International Journal of English Linguistics; Vol. 9, No. 3; 2019
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

Using an Estimate of Language Ability for Making Pass-or-Fail Decisions at an Intensive English Program in Saudi Arabia

Mohammed Shuaib Assiri[1]

[1] King Khalid University, Saudi Arabia

Correspondence: Mohammed Shuaib Assiri, King Khalid University, Saudi Arabia. E-mail: msasiri@kku.edu.sa

Received: March 22, 2019 Accepted: April 20, 2019 Online Published: May 20, 2019

doi:10.5539/ijel.v9n3p347 URL: https://doi.org/10.5539/ijel.v9n3p347

Abstract

The pass-or-fail decisions at an intensive English program in Saudi Arabia are often based on assumptions as to whether the learner has passed in all language skills. For instance; if a learner fails in one skill, he is treated as if he failed in all skills. Scores that sum up skill scores or average them out are marginalized in the making of a pass-or-fail decision. Learners who fail in one or two skills, usually have to repeat the whole course of study at the levels they were attending. Hence, the current study aims to prove the adequacy of reporting total average scores along with individual skill scores and using them to decide whether a learner should pass or fail. It employed score data from 644 learners' score reports at an intensive English program in Saudi Arabia. The results of factor analysis, linear regression, and correlation tests revealed that a total average score could serve both as an accurate estimate of language ability and as a basis on which a pass-or-fail decision could best be made. The study report concludes with practical implications that can go hand in hand with the implementation of such a research finding.

Keywords: assessment fairness, language ability, pass-or-fail decision, score reporting, the four skills

1. Introduction

In an intensive English program that is part of a public institute in Saudi Arabia, learners study at four levels of language proficiency: preparatory, elementary, intermediate, and advanced. At each level, they study five language skills: listening, oral, reading, writing, and grammar for a period of two months, referred to as a session. For each skill, learners are assessed by means of two short quizzes, one midterm test, and a final exam. Besides such modes of summative assessment, learners are assessed using formative modes that primarily comprise homework and participation. At the time of reporting learners' grades at the end of a session, a score report is issued for each learner in the form of a table incorporating four skill scores. To be exact, a score out of 100 is assigned to each of the reading, writing, and grammar skills in addition to a combination of the oral and listening skills (60 & 40 respectively). As such, these skills are labeled as the four-skill components throughout this report. This is to say that a score that sums up skill scores or averages them out is not included in the score report.

Because learners at the given language program generally exhibit varying levels of performance across the aforementioned language skills, there are learners who pass all skills and others who fail in one or more skills. A learner who fails in one skill is no different from a learner who fails in all skills. If he is eligible for another chance to repeat the same course of study he was taking, he can definitely benefit from such a chance; otherwise, he will be dismissed from the language program. Until recently, instructors who taught the five skills to the same group of learners would meet and decide whether or not to fail a learner whose score(s) in one or two of the four-skill components fell within a range of 50 to 60, with 60 being the cutoff score. Nowadays, such decisions can be made by individual instructors. The decision about the status of a learner whose scores are below the cutoff score is both crucial and consequential. There are learners who pass all skills and others who fail in one or more skills. Learners generally exhibit varying levels of performance across the aforementioned language skills. The fact that most EFL learners in Saudi Arabia achieve low levels of language proficiency might be the outcome of different factors, including the influence of L1 and low levels of motivation (Alrabai, 2016).

Given this state of affairs, basing a pass-or-fail decision on individual skill scores may be perceived as both unfair and unethical, at least among the failed learners and their relatives, which is absolutely against the premises of the program's mission. This is very important because testing experts (e.g., Banerjee, 2017; Chalhoub-Deville, 2015; Kane, 2013; Purpura, Brown, & Schoonen, 2015) are all in agreement that assessment

fairness is reflected by, besides facets of validity and reliability, the interpretations assigned to test scores, the formats used in reporting them, and the decisions made on their basis. Language assessors are accountable for providing all necessary information that justifies any decisions to be made on the basis of the assessments that have constructed (Bachman, 2007). Of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), *Standard 2.1.* states that "[F]or each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported" (p. 31). Evidently, such a standard points out the importance of adequate and full reporting of assessment scores or results.

