ATU PRESS

ISSUES IN LANGUAGE TEACHING (ILT)
VOL. 10, NO. 1, 1-36, June 2021
https://doi.org/10.22054/ilt.2021.52989.507 Document Type: Research Paper

L1-Based Elicitation as a Valid Measure of L2 Classroom Performance Assessment: A Multi-Method Mono-Trait Model of Validation

Ali Mohammadi Darabad
Ph.D. Candidate in TEFL, Faculty of Persian Literature and Foreign Languages, Islamic Azad University, South Tehran Branch, Tehran, Iran

Gholam-Reza Abbasian*
Associate Professor of TEFL, Imam Ali University & Islamic Azad University, South Tehran Branch, Tehran, Iran

Bahram Mowlaie
Assistant Professor, Faculty of Persian Literature and Foreign Languages, Islamic Azad University, South Tehran Branch, Tehran, Iran

Ali-Asghar Rostami Abusaeedi
Professor of Linguistics, English Department, Faculty of Literature and Humanities, Shahid Bahonar University of Kerman, Kerman, Iran

Received: July 03, 2020**; Accepted:** February 21, 2021

Abstract
Classroom performance assessment has gained prominence parallel to the multiplicity of the purposes ahead of the assessment. Of many, the major controversy, which was the motive behind this study, is the incorporation of L1-based elicitation as a valid measure of L2 performance assessment. To shed empirical light on this issue, this explanatory sequential mixed-methods research employed 87 Iranian intermediate EFL learners, whose L2 classroom performance was assessed through L1-based elicitation techniques. In order to validate this mechanism, the multi-method mono-trait model (namely, Pearson correlation, structural equations, exploratory and confirmatory factor analysis, composite reliability, and convergent validity) suggested by Henning and Mesick's Unitary Concept of validity were applied. The results from these multiple sources of evidence yield support to their common consensus that L1-based elicitation techniques are valid measures of L2 performance assessment. The findings then offer a legacy to the educational implications of L1-based mechanisms both in L2 instruction and assessment.

***Keywords*:** L1-based elicitation, Performance assessment, Speaking ability, Unitary concept of validity

***Corresponding author's email:** gabbasian@gmail.com

INTRODUCTION

There has been a significant move in language testing towards the development and utilization of performance tests. The premise for this move is the desire that such tests would assess a more valid construct of a real language. Language assessment and testing have consistently followed linguistic theories of the era. Consequently, the communicative period in the 1970s created a flood of criticism of the traditional non-communicative tests because they were considered as being restricted in their concept and as creating non-natural language. In later years, there was a move towards the development and utilization of tests resembling target language use (TLU) (Bachman, 1990) that necessitated test takers to present language that was performance-based, communicative, direct, and authentic. Thus, the performance turned out to be one characteristic among a number of others, such as authentic, functional, and direct, all of which were regarded as principles of communicative tests of that period.

On the other hand, the exclusive feature of the performance, according to Bachman, was that test-takers were assumed to reproduce the type of language used in situations other than testing. In this way, testing the performance, claimed mirroring TLU, exploits the tasks that empower the operators to exhibit what they know about a given topic (Flynn, 2008). As Flynn argued, as far as the assessment practices are concerned, too much emphasis is placed on assessing the content and little attention is given to knowledge, skills, and validity measures. Besides, many of the assessment practices conceptualize assessment as basically separate from instruction. However, if the curriculum, instruction, and assessment are integrated, the assessment itself becomes a valuable learning experience.

One of the main constraints in classroom performance assessment is the use of certain methods, e.g., elicitation techniques whereby to invite learners to engage in classroom interaction. Darn (2008) specifies that elicitation is a superior technique for encouraging the participation of learners in the class and for promoting a learner-centered classroom. This

would be more practical when the group elicitation method is involved. Group elicitation provides an alternative to the one-to-one interview method that many elicitation methods are based upon (Bloom, Critten, Johnson, & Wood, 2020).

More precisely, elicitation provides learners with opportunities to participate, thereby increasing student speaking time. In addition, Doff (1988, as cited in Suherdi, 2010) states that elicitation makes learners more active because their speaking time keeps them attentive, draws on what they already know, or provides chances for weaker learners to participate in the class and motivates them to learn.

Therefore, assessment and elicitation can be approached interwoven as the latter acts as a means to the former. But, elicitation may be exercised in a variety of ways, including resorting to the learners' L1 to elicit their L2 knowledge, contrary to what was banned in traditional approaches. However, the extent to which the L1-based elicitation technique can enjoy validity measures is controversial. This is the main rationale behind this study to investigate the extent to which L1-based elicitation is comparable to its L2-based counterpart in terms of its validity measures.

LITERATURE REVIEW

During the 1980s, testing the learners' performance became connected more with particular tasks and settings of professional provision and certification, typically in the work environment (Wesche, 1985). But continuously it turned into a typical type of assessment in the educational research settings predominantly dominated in the light of Discrete Point (DP) and then Integrative Testing (IN) techniques of structuralism and generativism, respectively.

Additional re-conceptualizations of language construct and its measurement approaches led to progress within the field. For instance, Oller (1979) supported the idea of language as a unitary element instead of a divisible construct. Furthermore, Canale and Swain (1980) tended to a

more comprehensive notion of the components of language which centered on the appropriate utilization of language in a particular setting. Bachman and Palmer (1983) revealed that language was not formed only of one general factor. In addition to a higher-order general factor, language is also composed of two trait factors. These factors are called grammatical and pragmatic competence. Jing (2016) valued pragmatic considerations in the development of language tests. Before that, Birjandi and Soleimani (2013) pointed to the absence of valid and reliable testing instruments for assessing the pragmatic knowledge of second language learners.

Within the light of such theoretical patterns, the tests tended to be authentic, communicative, direct, and functional with a particular focus on TLU-simulated performance. These performance definitions have since managed the process of test development: Needs analysis characterizes the purpose and context for a test; based on these analyses, the samples of the behavior in that particular context are defined; simulation tasks or authentic performances that elicit the performance are carefully chosen; the test-takers complete the tasks in real or simulated situations; the language samples are, then, elicited and weighed. Performance tests also have gained washback, high face validity, and user worthiness. Although competence must be deduced from observations of the performance of behaviors, these deductions are not frequently forthright. Then, the validity of such measures and deductions are uncertain.

In line with this trend and the resulting controversies, Messick (1994) made distinctions between constructs and tasks. The former, then, refer to theories of competence knowledge, communication, and skills underlying performance, while the latter implies performance. He then propounded the necessity of establishing construct validity through empirical evidence in a bid to distinguish competence from performance. Similarly, McNamara (1996) tried to associate language performance with other affective and cognitive areas that incorporate language knowledge with communication skills. Therefore, aspects of L1 performance are needed to be introduced to make the notion of communicative competence more comprehensive than

what is found in Hymes' (1979) model. The common core of all these speculations is the invalidity of performance tests in that they are not based on a fuller construct theory but, rather, on a narrow view of communication and devoid of additional components.

