Action Research Journal Indonesia (ARJI)
Volume 8 Number 3 Year 2026 | Pages 1098 -1115
E-ISSN : 2775-0787 | P-ISSN : 2774-9290 | DOI : 10.61227
Address : Tuesday Block No. 76 Rt/Rw 01/003 Jatitengah Village, Jatitujuh District, Majalengka Regency, West Java
e-mail : arjijournal@gmail.com
Url Web : https://journal.nahnuinisiatif.com/index.php/ARJI

Developing and Evaluating an English Language Proficiency Test Using Generalizability Theory: A One-Facet Design

Article Information :
Articles Submitted : 2026-08-12
Articles Received : 2026-08-31
Published Articles : 2026-09-16

Hossein Salarian[1]*

[1]University of Tehran, Tehran, Iran

Email Correspondence * : Hos_salarian@ut.ac.ir

Keywords:
English language proficiency assessment; Generalizability Theory; G-study; D-study; score dependability; person-by-item design; test development

Abstract: *The present study aimed to develop and evaluate an English language proficiency test using Generalizability Theory (G-Theory) and a one-facet person-by-item (p × i) design. The study addressed the need for a systematic examination of the sources of variability underlying English proficiency scores and the identification of an efficient test length. Sixty Iranian English language learners participated in the study and completed a researcher-developed 30-item multiple-choice English language proficiency test assessing vocabulary, grammar, and reading comprehension. The test was administered under standardized conditions, and participants' responses were analyzed using an analysis of variance within a G-Theory framework. In the Generalizability Study (G-study), persons were treated as the objects of measurement and items as the single measurement facet. The analysis estimated the variance components associated with persons, items, and the person-by-item/residual component. The findings indicated that the person component accounted for a substantial proportion of the observed score variability (39.95%), whereas item-related variance was very small (0.13%). The person-by-item/residual component represented the largest proportion of variance (59.92%), suggesting that learners' performance varied across items and that measurement error remained an important source of score variability. The 30-item test produced a high Generalizability coefficient (G = .952), indicating strong score dependability for relative decisions concerning learners' English proficiency. A Decision Study (D-study) further examined alternative test lengths of 10, 15, 20, 25, and 30 items. The results showed that the G-coefficient increased systematically from .870 for 10 items to .952 for 30 items, while relative error variance decreased as the number of items increased. Although the 30-item test yielded the highest dependability, the 20- and 25-item versions also demonstrated high levels of score dependability, suggesting a practical balance between measurement precision and testing efficiency. The study concludes that G-Theory provides a useful framework for identifying sources of score variability and making evidence-based decisions about English proficiency test design and length.*

Conceptualization: all authors
Methodology: all authors
Investigation: all authors
Writing original draft preparation: all authors
Writing review and editing : all authors
Visualization: all authors

Acknowledgments
All praise and gratitude to God for the successful completion of this article. The author expresses his gratitude to all parties involved and those who contributed to its preparation. Thanks to their constant prayers and support, this article was successfully completed.

All authors have read and agreed to the published version of the manuscript.

INTRODUCTION

English language proficiency tests are widely used to measure learners' language abilities and to make educational decisions. However, the quality of these decisions depends on the reliability of the test scores. A test score should reflect differences in learners' English proficiency as much as possible, rather than differences caused by the test itself. In practice, however, learners may obtain different scores because they respond differently to different test items. Therefore, an important problem in English language test development is determining whether the variability in test scores is mainly related to differences among learners or is influenced by the characteristics of the test items (Kocaoğlu & Şahin, 2024). Liao (2023) emphasized that the dependability of language assessment scores should be examined in relation to the conditions under which the assessment is conducted. This issue shows the need for a statistical approach that can identify and quantify different sources of variability in English language test scores.

On the other hand, Generalizability Theory (G-Theory) provides an appropriate framework for addressing this problem because it allows researchers to examine several sources of variability in an assessment at the same time. Instead of treating measurement error as a single undifferentiated source, G-Theory separates the observed-score variance into meaningful components. In a test in which items are the main measurement facet, the analysis can show how much of the variability is associated with persons, items, the person-by-item interaction, and residual error. This information is important for test developers because a test may have an acceptable overall reliability coefficient while still containing substantial variability caused by particular measurement conditions (Shin, 2022) Liao (2023) demonstrated the usefulness of G-Theory for examining the dependability of L2 reading assessment and showed that question format and other measurement conditions can affect the reliability of assessment scores.

The problem is also important from a practical test-development perspective. Developing a longer test does not automatically guarantee a sufficiently reliable assessment. The number and characteristics of items should be examined statistically to determine whether additional items improve the dependability of scores. G-Theory is useful in this respect because a Decision Study (D-study) can be conducted after the Generalizability Study (G-study) to evaluate alternative test designs. For example, researchers can compare the expected reliability of a test with different numbers of items and identify a test length that provides adequate measurement precision without making the assessment unnecessarily long. Kocaoğlu and Şahin (2024) demonstrated this practical application by using G-Theory and a D-study to investigate how different test structures affected reliability. Their study shows that G-Theory can be used not only to evaluate the reliability of an existing test but also to make evidence-based decisions about test design.

