Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 4, 2020, pp 442-452
https://doi.org/10.18510/hssr.2020.8443

PROPER SAMPLE SIZES FOR ENGLISH LANGUAGE TESTING: A SIMPLE STATISTICAL ANALYSIS

Faisal Mustafa[1], Roderick Julian Robillos[2*]
[1]Lecturer, Universitas Syiah Kuala, Banda Aceh, Indonesia; [2]Lecturer, Khon Kaen University, Khon Kaen, Thailand.
Email: [1]faisal.mustafa@unsyiah.ac.id, [2*]rodero@kku.ac.th

Article History: Received on 17th May 2020, Revised on 25th July 2020, Published on 13th August 2020

Abstract

Purpose of study: Small sample size is the most common limitation which restricts the generalization of research results, and this is true to many fields, including language testing. The current study is sought to show the predictive power of sample sizes over the population mean to decide what sample minimum size can be considered as a proper sample size for a language test.

Methodology: The data for this quantitative research was 5,250 paper-based TOEFL test scores considered as the population, which includes listening, structure, and reading tests, and it is the most familiar standardized test among EFL researchers. Due to its objective nature, it leaves little chance for bias scores. The score ranged between 30.7% of 417 in the TOEFL scale and 95.7% or 653. Standard error was used as the parameter in deciding the proper sample size. It was the cut-off point when the parameter did not show any obvious change when the sample size was added. We used hierarchical agglomerative clustering with three clusters, determined using 30 indices through the majority rule, in finding out the cut-off point.

Main Findings: It was found that the cut-off point is at the sample size of 52 with the range between 46 and 59. Therefore, it can be concluded that the minimum proper sample size for a research study involving a language test is n = 46.

Application of this study: The results of this study apply to the area of English language teaching and testing. However, it does not rule out the possibility that the study result applies to tests in other languages.

Novelty/Originality of this study: The result of this study should be treated as statistical evidence of the proper sample size to avoid inaccurate or conflicting research results in language teaching where a test is used for analysis.

Keywords: *Sample Size, Language Teaching, Testing, Standard Error, Hierarchical Agglomerative Clustering.*

INTRODUCTION

The sample size is the first most significant component in any quantitative research studies. It allows researchers to make generalizations of the research result to a wider context because an adequate sample size can represent a population (Stangor, 2011, p. 122; Tuckman & Harper, 2012, p. 425). There is a common practice by researchers to express limitations of their research in published research papers, and the small sample size is the most common limitation which is perceived to restrict the generalization of research results. However, there has not been any consensus among EFL researchers regarding what sample size is considered small. Experts in research methodology in social sciences have failed to provide or to reach an agreement of the proper sample size which is representative of a population (Doryei, 2007, pp. 95–96; Gravetter & Forzano, 2012, pp. 141–142). The not-large enough sample sizes contribute to, among others, conflicting results in similar research studies such as the study about the strength of the relationship between the use of metacognitive language-learning strategies and language-learning motivation by Wu (2007) and Goh and Foong (1997), and the empirical evidence on the effect of learning by teaching technique by Kasim, Muslem, and Mustafa (2020). Another example is differences in effect sizes of peer-assessment across 54 published research found by Double et al. (2019) because they did not consider sample sizes in selecting papers to be analyzed. With the p-value of 0.05, which is most preferred in social science research, the Type I error, i.e. rejecting the null hypothesis when it is actually true, is very sensitive to sample size (Agresti, 2019, p. 150). To obtain accurate research results, researchers need to obtain the data from a proper sample size. For that, researchers will be able to draw conclusions leading to the establishment of new theories because the advantage of a quantitative research study is that the results are generalizable (Doryei, 2007, p. 34). This is achieved through the power of statistical analysis toward the data. Does this statistical analysis require that the sample size to be large (French et al., 2013, p. 154; Kothari, 2004, p. 160), but again, how large is considered large? The present research explores, for the first time, the proper sample size which can be used in language testing. We utilized a simple, basic statistical analysis to answer this very fundamental question using real data of more than 5,000 scores. The findings should make an important contribution for researchers in determining the number of participants to be recruited for a quantitative study involving language tests.

The remaining part of the paper proceeds as follows: First, the literature on sample sizes is comprehensively reviewed to show that there is an urgency for the current research. This section is summarized with the presentation of an alternative procedure that can be used to determine the proper sample size for language tests. The next section is concerned with the methodology used for this study, describing how the data were obtained and prepared for analysis. The data analysis and how the conclusion was made are discussed in detail in this section. The result section which follows presents the result

of statistical analysis with detailed visualization, and the results are discussed comprehensively in the result section. The summary of the findings is concluded in the conclusion section, accompanied by the implication of the results.

LITERATURE REVIEW

This section describes literature regarding how sample size is calculated to represent the population and standardized sample size in statistics. The section is concluded with an alternative procedure of determining the optimum sample size which can represent the population.

Population and sample

The population is the parameter to which the quantitative research results are intended. Most of the time, the population consists of a large group of individuals, i.e. tens of thousands, thus it is inefficient to involve the whole population for a research study for the sake of time and money (Navarro, 2016, p. 301). Therefore, we can make inferences about the population based on a sample (Privitera, 2018, pp. 6–7), that is some individuals drawn from the population (Neuman, 2014, p. 246). The purpose of sample selection is to obtain a proper number of those individuals in a way that can represent the population from which they are drawn (Lock et al., 2017, p. 20). For that purpose, there are some techniques of sample selection, i.e. probability sampling, which includes several types of random sampling, and non-probability sampling, which covers convenience sampling and purposive sampling (Kothari, 2004, p. 59). However, the most recommended sample selection which are confidently-representative of populations is completely-randomized sample selection (Neuman, 2014, p. 49).