In high-stake language testing systems such as the TOEFL and the IELTS, score reports include a total score along with skill scores. Admission decisions for academic programs in English-speaking countries are usually based on a whole score as rendered by the standardized language proficiency test. Many publishers of language tests consider a score that sums up all the scores received by the test taker for his performance on sub-tests that are used for assessing language skills as an estimate of his language ability (Pearson Education, 2012). Educational Testing Service (see Baldwin, Fowles, & Livingston, 2005) as well as testing experts from the fields of educational measurement and evaluation (e.g., Ferrara & Way, 2016) advocate the use of total scores when making pass-or-fail decisions, which is often markedly associated with making sound judgments. In the context of an intensive English program in Saudi Arabia, the current study aims to prove the adequacy of reporting a total average score along with individual skill scores, for both estimating language ability and informing pass-or-fail decision making.

2. Literature Review

In most intensive English programs across the globe, there has been a change towards an integrated mode of language testing. Even on the world's leading standardized tests, certain skills are integrated in order to offer an authentic measure of language ability. Integrated testing of language skills is increasingly being used more than testing that focuses on language skills on an individual basis (Powers, 2010). This stresses the fact that, in any language assessment endeavor, accounting for language ability in the form of a whole or average score is indispensable.

Our reporting of test scores can benefit from the controversy over the nature of language ability; that is, is it unitary or divisible? As the initiator of the unitary view of language ability, Oller (1976) found research evidence suggesting that language ability was ideally represented by a general, higher-order factor and more specific, first-order factors. Such a significant finding has gained support from subsequent research efforts that applied factor analyses to scores obtained from a variety of language tests (e.g., Bachman et al., 1995; Carroll, 1983; Fouly, Bachman & Cziko, 1990; Oller, 1979; Scholz et al., 1980; Shin, 2005).

On the other hand, other researchers (e.g., Carroll, 1975; Gardner & Lambert, 1965; Pimsleur, Stockwell, & Comrey, 1962) have found evidence for the divisible nature of language ability. Across these studies, the results of factor analyses showed that language ability was represented by two or more factors. Therefore, in addition to the unitary hypothesis, a divisible hypothesis of language ability was proposed (Oller & Hinofotis, 1980). The divisible view of language ability has been labelled as the multidimensional view by Carrol (1965). It is worth noting that the research results that gave rise to the two contrasting views of language ability were attributed to the use of different factor analytical procedures, specifically principal component analysis versus principal factor analysis (see Gu, 2011). Taken both the unitary and divisible views in consideration while reporting test scores implies that a total score for the whole test as well as scores for the sections of the test that assess language skills should be reported. A total score can be replaced with a score that averages out the skill scores, which is simpler and more comparable to skill scores.

Research on large-scale tests (e.g., Bozorgian, 2012), including the TOEFL and the IELTS, has shown that scores on test sections that measure language skills were positively correlated and that the score on each sectionwas positively correlated with the overall test score. These findings suggest that on tests composed of the four skills, language ability draws on each skill and that a score that conveys overall test performance can serve as an estimate of language ability. Other studies (e.g., Gu, 2015; Sawaki, Stricker, & Oranje, 2009; Stricker & Rock, 2008) that examined test performance on the TOEFL have reported the same factor structure as proposed by Oller (1976). That is, across these studies, a major finding was that test performance represented a higher-order factor structure with first-order factors. The higher-order factor signified language ability while other first-order factors were for language skills.

The inclusion of skill scores in a score report is also necessary for informing decision-making efforts. It helps spot any weaknesses in specific language skills that may benefit from a remedial procedure later on. Test takers and score users desire to see which language skills are characterized by high versus low test performance, and typically, the extent of overall language development or attainment. Based on a survey of the TOEIC examinees' readiness to perform a variety of language tasks reflecting everyday activities at workplaces, Powers et al. (2009) found that each of the four skills formed an essential component of language ability. Another study examined the relationship between the four skills that make up the TOEIC and proposed that each skill assessed a certain component of language ability that could not be adequately assessed by another skill (Liu & Costanzo, 2013).

