INTERNATIONAL JOURNAL OF
Practical and Pedagogical Issues in English Education
ISSN: 2980-9533
Arad Institute of Higher Education

A Search for Accountability in Low-Stakes Tests:

An Evidence-based Approach to Validation of a Local Writing Test

Ali Ansari

Department of English Language, Tarbiat Modares University, Tehran, Iran

ali.ansary90@gmail.com

Abstract

Validation has always been a significant part of assessment in educational settings. However, due to its complexities, most officials prefer to use ready-made tests to avoid validating their own tests. This study validates a screening writing test in an EFL context using an evidence-based validation approach. This study reflects the researcher's attempts to collect evidence to justify the newly developed test use and interpretation. The participants of this study were 53 Iranian B1 English language learners. To collect the required evidence, the researcher employed a questionnaire to investigate the learners' cognitive processes that they went through to write the texts. Furthermore, the scores were correlated with the participants' scores on a First Certificate in English (FCE) writing task to attain the criterion-related validity evidence. Evidence related to scoring validity and consequential validity was also collected by employing simple statistical measures, such as correlation, t-test, and percentage, that are easy for teachers to run. The findings of the study showed the (cognitive, context, scoring, consequential, and criterion-related) validity of the developed test. The pieces of evidence collected in this study supported the validity of the test for the intended use. The present report can enlighten novice teachers as to how to collect a body of evidence to ensure different stakeholders about the validity of their tests. While validation might be perceived as a complicated process, which is not within teachers' ability, this study can provide a blueprint to guide teachers on how to show their accountability through easy-to-implement steps.

Keywords: Accountability, Evidence-based validity, Second language testing, Validation, Writing

Introduction

The language testing community has confronted an increasing demand for accountability with regard to examinations, ranging from classroom assessments to international assessments of language proficiency (Bachman & Palmer, 2010; Izadpanah et al., 2014). Different stakeholders in the process of assessment want test developers to be accountable (Norton, 1997). This accountability is defined as "being able to demonstrate to stakeholders that the intended uses of

our assessment are justified" (Bachman & Palmer, 2010, p. 92). To justify the intended uses or decisions and conclusions, a tester should go through the process of validating the test; however, the validation process has changed with regard to its content and the data collection procedure. The genesis of drastic changes dates back to the late 1980s and Messick's (1989) proposition of the unitary concept of validity. From then on, several validation models (e.g., Bachman & Palmer's Test Usefulness Framework, 2010; Kane's Argument-based Approach to Validation, 2006; Chapelle et al.'s Validity Framework for Language Tests, 2008; Van Dijk's Sociocultural Model, 2011; Weir's Socio-cognitive Framework, 2005), either in evidence-based or argument-based forms, have been proposed to assist language test developers in establishing the validity of their assessments in a systematic manner.

In regard to second language low-stakes (e.g., classroom or school) assessment, little attention is paid to validating tests that are developed and administered in a local context and for low-stakes purposes. This study attempts to depict the way validity evidence can be gathered by following a set of clear rules provided by a validation model. Moreover, it attempts to illustrate how rather simple statistical measures (e.g., descriptive statistics, correlation, t-test and percentage) can be employed to answer tough validity-related questions. To be more specific, the present study is an endeavor, among many others, to employ one of the proposed models to establish the validity of a newly developed writing assessment which includes a directed and a free writing task. As mentioned by Zainal (2012), the replacement of indirect tests of writing with direct tests behooves test developers to establish the validity of these new tasks meticulously. Informed by the evidentiary validation model of Shaw and Weir (2007), the present study established the validity of two direct writing tasks through the study of cognitive validity, context validity, scoring validity, consequential validity, and criterion-related validity of the assessment.