Although the purpose of randomized-sample selection is to obtain a limitless generalization of the research results, many populations are influenced by many characteristics such as age, educational background, gender, and the list can go on limitlessly (Stangor, 2011, p. 257). However, the generalization of research results should be limited to only characteristics which are controlled in the research (Gravetter & Forzano, 2012, pp. 236–237). These limitations are usually expressed by researchers in the last part of the discussion or conclusion section in their papers, and further research studies are usually advised to address these limitations (Peacock, 2002, p. 486; Ruiying & Allison, 2003, pp. 376–379).

Sample sizes

In addition to the proper way of sample selection, the number of individuals included in the sample is no less of significance. The sample sizes very much determine the representativeness of the population (Lock et al., 2017, p. 205). When the sample size is big enough, the sample statistic is very close to the population parameter (Lock et al., 2017, p. 204). However, in current practice, many quantitative and qualitative studies in language pedagogy involved only small sample sizes (Baese-Berk & Morrill, 2015; Latifi et al., 2011; Perakyla, 1997, pp. 295–296). As much as the sample sizes were too small, many researchers insisted on generalizing the results of their research (see Atai & Nazari, 2011; Baese-Berk & Morrill, 2015; Latifi et al., 2011). Some researchers admitted that the results of their research cannot be generalized to other contexts (Shieh & Freiermuth, 2010). Although they recognize the trade-off of their inadequate sample size, still their research is not very meaningful for the current knowledge. A research study is intended to understand how the world works, and without achieving that purpose, the research is less practical.

The fact that a small sample size does not represent the population is caused by the sample not being able to mimic all significant characteristics of the population, such as in Fageeh (2014, p. 15). For example, EFL students as the population consist of students who:

1) Learn English as a foreign language;
2) Come from under-developed, developing, and developed countries;
3) Are males and females;
4) Learn English at school only and lean English by using other media such as movies;
5) Are taught by good teachers or better teachers.

If the sample size is too small, not all characteristics of the population above, for example, are caught by the sample (Mendenhall et al., 2013, p. 297). If the sample only consists of EFL students from developed countries, the research results are not generalizable to EFL students in under-developed or developing countries. In addition, when the sample is broken down into groups such as to analyzed the effect of gender, the sample size becomes even smaller. As a result, Fageeh (2014) found that the effect of some types of writing practices on writing proficiency between students with different levels of English proficiency, which is unexpected and difficult-to-explain because this is not in line with the current well-established literature. Because Fageeh (2014) used a small sample size, it is less wise to use the research result to dispute the literature. Also, Coolican (2014, p. 52) claimed that there is a possibility of failing to reject the null hypothesis (H_0) even when the H_0 is not true when we rely on a small sample.

Commonly-used sample sizes in language tests

Many quantitative research studies in language teaching, particularly English language teaching, involve language tests as a measurement to be used as the data for analysis. Some research studies used excessive sample sizes such as those revealed by Wei et al. (2019, p. 3). They reported that some articles published in System utilized sample sizes of more than 300. On the other hand, most research in second language acquisition involves a small sample size, generally less than 20 (Gonulal, 2016, p. 7). Coxhead (2017, p. 84) mentioned that it was not easy to obtain a large sample for research in applied linguistics, where language testing is one of the subfields. One of the causes is the ideal class size for second and foreign language classes are small, for example, 16 in Turkey as reported by Gonulal (2016, p. 7). Small class size contributes to better success in language teaching (Harfitt, 2015); however, it is a threat for research. In addition, undergraduate and post-graduate students conducting research as the university requirement tend to select only one class for a time and cost efficiency without considering the generalizability of their research result (Fadiana et al., 2020; Nirwan, 2020; Setiawan & Wiedarti, 2020).

Other research has tried to manipulate moderately small sample size by using a robust statistical analysis called bootstrapped quantile regression (BQR) analysis, which is a method of randomly replicating the existing sample (Nikitina et al., 2019). However, this method does not guarantee that the sample represents the population, especially when the existing sample is small (Hesterberg, 2015, p. 384). In addition, this method involves a robust advanced statistic, which is not common in applied linguistics research (Khany & Tazik, 2019). In addition, the statistical literacy even among graduate students in the non-mathematics related field was limited in such inferential statistics (Gonulal, 2016).

Standardized sample sizes in statistics

Experts in statistics have proposed the maximum sample sizes considered small. The rules of thumb to analyze differences between groups by using a significant test such as t-test and ANOVA, the suggested sample sizes are no less than 29 (Lock et al., 2017, p. 419) or 30 (Mendenhall et al., 2013, p. 304; Privitera, 2018, p. 288). A smaller sample size is allowed for a Chi-square test, which is also often in language pedagogy research. VanVoorhis and Morgan (2007, p. 48) and Camilli and Hopkins (1978, p. 165) suggested that the overall sample size is 20 and no less than 5 for each variable. In addition, the sample size for correlation and regression analyses should be larger in order for the analysis results to be meaningful, i.e. 50 (VanVoorhis & Morgan, 2007, p. 48).