It is interesting to note that in another strand of research, it was found that assessing certain skills by the TOEIC could also assess other skills, but in an indirect fashion. For example, Wilson (1993) noticed that TOEIC test takers' scores on the listening and reading tests furnished an ancillary measure of their speaking skills that were assessed by means of proficiency interviews. However, such a fact should not lead us to limit proficiency assessment to two or three skills because each skill is a necessary element of language ability, as Liu and Costanzo's (2013) finding suggests. Taken all together, the findings from previous research clearly suggest that adequate assessment of language ability should incorporate the four skills (i.e., listening, speaking, reading, and writing) and that a total score ought to be reported in addition to individual skill scores.

The current study aims to prove the adequacy of reporting a total average score, in addition to scores of the individual skills, for both estimating language ability and informing pass-or-fail decision making. Accordingly, the following research questions were formulated:

- Would a higher-order factor structure best fit the score data?
- Would average scores of the four-skill components predict language ability?

3. Methods

Given its specific context, the current study made use of data in the form of total scores that learners attending the language program had attained throughout a whole session in the five skills. Each of the grammar, reading, and writing skills was assigned a score out of 100, and so was a combination of the oral and listening skills. To reiterate, since a learner's score report includes each of grammar, reading, and writing and both of oral and listening combined, all skills are referred to as the four-skill components in this report. The total number of learners whose scores were used in the study was 644, distributed among the four levels of the language program: 71 preparatory, 169 elementary, 80 intermediate, and 324 advanced students.

The scores were obtained with the permission of the program director and its coordinator. During data collection and analysis of the study, score data and any learner-related information were treated as private and confidential. In data preparation, the total scores of the four-skill components for each learner were averaged out to get a value out of 100, referred to as a total average score. It was assumed that the four-skill components comprise a whole test of language ability and that a total average score would represent the overall test performance for a given learner.

4. Results

To answer the first research question, score data were examined for their fit in a factor structure using an exploratory factor analysis (EFA). Such an analysis made use of the principal component analysis as the extraction method, and Varimax as the rotation method. Factor scores were generated using the regression method. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .724, above the commonly recommended value of .6, and the Bartlett test of sphericity was significant, $\chi^2(6) = 355.385$, $p = .000$. Such results suggested that the sample size of the study was adequate and that the EFA was an appropriate analytical procedure (see Child, 2006). Table 1 below shows that the resulting commonalities from the EFA procedure, which refer to the proportions of variance in the skill components that could be explained by language ability, were of proper values for a factor analysis to be pursued.

Table 1. EFA communalities

Skill component	Initial	Extraction
Grammar	1.000	.528
Writing	1.000	.471
Reading	1.000	.541
Oral and Listening	1.000	.477

Based on the EFA results, score data could best be accounted for by a unidimensional solution. The estimates of the total variance explained indicated that only one factor (or latent variable) was extracted or retained (Table 2). Such a factor can presumably be referred to as language ability. No estimates of sums of squared loadings associated with factor rotations are reported here because only one factor was extracted.

Table 2. Total variance explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.016	50.400	50.400	2.016	50.400	50.400
2	.744	18.606	69.006			
3	.658	16.456	85.462			
4	.582	14.538	100.000			

The unidimensional factor solution was also supported by a scree plot (Figure 1) which makes it clear that out of the four components in Table 2, only one factor (Component 1) with an eigenvalue above one could be retained. That is, the Guttman rule was applied when deciding about the number of factors to retain (see Guttman, 1954). Components 2, 3, and 4 had eigenvalues ranging from .582 to .744.

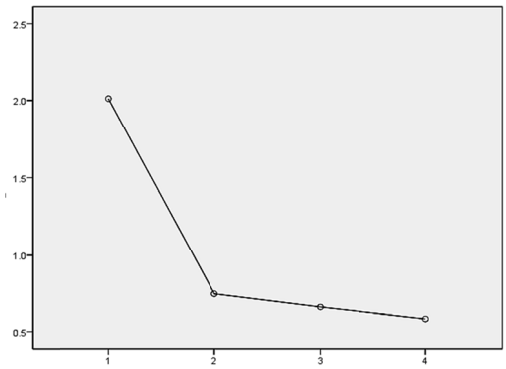

Figure 1. A scree plot of the eigenvalues associated with factor components

As shown in Figure 2, the loadings of the four-skill components on the extracted factor were all above (.5). This suggests that the extracted factor (language ability) contributed substantially to each of the four-skill components. The four-skill components represent the four first-order factors in the factor structure. It is also obvious that the loadings of both grammar and reading were very close as were the loadings of writing and the oral and listening skill component.