As a result, these theoretical movements brought about new Test Standards like in APA (1966), and AERA et al. (1999) mainly in the light of Messick's (1989) unitary belief of validity based on which construct validity is at the core of validity defined as an "elaborated theory and supporting methods" (p. 5). Construct validity includes different sources of evidence relevant to the validation, interpretation, and use of the test scores, including the sources of evidence-based on test content, response processes, internal structure, relations to other variables, empirical or criterion-related validity, and consequences, which suggest that there might be different validities equal in number to the number of sources of evidence, including the sources and techniques of elicitation of the very target construct.

Generally, L1 use in L2 classes either as a teaching or a testing means has been debated for decades. Since the 1900s, some methods have favored the target language (TL) exclusively in the classroom like the era of the Direct Method and L1 prohibition favored in Krashen's hypotheses (1982). However, this monolingual approach is still a question since the evidence is not convincing (Macaro, 2009). So, some investigators commenced to explore the role that L1 can play in the L2 classroom (e.g., Cook, 2001; Macaro, 2001; Turnbull, 2001), claiming that some careful utilization of L1 can be helpful to learners' second language acquisition (Bozorgian & Fallapour, 2015; Cummins, 2007; De la Campa & Nassaji, 2009; Lin, 2015; Lo, 2015; Macaro, 2001, 2009; Swain & Lapkin, 2013; Swain, Kirkpatrick & Cummins, 2011).

More recently, Vygotsky's cognitive and sociocultural hypothesis and Cummins' linguistic interdependence speculation (1991) support 'utilizing L1 within the L2 classroom'. As Cummins argued, L1 and L2 are not separated from each other and they are characterized as a 'Common

Underlying Proficiency' as both are merged in the mind so that they do not function independently. Cummins (2007) believed that this 'common underlying proficiency' makes it possible for the cognitive/academic or literacy-related proficiency to be transferred from one language to another.

As to the mechanisms of elicitation and assessment, Gass (2018) holds that ways of eliciting data to understand how languages are learned and the type and extent of L2 knowledge are limited to some extent only by the imagination of the researcher. Gass (2018) reviews two common elicitation tasks: judgments and elicited imitation. Plonsky et al. (2019) report four types of judgment tasks: magnitude estimation, grammaticality judgments, truth-value judgments, and preference judgments and emphasize the effectiveness of these elicitation types on L2 assessment. According to Gaillard and Tremblay (2016), various ways, especially elicited imitation, have been employed for measuring proficiency. Such use is grounded on the hypothesis that processing efficiency is a reflection of proficiency intervened by working memory capacity. In their studies, Gaillard and Tremblay (2016) and Wu and Ortega (2013) showed that elicited imitation was a worthy measure of general proficiency. The obtained result was also supported by a meta-analysis conducted by Yan, Maeda, Lv, and Ginther (2016). They analyzed 21 studies, which had been conducted for 45 years. Their results suggest that elicited imitation is a discrimination factor across proficiency levels. In addition, the picture story can also be used as a reliable and valid narrative eliciting tool for Persian data at the microstructure and macrostructure levels (Mojahedi Rezaeian, Ahangar, Hashemian, Mazaheri. 2020)

The concurrent validity of semi-direct and direct tests was examined in some languages, the results of which demonstrated high correlations between the two types of tests (Stansfield & Kenyon, 1992). The use of semi-direct tests was recommended as valid and practical substitutes for direct tests. The comparability of two versions of an oral interaction test, i.e., a tape-based (semi-direct) version, and a live interview (direct) version was investigated by Wigglesworth and O'Loughlin (1993). They showed

that the two versions were highly comparable.

The validity of semi-direct versus direct tests was explored by Shohamy (1994) using both quantitative and qualitative procedures. The correlational analyses revealed high concurrent validity of the two tests (Shohamy, Gordon, Kenyon & Stansfield, 1989; Shohamy & Stansfield, 1991); however, the tests differed in several aspects. Qualitative analyses specified that the differences were in the topics and number of functions employed in the elicitation tasks and the communicative strategies, i.e., more paraphrasing and self-correction on the semi-direct test, and more shifts to L1 resources on the direct test.

Nakatsuhara, Taylor, and Jaiyote (2018) investigated the role of the L1 in assessing L2 English proficiency. Their focus was mainly on tests of L2 English speaking ability. Weir's (2005) socio-cognitive framework was firstly taken into considerations for developing and validating speaking tests. The results of their study showed that when evaluating the role of the L1 in an L2 speaking test, all the validity components should be taken into considerations, as the L1 issue could theoretically affect all parts of the framework.

Even though some theoretical speculations and empirical findings are supporting the positive influence of L1 in L2 learning, some researchers (e.g., McMillan & Turnbull, 2009) warn language teachers not to overuse L1. As Rezaee and Fathi (2016) concluded in their study, the function of switching to the learners' L1 should be done cautiously as many of the participants in their study disagreed with instructors' explaining instructions in the learners' L1. Therefore, the degree to which L1 should be used in L2 classrooms is still a mystery, and it seems that more empirical research and studies are required before concluding this issue.

PURPOSE OF THE STUDY

Performance-based assessment has received more considerations from ELT professionals since the real performances generated by language

learners are evaluated in this sort of assessment. However, the assessment of language learners' performances involves more complicated strategies when compared with more conventional testing strategies. This study, therefore, points out crucial considerations in adopting this type of assessment in a language class. In this vein, the data elicitation technique also gains prominence when the performances of language learners are at stake. The data elicited based on the learners' L1, assessing the learners' L2 performance, and also validity measures are focused in this study. Given these facts and following Messick's unitary approach to validation, the present study aimed at investigating the feasibility and validity of L1-based elicitation in the assessment of L2 performance.

METHOD

Participants

The participants of this study were 86 out of 97 conveniently-selected Iranian intermediate EFL students of Translation Studies (18–25 years old) from Islamic Azad University (Science and Research Branch) in Tehran, who were selected as a homogeneous sample based on their performance on the 2015 version of Cambridge PET. According to the assessment guidelines for the Cambridge PET, a raw score ranging from 140-170 is categorized as intermediate.

Instrumentation

In addition to the Cambridge PET, the Interchange 2, 3rd Edition by Jack C. Richards (2018), Steps to Understanding by L. A. Hill (2017), and about 30 English stories (Levels 3-5) published by Oxford Publications for retelling purposes were used for this study. These stories were examined by a panel of experts as compatible for the intermediate level in terms of vocabulary load and structural complexity.

Elicitation techniques based on L1*:* The eliciting techniques include

defining, synonyms, paraphrasing, and asking multiple questions via the participants' L1 (Farsi) and L2 (English).

The self-designed rating scale: A descriptive scale, initially composed of seventeen items, was developed and evaluated to serve as the speaking assessment scale of the study. It included 3 major principles of speaking assessment, namely, interactive communication, language skills, and discourse management. A panel of experts in applied linguistics was asked to check the items in terms of content validity, ambiguity, and appropriateness. Following the experts' views and reviewing the literature (e.g., Assessing Speaking Performance – intermediate level – English Qualifications, Cambridge, 2008; Fulcher, 2003, 2010; O´Malley & Pierce, 1996; Underhill, 1987), an attempt was made to generate simple and short items. The Intermediate Assessment Scales is divided into five bands from 1 to 5, with 1 being the lowest and 5 the highest. The table then was modified and the sub-scales for each parameter were determined. Finally, a Mark Sheet was developed and reviewed by three experts for any probable inconsistencies before employing it for the scoring process (see appendix I).