Statistically, Kothari (2004, pp. 175–177) mentioned that the expected sample size can be determined using a method involving intended standard error and confidential interval. Using this method, when the population size is unknown, the sample size is calculated by using the following formula.

$$n = \frac{z^2 . \sigma_p}{e^2}$$

Where: n = sample size

z = the value of the standard variate at (1.96 for p = 0.05 and 2.5758 for p = 0.01)

σ_p = standard deviation of the population (to be estimated from past experience or on the basis of a trial sample)

e = standard error of the sample statistic (acceptable error)

When the population size is known to researchers, the formula to determine the sample size is as follows.

$$n = \frac{z^2 . N . \sigma_p^2}{(N-1)e^2 + z^2 . \sigma_p}$$

Where n = sample size

N = population size

z = the value of the standard variate at (1.96 for p = 0.05 and 2.5758 for p = 0.01)

σ_p = standard deviation of the population (to be estimated from past experience or on the basis of a trial sample)

e = standard error of the sample statistic (acceptable error)

As much as it is theoretically profound, the procedure above requires statistical skill, especially in determining the standard deviation of the population and acceptable error. Therefore, although "sample size determination is a prerequisite for statistical surveys" (Sadia & Hossain, 2014, p. 420), a very small number of researchers in language pedagogy reported that they determined the sample size based on any of those methods. Therefore, ready-to-use sample size is required in the field of language pedagogy.

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 4, 2020, pp 442-452
https://doi.org/10.18510/hssr.2020.8443

An alternative procedure

The optimum sample size can also be determined using the standard error for every estimated sample size. The optimum size is concluded when the standard error does not change after increasing the sample size. As it is required by the standard error formula, the standard deviation of the population can be calculated by using the formula given by Lock et al. (2017, p. 79) and Privitera (2018, p. 123).

$$s = \sqrt{\frac{\sum(x - \underline{x})^2}{n - 1}}$$

Where s = standard deviation of the sample

n = sample size

x = individual datum in the population

$\underline{x}$ = population mean

After the standard deviation is obtained, the standard error of sample mean (e) can be calculated based central limit theorem, as provided by Coolican (2014, p. 400) and Adams and Lawrence (2015, p. 229) in the following.

$$e = \frac{s}{\sqrt{n}}$$

Where e = standard error of the sample mean

s = standard deviation of the population

n = sample size

Based on the formula, it is expected that increasing sample size results in smaller standard error of sample mean, which is in line with Lock et al. (2017, p. 204), Mendenhall et al. (2013, p. 297), and Privitera (2018, p. 227).

The current study

Many books and research studies related to the methodology of research involving a language test have been published by experts, researchers, teachers, and other education practitioners. Our literature review section has shown that the proper sample size for a language test has not been empirically proposed in any previous publications. The current study used the alternative procedure discussed above, i.e. using the standard error for every estimated sample size, to fill this significant gap of literature. Therefore, the conclusion in this research was drawn based on robust statistical analyses and big data, which allows the generalization of the results.

METHODOLOGY

This section is the most essential section in this paper, where the utilization of basic inferential statistics is backed up with robust advanced inferential statistics to determine the predicted proper sample size. The section concludes by presenting how the optimum sample size was determined after the analysis.

DATA

The data used for this research consisted of 5,250 scores of Test of English as a Foreign Language (TOEFL) obtained from the Language Center of Syiah Kuala University in Indonesia. TOEFL scores were used in this research because it is the most familiar English language test among EFL educators and researchers (Elfiondri, et al., 2020). There are three versions of the test, i.e. international internet-based, international paper-based, and institutional paper-based TOEFL. Unlike the other versions of TOEFL, institutional TOEFL is easy to administer because it does not have a speaking or writing section. In this research, the authors used institutional paper-based TOEFL, well known as ITP TOEFL, because the absence of speaking and writing sections makes the results free of rater bias. The test takers were English as a foreign language (EFL) learners between 24 and 40 years of age. The scores selected range from 30.714% to 95.714, where the scores below 30.714% have been removed because, according to Mustafa and Anwar (2018), those scores do not represent learner's English proficiency. These scores are considered the population on which the parameter is based, and the size is large enough to be considered a population. The distribution of the scores is presented in Figure 1.

The scores in Figure 1 is a little right-skewed because there are more scores in lower ranges. The population mean is 49.65347, and the standard deviation is 8.33, with 95% of the data are from 43.571 to 55.

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 4, 2020, pp 442-452
https://doi.org/10.18510/hssr.2020.8443