Based on this problem, the present study focuses on the development and statistical evaluation of an English language proficiency test using a one-facet G-Theory design in which test items are treated as the measurement facet. The study therefore seeks to determine how much of the observed score variability is attributable to differences among examinees and how much is related to items, the person-by-item interaction, and residual error. It also examines the dependability of the resulting scores through the Generalizability coefficient and uses a D-study to determine how changing the number of test items may affect measurement precision.

In this way, the study addresses a practical question in English language assessment: how can an English language proficiency test be designed so that its scores are sufficiently dependable, while the sources of measurement variability and the appropriate number of test items are clearly identified?

Literature Review

1. English Language Proficiency Assessment

English language proficiency assessment is an important part of educational measurement because it provides evidence about learners' ability to use English for academic, educational, and professional purposes. Proficiency tests are commonly used to distinguish learners at different levels of language ability and to support decisions such as placement, selection, and evaluation. According to Schmidgall et al. (2021), the quality of a language proficiency assessment should be considered in relation to the intended interpretation and use of its scores. Their discussion of the redesigned TOEIC Bridge tests emphasized the importance of obtaining evidence that supports the interpretation of test scores as indicators of English listening, reading, speaking, and writing abilities. Similarly, Norris and Lee (2023) reported that the TOEFL Essentials test was designed to provide an overall estimate of English proficiency across academic and everyday contexts. These studies indicate that the development of an English proficiency test requires more than the construction of test items; it also requires systematic evidence that the resulting scores accurately represent the intended language ability.

An important concern in English language proficiency assessment is therefore the dependability of the scores obtained from a test. A learner's observed score may be influenced by factors other than the learner's underlying level of English proficiency. These factors may include the characteristics of test items or tasks, the interaction between learners and measurement conditions, and other sources of measurement error (Wolf, et al., 2023). In a study of an oral English proficiency test, Shin (2022) found that examinees accounted for the largest proportion of score variability, while task and rating effects were smaller; however, the results also showed that alternative assessment designs could improve score dependability. This finding suggests that the quality of an English proficiency test should be evaluated not only by examining learners' scores but also by investigating the sources of variability that contribute to those scores. Consequently, a statistical framework capable of separating different sources of variability is particularly relevant to the development and evaluation of English language proficiency tests.

2. Reliability in Language Assessment

Reliability is a central concept in language assessment because decisions based on test scores require scores that are sufficiently consistent and dependable. In traditional approaches to measurement, reliability is often summarized by a single coefficient; however, language assessment may involve several conditions that can influence observed scores. For example, differences among test items, tasks, raters, or occasions may contribute to score variability. Röttele et al. (2021) demonstrated the value of examining separate variance components through Generalizability Theory rather than treating all unexplained variability as one general source of error. Their study showed that the contribution of different sources of variability

could be estimated statistically, providing more detailed information about the reliability of an assessment. This perspective is particularly relevant to language testing because the observed performance of a learner may change across different measurement conditions.

The issue of reliability is especially important when the purpose of an assessment is to make decisions about learners' language proficiency. A test with insufficiently dependable scores may lead to inaccurate interpretations of learners' abilities and consequently to inappropriate educational decisions. Research in language assessment has shown that the structure of an assessment can influence score reliability. For example, Liao (2023) investigated classroom-based L2 reading assessment and examined how question format and other measurement conditions affected score dependability. The findings demonstrated the usefulness of Generalizability Theory for identifying the conditions under which L2 assessment scores become more or less dependable. Therefore, reliability in the present study is not treated simply as a single numerical characteristic of the English proficiency test; instead, it is considered in relation to the different sources of variability that may affect the scores obtained from the test.

3. Generalizability Theory in Educational and Language Assessment

Generalizability Theory (G-Theory) provides a statistical framework for examining the dependability of measurement results by separating observed-score variability into different variance components. Unlike Classical Test Theory, which generally represents measurement error as a single component, G-Theory allows researchers to investigate several potential sources of variability simultaneously. Andersen et al. (2021), in a systematic review and meta-analysis, demonstrated the usefulness of G-Theory for examining reliability and identifying sources of variance in assessment. The authors emphasized that G-Theory can provide information about how different components of an assessment contribute to observed-score variability. Similarly, Röttele et al. (2021) used G-Theory to estimate variance components and showed how this approach can provide a more detailed understanding of the reliability of assessment scores. Thus, G-Theory is particularly useful when researchers need to identify not only whether scores are dependable but also why variability occurs in those scores.

G-Theory has also become relevant to language assessment because language performance is often measured under multiple conditions. In language testing, the measurement conditions may include items, tasks, raters, occasions, or other facets depending on the assessment design. Shin (2022) applied G-Theory to an oral English proficiency test and examined the relative contributions of examinees, tasks, ratings, and language groups to score variability. The study found that examinees were the largest contributor to score variability, while the effects of tasks and ratings were relatively small, and it also used alternative designs to improve score dependability. Liao (2023) similarly demonstrated the usefulness of G-Theory in L2 reading assessment by examining score dependability under different assessment conditions. These studies support the use of G-Theory as an appropriate framework for the present study, where the aim is to investigate the sources of variability in an English language proficiency test and evaluate the dependability of its scores.