Data Collection Procedure

Conducting the main principles of descriptive research of exploratory nature, this study mainly aimed at exploring and describing certain characteristics of the target phenomenon, which were done quantitatively, but in terms of the design the major part of the analyses were conducted in the light of family of correlational analyses (i,e., factor, regression analyses, and SEM) commonly used for exploration of factors and their validation purposes. For the purpose of this study, 86 out of 97 participants were selected based on version 2015 of the Cambridge PET, and randomly assigned into two groups (L1-based elicitation group and L2-based elicitation group). The participants in both L1-based and L2-based groups were provided with elicitation techniques, including defining, synonyms,

paraphrasing, forgetting, and asking multiple questions via the participants' L1 (Farsi) and L2 (English). In L1-based group, while the students were retelling a story or explaining an English proverb, which referred to the similar situation and experience of the learners (as a task), the teacher provided them with the definitions of target materials in their L1 (Farsi), e.g., words, and asked them to come up with the matching word in English.

To add a natural taste to the elicitation process, the teacher would pretend to forget the word, the grammatical structure, pronunciation, etc. so that grounds could be intentionally paved for the students to supply the target answer. The teacher would ask questions in Farsi whose answers would require the students to use the target linguistic feature. Some grammar-eliciting techniques such as picture description, conversations, readings, retelling stories, examples, etc. were employed and the required explanations were also provided through shifting to the learners' L1. Headlines, words, pictures, proverbs, personal notes, and free-writing, etc. were also provided as a tool for eliciting the learners' ideas.

Speaking tests

As a formative performance assessment, three similar speaking tests were conducted with one-week intervals between the tests. The following procedure was followed for rating the learners' performances:

Preparation for the test

Five minutes were considered for each student to prepare for the test without any interval to the proceedings. The first student randomly chose a test sheet that also contained a topic. The student wrote his name on the test sheet, previewed the allocated topic, and wrote down the key points before he/she returned the sheet to the assessors and started speaking. At the same time that s/he started, the assessor would assign the next student his/her topic and the same procedure would be followed for the rest of the language learners.

Individual speaking task

Three minutes were allocated for each student for the individual speaking task. The test taker would speak on the selected topic in front of the class and the assessors. According to the speaking assessment guidelines, the assessors should subtract one mark from the score assigned for content if the test taker speaks for less than one minute.

Question and answer session

The speaking assignment for the individual students was followed by a question-and-answer session for 2 minutes. In this session, two follow-up questions were raised by the assessors for each topic. Moreover, students should be encouraged to ask questions related to the topic. This assignment assessed the individual's ability to interact with an audience and respond to questions using various interactive strategies and his/her ability to defend and support his/her opinion credibly.

Assessment and evaluation

The language learners were assessed according to the four specified key criteria: grammar and vocabulary, content, pronunciation, and interactive communication. The first four criteria were employed in both segments of the speaking assessment. Interactive communication was run as an additional criterion for the question and answer session. Two trained raters rated the test takers simultaneously each having the same mark-sheet developed prior to the tests. Each performance assessment session was tape-recorded for further analysis.

Data Analysis

Multi-method mono-trait (Henning, 1987) approach including Pearson correlation, structural equations, exploratory and confirmatory factor

analyses, composite reliability, and convergent validity were run for the purpose of the data analysis.

RESULTS

All these statistical methods assume a lack of univariate and multivariate outliers and univariate and multivariate normality. Lack of univariate outliers was checked through standardized scores (Z-scores). As displayed in Table 1, and except for interactive communication, the rest of the variables had Z-scores higher than -3. Eight participants, whose Z-scores were higher than -3, were omitted; i.e. ID 53, 56, 92, 103, 104, 108, 113, and 140.

Table 1: Descriptive Statistics of Standardized Scores

	N	Minimum	Maximum	Mean	Std. Deviation
Z-score (PET)	95	-4.108	1.867	0.000	1.000
Z-score (PETLC)	95	-3.300	1.393	0.000	1.000
Z-score (PETRC)	95	-3.277	1.759	0.000	1.000
Z-score (PETSP)	95	-4.911	1.275	0.000	1.000
Z-score (DM)	150	-3.157	1.578	0.000	1.000
Z-score (GV)	150	-3.193	0.855	0.000	1.000
Z-score (PR)	150	-3.863	1.198	0.000	1.000
Z-score (IC)	150	-2.446	1.416	0.000	1.000

Lack of multivariate outliers was checked by computing the Mahalanobis distances which were compared against the critical value of chi-square at 0.001 level for 7 degrees of freedom (Tabachnick & Fidell, 2014); i.e. 24.32. Table 2 shows the descriptive statistics for the Mahalanobis distances. Since the maximum value of 17.17 was lower than 24.32, it can be concluded that the present data did not suffer from multivariate outliers. It should also be noted that 7 more participants were omitted because they had missing data on four or more dependent variables.

Table 2: Descriptive Statistics of Mahalanobis Distances

	N	Minimum	Maximum	Mean	Std. Deviation
Mahalanobis Distance	92	1.723	17.178	6.923	3.626

Table 3 shows the skewness and kurtosis indices. Since the skewness and kurtosis indices were lower than +/- 2, it can be concluded that the assumption of univariate normality was retained (Bachman, 2005; Bae & Bachman, 2010). The Mardia index of multivariate normality of 0.65 was lower than 3; hence multivariate normality of the present data.

Table 3: Tests of Univariate and Multivariate Normality

Variable	Min.	Max.	Skewness	kurtosis
IC	2.000	5.000	-.661	.496
PR	2.500	5.000	-.463	-.102
GV	2.500	5.000	-1.018	.298
DM	2.500	5.000	-.152	.236
PETLC	17.000	25.000	-.521	.099
PETRC	90.000	119.000	-.109	-.517
PETSP	18.000	24.000	.049	-.986
Mardia				.652

The validity of L1-based elicitation as a measure of EFL learners' L2 performance was investigated through multiple methods. The inter-rater reliability indices (as shown in Table 4) for the two raters rating the participants' performance on L1-based and L2-based tests of discourse management, vocabulary and grammar, pronunciation and interactive communication indicated that there were significant agreements between the two raters on L1-based and L2-based tests of mentioned components. The Pearson correlations were compared (Lenhard & Lenhard, 2014) to probe if there were any significant differences between L1-based and L2-based ratings. The results indicated that;

- There was not any significant difference between the L1 and L2 ratings of discourse management ($Z = .787$, $p = 0.216$).
- L2 had a significantly higher inter-rater on grammar and vocabulary than L1 ($Z = 2.56$, $p = 0.005$).
- L2 had a significantly higher inter-rater on pronunciation than L1 ($Z = 2.45$, $p = 0.007$).

- L2 had a significantly higher inter-rater on interactive communication than L1 (Z = 2.09, p = 0.018).

