	0.54
21	0.64	0.69	0.44	0.41	0.48
22	0.57	0.59	0.34	0.32	0.35
23	0.61	0.63	0.39	0.38	0.40
24	0.48	0.53	0.25	0.23	0.28
25	0.64	0.69	0.44	0.41	0.47
26	-0.40	-0.43	0.17	0.16	0.18
27	0.78	0.78	0.61	0.60	0.61
28	0.59	0.62	0.37	0.35	0.39
29	0.61	0.67	0.41	0.38	0.45
30	0.65	0.68	0.44	0.42	0.46
31	0.63	0.66	0.42	0.40	0.43
32	0.66	0.70	0.46	0.43	0.49
33	0.68	0.74	0.50	0.47	0.54
34	0.71	0.74	0.53	0.50	0.55
35	0.79	0.81	0.64	0.63	0.66
36	0.60	0.67	0.40	0.36	0.44
37	0.74	0.77	0.57	0.55	0.59
38	0.58	0.67	0.39	0.34	0.44
39	0.63	0.69	0.43	0.40	0.47
40	0.59	0.64	0.38	0.35	0.41
41	-0.39	-0.43	0.17	0.15	0.19
42	0.66	0.68	0.45	0.43	0.46
43	-0.43	-0.45	0.19	0.18	0.20
44	0.60	0.68	0.41	0.36	0.46
45	0.64	0.71	0.45	0.41	0.50
46	-0.44	-0.46	0.20	0.20	0.21
47	0.54	0.62	0.33	0.29	0.38
48	0.53	0.62	0.33	0.28	0.39
49	0.61	0.68	0.41	0.37	0.46
50	-0.51	-0.49	0.25	0.26	0.24
51	0.66	0.72	0.48	0.44	0.52
52	0.54	0.59	0.32	0.29	0.35
53	0.69	0.74	0.51	0.47	0.55
54	-0.55	-0.55	0.30	0.30	0.30
55	0.69	0.75	0.52	0.47	0.56
56	0.51	0.57	0.29	0.26	0.32
57	0.48	0.57	0.27	0.23	0.33
58	0.65	0.69	0.45	0.42	0.48
59	-0.48	-0.48	0.23	0.23	0.23
60	0.41	0.49	0.20	0.17	0.24
61	-0.57	-0.58	0.33	0.32	0.34
62	0.65	0.71	0.46	0.42	0.51

(continued)

Factor Loading of English Test in Male and Female Examinees Groups (continue)

Item	Female (X)	Male (Y)	XY	X^2	Y^2
63	-0.55	-0.56	0.31	0.30	0.32
64	0.65	0.69	0.45	0.42	0.48
65	0.48	0.57	0.27	0.23	0.32
66	0.56	0.61	0.34	0.31	0.37
67	0.45	0.51	0.23	0.20	0.26
68	-0.61	-0.60	0.36	0.37	0.36
69	0.65	0.69	0.45	0.42	0.48
70	0.73	0.76	0.55	0.53	0.57
71	-0.59	-0.58	0.34	0.34	0.34
72	-0.44	-0.46	0.20	0.19	0.21
73	0.67	0.66	0.44	0.44	0.43
74	0.54	0.57	0.31	0.29	0.32
75	0.63	0.67	0.42	0.40	0.45
76	-0.39	-0.37	0.15	0.15	0.14
77	0.60	0.65	0.39	0.36	0.42
78	0.50	0.56	0.28	0.25	0.32
79	0.59	0.60	0.35	0.34	0.36
80	-0.49	-0.50	0.24	0.24	0.25
81	-0.34	-0.37	0.13	0.11	0.14
82	-0.40	-0.40	0.16	0.16	0.16
83	-0.47	-0.46	0.22	0.22	0.21
84	0.70	0.70	0.49	0.48	0.49
85	0.48	0.53	0.25	0.23	0.28
86	0.48	0.53	0.25	0.23	0.28
87	0.50	0.54	0.27	0.25	0.29
88	-0.27	-0.27	0.07	0.07	0.07
89	-0.42	-0.42	0.17	0.17	0.17
90	0.58	0.60	0.35	0.34	0.36
91	-0.44	-0.44	0.19	0.19	0.19
92	0.66	0.68	0.45	0.44	0.46
93	-0.39	-0.40	0.16	0.15	0.16
94	-0.38	-0.40	0.15	0.15	0.16
95	-0.53	-0.53	0.28	0.28	0.28
96	-0.44	-0.49	0.22	0.19	0.24
97	-0.14	-0.18	0.03	0.02	0.03
98	-0.50	-0.50	0.25	0.25	0.25
99	0.69	0.72	0.50	0.48	0.52
100	0.56	0.60	0.33	0.31	0.36
Total			34.06	32.31	35.98