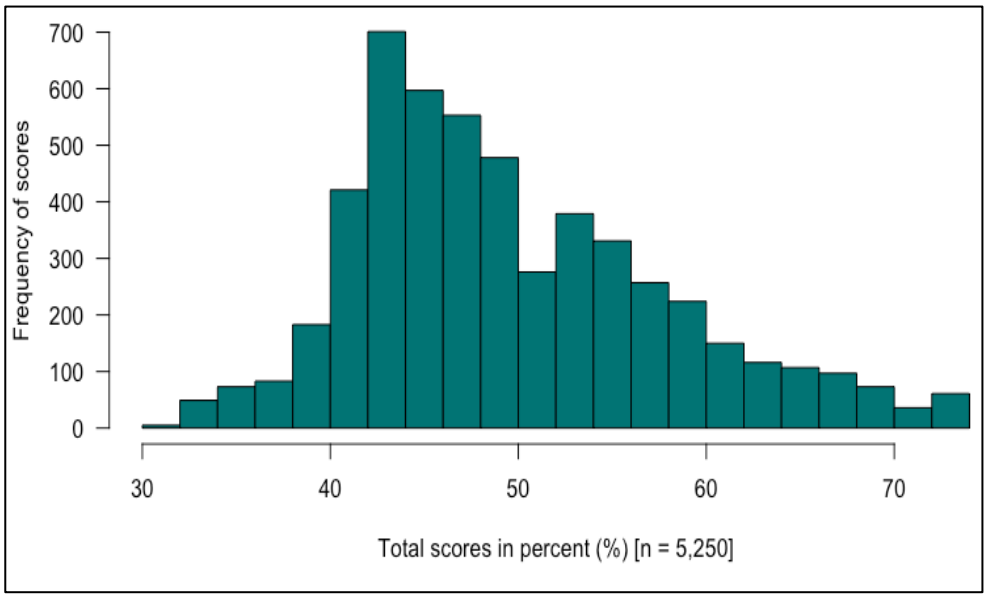

Figure 1: Summary of the primary population

Procedures of analysis

The sample were selected randomly from the population, i.e. 5,250 scores, with the size starting from 0.25% of the population size (13 scores), and the size increased 0.25% for the following score until it reaches 20% (1,050 scores). To randomly select the sample, the sampling with replacement was performed automatically by using R, an opensource statistical computer software, using the R code from the mosaic package, as shown in Figure 2.

```
235  # random sample selection
236  sample(x, size, replace = TRUE)
```

Figure 2: Codes for random sample selection

In Figure 2, *x* in the code is the data frame of the population, and *size* is the sample size to be randomly selected from the population *x*. For the confidential estimation of the optimum sample size, this process was repeated from random selection to produce 100 samples for each size. The standard error of the mean for each sample was treated as a single datum.

For clustering purposes, the mean for each standard error was used. First, the number of optimum clusters was determined using 30 indices. According to Charrad et al. (2014, p. 3), this method "offers the best scheme from different results." We used the majority rule to decide the best number of clusters, meaning that the number of clusters resulted from the majority of indices is considered as the best number of clusters. The following code was used to calculate the number of clusters simultaneously for all indices.

```
32  #optimum number of clusters
33  NbClust(data = my_data, diss = NULL, distance = "euclidean",
34          min.nc = 2, max.nc = 15, method = "kmean",
35          index = "all")
```

Figure 3: Codes to determine the optimum number of clusters

After obtaining the optimum number of clusters. The data were clustered using hierarchical agglomerative clustering or known as Agglomerative Nesting (AGNES). In this approach, "each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy" (Baby & Sasirekha, 2013, p. 1). We selected a hierarchical agglomerative clustering technique for our analysis because Ah-Pine (2018, p. 1) claimed that the technique is generic, efficient and effective. The following R codes are used to execute the hierarchical agglomerative clustering formula and create the dendrogram, a chart that visualizes generated clusters, are as follows.

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 4, 2020, pp 442-452
https://doi.org/10.18510/hssr.2020.8443

```
# Calculating clusers by agglomerative clustering
agnes_mean <- agnes(hie_cluster_mean_manual, metric = "euclidean",
                    stand = FALSE, method = "average")

#Visualizing clusters in a dendrogram
pltree(agnes_mean, cex = 0.6, hang = -1, main = "SE clusters")
```

Figure 4: Codes to create and visualize clusters

The purpose of these analyses was to cluster standard error for each sample size in order to decide the sample size that is so optimum that adding more data does not result in decreasing the standard error. Therefore, the starting point of the last cluster was considered as the point where the sample size is considered optimum. The lowest starting point is the midpoint between the starting point of the last cluster and the endpoint of the previous cluster.

RESULTS

The purpose of the current research was to find out the smallest proper sample size which is not considered small in language testing by utilizing standard error. The analysis started by randomly selected the sample with the sizes starting from 0.25% of the population size with 0.25% increase in the next size until the size reached 25%. Then the standard error was calculated for every sample size. This process was repeated 99 more times, and the result of the calculation is presented in the following tables. For space availability, only the first and last 10 sample sizes and 10 samples were presented.

Table 1: Standard error of means for ten samples for the first ten sample sizes

	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
1	2.143	4.772	2.744	2.623	1.915	1.245	1.872	1.016	1.120	2.060
2	8.571	0.942	4.524	2.178	2.214	1.649	1.625	1.055	2.010	1.749
3	0.357	0.350	1.988	4.394	2.359	2.545	1.583	1.414	1.705	2.071
4	8.929	3.436	2.277	2.421	1.894	1.243	1.470	1.856	1.750	1.795
5	5.000	3.183	2.352	1.447	0.951	1.380	2.449	1.024	1.026	2.097
6	3.214	3.017	2.182	2.975	2.141	1.755	0.845	2.156	1.631	1.925
7	0.714	1.385	1.510	1.409	2.173	1.882	1.691	0.943	1.658	1.281
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
98	0.357	3.300	2.347	2.696	2.981	1.879	1.295	1.702	1.761	1.200
99	2.143	1.827	3.474	1.236	4.209	1.816	1.721	1.785	1.465	1.227
100	1.786	4.407	3.591	3.145	1.652	1.311	2.526	2.096	2.102	1.439

Table 2: Standard error of means for the last ten sample sizes

	22.75	23.00	23.25	23.50	23.75	24.00	24.25	24.50	24.75	25.00
1	0.478	0.445	0.511	0.468	0.494	0.489	0.480	0.436	0.495	0.451
2	0.488	0.538	0.504	0.501	0.493	0.458	0.462	0.474	0.473	0.460
3	0.478	0.501	0.479	0.532	0.489	0.471	0.482	0.440	0.443	0.501
4	0.519	0.504	0.445	0.520	0.518	0.455	0.494	0.411	0.461	0.472
5	0.521	0.470	0.503	0.470	0.485	0.529	0.459	0.396	0.473	0.495
6	0.469	0.511	0.514	0.509	0.523	0.461	0.497	0.511	0.483	0.515
7	0.496	0.462	0.488	0.472	0.467	0.475	0.456	0.474	0.471	0.460
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
98	0.505	0.488	0.451	0.533	0.423	0.430	0.474	0.482	0.487	0.419
99	0.452	0.487	0.499	0.478	0.449	0.509	0.507	0.536	0.463	0.441
100	0.489	0.472	0.443	0.465	0.428	0.482	0.487	0.454	0.459	0.407