Table 4: Comparing Inter-Rater Reliability Indices

	L1	L2	Z-Value	P
Discourse Management	.916	.884	.787	.216
Grammar and Vocabulary	.779	.921	2.56	.005
Pronunciation	.813	.931	2.45	.007
Interactive Communication	.867	.944	2.09	.018

Criterion-referenced Validity

The criterion-referenced validity of the L1-based and L2-based tests was investigated by computing their correlations with the PET and its sub-sections. Based on the results displayed in Table 5, it can be concluded that L1-based tests enjoyed significantly higher criterion-referenced validity indices than the L2-based test.

Table 5: Pearson Correlations; L1 vs. L2-Based Tests with PET

		L1	L2
DM	**Pearson Correlation**	.322*	.169
	Sig. (2-tailed)	.027	.267
	N	47	45
GV	**Pearson Correlation**	.528**	.257
	Sig. (2-tailed)	.000	.089
	N	47	45
PR	**Pearson Correlation**	.493**	.241
	Sig. (2-tailed)	.000	.111
	N	47	45
IC	**Pearson Correlation**	.630**	.287
	Sig. (2-tailed)	.000	.056
	N	47	45

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

An exploratory factor analysis (EFA) was run to probe the underlying constructs of the L1-based tests and sub-sets of PET. Before discussing the results, it should be noted that the assumption of sampling adequacy (KMO = .666 > .60), lack of singularity ($\chi 2$ (21) = 86.859, p = .000) and lack of multi-collinearity (Determinant = .132 > .00001) were retained.

The varimax rotation resulted in a two-factor solution (Table 6) which accounted for 58.08 percent of the variance. That is to say; the four L1-based components of tests and three components of PET measured two factors with an accuracy of 58.08 percent.

Table 6: Total Variance Explained; L1-Based Tests and Sub-Sets of PET

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.806	40.081	40.081	2.806	40.081	40.081	2.133	30.473	30.473
2	1.260	18.005	58.086	1.260	18.005	58.086	1.933	27.613	58.086
3	.965	13.789	71.875						
4	.831	11.866	83.742						
5	.561	8.019	91.760						
6	.310	4.426	96.186						
7	.267	3.814	100.000						

Table 7 displays the factor loadings of the tests under the two extracted factors. Pronunciation and grammar and vocabulary together with the listening sub-set of PET loaded under the first factor. Interactive communication and discourse management loaded under the second factor with reading and speaking sub-sets of PET.

Table 7: Rotated Component Matrix; L1-Based Tests and Sub-Sets of PET

	Component	
	1	**2**
Pronunciation	.866	
Grammar & Vocabulary	.820	
Listening (PET)	.780	
Interactive Communication		.870
Reading (PET)		.830
Discourse Management		.485
Speaking (PET)		.372
Composite Reliability	.863	.750
Convergent Validity	.677	.455

Both factors enjoyed acceptable composite reliability indices. The composite reliability of the first (.863) and second (.677) were higher than .70. The first factor also enjoyed acceptable convergent validity (.677>.50). However, the second factor did not enjoy convergent validity. Its validity index of .455 was lower than .50 which is considered as the minimum acceptable convergent validity.

Structural Equation Modeling

A multi-group SEM was run to probe the underlying constructs of the PET and L1-based and L2-based tests. Conceptual Model 1, as shown in Figure 1, displays the model being tested. On the left side of the model, the PET test and its three components of listening (LC), reading (RC), and speaking (SP) measure a latent variable labeled as "Proficiency". On the right side, the four components of L1-based and L2-based tests measured EFL learners' "performance" on discourse management (DM), grammar and vocabulary (GV), pronunciation (PR), and interactive communication (IC).

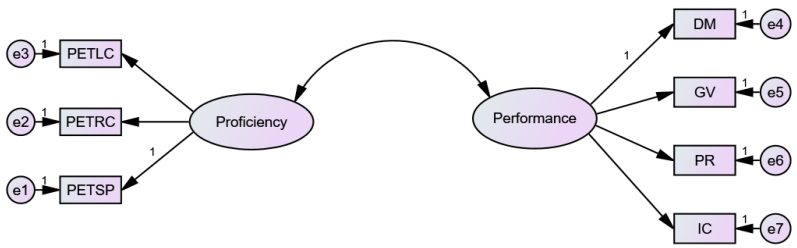

Figure 1: Conceptual Model 1, Multi-Group Model of PET & L1, L2-Based Test

The model did not show a good fit. The significant results of chi-square ($\chi 2$ (39) = 64, p = .000) indicated that the model did not enjoy a good fit; even though all other fit indices proved the fit of the model; except for TLI (.886 < .90) and REMSEA (.060 > .05). The Hoelter index of 157 was lower than 200 indicating that the present sample size was not

adequate for running SEM. Therefore, the researchers had to modify the indices in the model.

The modification indices, shown in Figure 2, were checked to find a solution to increase the fit of the model. They suggested two-way relationships be established between discourse management (DM) and interactive communication (IC), and IC and pronunciation (PR); as displayed through the second model.

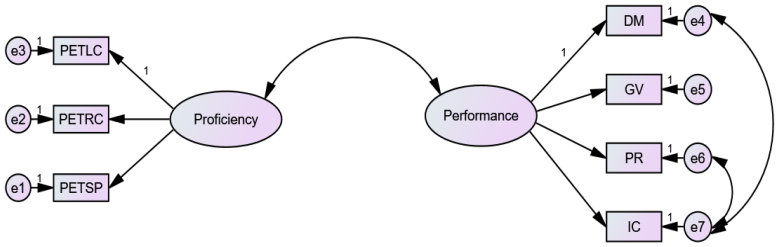

Figure 2: Conceptual Model 2, Modified Multi-Group Model of PET & L1, L2-Based Test

Before discussing the fit of the modified model, it should be noted that the assumptions of univariate and multivariate normality were retained. The skewness and kurtosis indices were lower than +/- 2 and the Mardia's index of multivariate normality was lower than 3 (Bachman 2005, and Bae & Bachman 2010).

The modified model showed a good fit. The non-significant results of chi-square (χ^2 (33) = 40.18, p = .182) indicated that the model enjoyed a good fit. All other fit indices proved the fit of the model. The Hoelter index of 216 was higher than 200 indicating that the present sample size was adequate for running SEM.

Table 8 displays the standardized and unstandardized regression weights for the total sample. They are analogous to B and beta values in an ordinary regression analysis. For example; the unstandardized regression weight from proficiency to speaking is .671. That is to say if proficiency increases one unit, speaking increases by .671 units. Its standardized index can be interpreted in terms of standard deviations. If proficiency increases

one standard deviation, speaking increases 0.429 standard deviations. The standardized value higher than 0.30 is considered as "moderate", and hence statistically significant, as it can be seen in Table 8. Structural Model 3, shown in Figure 3, displays the standardized relationships between the variables. The results indicated that all indicators significantly contributed to their latent variables.