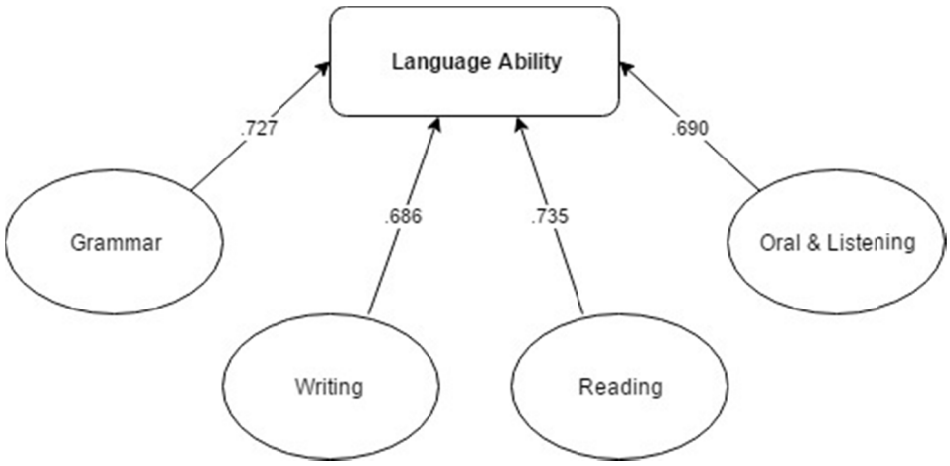

Figure 2. Factor structure of the four-skill components on language ability.

Factor scores that resulted from the EFA procedure were used in a linear regression analysis to answer the second research question. Such an analysis sought to determine the extent to which language ability could predict the learners' total average scores. Factor scores specified the learners' relative positions or placements along the continuum of the latent variable (language ability), ranging from -2.430 to +2.211. Total average scores (TASs) were obtained by adding up a learner's scores across the four-skill components and dividing the sums by four such that the resulting values were each out of 100. It was assumed that a given TAS would be the best estimate of the language ability of the learner who had such a TAS. Therefore, TASs were regressed on factor scores. It was postulated that if TASs were regressed on factor scores, the regression result would indicate the extent to which TASs could be used to estimate language ability.

The fit of the regression model was determined in terms of the correlation between TASs and factor scores, $r(642) = 1.000$, $p = .000$, and the standard error of estimate, $\sigma_{est} = 0.0787$, which is almost negligible. Factor scores explained a significant proportion of variance in TASs, $R^2 = 1.000$, $F(1, 642) = 6325933.240$, $p = .001$. As illustrated in Figure 3, factor scores significantly predicted TASs, $\beta = 1.000$, $t(643) = 2515.141$, $p = .000$. As suggested by the scale to the right side of the graph, 72 learners had TASs from 50.00 to 59.75, 261 learners had TASs from 60.00 to 69.75, 247 learners had TASs from 70.00 to 79.75, and 64 learners had TASs from 80.00 to 86.75.