The tables above show that the standard errors were higher for smaller sample sizes (Table 1) than those for larger sample sizes (Table 2). In addition, the standard errors are less stable across samples for smaller sample sizes, where some are very high, such as 9, and the others are lower such as 0.3. However, for larger sample sizes, the standard errors

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 4, 2020, pp 442-452
https://doi.org/10.18510/hssr.2020.8443

are not very different between one and another. To make a conclusion regarding proper sample size, the means for each standard error were calculated and the summary is presented in the following figure.

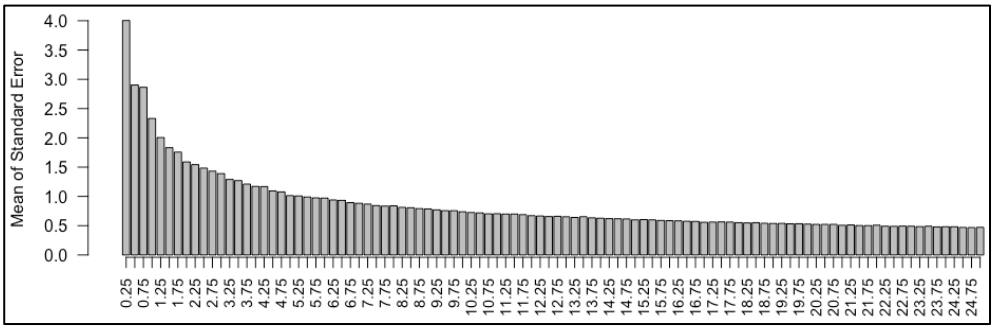

Figure 5: Means of standard error

Figure 6 shows standard errors which are gradually decreasing as the sample sizes are bigger. There are some fluctuations when the sample sizes reach more than 10%. However, these fluctuations are small dan hardly visible in the figure. For smaller sample sizes, the decrease is rather sharp compared to smaller ones.

Before the means of standard error were categorized into clusters, we determined the number of optimum clusters using 30 indices, from which the decision was drawn using the majority rule, and the result is shown in the following figure.

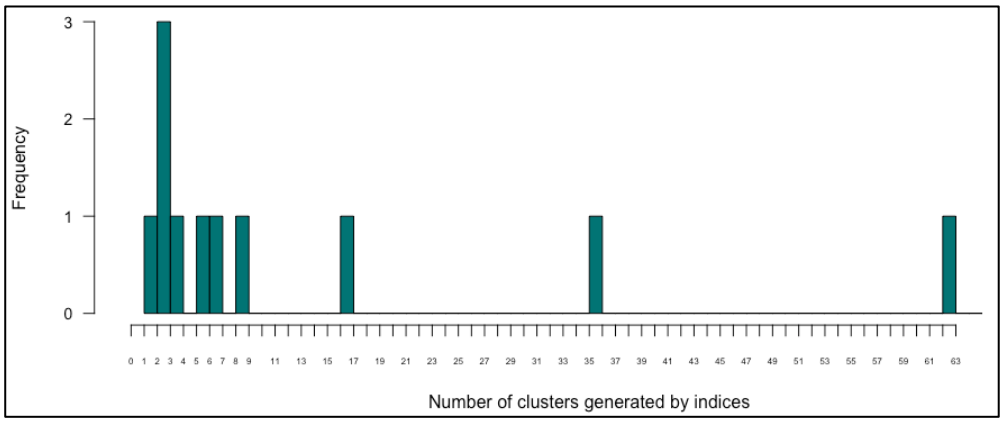

Figure 6: Optimum number of clusters

In Figure 6, only positive numbers greater than one were included, i.e. 11 possible proposed clusters. The other 13 indices proposed the numbers of clusters less than 1, and thus excluded from the histogram plot above. In addition, the other six indices did not propose any result. Among all indices included in Figure 6, three indices proposed 3 as the best number of clusters. Therefore, we picked three clusters in the following dendrogram.

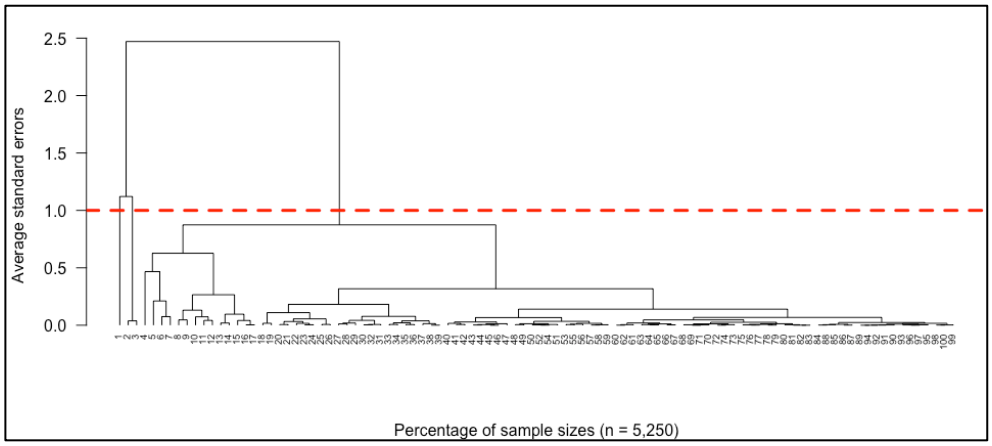

Figure 7: Dendrogram showing clusters for means of standard error

The dendrogram shown in Figure 7 is a result of clustering using hierarchical agglomerative clustering calculated and generated in R. The red horizontal dash line divides the data into three clusters, as determined in Figure 6. The last cluster starts at 4, that is the point of 1% of the total population. Since there is a distance between 0.75% and 1%, and