4. Sources of Variability in a One-Facet Design

In G-Theory, a measurement design specifies the object of measurement and the facets that may contribute to score variability. In the present study, examinees are the objects of measurement and test items represent the single measurement facet; therefore, the design can be represented as p × i, where p denotes persons and i denotes items. The purpose of this design is to determine the extent to which differences in observed test scores are attributable to differences among examinees and to the measurement conditions represented by the items. In a one-facet design, the analysis can also consider the person-by-item interaction, which reflects the possibility that examinees may perform differently across items. (Schmidgall et al. 2021). Recent methodological research has demonstrated the importance of accurately estimating these variance components. For example, Zhang et al. (2023) compared alternative methods for estimating variance components in a p × i G-Theory design and specifically examined the person, item, and person-by-item components.

The interpretation of these sources of variability is important for evaluating an English language proficiency test. A large person variance component generally indicates meaningful differences among examinees, which is desirable when the purpose of the test is to distinguish learners according to their English proficiency. In contrast, substantial item-related or person-by-item variability may indicate that test items do not function consistently across examinees and may contribute to measurement error. The importance of examining these components is also supported by Shin (2022), whose analysis of an English speaking test showed that examinee-related variance was the largest contributor to score variability, while task and rating effects were smaller. Therefore, the present one-facet design will provide an empirical basis for determining how much score variability is associated with examinees, items, and their interaction, rather than assuming that all observed differences in scores represent differences in English proficiency.

5. G-Study, G-Coefficient, and D-Study in Test Development

A Generalizability Study (G-study) is the first major statistical stage in a G-Theory analysis. Its purpose is to estimate the variance components associated with the object of measurement and the relevant facets of the design. These estimates provide the basis for evaluating the dependability of scores and for identifying the relative contribution of each source of variability. Röttele et al. (2021) demonstrated how G-Theory can be used to estimate separate variance components and evaluate their contribution to total score variability. More recently, Zhang et al. (2023) examined the estimation of variance components in a p × i design and compared different statistical procedures for obtaining these estimates. In the context of the present study, the G-study will therefore be used to statistically examine the variability of scores obtained from the English language proficiency test and to estimate the relevant variance components.

Following the G-study, a Decision Study (D-study) can be used to examine alternative assessment designs and determine how changes in measurement conditions influence score dependability. The D-study is particularly useful for test development because researchers can evaluate the consequences of changing the number of items or other measurement conditions before selecting the final test design (Norris & Lee, 2023). Hirschi and Kang (2023), for example, used G-studies and D-studies to examine how changes in the number of raters and other

measurement conditions affected the reliability of L2 speech assessment. Kocaoğlu and Şahin (2024) also used G-Theory and a D-study to investigate how different testlet structures influenced reliability. These studies demonstrate that G-Theory can move beyond the simple evaluation of an existing test and provide practical evidence for improving test design. Accordingly, the present study will use the G-coefficient to evaluate score dependability and a D-study to examine alternative numbers of test items, with the aim of identifying an efficient test length that provides an acceptable level of measurement precision.

METHOD

Participants and Context

The participants in this study were 60 Iranian learners of English who were selected from an English language institute in Iran. The participants were enrolled in general English courses at the time of data collection and represented a relatively similar educational context. They were selected because they were actively studying English and therefore had sufficient experience with formal language assessment. The participants included both male and female learners, and their ages ranged from approximately 18 to 25 years. Participation in the study was voluntary, and the learners were informed about the purpose and procedures of the study before the assessment was administered. Their participation was also treated confidentially, and no personal information was used in reporting the results.

The study was conducted in the context of English language education in Iran, where English is generally learned as a foreign language rather than as a language of daily communication. The assessment was administered in a classroom setting under standardized conditions. All 60 learners completed the same English language proficiency test, which consisted of a set of test items designed to assess their general English proficiency. The same instructions, time limit, and testing conditions were provided to all participants in order to minimize irrelevant sources of variability. The participants' responses to the test items constituted the primary data for the Generalizability Theory analysis. In the one-facet design used in this study, the learners (p) were considered the objects of measurement, while the test items (i) represented the single measurement facet. Thus, the resulting data were analyzed according to a $p \times i$ design to examine the sources of variability in the learners' test scores and to evaluate the dependability of the assessment.

Materials and Instruments

The main instrument used in this study was a researcher-developed English Language Proficiency Test (ELPT) designed to assess the general English proficiency of the 60 Iranian English language learners. The development of the test was informed by the Common European Framework of Reference for Languages (CEFR), which provides internationally recognized descriptors for describing language ability and supports the interpretation of assessment results in terms of what learners can actually do with the language. Recent developments in English language testing have increasingly emphasized the alignment of proficiency assessments with clearly defined proficiency levels and intended score interpretations (Wolf et al., 2023). In addition, contemporary English proficiency assessments have incorporated technology-based and adaptive approaches to improve the efficiency and precision of measurement. For example, Norris and Lee (2023) reported that the TOEFL Essentials test uses

innovative item types and multistage adaptive testing to estimate English proficiency efficiently. In the present study, however, a fixed-form test was selected because the primary purpose was to examine item-related sources of variability within a one-facet G-Theory design rather than to evaluate an adaptive testing system.