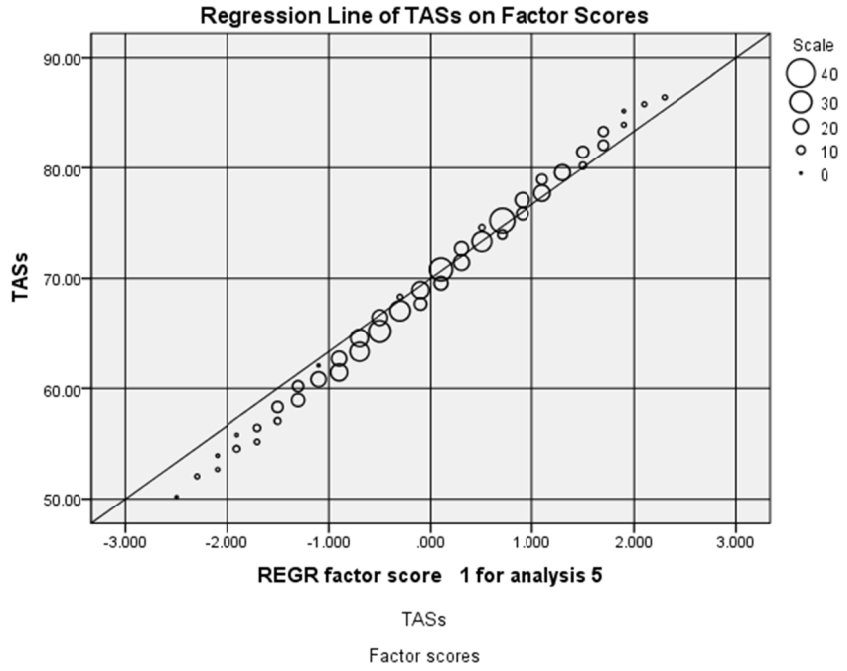

Figure 3. Regression line of TASs on factor scores. The scale on the right side indicates the number of cases

Then, a Pearson correlation analysis was run so as to measure the associations between TASs and the number of failed-skill components (FSCs). The result pointed out a strong, negative correlation between TASs and FSCs, $r(644) = - 0.829$, $p = 0.000$. Figure 4 demonstrates the relationship between TASs and FSCs in a manner suggesting what follows:

- Learners who failed in all four-skill components had a TAS range from 50.50 to 56.75
- Learners who failed in three-skill components had a TAS range from 53.75 to 64.75
- Learners who failed in two-skill components had a TAS range from 56.00 to 68.75
- Learners who failed in one-skill component had a TAS range from 59.50 to 79.50
- Learners who did not fail in any skill component had a TAS range from 62.00 to 86.75

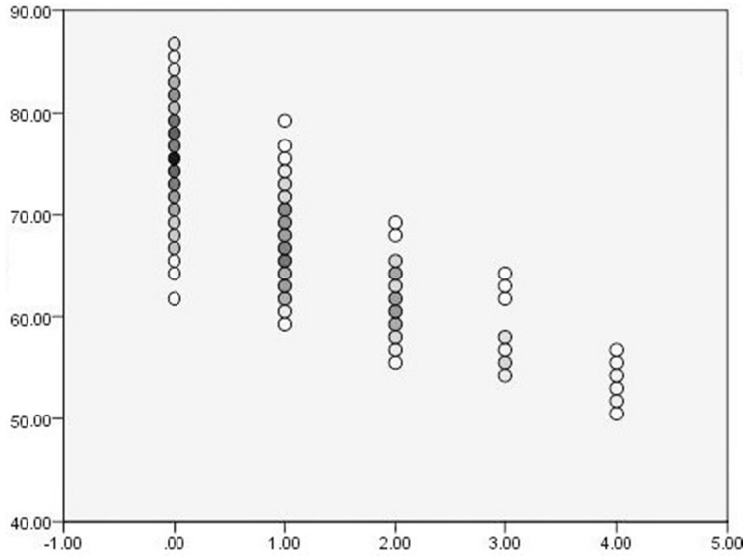

Figure 4. Relationship between TASs and FSCs. The colour levels in the circles reflect the number of cases

In this section, three groups of learners were selected to represent three scoring levels as determined by TASs (Table 3). The scores that the cases earned on the four-skill components, each out of 100, as well as the number of skill components they failed in (FSCs), if any, are included. The selected cases are assigned numbers from 1 to 15 for ease of presentation and discussion. Two pass-or-fail decisions are made about each case. Decision 1 (D1) is informed by the current policy: If a learner fails in one skill, he fails in all of his course of study at his given level. Decision 2 (D2) is made on the basis of whether the learner's TAS is 60 or above, or below, with the assumption that his TAS is an estimate of his language ability.

Table 3. TASs and FSCs of three scoring levels and pass-or-fail decisions

Group	Case	G	W	R	OandL	TAS	FSCs	D1	D2
low	1	50	51	50	51	51	4	Fail	Fail
	2	50	50	56	51	52	4	Fail	Fail
	3	50	51	57	50	52	4	Fail	Fail
	4	55	52	51	50	52	4	Fail	Fail
	5	52	50	50	57	52	4	Fail	Fail
middle	6	55	59	86	64	66	2	Fail	Pass
	7	71	67	68	58	66	1	Fail	Pass
	8	62	80	59	64	66	1	Fail	Pass
	9	71	68	56	70	66	1	Fail	Pass
	10	67	69	60	69	66	0	Pass	Pass
high	11	90	90	86	79	86	0	Pass	Pass
	12	87	80	90	89	87	0	Pass	Pass
	13	84	87	88	88	87	0	Pass	Pass
	14	80	90	89	88	87	0	Pass	Pass
	15	81	86	90	90	87	0	Pass	Pass

Note. G = grammar, W = writing, R = reading, and OandL = oral and listening.