Humanities & Social Sciences Reviews
eISSN: 2395-6518, Vol 8, No 4, 2020, pp 442-452
https://doi.org/10.18510/hssr.2020.8443

between 1% and 0.25%, the midpoint between the distance was considered the border separating both percentages. The detail is shown in Figure 8.

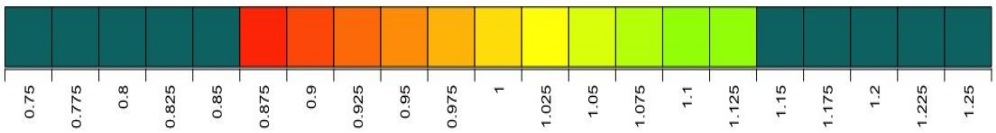

Figure 8: Midpoints of cluster 3 starting point

In figure 8, the midpoint between 0.75% and 1% is at 0.875, therefore 0.875% is considered the lowest percentage which belongs to 1%. In addition, the midpoint between 1% and 1.25% is 1.125%. Therefore, the range of the lowest point of cluster 3 is between 0.875% (46 scores) and 1.25% (59 scores). Therefore, the sample sizes at these points are used as the optimum sample size for a language test, which is comprehensively discussed in the following section.

DISCUSSION

The objective of this study was to find out the proper minimum sample size for language testing by utilizing standard errors from 5,250 TOEFL scores. The sample size was considered proper when adding more data did not significantly lower the standard error. Hierarchical agglomerative clustering was used to divide standard errors into clusters and the number of optimal clusters was calculated using 30 indices. The starting point of the last cluster was considered the smallest proper sample size for language testing. The analysis result shows that the last cluster, i.e. cluster 3, started from 1% (44 scores). By considering 0.75% as the previous point before 1% and 1.25% as the next point, 0.125 was taken as the interval of 1%; therefore, the lowest proper sample size was 0.125%±1%, i.e. 0.875% (46 scores) and 1.25% (59 scores).

This result is very significant for researchers in language pedagogy who conduct research involving language test for quantitative analysis. The purpose of a quantitative research study was to make generalization. There have been countless published studies that have used sample size smaller than 38 but forced generalization based on the results of the study. Among others are Razaghi, Bagheri, and Yamini (2019), Farvardin (2019), Slim and Hafedh (2019). Many of those previous studies have proposed to generalize the results of their researches in spite of improper sample size. Such over generalization should be revisited for validation. There is a chance that the results may deviate from previously obtained results. According to Faber and Fonseca (2014, p. 28), using a small sample size increases a chance of assuming a false premise as true. In addition, we often fail to reject the null hypothesis, which is a type II error in statistics, when the sample size is too small (Lock et al., 2017, p. 312).

The result that 38 scores are considered proper sample size for language testing provides *good news* for researchers in language pedagogy involving language testing. In statistics, the proper sample size suggested without considering the population size was at least 362 participants (Bhalerao & Kadam, 2010). For population size similar to our study, i.e. 5,250, *Raosoft Survey*, which is one of the most used sample size calculator recommended by Vaux and Briggs (2006), estimates 359 as the sample size. These numbers are *too impossible* for most researchers, especially undergraduate students who conduct research for their degree requirements, with limitations of time and financial support. Therefore, many researchers reported the combined sample size although the sample was split in their analysis, such as splitting into experimental and control groups. In addition, researchers in language pedagogy rarely cited literature to provide a reason for using a certain sample size because they could not follow the sample size suggested in the literature. With the result of our study, it is possible for most researchers, including undergraduate and postgraduate students, to obtain 46 participants for their quantitative research. Therefore, they can use the result of our study to justify the adequacy of their sample sizes.

CONCLUSION AND IMPLICATION

This study was intended to provide a reference for researchers in language pedagogy regarding the proper sample size in language testing. With the use of simple statistical analyses backed up with robust inferential statistics based on standard errors as the parameter of estimate and hierarchical agglomerative clustering as the procedure of estimation, our study has revealed that the smallest sample size which can be used to predict the true mean or population mean was between 46 and 59 participants. The results of the analyses show that increasing the sample size does not improve the accuracy of results, where the mean of the sample does not differ much to the mean of the population. Therefore, for time and cost efficiency, research involving language tests can generalize the results based on the sample size of as small as 46 participants.

This study has proposed an alternative procedure for calculating proper sample sizes. In statistics, the sample size is ideally determined by considering the requirements such as intended effect size, confidence level, the margin of error, and standard error, and input the requirement into a sample size formula. In contrast, we proposed to estimate the sample sizes by testing standard error produced by each sample size to the mean of the population from real-world data. As an implication, this procedure can be replicated to estimate proper sample sizes in other fields. For that purpose, big data in the form of scores from a test in any discipline can be collected, and the researcher can use the procedure that has been

demonstrated in our analysis to find out the proper sample size in that particular discipline. Afterward, previous research that utilized sample sizes smaller than the proper sample size should be replicated to validate the results and draw more accurate conclusion, which is significant for the development of knowledge. Therefore, the result of our study would be helpful for researchers across disciplines.