Table 8: Standardized and Unstandardized Regression Weights (Total Sample)

			Unstandardized	S.E.	C.R.	P	Standardized
PETSP	<---	Proficiency	.671	.232	2.893	.004	.429
PETRC	<---	Proficiency	5.294	1.635	3.237	.001	.742
PETLC	<---	Proficiency	1.000				.569
DM	<---	Performance	1.000				.440
GV	<---	Performance	2.174	.557	3.901	.000	.933
PR	<---	Performance	1.897	.465	4.082	.000	.787
IC	<---	Performance	1.488	.364	4.086	.000	.542

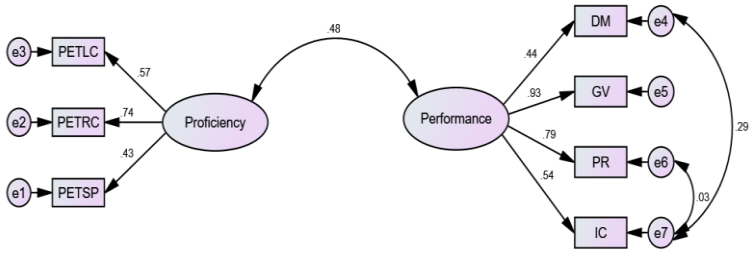

Figure 3: Structural Equation Model 3, Standardized Regression Weights (total sample)

Table 9 displays the two-way relationship between proficiency test and performance on L1-based and L2-based tests for the total sample. The covariance between the two latent variables is .125, and its ratio over the standard error was 2.259. That is to say, the relationship between the two latent variables was 2.259 standard errors above zero.

Table 9: Two-Way Relationships between Latent Variables and Error Terms (Total Sample)

			Covariance	S.E.	C.R.	P	Correlation
Proficiency	<-->	Performance	.125	.055	2.259	.024	.479
e4	<-->	e7	.095	.039	2.413	.016	.285
e6	<-->	e7	.007	.034	.204	.838	.029

Table 10 displays the standardized and unstandardized regression weights for the L1-Based group. All indicators had significant contributions to the latent variables; except for the speaking sub-section of PET. Structural Model 4, shown in Figure 4, displays the standardized relationships between the variables for the L1-Based group.

Table 10: Standardized and Unstandardized Regression Weights (L1-Based Group)

			Unstandardized	S.E.	C.R.	P	Standardized
PETSP	<---	Proficiency	.355	.209	1.701	.089	.187
PETRC	<---	Proficiency	3.726	1.161	3.209	.001	.466
PETLC	<---	Proficiency	1.000				.443
DM	<---	Performance	1.000				.339
GV	<---	Performance	2.097	.996	2.106	.035	.714
PR	<---	Performance	2.522	1.188	2.123	.034	.760
IC	<---	Performance	1.859	.927	2.005	.045	.554

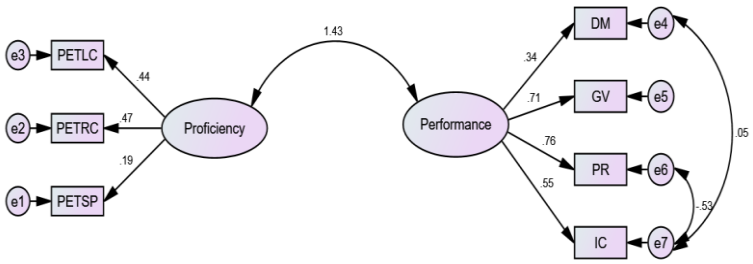

Figure 4: Structural Equation Model 4, Standardized Regression Weights (L1-Based Group)

Table 11 displays the two-way relationship between proficiency test and performance on L1 and L2-based tests for the L1-based group. The

covariance between the two latent variables is .197, and its ratio over the standard error was 1.887. That is to say, the relationship between the two latent variables was 1.887 standard errors above zero.

Table 11: *Two-Way Relationships between Latent Variables and Error Terms (L1-Based Group)*

			Covariance	S.E.	C.R.	P	Correlation
Proficiency	<-->	Performance	.197	.105	1.887	.059	1.433
e4	<-->	e7	.011	.034	.324	.746	.050
e6	<-->	e7	-.089	.033	-2.699	.007	-.527

Exploratory Factor Analysis; Components of L1-Based Tests

An exploratory factor analysis (EFA), shown in Table 12, was run to probe the underlying constructs of the components of L1-based tests. Before discussing the results, it should be noted that the assumption of sampling adequacy (KMO = .551 < .60) was not retained. However, the assumption of lack of singularity (χ^2 (6) = 33.480, p = .000) and lack of multi-collinearity (Determinant = .466 > .00001) were retained.

Table 12: KMO and Bartlett's Test; Components of L1-Based Tests

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.551
Bartlett's Test of Sphericity	Approx. Chi-Square	33.480
	df	6
	Sig.	.000
Determinant		.466

The results extracted two factors (Table 13) which accounted for 45.51 percent of the variance. That is to say; the four components of L1-based tests measured two factors with an accuracy of 45.51 percent.

Table 13: Total Variance Explained; Components of L1-Based Tests

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	1.892	47.308	47.308	1.500	37.495	37.495	1.321	33.025	33.025
2	1.031	25.777	73.085	.365	9.124	46.619	.544	13.593	46.619
3	.764	19.103	92.188						
4	.312	7.812	100.000						

Table 14 displays the factor loadings of the tests under the two extracted factors. Pronunciation, and grammar, and vocabulary loaded under the first factor. Interactive communication and discourse management loaded under the second factor.

Table 14: Rotated Component Matrix; Components of L1-Based Tests

	Component	
	1	2
Pronunciation	.846	
Grammar & Vocabulary	.758	
Interactive Communication		.537
Discourse Management		.426
Composite Reliability	.784	.377
Convergent Validity	.645	.235

The composite reliability of the first factor (.784) was higher than .70, and its convergent validity of .645 was higher than .50. Thus it can be claimed that the first factor enjoyed acceptable composite reliability and convergent validity. However, the second factor failed to meet the minimum requirements for acceptable composite reliability and convergent validity. Its composite reliability of .377 was lower than .70, and its validity index of .235 was lower than .50, which is considered as the minimum acceptable convergent validity.

Polytomous Item Response Model; L1-Based Test

A Polytomous IRT using a graded response model was run to analyze the

L1-based tests. Based on the results displayed in Table 15 and Figure 5, it can be concluded the discourse management was the most discriminating (b=4.79) component of L1-based tests. This was followed by grammar and vocabulary (b=1.819, pronunciation (b=1.165), and interactive communication (b=.385). As displayed in Figure 5, discourse management has the steepest curve. This was followed by grammar and vocabulary, pronunciation, and interactive communication. The latter one showed an almost flat line.

Table 15: Thresholds and Discrimination Indices; L1-Based Tests

	Threshold 1	Threshold 2	Discrimination
DM	-1.946	-0.185	4.793
GV	0.598	1.528	1.819
PR	-0.636	1.153	1.165
IC	-2.552	3.551	0.385

Regarding the thresholds, it can be concluded that interactive communication was the most difficult test with a threshold of 3.55. An inspection of its curve reveals the fact that even able respondents failed to get higher marks on this test. The probability of answering this test was lower than .50 on the right side of the graph. Grammar and vocabulary was the second most difficult test. This was followed by pronunciation and discourse management. The latter was the easiest test.