As a proper alternative to traditional methods of testing, integrating components of students' engagements, constant feedback and formative assessment can pave the way to emerge a new way of testing (Fattah, 2024). Validation of tests has been at the center of testing researchers' agenda since the mid-1960s. Following the lead of Cronbach and Meehl (1955), testing researchers tackled the issue as *validities* and not *validity*. Here validities refer to the distinct validity types, namely, construct, content, and criterion-related validity, which were studied separately, and could be employed interchangeably. Around two decades later, Messick's ideas on taking validity as a unitary concept became the fashionable view of validity. In the traditional view, validity deals with the test or assessment scores; however, Messick's view has to do with the inferences made from test scores. Messick (1989) defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Validity, to Messick, deals with justifications which are of two types of evidential and consequential bases. At the same time, the function and outcome of testing, interpretation, and use are taken into account.

Unlike the pre-Messick era, the context of test use, the value system, the effects of the test on society, and ethical issues gained momentum in the 1990s. The model provided by Messick can be taken as the most prominent departure from the positivist approach to validity. This model urged test developers to collect evidence of different kinds, both evidential and consequential, to justify the validity of the test. These arguments may be founded on scientific arguments or pragmatic

concerns. Several evidence-based and argument-based frameworks for validation have been proposed since the heydays of Messick's model. Kane (2006), Mislevy et al. (2003), Bachman (2005), weir (2005), Chapelle et al. (2008), and Bachman and Palmer (2010) are different evidence- and argument validation frameworks. An extension of Weir's (2005) framework, Shaw and Weir (2007), which is employed to validate second language writing tests is at the heart of the present study.

Shaw and Weir's (2007) Socio-Cognitive Validity Model of Writing

Test taker characteristics, context validity, cognitive validity, scoring validity, criterion-related validity, and consequential validity are different components of this framework. Introductory accounts along with the specifications provided by the model are reviewed here in brief.

Test Taker Characteristics

Second language learners have different characteristics that are irrelevant to second language ability but influence their performance in a language test (Bachman, 1991). There are a plethora of variables of test takers such as short-term memory span, attitude towards the test (e.g., Bukta, 2013; Mehrpour et al., 2023), gender (e.g., Sunderland, 2000), language background (e.g., Ginther & Grant, 1997; Hoomanfard & Meshkat, 2015), topical knowledge (e.g., Jensen & Hansen, 1995) that can affect learners' performance and should be considered by a test developer to gain valid results.

Test taker characteristics, according to Weir (2005), can be categorized into three categories physical/physiological, psychological, and experiential. The first category is the most transparent variable which affects the performance of the test takers systematically. Special arrangements, for example, are made to give testees with permanent long-term, or temporary disabilities an equal chance to understand and answer the items. The use of enlarged print papers and Braille papers are examples of these arrangements. The psychological characteristics of the test taker, too, affect his performance. The conditions of the test should be in a way that provides students with a relaxed atmosphere suitable for demonstrating their abilities. Although students with different learning styles and personality types will perform differently, it is assumed that the same effect of individual psycholinguistic factors exists in real-life writing tasks. Another subcomponent of test taker characteristics deals with the familiarity of testees with the conditions of the environment and the examination format. It is believed to provide the potential test takers with a previously used exam that mirrors the format of the exam that they are supposed to take.

Cognitive Validity

A section that has recurred in different validity models has to do with the involvement of learners' characteristics in performing the tasks. The extent to which a test task elicits the resources and executive processes has been labeled as interactional authenticity (Bachman, 1990), Interactiveness (Bachman & Palmer, 2010), theory-based validity (Weir, 2005), and cognitive validity (Shaw & Weir, 2007). Bachman and Palmer (2010, p. 25) define interactiveness as "the

ways in which the test taker's areas of language knowledge, metacognitive strategies, topical knowledge, and affective schemata are engaged by the test task". This type of validity deals with the cognitive processes that a test task requires test takers to undergo before, during, and after writing a piece of writing. Shaw and Weir (2007, p.42) state, "from a cognitive perspective, a valid writing test would involve candidates engaging in all the processing components described above".

This type of validity has to do with the extent to which the writing task requires the same cognitive processes that are involved in the real-world context. Shaw and Weir (2007) have proposed a model of cognitive processes that encompasses six major processes which are reviewed here.