First, a look at the first group of cases, or the low-scoring group, suggests that the two decisions correlate and that each case should fail. That is because across the five cases in this group, TASs are below 60 and FSCs are 4 each. The score range of this group across the four-skill components is (50–57). This, in effect, demands that each case repeat his course of study at the level he was attending.

Next, a look at the middle-scoring group reveals that each one of the first four cases deserves a fail (Decision 1) because each one fails in at least one-skill component although the cases' TASs are all above 60. However, for Decision (2), each one of the five cases in this group deserves a pass because their TASs are above 60. For

instance, although Case 6 fails in two-skill components, he can still pass because his TAS is 6 points above the cut-off score. The score range of this group across the four-skill components is (55–86) which is noticeably higher than that of the first group.

Last, a look at the five cases in the high-scoring group shows that each case deserves a pass according to both decisions. In satisfaction of Decision (1), none of the cases in this group fails in any skill component; and for Decision (2), the cases in this group have TASs that are 27 points on average above the cut-off score. The score range of this group across the four skill components is (79–90) which is remarkably higher than those of the two other groups.

5. Discussion

The EFA results in this study were consistent with those of the previous studies that explored the nature of language ability on the basis of skill-based test performances (incl., the TOEFL, IELTS, TOEIC … etc.). For example, Bachman et al. (1995) identified a factor structure for language ability in the shape of a higher-order model with three distinct first-order factors that involved speaking, listening, and writing skills. Fouly, Bachman, and Cziko (1990) found that language ability comprised a higher-order model with three first-order factors that consisted of structure and reading skills, oral and aural skills, and discourse skills. In two other studies, Stricker and Rock (2008) and Sawaki, Stricker and Oranje (2009) found a higher-order factor structure for the TOEFL iBT with first-order factors representing the four skills (i.e., listening, speaking, reading, and writing). Also, Gu (2014) observed a factor structure of language ability among young EFL learners that was made up of four first-order factors corresponding to the four skills subsumed under a higher-order factor denoting language ability.

The fact that grammar and reading had nearly identical loadings on language ability in this study is supported by the finding from a study by Sang et al. (1986), according to which grammar and reading skills made up one of three main elements in a three-dimensional model of language ability. On the other hand, the fact that writing and the oral and listening skill component had almost the same loadings on language ability is supported by the finding from Powers et al.'s (2009) study in which test takers' scores on both of the TOEIC speaking and writing tests equally predicted their levels of language ability. Also, in Bozorgian's (2012) study, scores on the TOEIC writing, listening, and speaking skills had strong, positive correlations with one another. Therefore, in this given context of English learning, a learner's performance on tests that measure the four-skill components can be considered an indication of his language ability. In view of that, the relationship between language ability and the four-skill components can be represented using a factor structure with a higher-order factor and four first-order factors. Such a factor model suggests that a learner's language ability draws on the four-skill components in a balanced fashion.

The results of the linear regression analysis suggested that language ability significantly predicted test performances on the four-skill components, as measured by TASs. This finding is obvious evidence that a learner's TAS can provide an adequate estimate of his language ability. Also, it endorses the findings from the factor analysis in the answer to the first research question. That is, a learner's overall performance on the tests assessing the four-skill components, as measured by TASs, can be considered a reflection of his language ability. Such converging evidence implies that if a learner's test performances on the four-skill components are high, his language ability must be high enough to allow for such high test performances, and the reverse is true. Because of the methodological complexities of obtaining factor scores to represent language ability, the staff in charge of test administration at the program can use TASs instead. Moreover, factor scores typically have minus or negative values, which renders them inadequate for score reporting and pass-or-fail decision making.