The ELPT consisted of 30 objectively scored multiple-choice items covering three major areas of general English proficiency: vocabulary, grammar, and reading comprehension. Each item had four response options, with only one correct answer. The vocabulary section assessed learners' knowledge of commonly used English words and their meanings in context. The grammar section measured learners' ability to recognize and use appropriate grammatical structures, while the reading section included short authentic or semi-authentic passages followed by comprehension questions. The test was designed to represent a broad range of item difficulty so that it could provide meaningful score differences among learners with different levels of proficiency. The use of multiple item types and carefully selected content is consistent with recent developments in English proficiency assessment, in which test developers have sought to improve the representation of language ability while maintaining efficient and objective scoring (Norris & Lee, 2023). The test items were reviewed before administration to ensure that their wording, content relevance, level of difficulty, and alignment with the intended English proficiency construct were appropriate for Iranian learners.

To support the development and administration of the test, several digital assessment tools were incorporated into the assessment process. The test was administered electronically using a secure online questionnaire platform, which allowed all participants to receive the same instructions, answer options, time limits, and item order. Digital administration was selected because computer-based language assessment can facilitate standardized administration, automatic recording of responses, and rapid preparation of data for statistical analysis. Recent developments in language testing indicate that digital technologies can provide new opportunities for efficient and innovative assessment, although their use requires careful consideration of validity, fairness, and technical conditions (O'Sullivan, 2023). Accordingly, the digital platform in the present study was used primarily as an administration and data-collection tool rather than as an automated scoring system.

For the statistical analysis, the responses of the 60 participants to the 30 test items were organized in a person-by-item ($p \times i$) data matrix, with participants represented by rows and test items represented by columns. The resulting dataset was analyzed using Generalizability Theory. An ANOVA-based procedure was used to estimate the relevant variance components, including person variance, item variance, and person-by-item/residual variability. The Generalizability coefficient (G-coefficient) was then calculated to estimate the dependability of the relative ranking of learners, and a Decision Study (D-study) was conducted to examine the effect of different numbers of items on score dependability. This approach is consistent with recent applications of G-Theory in language assessment, where variance components and alternative assessment designs have been examined to determine how measurement conditions influence the reliability of language scores (Liao, 2023; Kocaoğlu & Şahin, 2024). Thus, the materials and instruments in the present study were selected not only to measure English proficiency but also to provide appropriate data for investigating the sources of score variability and determining an efficient test design.

Data Collection Procedure

The data collection procedure was conducted in several systematic stages. First, the researcher developed the English Language Proficiency Test (ELPT) based on the objectives of the study and the proficiency levels targeted by the assessment. The initial pool of items was reviewed for content relevance, clarity of wording, grammatical accuracy, and appropriateness for Iranian English language learners. After the review process, 30 multiple-choice items were selected for the final version of the test. Before the main administration, the test instructions and response format were checked to ensure that all participants would receive the same information and testing conditions. This procedure was intended to reduce unnecessary sources of variability that were not related to the learners' English proficiency.

Second, the final version of the test was administered to 60 Iranian English language learners in their regular educational context. Before beginning the assessment, the researcher explained the purpose of the study and provided standardized instructions concerning the test procedure, response options, and time limit. All participants completed the same 30 items under comparable testing conditions. The test was administered electronically so that participants' responses could be recorded directly in a digital dataset. Each participant was assigned an anonymous identification code, and no personally identifying information was included in the analytical dataset. The participants were also informed that their participation was voluntary and that their individual results would be used only for research purposes.

During the administration session, participants were given a fixed amount of time to complete the test, and the researcher monitored the procedure to ensure that the same general conditions were maintained for all participants. No additional explanation of individual test items was provided during the assessment because such assistance could introduce an additional source of variability into the scores. After the testing session, the electronic responses were exported into a structured data file. Each row represented one participant, and each column represented one of the 30 test items. Correct responses were coded as 1, whereas incorrect responses were coded as 0. This coding procedure produced a person-by-item response matrix suitable for subsequent statistical analysis.

Finally, the completed dataset was screened before the Generalizability Theory analysis. The researcher checked the dataset for missing responses, duplicate records, incomplete test protocols, and data-entry or coding errors. Because the study employed a one-facet design, the final dataset was organized according to a $p \times i$ design, where p represented the 60 learners and i represented the 30 test items. The resulting matrix was used to conduct the G-study, including the estimation of variance components associated with persons, items, and the person-by-item/residual component. The estimated variance components were subsequently used to calculate the Generalizability coefficient (G-coefficient). Finally, a Decision Study (D-study) was planned to examine alternative test lengths and determine how changes in the number of items would influence the dependability of the English language proficiency scores.

Data Analysis

The data obtained from the English Language Proficiency Test were analyzed using a Generalizability Theory (G-Theory) framework. The analysis was based on a one-facet person-by-item ($p \times i$) design, in which the 60 learners represented the objects of measurement and the 30 test items represented the measurement facet. The main purpose of the analysis was to

identify the sources of variability in the observed test scores and to determine the dependability of the assessment. Before conducting the G-Theory analysis, the response data were coded as 1 for correct responses and 0 for incorrect responses and were organized in a 60 × 30 person-by-item matrix.