The organization is another step in writing that deals, primarily, with the content of the text in the mind of the writer. The subcomponents of this process are reported as "ordering the ideas; identifying relationships between them; determining which are central to the goals of the text and which are of secondary importance" (Shaw & Weir, 2007, p. 34).

The next step of writing is micro-planning, which, unlike macro-strategy, has to do with the formulation of ideas to be put in each paragraph. What to include in the paragraph to be written, and the planning for what and how to write the upcoming sentence are subcomponents of this step. The outcome of this process is reported to be the abstract decisions in the paragraph or sentential level.

The next process which is of a conversion nature is called translation. It is the stage in which the abstract form of content will be converted into linguistic form. The language resources are employed to perform the task of translation. This leads to higher levels of cognitive demands on the writer to deal with both the content and the language resources. Usually, applying communication strategies to compensate for the incompetency in lexis and syntax makes translation more cumbersome (Yakut & Bada, 2022). Those general decisions on the generic and other discoursal features that were made in macro-planning, now, are subjected to meticulous scrutiny which should lead to specific decisions.

The next stage is monitoring; at the basic level, syntax, and mechanical accuracy of spelling and punctuation are checked, and at the advanced level, the text is examined to check whether the writer's intentions are reflected in the text and whether the organization is suitable for the writer's purpose (Granena, 2023). This monitoring, according to Field (2005), might be performed after writing a word, a sentence, a paragraph, or the whole text. The cognitive load of this stage is noticeable, especially when it is performed at the same type the writer is producing the text (Wang & Lajoie, 2023).

Revising is the last stage of the writing process; in this stage, which is the result of the monitoring process, those sections, ranging from spelling to rhetorical structure, which are taken as unsatisfactory by the writer will be revised. Monitoring and revising can lead to a second time macro-planning which urges the other stages too.

Context Validity

The examination of the contextual variables of tasks and their administration conditions is called context validity (Shaw & Weir, 2007). Bachman and Palmer (2010, p. 23) call this feature authenticity and define it as "the degree of correspondence between the characteristics of TLU tasks and those of the test task". Context validity deals with the contextual factors which have a reciprocal relationship with cognitive processes. As Shaw and Weir (2007, p. 63) state, "Context validity relates to the linguistic and content demands that must be met for successful task realization and to features of the task setting that serve to describe the performance required". Context validity is required to achieve situational authenticity of the intended test task.

Shaw and Weir (2007) propose three aspects of context validity which are task, administration, and linguistic demands. With regard to task, response format, clarity of the purpose of the test, knowledge of criteria and their weighting, text length required, time limitation, and the relationship between writer and reader should be attended.

Administration of the writing task which is the last part of the context validity can affect almost all other validity components, especially scoring validity (Brunfaut, 2023). Physical conditions should be checked concerning materials, air conditioning system, enough chairs, suitable pathways for those with walking disabilities, etc. Temporal and physical conditions of different administrations of a test should be the same because a lack of consistency will result in unreliable results. The uniformity of conditions and procedures boosts the fairness of the test since all test takers would have equal conditions to demonstrate their performance.

Linguistic Demands is another category that chiefly deals with the task input and output. Lexical, structural, and functional resources are the main components of the linguistic demands that are required by direct writing tasks. In addition to these components, the discourse mode selected by an examiner for a specific task can affect the quality of the final product (Kim et al., 2021).

Scoring Validity

Scoring validity is used as an umbrella term for all reliability-related issues. Shaw and Weir (2007) conceptualize scoring validity as

> the extent to which test scores are based on appropriate criteria, exhibit consensual agreement in their marking, are as free as possible from measurement error, stable over time, consistent in terms of their content sampling and engender confidence as reliable decision-making indicators. (p. 6)

Selecting the fittest rubric among holistic, primary trait, and multiple traits models of scoring is necessary, but not sufficient. Rater characteristics is a variable that can affect the way an assessor scores a performance (Jia & Zhang, 2023). A suggestion which is given for the uniformity of the scoring process is to have an examiner meeting. This examiner meeting might be at the same time as a training session, where things get clear for everybody in the

group. The score which is the outcome of the scoring procedure can be used for evaluating consequential and criterion-related validity.