Nijerya'da İngilizce Sınavına Katılan Öğrenci Puanlarının Güvenilirliğinin Genellenebilirlik Kuramı ile İncelenmesi

Giriş

Genellenebilirlik Kuramı, yazılı veya bilgisayar tabanlı gerçekleştirilen bilgi testlerinin, derecelendirme ölçeklerinin veya öz değerlendirme ölçeklerinin ve kişilik testleri gibi performans testlerinin vb. psikometrik testlerin sonuçlarını analiz etmek için kullanılan istatistiksel bir yöntemdir (Breithaupt, 2011). Tek bir ölçüm sürecinde eşzamanlı olarak ortaya çıkan ve sonuçlara karışan birden çok hata kaynağını ayrıştırdığı için genellenebilirlik kuramı, (G-Kuramı) gerçek sonuçlara ulaşmayı hedefler. Gözlemlenen bir puanı, evren puanı için bir bileşene ve bir veya daha fazla hata bileşenine ayrıştırılarak eldeki bir gözlemden uygun gözlem evrenine genelleme yapılması amaçlanır. Klasik Test Teorisinde (KTT) imkânsız olan çeşitli olası etkileşimlerden kaynaklanan tutarsızlıkların yanı sıra hem değerlendiriciler arası hem de görevler arası tutarsızlıkları eş zamanlı olarak hesaba katarken her bir sınava giren kişi için ortalama derecelendirmenin güvenilirliğini tahmin edebilmesi açısından da avantajlıdır (Brennan, 2001). G Kuramında ölçümün gerçek değerinden uzaklaşmasına neden olan çeşitli hata kaynakları araştırılır. Ölçümün hata bileşenlerini çözme fırsatı verir ve ayrıca davranışsal ölçümün güvenilirliği veya güvenilirliği ile ilgilendiği için ölçme ve değerlendirme yönteminin kalitesini değerlendirmede ve kesinliğini geliştirmede değerli bir araçtır.

Tüm test puanları, diğer tüm ölçümler gibi, test puanlarının güvenilirliğini etkileyen bazı hatalar içerir. Aynı koşullar altında ölçümde farklılıklar olduğunda hata devreye girer. Ölçümde hata, kişinin gözlenen puanı ile gerçek puanı arasındaki fark olarak tanımlanabilir. Breithaupt (2011), maddelere ve test puanlarına karışan iki tür ölçüm hatası tanımlamıştır: rastgele ve sistematik hata. Elde edilen puanlardaki çeşitlilik genellikle hata kaynaklarına atfedilir ve bu nedenle bir testin psikometrik özelliğini belirleme zorluğunu ortaya çıkar. Psikometrik analizin amacı, gözlemlenen puanın (X) gerçek puanın (T) iyi bir ölçüsü olması için, mümkünse hata varyansını tahmin etmek ve en aza indirmektir. Gerçek değer tam olarak bilinmese de ideal değer bilinmeye çalışılabilir. KTT, gözlemlenen herhangi bir puan, çeşitli hata kaynaklardan gelse bile gerçek bir bileşen ile rastgele bir hata bileşeninin birleşimi olarak görülür. Bununla birlikte herhangi bir zamanda yalnızca tek bir ölçüm hata kaynağı incelenebilir. Genellenebilirlik Kuramı aynı zamanda evren puanına veya ölçüm sürecindeki tüm olası varyasyonlarda (örneğin farklı puanlayıcılar, formlar veya maddeler) beklenen ortalama puana odaklanır. Bu evren puanının, ölçüm nesnesi için belirli bir özelliğin değerini temsil ettiğine inanılır (Crocker & Algina, 2008). G Kuramı, birden fazla ölçüm hatası varyansının eşzamanlı etkisine odaklandığından araştırmacılara daha fazla geri bildirim sağlamaktadır.