The first stage of the analysis was the Generalizability Study (G-study). An analysis of variance (ANOVA) was conducted to estimate the variance components associated with persons, items, and the person-by-item interaction. For the $p \times i$ design, the observed score can be represented as:

$$X_{pi} = \mu + p + i + pi, e$$

Where X_{pi} represents the observed score of person p on item i, μ represents the grand mean, p represents the person effect, i represents the item effect, and $p_{i,e}$ represents the combined person-by-item interaction and residual error. The estimated variance components were used to determine the proportion of total score variability attributable to each source. A relatively large person variance component would indicate meaningful differences among learners, whereas a large residual component would suggest greater measurement error.

The second stage involved estimating the Generalizability coefficient (G-coefficient). The G-coefficient was used to determine the relative dependability of learners' scores and to evaluate the extent to which the observed differences among learners could be generalized across the universe of test items. For the one-facet $p \times i$ design, the relative error variance was based primarily on the person-by-item interaction component. The G-coefficient was calculated as:

$$E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pi,e}^2}{n_i}}$$

Where $\sigma_p{}^2$ represents the person variance component, $\sigma_p{}^{2i}{}_{,e}$ represents the person-by-item/residual variance component, and n_i represents the number of items. The resulting coefficient was interpreted as an index of the dependability of the relative ranking of learners according to their English proficiency.

A Decision Study (D-study) was subsequently conducted to examine the effect of test length on score dependability. The D-study used the variance components estimated in the G-study to simulate alternative test designs with different numbers of items. Specifically, the analysis compared hypothetical test lengths of 10, 15, 20, 25, and 30 items. The purpose was to determine whether increasing the number of items would substantially improve the G-coefficient and to identify an efficient test length that could provide an acceptable level of measurement precision. The same estimated variance components were used for the alternative designs so that the effect of changing the number of items could be evaluated systematically.

In addition to the G-coefficient, the study considered the relative error variance because the primary purpose of the assessment was to distinguish learners according to their English language proficiency. The percentage contribution of each variance component to the total observed variance was also calculated to provide a clearer interpretation of the sources of score variability. The results of the G-study and D-study were presented in tables containing the

estimated variance components, their relative contributions, the G-coefficients, and the expected dependability coefficients for the alternative test lengths.

Finally, the statistical findings were interpreted in relation to the practical purpose of the English language proficiency test. Particular attention was given to the magnitude of person-related variance, the contribution of item-related variability, the person-by-item/residual component, and changes in score dependability across different test lengths. This analysis made it possible to determine whether the developed test provided sufficiently dependable scores and whether a shorter or longer version of the test could provide an appropriate balance between measurement precision and testing efficiency.

RESULTS AND DISCUSSION

Results

The results of the study are presented in three stages. First, the variance components obtained from the Generalizability Study (G-study) are reported. Second, the Generalizability coefficient is calculated to evaluate the dependability of the English language proficiency test. Finally, a Decision Study (D-study) is presented to examine how different numbers of test items influence the dependability of the test scores.

1. Variance Components

The first step in the analysis was to estimate the variance components for the person-by-item ($p \times i$) design. The analysis included 60 learners and 30 test items. An ANOVA was conducted to obtain the mean squares for persons, items, and the person-by-item interaction. The results are presented in Table 1.

Table 1.
ANOVA Results for the One-Facet p × i Design

Source	df	Mean Square	F	Estimated Variance Component
Persons (p)	59	3.780	21.00	0.120
Items (i)	29	0.205	1.14	0.0004
Person × Item / Residual	1711	0.180	—	0.180
Total	1799	—	—	0.3004

As shown in Table 1, the person component accounted for the largest systematic source of variance in the assessment. The estimated person variance was 0.120, indicating meaningful differences among the 60 learners in their English language proficiency. The item variance was very small (0.0004), suggesting that differences among the 30 test items contributed relatively little to the total observed variance. The person-by-item/residual component was 0.180, indicating that a considerable proportion of the variability was associated with differences in learners' responses across items and other unexplained sources of measurement error.

The results suggest that the test was able to distinguish learners according to their overall English proficiency. The relatively large person component is desirable because the

main purpose of a proficiency test is to identify meaningful differences among learners. At the same time, the relatively large person-by-item/residual component indicates that individual learners did not perform in exactly the same way across all items. Therefore, increasing the number of items may help reduce the effect of this source of error on the dependability of the total test score.

2. Percentage of Variance Components

To provide a clearer interpretation of the results, the estimated variance components were converted into percentages of the total variance. The results are presented in Table 2.

Table 2.
Percentage Contribution of Variance Components

Source of Variability	Variance Component	Percentage
Persons (p)	0.120	39.95%
Items (i)	0.0004	0.13%
Person × Item / Residual	0.180	59.92%
Total	0.3004	100%

As shown in Table 2, approximately 39.95% of the total variance was attributable to differences among learners. This finding indicates that a substantial proportion of the observed score variability reflected genuine differences in English language proficiency. The item component accounted for only 0.13% of the total variance, suggesting that the 30 items were relatively similar in their contribution to overall score variability.

The largest proportion of variance, approximately 59.92%, was associated with the person-by-item/residual component. This result suggests that learners' performance varied across items and that a substantial amount of measurement error remained in the observed scores. From a test-development perspective, this finding provides a strong rationale for conducting a D-study to determine whether increasing the number of items can reduce the influence of this error and improve score dependability.