LIMITATION AND STUDY FORWARD

The analysis in the current article was so simple that it might be seen unnecessary by those who have advanced statistical sense. However, it was appropriate and adequate to arrive at conclusions that make sense for most non-statistician researchers. However, the generalizability of this research results is subject to some limitations. First, the data used were real-world data in non-productive language proficiency, i.e. listening comprehension, structure, and reading comprehension. Therefore, the results of our analysis only apply to non-productive language test scores. Productive language scores such as writing and speaking need to be analyzed separately to find the optimum sample size. We do not currently have access to such data in a large quantity. Second, we only consider mean in deciding the sample size, and thus we believe that the result of our study would apply for any statistical tests. However, some statistical analysis such as correlation analysis and linear regression analysis might need a larger sample size for their accurate analysis. Third, the data that we used were numerical data; therefore, the results might not be true for categorical data. Finally, we did not validate the usability of our proper sample sizes for other populations. Therefore, we do not know the level of its effectiveness. Further research is recommended to validate the correctness of our predicted proper sample size. This research paper has already claimed the correctness of the predicted result. So future recommendation can be to validate the correctness of the prediction rather than testing the correctness as mentioned in the last line. It is desired that the researcher will make it clear about the non-productive and productive language scores. Because it is a vital part of the methodology of the paper.

ACKNOWLEDGEMENT

We would like to express their gratitude to Mr. Novi Reandy Sasmita and Mr. Samsul Anwar for their insightful discussions regarding data analysis for our research. Great appreciation is also addressed to Prof. Kevin Woods, in whose class the original idea for this paper was emerged. The data for this research was obtained from the Language Center of Syiah Kuala University, and thus we would like to express our gratitude for their agreement to use the data for our research.

AUTHORS CONTRIBUTION

Faisal Mustafa dealt with data collection, data analysis, and results and wrote the result section, and revised the method section in the article. He also converted all the references to the format required by the journal and prepared the conclusion section.

Roderick Julian Robillos dealt with the introduction section prepared the first draft of the research methodology and discussed the results of the study.

REFERENCES

1. Adams, K. A., & Lawrence, E. K. (2015). *Research methods, statistics, and applications*. Sage Publications.
2. Agresti, A. (2019). *An introduction to categorical analysis* (3rd ed.). John Wiley & Sons, Inc.
3. Ah-Pine, J. (2018). An efficient and effective generic agglomerative hierarchical clustering approach. *Journal of Machine Learning Research*, *19*, 1–43.
4. Atai, M. R., & Nazari, O. (2011). Exploring reading comprehension needs of Iranian EAP students of health information management (HIM): A triangulated approach. *System*, *39*(1), 30–43. https://doi.org/10.1016/j.system.2011.01.015
5. Baby, P., & Sasirekha, K. (2013). Agglomerative hierarchical clustering algorithm- A review. *International Journal of Scientific and Research Publications*, *3*(3), 1–3.
6. Baese-Berk, M. M., & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *The Journal of the Acoustical Society of America*, *138*(3), 223228. https://doi.org/10.1121/1.4929622
7. Bhalerao, S., & Kadam, P. (2010). Sample size calculation. *International Journal of Ayurveda Research*, *1*(1), 55–57. https://doi.org/10.4103/0974-7788.59946
8. Camilli, G., & Hopkins, K. D. (1978). Applicability of Chi-square to 2 X 2 contingency tables with small expected cell frequencies. *Psychological Bulletin*, *85*(1), 163–167.
9. Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36.
10. Coolican, H. (2014). *Research methods and statistics in psychology* (6th Ed.). Psychology Press. https://doi.org/doi:10.4324/9780203769669
11. Coxhead, A. (2017). Dealing with low response rates in quantitative studies. In J. McKinley & H. Rose (Eds.), *Doing research in applied linguistics: realities, dilemmas and solutions* (pp. 81–90). Routledge Taylor & Francis Group. https://doi.org/10.4324/9781315389608-8
12. Doryei, Z. (2007). *Research Methods in Applied Linguistics: Quantitative, qualitative, and mixed*