Bazı araştırmalar, (Mushquash & O'Connor, 2006; Webb, Shavelson, & Haertel, 2006) çeşitli varyans kaynaklarının etkilerinin, belirli bir zamanda yalnızca tek bir ölçüm hatası kaynağının incelenmesinin mümkün olduğu KTT modelleri kullanılarak test edilebileceğini belirtmişlerdir. Ancak farklı hata kaynakları arasında meydana gelen etkileşim etkilerini incelemek mümkün değildir. Genellenebilirlik Kuramı, araştırmacılara özellikle bu konuda katkı sağlamaktadır; ölçüm durumunun her bir başarısı, test puanlarında ve onun adlandırılmış boyutunda bir hata kaynağıdır. Bu nedenle, birçok yazarın işaret ettiği gibi (Brennan, 2001; Johnson & Johnson, 2009) çok sayıda hata kaynağının açıklanamaması ve araştırmacıların KTT'nin olası hata kaynaklarını belirleyememesi ve aynı anda inceleyememesi G Kuramı'nın gelişmesini sağlamıştır. G Kuramı iki tür çalışmayı içerir: Genellenebilirlik çalışması (G-çalışması) ve Karar çalışması (D çalışması). Bir G-çalışmasının temel amacı, çeşitli kaynaklarla ilişkili puan varyansının bileşenlerini tahmin etmektir. D-çalışması ise bu tahmin varyans bileşenlerini kullanarak sonraki ölçüm için alternatifleri değerlendirerek optimal sonuca ulaşmaktır.

Alkharusi (2012) herhangi bir öğrenci için bazı ölçüm prosedürleriyle elde edilen gözlenen puanın gerçek puana ve tek bir hataya ayrıştırılabileceğini açıklamıştır. Ulusal Sınav Konseyi (NECO) Kıdemli Okul Sertifika Sınavında (Senior School Certificate Examination-SSCE) öğrencilerin

performansları toplam puanlarının yani KTT'nin toplamına dayandığından, karar alırken testin psikometrik özelliklerinin (zorluk, ayırt edicilik, güvenirlik, geçerlik) için testin yapısı, yönetimi ve analizine yönelik iyileştirme çalışmaları için adayların gözlemlenebilir performansının dikkate alınmasına ihtiyaç vardır. Güvenilirlik ve geçerlik, testlerin kalitesini ve kullanışlılığını ve ayrıca sınavlar için test maddelerinin oluşturulmasında dikkate alınması gereken ana faktörleri gösteren iki psikometrik özelliktir. Junker (2012) güvenilirliği, testin aynı koşullar altında tekrar uygulandığında tutarlı sonuçlar üreteceği kapsam olarak tanımlamıştır. Rastgele ölçüm hatası yoksa birey her seferinde aynı test puanını, yani gerçek puanı alacaktır. Güvenilir bir ölçüm geçerli olmayabileceğinden her biri için ayrı ayrı kanıt toplanması gerekmektedir. Ayrıca güvenirlik, geçerlik bir ön koşul olduğundan ölçüm sonuçlarının öncelikle güvenirliğine yönelik kanıtlar toplanabilir. Ölçme sonuçlarına karışan hatalar, öncelikle güvenilirliği etkiler ancak hatalar, geçerliği de tehdit eder. Bu nedenle ilgilenilen yapının yeterli ölçümünü sağlamak için her ikisine yönelik kanıtların toplanmasına ihtiyaç vardır. Öğrenci notları, ölçümün doğruluğunu azaltan çeşitli hata türlerinden etkilenir. NECO tarafından yapılan tek uygulamalı bir testin belirli bir formunda bir seferde elde edilen tek bir puan tamamen güvenilir değildir çünkü o kişinin tüm kabul edilebilir durumlar, test formları ve uygulamalardaki ortalama puanıyla eşleşmesi olası değildir. Bir kişinin puanı genellikle diğer durumlarda, test formlarında veya farklı yöneticilerle farklı olacaktır. En ciddi tutarsızlık veya hata kaynakları hangileridir? Mümkün olduğunda, tanımlanan her bir kaynaktan kaynaklanan hata varyanslarının tahmin edilmesi beklenir. G-Kuramının güçlü yönlerinden bağımsız olarak, Nijerya'da ortaokul sınavlarındaki öğrencilerin puanlarının güvenilirliğini tahmin etmek için özel olarak geniş çapta uygulanmamıştır.