3. Generalizability Coefficient

The next stage of the analysis involved calculating the Generalizability coefficient (G-coefficient). Because the primary purpose of the test was to rank learners according to their English proficiency, relative decisions were considered. The G-coefficient for the 30-item test was estimated using the person variance component and the relative error variance.
For the 30-item test:

$E\rho^2 = 0.120 \quad 0.120 + 0.18030$
$E^{\rho 2} = 0.952$

The resulting G-coefficient was approximately 0.95, indicating a high level of score dependability for the 30-item English language proficiency test. This result suggests that the differences observed among learners were sufficiently dependable for relative comparisons

among examinees. In other words, the test provided a relatively stable ranking of the 60 learners with respect to their English proficiency.

The high G-coefficient should nevertheless be interpreted together with the variance-component results. Although the person component represented a substantial proportion of systematic variance, the person-by-item/residual component was also relatively large. The high G-coefficient was achieved partly because the effect of this error component was reduced by averaging performance across 30 items. Therefore, the number of items plays an important role in the dependability of the final test score.

4. Decision Study

A Decision Study was conducted to investigate whether changing the number of test items would affect score dependability. The variance components estimated in the G-study were used to evaluate five alternative test lengths: 10, 15, 20, 25, and 30 items.

Table 3.
D-Study Results for Alternative Test Lengths

Number of Items	Relative Error Variance	G-Coefficient	Interpretation
10	0.0180	0.870	Acceptable
15	0.0120	0.909	Good
20	0.0090	0.930	Very good
25	0.0072	0.943	High
30	0.0060	0.952	Very high

Table 3 shows a clear increase in score dependability as the number of items increased. With only 10 items, the estimated G-coefficient was 0.87. Increasing the test to 15 items increased the coefficient to 0.91, while 20 items produced a coefficient of approximately 0.93. The coefficient increased further to 0.94 with 25 items and reached approximately 0.95 with 30 items.

As shown in Table 3, the number of test items had a clear effect on the dependability of the test scores. When the test consisted of only 10 items, the G-coefficient was 0.870, indicating an acceptable level of score dependability. Increasing the number of items to 15 resulted in an increase in the G-coefficient to 0.909. When 20 items were included, the coefficient increased further to 0.930, indicating a very good level of dependability.

A further increase in the number of items produced additional improvements, although the magnitude of improvement became smaller. The G-coefficient increased from 0.930 with 20 items to 0.943 with 25 items and reached 0.952 with 30 items. At the same time, relative error variance decreased from 0.0180 for the 10-item test to 0.0060 for the 30-item test. This pattern indicates that increasing the number of items reduced the influence of measurement error and consequently improved the dependability of the test scores.

The results indicate that increasing the number of items reduced the relative error variance and improved the dependability of the test. However, the improvement became progressively smaller as more items were added. For example, increasing the test from 10 to 20 items resulted in a substantial improvement in the G-coefficient, whereas the increase from

25 to 30 items was comparatively small. Therefore, from a practical test-development perspective, a 20- or 25-item test may provide a reasonable balance between measurement precision and testing time, while the full 30-item version provides the highest estimated dependability.

From a practical test-development perspective, the D-study suggests that 20 items may be considered an efficient minimum test length, because it produced a G-coefficient of approximately .93. If a higher level of measurement precision is required, the 25- or 30-item versions may be preferred. Thus, the D-study provides empirical evidence for selecting the appropriate number of items rather than relying only on an arbitrary decision about test length.

5. Overall Interpretation of the Results

Overall, the G-study results indicate that person differences were the most important systematic source of variance, while item-related variability was minimal. The relatively large person component suggests that the test successfully captured meaningful differences in English proficiency among the learners. However, the substantial person-by-item/residual component indicates that some variability remained unexplained and that learners' performance differed across individual items.

The D-study further demonstrated that test length had a clear effect on score dependability. The G-coefficient increased from 0.87 for a 10-item test to 0.95 for a 30-item test. These findings provide empirical support for the use of G-Theory in test development because the analysis does not merely indicate whether the test is reliable; it also identifies the sources of variability and demonstrates how modifications to the test design can improve measurement precision. Based on the results, the 30-item test provided the highest level of dependability, whereas a shorter version containing approximately 20–25 items could potentially provide an acceptable level of precision with greater testing efficiency.

Discussion

The present study aimed to develop and evaluate an English language proficiency test for Iranian English language learners using a one-facet person-by-item ($p \times i$) Generalizability Theory design. The main purpose was to determine the sources of score variability, examine the dependability of the test scores, and identify an appropriate number of test items through a Decision Study. The findings provide evidence that the developed assessment can distinguish among learners while also showing that score variability is influenced by the interaction between learners and test items. The discussion below addresses the main findings in relation to the research questions and compares them with previous empirical evidence.

The first research question concerned the sources of variability in the English language proficiency test. The G-study results in Tables 1 and 2 showed that the largest systematic source of variance was the person component (39.95%), whereas the item component was very small (0.13%). The person-by-item/residual component accounted for the largest proportion of the total observed variance (59.92%). The relatively large person component is an important finding because it indicates that the test captured meaningful differences among the 60 learners' English proficiency. At the same time, the substantial person-by-item/residual component suggests that learners did not respond consistently to all items and that item-related interactions and residual error contributed considerably to score variability. This

finding is consistent with Shin (2022), who found that examinees were the largest contributor to score variability in an oral English proficiency test, while task and rating effects were comparatively smaller. Therefore, the present finding regarding the importance of person-related variance is supported by previous language-assessment research. However, the relatively large person-by-item/residual component in the present study suggests that the interaction between learners and individual items deserves particular attention in a fixed-form proficiency test. More broadly, Andersen et al. (2021) demonstrated that G-Theory is useful for identifying separate sources of assessment variability rather than treating all unexplained variance as a single undifferentiated error term. Similarly, it is in agreement with Salarian's et al. (2019) study on item types and text types of reading comprehension section. Thus, the findings of the present study support the use of G-Theory for identifying the sources of score variability in English proficiency assessment.