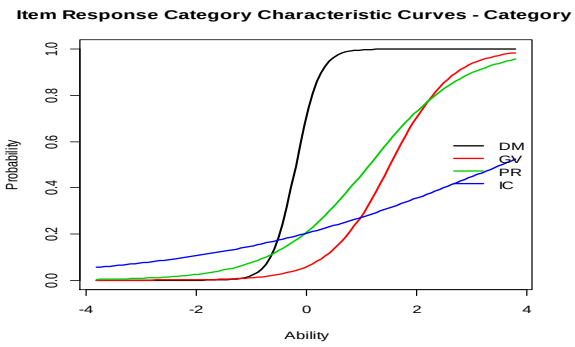

Figure 5: Item response category characteristic curve of L1-based tests

DISCUSSION

The present study aimed at investigating whether the L1-based elicitation technique could be a valid measure of assessing L2 performances. To achieve this objective, Messick's Unitary Concept of validity in which the construct validity is taken as core shaped the theoretical framework of the study.

To investigate the first evidence of Messick's model, a descriptive scale, as described in section 3.2, which resulted in the formation of a mark sheet, was developed and evaluated to serve as the speaking assessment scale of the study. A Pearson correlation coefficient was computed to assess the relationship between the components of speaking performance, namely, discourse management, grammar and vocabulary, pronunciation, and interactive communication and the components of PET. According to the results, there was a high correlation between DM and PET ($r=0.322$, $n=47$, $p=0.027$), GV and PET ($r=0.528$, $n=47$, $p=0.000$), PR and PET ($r=0.493$, $n=47$, $p=0.000$), and IC and PET ($r=0.630$, $n=47$, $p=0.000$), meaning that L1-based tests enjoyed significantly high criterion-referenced validity indices. Therefore, the fourth source of evidence in Messick's model is confirmed.

In a similar study conducted by Nakatsuhara and Jaiyote (2015) on the paired speaking task (collaborative task) of the First Certificate in English (FCE), the degree to which the shared or non-shared L1 partner would affect the language learners' performance on the task was explored. Two paired speaking tests were administrated among all participants. The first speaking test was run with a shared-L1 partner and the other one with a non-shared-L1 partner. A monologic speaking test was also administrated among the participants. The monologic and paired speaking tests were double-marked by two trained assessors. A relatively high inter-rater reliability was obtained between the two raters. The obtained results indicate that there was not any statistically significant difference for any of the analytic categories in the two types of pairing. This suggests that the

type of pairing does not affect test-takers' paired test scores. The strength of the correlations between the analytic categories in the two types of pairing was examined using Spearman correlations. The analysis shows that while none of the correlations between listening and monologic test scores was statistically significant, positive significant correlations were found between listening and paired speaking scores for Grammar and vocabulary (rho=0.32, p=0.04) and Discourse management (rho=0.35, p=0.03).

In another similar study conducted by Muñoz et al. (2003), the correlation between speaking performance and PET was assessed using the Pearson correlation coefficient. The results indicate that there was not any significant difference between the scores of the four Language Center evaluators assessing speaking and the PET evaluators at the 90% confidence level (P-value = 0.0722). However, the PET evaluator tends to assign higher marks (mean = 3.5) than the LCEs (highest mean = 3.2). Furthermore, there are no statistically significant differences between the mean scores of the Language Center evaluators at the 95% confidence level. Consequently, there were high correlations between the components of the speaking test and PET.

On the other hand, an exploratory factor analysis (EFA) was run to probe the underlying constructs of the components of L1-based tests. The results extracted two factors which accounted for 45.51 percent of the variance, meaning that the four components of L1-based tests measured two factors with an accuracy of 45.51 percent. The results also indicated that the components of pronunciation and grammar and vocabulary loaded under the first factor and the components of interactive communication and discourse management loaded under the second factor. The Composite Reliability of the first factor (0.748) was higher than 0.70, and its convergent validity of 0.645 was higher than 0.50. Accordingly, the first factor enjoyed acceptable composite reliability and convergent validity. However, the composite reliability of the second factor was 0.377, and its validity index was 0.235, both of which are considered as the minimum

acceptable convergent validity. Therefore, the third source of evidence in Messick's model is confirmed.

A Polytomous Item Response Theory (IRT) using a graded response model was run to analyze the L1-based tests. The obtained results indicated that discourse management was the most discriminating (b = 4.79) component of L1-based tests. It was followed by grammar and vocabulary (b = 1.819), pronunciation (b = 1.165) and interactive communication (b = .385). Discourse management has the steepest curve. It was followed by grammar and vocabulary, pronunciation, and interactive communication. The latter showed an almost flat line. The thresholds and discrimination indices for L1-based tests showed that interactive communication was the most difficult test with a threshold of 3.55. An inspection of its curve reveals the fact that even able respondents failed to get higher marks on this test. Grammar and vocabulary, pronunciation, and discourse management were the next difficult tests, respectively. And this is the consequence evidence of Messick (the fifth component) referring to the intended and unintended use of an instrument and how its unintended use weakens score inferences.

In a similar study, Zhou (2016) constructed a study for investigating the construct validity of communicative proficiency in the Test of English Proficiency (Oral) to assess university students' proficiency in speaking. The study examined the construct validity of TEP (Oral). The high internal consistency of reliability, the high inter-rater consistency, as well as the results from Factor Analysis, proved the construct validity of TEP (Oral) in that the five categories and 1-5 points in the rating scales were homogeneous in contributing to the assessment of the components of communicative performance.

Messick's framework for guiding the validation of performance assessment is a valuable practice in our context. Ruhe (2002b) showed how the framework performed when used to validate assessment tasks in a distributed, multimedia foreign language course. Messick’s framework has also been used for evaluating the distance education program (Ruhe,

2002a). Bunderson (2003) also adapted the framework in his validity-centered design; however, he did not provide any empirical evidence of the kinds of issues that emerged from applying the framework to authentic data. Messick's framework was employed by Chapelle et al. (2003) to lead the validation of web-based English for Second Language test, but the study was limited to intended impact, construct validity, interactivity, and authenticity. The evidential basis of Messick's framework was only investigated, which is considered a classical approach to validation. They also predicted more research on the theory, the argument, and the unintended consequences of assessment in distance contexts. Nakatsuhara, Taylor, and Jaiyote (2018) argued the role of the L1 in assessing L2 English proficiency. They particularly focused on tests of L2 English-speaking ability using Messick's unitary conceptualization of test validity.

The differences between the participants' L1-based performances were examined in relation to their PET performances. The L1-based ratings were also examined. Finally, the Item Response Theory between L1 and L2 performances was explored.

The inter-rater reliability indices for the two raters indicated that there were significant agreements between the two raters on L1-based tests of discourse management, vocabulary and grammar, pronunciation, and interactive communication. Therefore, the inter-rater reliability indices between the four components of the speaking test shows that the rating mechanisms for L1-based performances were reliable ($r_{DML1} = 0.916$; $r_{GVL1} = 0.779$; $r_{PRL1} = 0.813$; $r_{ICL1} = 0.867$).

The criterion-referenced validity of the L1-based tests was investigated by computing their correlations with the PET test and its sub-sections. Based on the results, it can be concluded that L1-based tests enjoyed significantly higher criterion-referenced validity indices. Comparing to other components of the L1-based speaking test, the Interactive Communication (IC) had the highest significant correlation with PET test (r=0.630, n=47, p=0.000). Following IC are Grammar and Vocabulary (GV) (r=0.528, n=47, p=0.000), Pronunciation (PR) (r=0.493,

n=47, p=0.000), and Discourse Management (DM) (r=0.322, n=47, p=0.027). Therefore, comparing the L1-based correlation coefficients with PET test and its sub-sections revealed that L1-based tests enjoyed significant validity indices.