Nijerya'da, ortaokul eğitiminin sonunda, öğrencilerin Batı Afrika Sınav Konseyi (WAEC) ve NECO tarafından yürütülen SSCE veya Ulusal Sınavlar gibi sertifika sınavları yazmaları beklenir. Ulusal İş ve Teknik İnceleme Kurulu (NABTEB) tarafından yürütülen İşletme ve Teknik Sertifika Eğitimi (NBTCE). NECO, SSCE'yi her yıl Haziran/Temmuz ve Kasım/Aralık aylarında yürütür. 1999 yılında WAEC'in iş yükünü azaltmak, özellikle çok sayıda adayı test etme yükünü azaltmak amacıyla kurulmuştur.

İngilizce eğitimi, Nijerya mirasına derinlemesine yerleşmiş ve mevcut durumda vazgeçilmez hâle gelen bir mirastır. Dil eğitimi; ekonomik kalkınmanın yanı sıra sosyo-kültürel ve politik entegrasyon için mükemmel bir araç olarak kabul edilmektedir. Eğitim dilinin yanı sıra İngilizcenin ülkede ikinci bir dil olarak kullanılması, bilim ve teknolojideki modern gelişmelere hızlı bir erişim sağlamıştır (Olusoji 2012). Söz konusu nedenlerden dolayı, ülke çapında ve ülke eğitim sisteminin tüm seviyelerinde İngilizce eğitimine büyük önem verilmektedir.

Bu nedenle, İngilizcenin uluslararası bir dil olarak önemi ve Nijeryalı ortaokul öğrencilerinin performansı üzerindeki etkisi göz önüne alındığında, ortaokul sınavlarının güvenilirliğini incelemek için genellenebilirlik kuramının kullanılması önem taşımaktadır.

Yöntem

Bu araştırma betimsel araştırma yöntemine dayalı olarak yürütülmüştür. Betimsel araştırmalar, mevcut ilişkilerin herhangi bir manipülasyonu olmaksızın, mevcut değişkenlerin seçilmesi ve gözlemlenmesi yoluyla neden ve sonucu ilişkisini incelemektedir. Nijerya'da 2017 yılında NECO SSCE İngilizce Dil Sınavı'na giren toplam 1,037,129 öğrenci bulunmakta olup sınava giren 311,138 aday, çalışmanın örneklemini oluşturmuştur. Örneklem seçkisiz örnekleme yöntemlerinden tabakalı örnekleme tekniği kullanılarak seçilmiştir. Her eyaletten adayların yüzde otuzu rastgele seçilerek çalışma yürütülmüştür. Çalışmada kullanılan veriler, NECO ofisinden alınan OMR sayfalarında belirtildiği gibi Nijerya'da NECO Haziran/Temmuz 2017 İngilizce SSCE yazan adayların (100 maddelik çoktan seçmeli teste) verdiği yanıtlardır. Verilerin analizinde G Kuramına dayalı olarak öncelikle G-çalışması, ardından D-çalışması yürütülmüştür.