Consequential Validity

According to Shaw and Weir (2007), consequential validity can be categorized into two main sections which are the impact and washback effect and the avoidance of test bias. In their conceptualization, the impact can be equated with what Messick (1989) calls consequential validity of the test which requires the assessor to take inner (linguistic and psychological) and outer (social) elements of the test, test takers, and the setting into consideration throughout the process of testing.

Bias, which is a part and parcel factor in validating a test, is also dealt with here. The first question to ask is whether the participants find the assessment a fair test. The results should also be examined to check if the assessment functions for or against a particular group due to any construct-irrelevant or construct under-representation. The grouping might be based on gender, first language, background knowledge, cultural background, etc. One of the best ways found to statistically make sure that no certain group is favored is called differential validity.

Criterion-related Validity

The last type of evidence that should be gathered to validate a test, according to Shaw and Weir (2007), deals with the extent to which the test can demonstrate a relationship between test scores and an external criterion which is taken as a measure of the same construct. This type of validity which is called criterion-related validity can be achieved through three different measures which are cross-test comparability, comparison with different versions of the same test, and comparison with external standards.

This study is significant because, as noted earlier, the limited number of studies investigating local tests using an evidence-based approach suggests that there is room for contributions like the present one. While there have been a few studies applying evidence- or argument-based approaches to validate local or classroom tests, they remain scarce. For instance, Zainal (2012) did not collect evidence related to consequential validity and did not use specific instruments, such as questionnaires or think-aloud protocols, to assess the suitability of tasks in relation to the cognitive processes of participants. This study aims to add to the literature on second language testing by validating a local writing test through an improved research design.

Studies like the one by Zainal (2012) and the present study are designed to uncover the nuts and bolts of local assessment validation; these studies can indicate the way adequate validity evidence can be generated by teachers or local officials. In addition, they can expose how different qualitative and quantitative data collection and analysis procedures that are at teachers' disposal are employed based on the significance of tests and their intended decisions.

The present study is an attempt to probe into the context, cognitive, scoring, consequential, and criterion-related validity of the two writing task types proposed to be included in a screening test for upper intermediate students to go to the advanced level in a major language institute in

Iran. In order to check the theory-based validity, the students were requested to complete a questionnaire after finishing each task. Further evidence with regard to test takers, context, scoring, consequential, and criterion-related validity is also provided to indicate how a priori and posteriori data collection of a second language writing test can be employed to justify the intended use of a test. Thus, this study was conducted to answer the following three research questions:

1. Do the participants employ cognitive processes that are intended by the test developers?

2. Do the contextual characteristics of the tasks and administration approve the suitability of the tasks for the participants?

3. Does the scoring, consequential, and criterion-related validity evidence approve the suitability of the tasks?

Method

Design and Test Takers' Characteristics

This study follows an evidence-based validation approach to justify the newly developed test use and interpretation. As Shaw and Weir (2007, p. 17) state, "test-taker, rather than the test task, is at the heart of the assessment event", so a thorough examination of the test takers seems crucial. Since the test takers of the present study were from the same city, the existence of individual differences with regard to their native language, cultural background, and second language learning facilities that might jeopardize the fairness of the test was minimized. Sixty-four test takers (19 males & 34 females) took the writing section of the test. The participants took the Oxford Placement Test (Allan, 2004). The results indicated that the scores of 53 participants were between 135 and 149 which were equal to B2 in the Common European Framework (Allan, 2004) and were regarded as upper-intermediate. Out of 53 participants, 33 had attended the same institute's intermediate-level courses before the exam, and 20 applicants were newcomers.