The findings from the current study point to the suitability of using a learner's TAS when deciding whether he should pass or fail in his course of study at a given level of the program. In so doing, there is little room that he will fail because of a score that is 1 to 10 points below the cut-off score in one-skill component. Obviously, the higher a learner's TAS, the less the likelihood that he failed in any skill component, and the opposite is true. This finding furnishes *prima facie* evidence in support of the findings associated with the first research question; to be exact, the learners' TASs could serve as accurate estimates of their levels of language ability. Of course, if a learner has scores in two or more skills that are below the cutoff score within the range from 1 to 10, he is likely to end up having a failing TAS. Decision makers can use a learner's TAS to assign him a pass-or-fail status in a manner that highly correlates with his level of language ability. To reiterate, assessment fairness is violated when a leaner is coerced to repeat a whole course of study involving all skills at his given level because he failed in one- or two-skill components.

6. Conclusions

The current practice at the intensive English program where this study was pursued requires learners to repeat a whole course of study at their given levels for failing in one or two skills. However, such a practice has had undesirable consequences over the years. The failing learners who do not have any more chance of repeating will have to leave the program. The repeating learners, on the other hand, will definitely experience high levels of boredom for having to study the same language skills again. There is a noticeable tendency among the repeating learners to exhibit conspicuous mental or, even worse, physical absence. There are incidents of troublemaking behaviours on the part of repeating learners towards their teachers and classmates. Furthermore, there is still a possibility that repeating learners may fail in skills that they passed in previous sessions.

Language ability can be seen as the outcome of developing language skills as well as the determinant of how successful a learner can be in his pursuit of a native-like mastery of the target language. Therefore, in addition to skill scores, a total average score for each learner should be reported. This approach of score reporting helps ensure ample reflection of the learner's language ability in addition to his strengths and weaknesses across various skills. It also helps promote washback in the light of which teachers and learners do not focus on skills on an individual basis; rather, they deal with them in an integrated fashion. This accords with the current trend towards more comprehensive and integrated testing of language skills.

A learner's ability to use the target language effectively can best be measured by assessing all language skills, and not focusing on one or two. This is because language skills share certain sub-skills, for example, vocabulary is an important sub-skill in listening, reading, speaking, and writing. Therefore, reporting the results of all assessments of the four skills by using a total average score can provide a more adequate and valid indicator of the learner's language ability than just using skill scores. Such an assumption gains support from correlational studies (Powers, 2010). Language ability is highly essential to adequate performance in academia and at workplace. Scholars (e.g., Liu & Costanzo, 2013) have stressed the importance of the four-skill paradigm in language learning and assessment so as to embrace the key components of language ability. Consequently, the final outcome of a language learning and assessment experience should take the form of a score that reflects language ability.

Some practical implications that can prove workable alongside the use of total average scores in the pass-or-fail decision making are presented here, with the specific setting of this research in mind. Skill classes at levels two, three, and four of the program should start with at least five-hour reviews (i.e., almost 15% of the total class hours required per session) of what was covered at the previous level with respect to skill-related basics and abilities. This is intended to help learners who scored below 60 in certain skills at the previous level have proper amounts of remediation, and so, be able to match up their superior classmates in terms of competency across language skills. Towards such a goal, instructors should be made aware of any learners whose language skills are below average and ways of how to mentor and assist them. Averaging up scores that are based on performance on a variety of assessment formats can lead to trustworthy decisions as to whether or not a learner should repeat a whole course of study at a given level.

References

Alrabai, F. (2016). Factors underlying low achievement of Saudi EFL learners. *International Journal of English Linguistics*, *6*(3), 21–37. https://doi.org/10.5539/ijel.v6n3p21

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bachman, L. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment.

Bachman, L., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge- TOEFL comparability study*. New York, NY: Cambridge University Press.

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed-response and other performance assessments*. Princeton, NJ: Educational Testing Service.

Banerjee, H. (2017). Test fairness in second language assessment. *Working Papers in TESOL and Applied Linguistics*, *16*(1), 54–59.

Bozorgian, H. (2012). Listening skill requires a further look into second/foreign language learning. *ISRN Education*, 1–10. https://doi.org/10.5402/2012/810129

Carroll, J. (1965). Fundamental considerations in testing for English language proficiency of foreign learners. In H. B. Allen (Ed.), *Teaching English as a second language: A book of readings* (pp. 364–372). New York, NY: McGraw-Hill.

Carroll, J. (1975). *The teaching of French as a foreign language in eight countries*. New York, NY: Wiley.

Carroll, J. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 80–107). Rowley, MA: Newbury House.