An investigation into the results obtained from the Polytomous item response model for L1-based tests revealed that Interactive Communication was the most difficult test among the four components of the speaking test in L1-based performance test. Grammar and Vocabulary, Pronunciation, and Discourse Management were the next difficult tests, respectively.

CONCLUSION AND IMPLICATIONS

Considering the use of L1, the prevailing body of research in L2 learning has supported the constructive influence of L1 use over provoking L2 learning (e.g., Brooks-Lewis, 2009; Cheng, 2013, Grim, 2010; Lee & Macaro, 2013). Nevertheless, utmost utilization of L2 is essential since language classroom is considered as the only setting for most language learners who are studying English as their second/foreign language to the fact that for most L2 learners language classroom is the only context they have at their disposal for L2 exposure (Littlewood & Yu, 2011).

By examining various components of Messick's unitary model of validity, it can be claimed that the L1-based elicitation technique was a valid measure of L2 performance assessment. The findings, along with the work of the previously cited authors (e.g., Bunderson, 2003; Chapelle et al. 2003; Muñoz et al., 2003; Nakatsuhara & Jaiyote, 2015; Nakatsuhara, Taylor & Jaiyote, 2018; Ruhe, 2002; Zhou, 2016), represent a practice, which is, using Messick’s framework to guide validation practice in performance-based assessment. Our findings are supported by Messick's framework in the validity of performance assessment in closing the gap between validity theory and validation practice. In the future, more studies should be conducted regarding the application of this framework in diverse

performance assessment, and take the other sources of evidence into considerations. L1-based elicitation techniques call for a comprehensive approach to validity and validation based on evidence, values, and consequences in other contexts.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Ali Mohammadi Darabad http://orcid.org/0000-0003-2120-9500
Gholam-Reza Abbasian http://orcid.org/0000-0003-1507-1736
Bahram Mowlaie http://orcid.org/0000-0001-6153-6050
Ali-Asghar Rostami-Abusaeedi http://orcid.org/0000-0001-8423-7272

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: American Psychological Association.

Bachman, L. F. (2005). *Statistical analysis for language assessment.* (2nd ed.). New York, NY: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental consideration in language testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1983). *Oral interview test of communicative proficiency in English*. Urbana, IL: University of Illinois Press.

Bae, J., & Bachman, L. F. (2010). An investigation of four writing traits and two tasks across two languages. *Language Testing 27*, 213–230.

Birjandi, P., & Soleimani, M. M. (2013). Assessing language learners' knowledge of speech acts: A test validation study. *Issues in Language Teaching, 2*(1), 1–26.

Bloom, A., Critten, S., Johnson, H., & Wood, C. (2020). A critical review of methods for eliciting voice from children with speech, language and communication needs. *Journal of Research in Special Educational Needs, 20*(4), 308–320.

Bozorgian, H., & Fallahpour, S. (2015). Teachers' and students' amount and purpose of L1 use: English as foreign language (EFL) classrooms in Iran. *Iranian Journal of Language Teaching Research, 3*(2), 67–81.

Brooks-Lewis, K. A. (2009). Adult learners' perceptions of the incorporation of their L1 in foreign language teaching and learning. *Applied Linguistics, 30*(2), 216–235.

Bunderson, C. V. (2003). *On the validity-centered design and continuing validation of learning progress measurement systems*. Unpublished manuscript.

University of Cambridge (2008). *Assessing speaking performance – intermediate level – English qualifications*. Cambridge: Cambridge University Press.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*, 1–47.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing, 20*(4), 409–439.

Cheng, T. P. (2013). Codeswitching and participants orientations in a Chinese as a foreign language classroom. *The Modern Language Journal, 97*(4), 869–886.

Cook, V. (2001). Using the first language in the classroom. *Canadian Modern Language Review, 57*(3), 402–423.

Cummins, J. (2007). Rethinking monolingual instructional strategies in multilingual classrooms. *Canadian Journal of Applied Linguistics, 10*(2), 221–240.

Cummins, J. (1991). Interdependence of first- and second-language proficiency in bilingual children. In Bialystok, E. (Ed.), *Language processing in bilingual children* (pp. 70-89). Cambridge: Cambridge University Press.

Darn, S. (2008). Asking questions. The BBC and British Council. Retrieved from: http://www.teachingenglish.org.uk/articles/asking-questions

De la Campa, J. C., & Nassaji, H. (2009). The amount, purpose, and reasons for

using L1 in L2 classrooms. *Foreign Language Annals, 42*(4), 742–759.

Flynn, L. A., (2008). In praise of performance-based assessments. *Science and Children, 45*(8), 32–35.

Fulcher, G. (2003). *Testing second language speaking*. Edinburgh: Pearson Education Limited.

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning, 66*(2), 419–447.

Gass, S. (2018). SLA elicitation tasks. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 313-337). East Lansing, MI: English Language Center, Michigan State University.

Grim, F. (2010). L1 in the L2 classroom at the secondary and college levels: A comparison of functions and use by teachers. *Electronic Journal of Foreign Language Teaching, 7*(2), 193–209.

Henning, G. (1987). A guide to language testing: Development - evaluation - research. Rowley, MA: Newbury House.

Hill, L. A. (2017). *Steps to Understanding*. Oxford: Oxford University Press.

Jing, X. (2016). Pragmatic language testing and its application in testing oral proficiency. *International Conference on Humanities Science, Management and Education Technology* (pp. 66–70). Beijing, China January 23-24.

Krashen, S. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon.

Lee, J. H., & Macaro, E. (2013). Investigating age in the use of L1 or English-only instruction: Vocabulary acquisition by Korean EFL learners. *The Modern Language Journal, 97*(4), 887–901.

Lenhard, W., & Lenhard, A. (2014). *Hypothesis tests for comparing correlations*. Available: https://www.psychometrica.de/correlation.html. Bibergau: Psychometrica.

Lin, A. M. (2015). Conceptualizing the potential role of L1 in CLIL. *Language, Culture and Curriculum, 28*(1), 74–89.

Littlewood, W., & Yu, B. (2011). First language and target language in the foreign language classroom. *Language Teaching, 44*(1), 64–77.

Lo, Y. Y. (2015). How much L1 is too much? Teachers' language use in response

to students' abilities and classroom interaction in content and language integrated learning. *International Journal of Bilingual Education and Bilingualism, 18*(3), 270–288.

Macaro, E. (2009). Teacher use of codeswitching in the second language classroom: Exploring 'optimal' use. In M. Turnbull & J. Dailey-O'Cain (Eds.), *First language use in second and foreign language learning* (pp. 35-49). Bristol: Multilingual Matters.

Macaro, E. (2001). Analyzing student teachers' codeswitching in foreign language classrooms: Theories and decision making. *The Modern Language Journal, 85*(4), 531–548.

McMillan, B., & Turnbull, M. (2009). Teachers' use of the first language in French immersion: Revisiting a core principle. In M. Turnbull, & J. Dailey-O'Cain (Eds.), *First language use in second and foreign language learning* (pp. 15-34). Bristol: Multilingual Matters.