Test Description

The present writing test is a part of a screening test that includes four subsections which are listening, reading, writing, and speaking skills. This test is used to assess the English language ability of both those students who have passed intermediate levels in the same institute and those who come from other institutes and want to get into the advanced levels of this language institute. Due to the nature of the exam, the test tasks do not address certain items taught in the intermediate books of this institute; the tasks are based on a theory of language proficiency (Bachman, 1990). This leads to equality for the students of the same institute and newcomers with regard to the experience of practicing the items. However, the study of five main textbooks taught in institutes in Iran indicated the familiarity of the learners with task types. The tasks were given to ten PhD holders and candidates for expert judgment about the face validity and the complexity of the input and their ideas were employed to modify the items.

Both tasks were of a direct nature. The testees were required to produce two compositions, each around 250 words. The first task was a directed task which provided the writers with a set of guidelines. The second one was a free composition and the participants had a choice between two topics. These topics were selected since they were related to students' everyday lives, they most possibly had enough background knowledge to accomplish the tasks. The participants had to select one of them to write about.

The first task type which was of a directed nature was presented as follows:

This is a report to the principal of your school. You are very dissatisfied with the services provided by your school lunchroom. You decide to write a report to inform your school principal about the terrible conditions and services of the lunchroom. These would include:

- Limited time period for serving food
- The low quality of food
- The high price of food
- Hygienic concerns
- Lack of variety in the weekly menu

- Give your report a title
- Include all of these points in your report
- Add related details
- You have 40 minutes to write this task.

The second task is of a free nature; the task is provided as follows:

Decide one of the topics provided below and write on it. (Time: 45 minutes)

- Study abroad
- My ideal teacher

Instruments

Shaw and Weir's Model of Writing (2007)

According to Shaw and Weir's model (2007), the second language writing construct deals with the interaction of the trait, context, and score. Shaw and Weir (2007) assert that unlike the trait-based approach to assessment which was compatible with TBLT, their model is of an interactionalist nature which proposes the mutual effects of the cognitive ability and the context. This model proposes a temporal frame that dictates the collection of validity evidence at different stages. This socio-cognitive validation model encompasses both a priori and a posteriori validation

component. The former includes theory and context validity, and the latter deals with the scoring, criterion, and consequential validity.

Cognitive Processing Questionnaire

After writing each writing task, the participants were given the questionnaire (developed by Weir, et al., 2007) which is intended to explore the cognitive processes that the writers go through to complete a writing task. The questionnaire which is in the form of a Liker-scale has 38 items and is presented in the cognitive validity section, below. The Cronbach alpha reliability index for the directed task was .78, and for the free writing task was .81. Weir et al. (2007) have validated this questionnaire by providing substantial evidence.

Writing Scoring Rubric

To assign scores to students' essays, Jacobs et al. (1981) rating scale which is a widely accepted analytical rating scale was utilized. This scale examines writing ability in 5 dimensions. It allocates 30 points to the content, 20 to the organization, 20 to the vocabulary, 25 to the language use, and 5 points to the mechanics. The maximum score that a student can obtain is 100.

Generation of Validity Evidence

Test Takers

In order to avoid any sort of bias against even a single test taker, some measures were taken before and during the administration of the test. The participants were asked if they were left-handed so that suitable chairs could be prepared for the examination day. The participants were asked about any sort of vision disability which required special printing. Although all of the participants were healthy, two assistants, assigned by the institute, were on standby on the examination day to help those who might have been unable to write. A pack of painkillers was also prepared for the examination day to be taken by those who might have had a headache to help them get over short-term disabilities that can affect their real abilities. Measures were also taken for those who used wheelchairs, but none of the participants of this administration required such assistance.

With regard to the psychological features, the first issue was the task topics. The topics were attempted to generate positive feelings about the task. The topics did not go for or against any religious, ethnic, or political opinions to avoid producing negative feelings that could affect test takers' performance adversely. Furthermore, for the second topic, two choices were provided so that the participants could select the one that they found more compatible with their repertoire.