Chalhoub-Deville, M. (2015). Validity theory: Reform, policies, accountability testing, and consequences. *Language Testing*, 33(4), 453–472. https://doi.org/10.1177/0265532215593312

Child, D. (2006). *The essentials of factor analysis*. London: Continuum.

Ferrara, S., & Way, D. (2016). Design and development of end-of-course tests for learner assessment and teacher evaluation. In H. Braun (Ed.), *Meeting the challenges to measurement in an era of accountability* (pp. 11–48). New York, NY: Routledge.

Fouly, K., Bachman, L., & Cziko, G. (1990). The divisibility of language competence: A confirmatory approach. *Language Learning*, *40*, 1–21. https://doi.org/10.1111/j.1467-1770.1990.tb00952.x

Gardner, R., & Lambert, W. (1965). Language aptitude, intelligence, and second language achievement. *Journal of Educational Psychology*, *56*(4), 191–199. https://doi.org/10.1037/h0022400

Gu, L. (2011). *At the interface between language testing and second language acquisition: Communicative language ability and test-taker characteristics* (Doctoral dissertation). Available from ProQuest dissertation and theses database (Document ID 3461133).

Gu, L. (2014). At the interface between language testing and second language acquisition: Language ability and context of learning. *Language Testing*, *31*(1), 111–133. https://doi.org/10.1177/0265532212469177

Gu, L. (2015). Language ability of young English language learners: Definition, configuration, and implications. *Language Testing*, *32*(1), 21–38. https://doi.org/10.1177/0265532214542670

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161. https://doi.org/10.1007/BF02289162

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Liu, J., & Costanzo, K. (2013). *The relationship among TOEIC listening, reading, speaking, and writing skills*. (TOEIC Compendium Study). Princeton, NJ: Educational Testing Service.

Oller, J. W. (1976). Evidence of a general language proficiency factor: An expectancy grammar. *Die Neuren Sprachen*, *75*(2), 165–174.

Oller, J. W. (1979). The factorial structure of language proficiency: Divisible or not? In J. W. Oller (Ed.), *Language tests at* school*: A pragmatic approach* (pp. 423–458). London: Longman.

Oller, J. W., & Hinofotis, F. (1980). Two mutually exclusive hypotheses about second language ability: Factor analytic studies of a variety of language subtests. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 13–23). Rowley, MA: Newbury House.

Pearson Education. (2012). *PTE Academic: Score Guide* [pdf Document]. Retrieved from http://pearsonpte.com/wp-content/uploads/2014/07/PTEA_Score _Guide.pdf

Pimsleur, P., Stockwell, R., & Comrey, A. (1962) Foreign language learning ability. *Journal of Educational Psychology*, *53*(1), 15–26. https://doi.org/10.1037/h0044336

Powers, D. (2010). *The case for a comprehensive, four skills assessment of English language proficiency*. (TOEIC Compendium Study). Princeton, NJ: Educational Testing Service.

Powers, D., Kim, H., Yu, F., Weng, V., & VanWinkle, W. (2009). *The TOEIC speaking and writing tests: Relations to test-taker perceptions of proficiency in English* (Research Report No. RR-09-18). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02175.x

Purpura, J., Brown, J., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, *65*(1), 36–73. https://doi.org/10.1111/lang.12112

Sang, F., Schmitz, B., Vollmer, H., Baumert, J., & Roeder, P. (1986). Models of second language competence: A structural equation approach. *Language Testing*, *3*(1), 54–79. https://doi.org/10.1177/026553228600300103

Sawaki, Y., Stricker, L., & Oranje, A. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing*, *26*(1), 5–30. https://doi.org/10.1177/0265532208097335

Scholz, G., Hendricks, D., Spurling, R., Johnson, M., & Vandenburg, L. (1980). Is language ability divisible or unitary? A factor analysis of twenty-two English proficiency tests. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 24–33). Rowley, MA: Newbury House.

Shin, S. K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing*, *22*(1), 31–57. https://doi.org/10.1191/0265532205lt296oa

Stricker, L., & Rock, D. (2008). *Factor structure of the TOEFL Internet-based test across subgroups* (Research Report No. RR-08-66). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02152.x

Wilson, K. (1993). *Relating TOEIC scores to oral proficiency interview ratings* (Research Summary No. TOEIC-RS-01). Princeton, NJ: Educational Testing Service.