Sonuç ve Tartışma

Bu çalışmada öncelikle bir adayın 2017 NECO İngilizce dil sınavındaki puanından bir evren puanına genellemede hatasının büyüklüğü incelenmiştir. Adaylar için varyans bileşeni, tüm varyansın daha küçük bir yüzdesini oluşturmaktadır. Sınava girenlerin aldığı puanların benzer olduğu bulunmuştur. Doğru cevaplandırılan maddelerin oranı, maddeler tarafından açıklanan yüksek çeşitlilik yüzdesini doğrulayan birçok farklılaşmayı yansıtır. Büyük artık varyans, hem kişi bazında madde etkileşimini hem de çözülemeyen rastgele hatayı göstermektedir. Araştırmanın sonuçları, de Vries'in (2012) inceleme için hata varyansının çoğunluğunun kişilerin maddelerle etkileşiminden kaynaklanabileceği ve bu varyansın düşürülmesinin güvenilirlikte bir artışa yol açacağı yönündeki bulgularıyla uyumludur.

Araştırma kapsamında NECO'ya katılan öğrencilerin cevapları doğrultusunda göreceli kararlar ve rastgele etkiler tasarımı için genellenebilirlik katsayısının oldukça yüksek olduğu tespit edilmiştir. Ölçüm prosedürünün muayenenin güvenilirliğine katkısını yansıtan bir indeks olan güvenilirlik katsayısı Φ da güvenilir bulunmuştur. Bu sonuçlar Gugiu, Gugiu ve Baldus (2012), Fosnacht ve Gonyea (2018), Taşdelen-Teker, Şahin ve Baytemir (2016), Nalbantoğlu-Yılmaz (2017), Kamış ve Doğan (2018) ve Rentz'in (1987) güvenilirlik için kabul edilebilir standartların ≥ .70 olması gerektiği bulgusuyla tutarlıdır.

Çalışma aynı zamanda Uzun, Aktaş, Aşiret ve Yorulmaz (2018), de Vries (2012) ve Solano-Flores ve Li (2006), her test maddesinin bir dizi dilsel zorluk oluşturduğunu ve her öğrencinin dilsel olarak güçlü ve zayıf yönlerini ortaya koymaktadır. Bu nedenle, güvenilir puanlar elde etmek için belirli sayıda maddeye ihtiyaç duyulacaktır. Uzun ve diğerleri (2018) ve de Vries (2012) ayrıca, puanlayıcı ve durum birer faktör olarak ele alındığında puanlayıcı veya durum sayısının artırılmasının puan güvenilirliğini artıracağına dikkat çekmiştir. Li, Shavelson, Yin ve Wiley (2015) madde sayısını artırmanın hata varyansını azalttığını ve hem G hem de phi katsayılarını artırdığını doğrulamıştır. Araştırma sonuçları, bu bulgularla tutarlıdır.

Erkek ve kadın adayların örneklemleri altında tahmin edilen faktörlerin benzerliklerine ilişkin Tucker uyum indeksi (0.99) sonucuna göre, erkek adayların performansının altında yatan faktör, kadın adayların performansının altında yatan faktör ile hemen hemen aynı bulunmuştur. Sonuç, bir madde için faktör yükünün .50'ye eşit veya daha yüksek olması gerektiğini bildiren Zainudin (2012) ile uyumludur. Ayrıca Lorenzo-Seva ve Ten Berge (2006), .85-.94 aralığındaki bir değerin makul bir benzerliğe karşılık geldiğini, ancak .95'ten yüksek bir değerin karşılaştırılan iki faktör veya bileşenin eşit kabul edilebileceğini ima ettiğini öne sürmüşlerdir.

Çalışma, güvenilirliğin yüksek olduğunu yansıtmakta ve bu da adaylara verilen puanların güvenilir ve genellenebilir olduğunu ortaya koymaktadır. Ayrıca, maddelerin güvenilirliğinin altını çizen temel yapıyı ölçtüğü için öğe geçerliliği yüksek hesaplanmıştır. Sonuçlar, G-Kuramı ile kestirilerek sonuçlar üzerinde yorumlar yapılmıştır.