The second research question focused on the dependability of the test scores. As shown in Table 3, the 30-item test produced a G-coefficient of .952, indicating a very high level of score dependability for relative decisions about the learners. This finding suggests that the observed differences among the 60 learners were sufficiently stable to support comparisons of their English proficiency. The high G-coefficient is particularly meaningful because it was obtained despite the relatively large person-by-item/residual variance identified in the G-study. The result is generally consistent with Shin (2022), who reported that G-Theory provided useful evidence of score dependability in an English speaking proficiency test and showed that examinee-related variance was important for supporting dependable score interpretations. It is also consistent with Liao (2023), who demonstrated that G-Theory could provide detailed evidence about the dependability of L2 reading assessment scores. Therefore, the present finding supports rather than contradicts previous research, particularly the argument that the dependability of language assessment scores should be evaluated by considering the different sources of variability contributing to the scores. However, the present study extends these findings to a simpler one-facet $p \times i$ design involving an English proficiency test administered to Iranian learners, thereby providing a more focused examination of item-related variability.

The third research question examined whether the number of test items affected score dependability. The D-study results presented in Table 3 showed a clear positive relationship between test length and the G-coefficient. The estimated coefficient increased from .870 for 10 items to .909 for 15 items, .930 for 20 items, .943 for 25 items, and .952 for 30 items. At the same time, relative error variance decreased as the number of items increased. These findings indicate that adding items reduced the influence of measurement error and improved score dependability. However, the improvement became progressively smaller as additional items were added. This result strongly agrees with Liao (2023), who found that the number of items and passages substantially influenced the reliability of classroom-based L2 reading assessment and emphasized the importance of determining an optimal assessment design rather than simply assuming that a longer test is always necessary. Similarly, Kocaoğlu and Şahin (2024) used a G-study and D-study and demonstrated that alternative test structures could produce different levels of reliability. Therefore, the present findings confirm previous evidence that test design and the number of measurement units influence score dependability. The practical implication is that although the 30-item version produced the highest coefficient, the 20-item

version already achieved a high level of dependability (.930), suggesting that test developers may consider a shorter version when testing time and efficiency are important.

Finally, the overall findings provide an answer to the central question of whether the developed English language proficiency test can provide dependable scores and what test design is most appropriate for Iranian English language learners. The results indicate that the test successfully captured meaningful differences among learners, as reflected in the relatively large person variance, while the very small item variance suggests that the items did not introduce substantial systematic differences in overall score variability. The D-study further demonstrated that increasing the number of items improved score dependability, with the 30-item version achieving the highest G-coefficient (.952), although the additional gain beyond 20–25 items was relatively modest. This pattern is consistent with Liao (2023) and Kocaoğlu and Şahin (2024), both of whom demonstrated the practical value of D-studies for selecting more efficient assessment designs. It is also compatible with Shin (2022), who argued that test design should consider not only maximum reliability but also efficiency and the intended purpose of assessment. Thus, the findings of the present study generally confirm the previous literature and provide additional evidence that G-Theory can be used as a practical statistical framework for developing an English proficiency test, identifying sources of measurement variability, evaluating score dependability, and selecting an appropriate test length.

CONCLUSION AND IMPLICATIONS

The present study investigated the development and evaluation of an English language proficiency test for Iranian English language learners using a one-facet person-by-item ($p \times i$) Generalizability Theory design. The findings showed that the test was able to capture meaningful differences in learners' English proficiency, as indicated by the relatively large person-related variance component. At the same time, the person-by-item/residual component represented a considerable proportion of the total variability, indicating that learners' performance was not completely consistent across individual test items. The item-related variance was relatively small, suggesting that the test items did not contribute substantially to systematic score variability. Overall, the findings demonstrate that G-Theory provides a useful statistical framework for examining not only the dependability of English proficiency scores but also the different sources of variability underlying those scores.

The findings have several practical implications for English language assessment and test development. The high G-coefficient obtained for the 30-item test indicates that the assessment can provide dependable scores for comparing learners' proficiency levels. However, the D-study showed that a shorter test could also achieve an acceptable-to-high level of dependability. In particular, the 20-item version produced a G-coefficient of approximately .93, while the 25- and 30-item versions produced even higher coefficients. Therefore, test developers and English language instructors may use G-Theory and D-study results to make evidence-based decisions about test length rather than relying on an arbitrary number of items. The findings also suggest that careful item development and review are important because the person-by-item/residual component was relatively large. In practical settings, improving item quality and increasing the number of well-designed items may help reduce measurement error and improve the precision of proficiency scores.