McNamara, T. (1996). Measuring second language performance. London: Addison Wesley Longman Ltd.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). New York, NY: Macmillan.

Mojahedi Rezaeian S., Ahangar A., Hashemian P., & Mazaheri M. (2020) Assessing an eliciting narrative tool used for studying the development of Persian-speaking children's narrative discourse skills. *Journal of Modern Rehabilitation, 14*(1), 55–68.

Muñoz Restrepo, A. P., & Álvarez Villa, M. E. (2003). Estimating the validity and reliability of an oral assessment instrument. *REVISTA Universidad EAFIT, 39*(132), 65–75.

Nakatsuhara, F., & Jaiyote, S. (2015). Exploring the impact of test-takers' L1 backgrounds on paired speaking test performance: How do they perform in shared and non-shared L1 pairs? *Paper presented at the BAAL/CUP Applied Linguistics Seminar*. York, St. John University, UK, 24-26/06/2015.

Nakatsuhara, F., Taylor, L., & Jaiyote, S. (2018). The role of the L1 in testing L2 English. In C. J. Hall, & R. Wicaksono (Eds.), *Ontologies of English conceptualizing the language for learning, teaching, and assessment* (pp. 1-18). Cambridge: Cambridge University Press.

Oller, J. W. Jr. (1979). *Language tests at school: A pragmatic approach*. London: Longman.

O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. Reading, MA: Addison-Wesley Pub. Co.

Cambridge University Press. (2015). *Preliminary English test*. Cambridge: Cambridge University Press.

Plonsky, L., Marsden, E., Crowther, D., Gass, S., & Spinner, P. (2019). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research, 35*(1), 1–39.

Rezaee, A. A., & Fathi, S. (2016). The perceptions of language learners across various proficiency levels of teachers' code-switching. *Issues in Language Teaching, 5*(2), 233–254.

Richards, J. C. (2018). *Interchange 2* (3rd ed.). Oxford: Oxford University Press.

Ruhe, V., & Zumbo, B. D. (2006). Using Messick's framework to validate assessment tasks in online environments: A course in writing effectively for UNHCR. In D. D. Williams, S. L. Howell, & M. Hricko (2006), *Online Assessment, Measurement, and Evaluation: Emerging Practices* (pp. 203-220). USA: Information Science Publishing.

Ruhe, V. (2002a). Applying Messick's framework to the evaluation data of distance/distributed instructional programs. (Doctoral dissertation, University of British Columbia). *Dissertation Abstracts International*.

Ruhe, V. (2002b). *A course in writing effectively for UNHCR: Evaluation report*. Vancouver, BC: Commonwealth of Learning.

Shohamy, E., Gordon, C., Kenyon, D. M., & Stanfield, C. W. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Hebrew Higher Education, 4*, 4–9.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing 11*, 99–124.

Shohamy, E., & Stansfield, C. (1991). The Hebrew oral test: An example of international cooperation. *AILA Bulletin 7*, 79–90.

Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal, 76*, 129–141.

Suherdi, D. (2010). *The practice of eliciting techniques in EFL classroom interaction (A descriptive study of techniques at one of the senior high school*

in Bandung). (Unpublished BA thesis). Bandung: Universitas Pendidikan Indonesia.

Swain, M., Kirkpatrick, A., & Cummins, J. (2011). *How to have a guilt-free life using Cantonese in the English class: A handbook for the English language teacher in Hong Kong*. Hong Kong: Research Centre into Language Acquisition and Education in Multilingual Societies, Hong Kong Institute of Education.

Swain, M., & Lapkin, S. (2013). A Vygotskian sociocultural perspective on immersion education: The L1/L2 debate. *Journal of Immersion and Content-based Language Education, 1*(1), 101–129.

Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (7th ed.). Boston, MA: Pearson.

Turnbull, M. (2001). There is a role for the L1 in second and foreign language teaching, but…. *Canadian Modern Language Review, 57*(4), 531–540.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Wesche, M. B. (1985). Introduction. In P. C. Hauptman, R. LeBlanc, & M. B. Wesche (Eds.), *Second language performance testing* (pp. 1-12). Ottawa, ON: University of Ottawa Press.

Wigglesworth, G., & O'Loughlin K. (1993). An investigation into the comparability of direct and semi-direct versions of an oral interaction test. *Paper presented at the 15th Language Testing Research Colloquium*, Cambridge, August.

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals, 465*(4), 680–704.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497–528.

Zhou, W. (2016). Investigating the construct validity of communicative proficiency in TEP (Oral) at Level B. *Journal of Language Teaching and Research, 7*(4), 690–699.

Appendix I

Speaking Assessment

Key Criteria	1 mark	2 marks	3 marks	4 marks	5 marks	Score
Grammar and Vocabulary	Serious grammatical errors and very limited vocabulary resulting in incoherence	Generally poor grammar and vocabulary usage with a high frequency of errors but they do not affect coherence	Accurate grammar and vocabulary usage about half the time with a few major errors	Accurate grammar and vocabulary usage most of the time with occasional errors	Exceptional use of grammar and vocabulary throughout with minimal or no errors	
Content	Content is unrelated to the prompt and completely irrelevant or incomprehensible	Content is related to the prompt but it lacks clarity and relevance most of the time	Content is relevant about half the time but the ideas are not well- organized	Content is relevant and cohesive most of the time but it lacks consistent focus with occasional digressions	Content is coherent, relevant and well-organized with consistent focus on the prompt	
Pronunciation	Pronunciation is completely incomprehensible making it difficult to grasp the content	Pronunciation is poor with a high frequency of errors and occasionally unintelligible	Pronunciation is accurate and clear about half of the time with a few major errors	Pronunciation is accurate and clear most of the time with occasional errors	Pronunciation is accurate and clear throughout with correct stress and appropriate intonation	
Interactive Communication (Q & A Session)	Fails to answer any questions or gives incomprehensible or totally irrelevant answers or unable to interact with the audience	Tries to respond but fails to answer most of the questions, interaction with the audience is minimal	Responds to some of the questions but lacks the necessary language skills or knowledge of the content to sustain interaction with the audience	Responds to all questions but responses are not always convincing, interacts with the audience but lacks confidence and conviction	Responds to all questions effectively and interacts with the audience with confidence and conviction	
					Total Mark (20)	

Speaking Test – Mark Sheet
Name: Class:

Part I General conversation
Task: saying who you are, asking for and giving individual information, spelling

Part II Responding to elicitation prompts
Task: describing and interpreting a picture or a photograph, talking about likes and dislikes

Part III Simulated situation
Task: making and responding to suggestions, agreeing and disagreeing, making choices

Assessor

Interactive communication
(initiating and responding, hesitation, turn-taking)

0	1	2	3	4	5

_____
_____

Language skills

Pronunciation

0	1	2	3	4	5

_____
_____

Grammar and vocabulary

0	1	2	3	4	5

_____
_____

Discourse management

0	1	2	3	4	5

_____

_____

Interlocutor (GENERAL IMPRESSION)

Interactive communication (Parts I – III)	(max. 5)	
Language skills (Parts I – III)	(max. 5)	

Assessors' score
Total (Max. 20)