methodologies. Oxford University Press. https://doi.org/10.1017/S0272263110000094
13. Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2019). The Impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*, *31*(4), 967–989. https://doi.org/10.1007/s10648-019-09510-3
14. Elfiondri, Kasim, U., Mustafa, F., Putra, T. M. (2020). Reading comprehension in the TOEFL PBT: Which sub-skill deserves more intensive training? *TESOL International Journal*, *15*(1), 53–64.
15. Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.EBO
16. Fadiana, D., Bahri Ys, S., & Inayah, N. (2020). Teaching vocabulary by using total physical response. *Research in English and Education (READ)*, *5*(1), 1–6.
17. Fageeh, A. I. (2014). The use of journal writing and reading comprehension texts during pre-writing in developing EFL students' academic writing. *Studies in Literature and Language*, *9*(3), 1–18.
18. Farvardin, M. T. (2019). Effects of spacing techniques on EFL learners' recognition and production of lexical collocations. *Indonesian Journal of Applied Linguistics*, *9*(2), 395–403. https://doi.org/10.17509/ijal.v9i2.20237
19. French, B. F., Immekus, J. C., & Yen, H.-J. (2013). Logistic regression. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (pp. 145–165). Sense Publishers.
20. Goh, C. C. M., & Foong, K. P. (1997). Chinese ESL students' learning strategies: A look at frequency, proficiency, and gender. *Hong Kong Journal of Applied Linguistics*, *2*(1), 39–53. http://eric.ed.gov/?id=EJ597324
21. Gonulal, T. (2016). *Statistical literacy among second language acquisition graduate students*. Michigan State University.
22. Gravetter, F., & Forzano, L.-A. (2012). *Research Methods for The Behavioral Sciences* (4th ed.). Wadsworth, Cengage Learning.
23. Harfitt, G. J. (2015). *Class size reduction: Key insights from secondary school classrooms*. Springer. https://doi.org/10.1007/978-981-287-564-8
24. Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *American Statistician*, *69*(4), 371–386. https://doi.org/10.1080/00031305.2015.1089789
25. Kasim, U., Muslem, A., & Mustafa, F. (2020). Empirical evidence on the effectiveness of Learning by Teaching technique among English as a foreign language university students. *Journal of Language and Education*, *6(4).*
26. Khany, R., & Tazik, K. (2019). Levels of statistical use in applied linguistics research articles: From 1986 to 2015. *Journal of Quantitative Linguistics*, *26*(1), 48–65. https://doi.org/10.1080/09296174.2017.1421498
27. Kothari, C. R. (2004). *Research methodology: Methods and techniques* (2nd Ed). New Age International (P) Ltd.
28. Latifi, M., Mobalegh, A., & Mohammadi, E. (2011). Movie subtitles and the improvement of listening comprehension ability: Does it help? *The Journal of Language Teaching and Learning*, *1*(2), 18–29.
29. Lock, R. H., Lock, P. F., Morgan, K. L., Lock, E. F., & Lock, D. F. (2017). *Statistics: Unlocking the power of data* (2nd ed.). John Wiley & Sons, Inc.
30. Mendenhall, W. I., Beaver, R. J., & Beaver, B. M. (2013). *Introduction to Probaility and Statistics*. https://doi.org/10.1017/CBO9781107415324.004
31. Mustafa, F., & Anwar, S. (2018). Distinguishing TOEFL score: What is the lowest score considered a TOEFL score? *Pertanika Journal of Social Sciences and Humanities*, *26*(3), 1995–2008.
32. Navarro, D. (2016). *Learning statistics with R: A tutorial for psychology students and other beginners*. University of New South Wales.
33. Neuman, W. L. (2014). *Social research methods: Qualitative and quantitative approaches* (7th ed.). Pearson Education Limited.
34. Nikitina, L., Paidi, R., & Furuoka, F. (2019). Using bootstrapped quantile regression analysis for small sample research in applied linguistics: Some methodological considerations. *PLoS ONE*, *14*(1), 1–19. https://doi.org/10.1371/journal.pone.0210668
35. Nirwan, N. (2020). Using KWL (know-want to know-learned) strategy in improving students' reading comprehension. *English Education Journal*, *11*(2), 199–214.
36. Peacock, M. (2002). Communicative moves in the discussion section of research articles. *System*, *30*, 479–497.
37. Perakyla, A. (1997). Reliability and validity in research based on naturally occurring social interaction. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (2nd Ed., pp. 283–304). Sage Productions.
38. Privitera, G. J. (2018). *Statistics for the behavioral sciences* (3rd Ed). Sage Production.
39. Razaghi, M., Bagheri, M. S., & Yamini, M. (2019). The impact of cognitive scaffolding on Iranian EFL learners' speaking skill. *International Journal of Instruction*, *12*(4), 95–112. https://doi.org/10.29333/iji.2019.1247a
40. Ruiying, Y., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes*, *22*(4), 365–385. https://doi.org/10.1016/S0889-4906(02)00026-1
41. Sadia, F., & Hossain, S. S. (2014). Contrast of Bayesian and classical sample size determination. *Journal of Modern Applied Statistical Methods*, *13*(2), 420–431. https://doi.org/10.22237/jmasm/1414815720

42. Setiawan, M. R., & Wiedarti, P. (2020). The effectiveness of Quizlet application towards students' motivation in learning vocabulary. *Studies in English Language and Education*, *7*(1), 83–95. https://doi.org/10.24815/siele.v7i1.15359
43. Shieh, W., & Freiermuth, M. R. (2010). Using the DASH Method to Measure Reading Comprehension. *TESOL Quarterly*, *44*(1), 110–128. https://doi.org/10.5054/tq.2010.217676
44. Slim, H., & Hafedh, M. (2019). Social media impact on language learning for specific purposes: A study in English for business administration. *Teaching English with Technology*, *19*(1), 56–71.
45. Stangor, C. (2011). *Research methods for the behavioral sciences* (4th ed.). Wadsworth, Cengage Learning.
46. Tuckman, B. W., & Harper, B. E. (2012). *Conducting educational research* (6th Ed). Rowman & Littlefield Publishers, Inc.
47. VanVoorhis, C. R. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 43–50.
48. Vaux, A., & Briggs, C. S. (2006). Conducting mail and internet surveys. In F. T. L. Leong & J. T. Austin (Eds.), *The Psychology Research Handbook: A Guide for Graduate Students and Research Assistants* (pp. 186–209). SAGE Publications, Inc. https://doi.org/10.4135/9781412976626.n13
49. Wei, R., Hu, Y., & Xiong, J. (2019). Effect size reporting practices in applied linguistics research: A study of one major journal. *SAGE Open*, *9*(2). https://doi.org/10.1177/2158244019850035
50. Wu, M. M. (2007). The relationships between the use of metacognitive language-learning strategies and language-learning motivation among Chinese-speaking ESL learners at a vocational education institute in Hong Kong. *Asian EFL Journal*, *9*(3), 93–117.