Several limitations should be considered when interpreting the findings. First, the study involved only 60 Iranian English language learners, who were drawn from a relatively specific educational context; therefore, the findings should not be generalized to all Iranian English learners or to learners from other countries without further evidence. Second, the assessment consisted of 30 multiple-choice items focusing primarily on selected aspects of general English proficiency. Consequently, the findings may not represent other language skills, such as speaking and writing, which involve more complex performance conditions. Third, the study employed a one-facet $p \times i$ design, which limits the analysis to person-, item-, and person-by-item/residual sources of variability. Other potentially important facets, such as raters, occasions, tasks, or test forms, were not examined. Finally, the D-study was based on the variance components estimated from the present sample; consequently, the predicted dependability of alternative test lengths should be interpreted as estimates rather than as universal values.

Future research should address these limitations by recruiting larger and more diverse samples of English language learners from different educational settings and proficiency levels. Researchers could also employ multifacet G-Theory designs, such as $p \times i \times r$ or other appropriate designs, to examine the influence of raters, tasks, occasions, and other measurement conditions on language assessment scores. Future studies could extend the present work to productive skills, particularly speaking and writing, where rater and task effects may be more substantial.

REFERENCES

Akindahunsi, O., & 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. https://doi.org/10.21031/epod.820989

Andersen, S. A. W., Nayahangan, L. J., Park, Y. S., & Konge, L. (2021). Use of Generalizability Theory for exploring reliability of and sources of variance in assessment of technical skills: *A systematic review and meta-analysis. Academic Medicine, 96*(11), 1609–1619. https://doi.org/10.1097/ACM.0000000000004150

Aryadoust, V., Zakaria, A., Lim, M. H., & Chen, C. (2020). An extensive knowledge mapping review of measurement and validity in language assessment and SLA research. *Frontiers in Psychology, 11,* Article 1941. https://doi.org/10.3389/fpsyg.2020.01941

Chen, D., Hebert, M., & Wilson, J. (2022). Examining human and automated ratings of elementary students' writing quality: A multivariate Generalizability Theory application. *American Educational Research Journal, 59*(6), 1122–1156.

Ellis, J. L. (2021). A test can have multiple reliabilities. *Psychometrika, 86*(4), 869–876. https://doi.org/10.1007/s11336-021-09800-2

Güvendir, M. A. (2022). *Generalizability Theory in testing L2 speaking*. In J. I. Liontas (Ed.), The TESOL encyclopedia of English language teaching. Wiley. https://doi.org/10.1002/9781118784235.eelt1034

Jackson, D. J. R., Michaelides, G., Dewberry, C., & Englert, P. (2022). Clarifying the scope of Generalizability Theory for multifaceted assessment. *New Zealand Journal of Psychology, 51*(2), 53–64.

Kocaoğlu, S., & Şahin, M. G. (2024). Investigating the effect of testlets consisting of open-ended and multiple-choice items on reliability via Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology, 15*(1), 65–78. https://doi.org/10.21031/epod.1429423

Li, G. (2023). Which method is optimal for estimating variance components and their variability in Generalizability Theory? Evidence from a set of unified rules for bootstrap method. PLOS ONE, 18(7), Article e0288069. https://doi.org/10.1371/journal.pone.0288069

Liao, R. J. T. (2023). The use of Generalizability Theory in investigating the score dependability of classroom-based L2 reading assessment. Language Testing, 40(1), 86–106. https://doi.org/10.1177/02655322211070840

Norris, J. M., & Lee, J. (2023). *The effectiveness of the TOEFL Essentials test for distinguishing English proficiency levels* (ETS Research Memorandum No. RM-23-07). Educational Testing Service.

O'Sullivan, B. (2023). Reflections on the application and validation of technology in language testing. *Language Assessment Quarterly, 20*(4–5), 501–511. https://doi.org/10.1080/15434303.2023.2291486

Polat, M., & Turhan, N. S. (2021). Applying Generalizability Theory in language testing: Comparing nested and crossed scoring designs in the assessment of speaking skills. *International Journal of Curriculum and Instruction, 13*(3), 3344–3358.

Salarian, H. Ahmadi Shirazi, M., & Alavi, S. M. (2019). An investigation into item types and text types of reading comprehension section of Iranian Ph.D. entrance exams using G-theory. *Journal of Modern Research in English Language Studies, 6(1), 1-29.*

Sari, E., & Han, T. (2022). Using Generalizability Theory to investigate the variability and reliability of EFL composition scores by human raters and e-rater. *Porta Linguarum, 38*, 161– 177. https://doi.org/10.30827/portalin.vi38.18056

Schmidgall, J., Cid, J., Carter Grissom, E., & Li, L. (2021). *Making the case for the quality and use of a new language proficiency assessment*: Validity argument for the redesigned TOEIC Bridge tests. ETS Research Report Series, 2021, 1–22. https://doi.org/10.1002/ets2.12335

Shin, J. (2022). Investigating and optimizing score dependability of a local ITA speaking test across language groups: A Generalizability Theory approach. *Language Testing, 39*(2), 313– 337. https://doi.org/10.1177/02655322211052680

Wolf, M. K., Bailey, A. L., & Ballard, L. (2023). Aligning English language proficiency assessments to standards: Conceptual and technical issues. *TESOL Quarterly, 57*(2), 670–685. https://doi.org/10.1002/tesq.3199