In regard to experiential characteristics, familiarity with the building and the atmosphere of the examination hall was the first measure related to physiological/ physical issues. To ensure fairness, test takers were invited to visit the institute and examination hall a week before the exam,

INTERNATIONAL JOURNAL OF
Practical and Pedagogical Issues in English Education
ISSN: 2980-9533
Arad Institute of Higher Education

allowing them to become familiar with the environment just as participants who had previously attended the institute were.

As indicated by the participants before the exam, all of them had taken direct writing tasks either in the form of learning or assessment tasks, thus the format of the exam was not for or against a group of participants. Furthermore, a paper containing ten tasks similar to those of the examination was given to the participants a week before the exam to make sure that they were cognizant of the format of the tasks.

Research question one: Do the participants employ cognitive processes that are intended by the test developers?

In order to ensure the cognitive validity of the tasks, and to have expert judgment, the tasks were given to ten TEFL PhD holders and students. The experts found the tasks suitable for intermediate and upper-intermediate level students. In order to examine the cognitive validity, the cognitive processing questionnaire (CPQ) was responded by all participants.

Table 1

Level of Participants' Agreement with the Task Writing Stages

	Writing Stages	Agreement %	
		Directed	Free
1	I FIRST read the title very slowly considering the significance of each word in it.	63%	78%
2	I thought of WHAT I was required to write after reading the title and instructions.	71%	66%
3	I thought of HOW to write my answer so that it would respond well to the title.	58%	63%
4	I thought of HOW to satisfy readers or examiners.	76%	62%
5	I was able to understand the instructions for this writing test completely.	62%	68%
6	I know A LOT about this topic, i.e. I have enough ideas to write about this topic.	79%	82%
7	I felt it was easy to produce enough ideas for the essay from memory.	53%	62%
8	I know A LOT about this type of essay, ie an argumentative essay.	87%	84%
9	I know A LOT about other types of essays, eg descriptive, narrative.	91%	94%
10	Ideas occurring to me at the beginning tended to be COMPLETE.	66%	56%
11	Ideas occurring to me at the beginning were well ORGANISED.	58%	63%
12	I planned an outline on paper or in my head BEFORE starting to write.	92%	74%

INTERNATIONAL JOURNAL OF
Practical and Pedagogical Issues in English Education
ISSN: 2980-9533
Arad Institute of Higher Education

13	I thought of most of my ideas for the essay BEFORE planning an outline.	85%	72%
14	I thought of most of my ideas for the essay WHILE I planned an outline.	76%	82%
15	I thought of the ideas only in ENGLISH.	64%	69%
16	I was able to prioritize the ideas.	71%	66%
17	I was able to put my ideas or content in good order.	68%	72%
18	Some ideas had to be removed while I was putting them in good order.	73%	69%
19	I felt it was easy to put ideas in good order.	78%	65%
20	I felt it was easy to express ideas using the appropriate words.	79%	73%
21	I felt it was easy to express ideas using the correct sentences.	65%	69%
22	I thought of MOST of my ideas for the essay WHILE I was actually writing it.	61%	84%
23	I was able to express my ideas by using appropriate words.	56%	53%
24	I was able to express my ideas using CORRECT sentence structures.	68%	67%
25	I was able to develop any paragraph by putting sentences in logical order in the paragraph.	66%	68%
26	I was able to CONNECT my ideas smoothly in the whole essay.	68%	65%
27	I tried NOT to write more than the required number of words in the instructions.	72%	68%
28	I reviewed the correctness of the contents and their order WHILE writing this essay.	63%	74%
29	I reviewed the correctness of the contents and their order AFTER finishing this essay.	78%	75%
30	I reviewed the appropriateness of the contents and their order WHILE writing this essay.	65%	69%
31	I reviewed the appropriateness of the contents and their order AFTER finishing this essay.	74%	76%
32	I reviewed the correctness of sentences WHILE writing this essay.	87%	81%
33	I reviewed the correctness of sentences AFTER finishing this essay.	78%	82%
34	I reviewed the appropriateness of words WHILE writing this essay.	73%	69%
35	I reviewed the appropriateness of words AFTER finishing this essay.	63%	68%
36	I was able to write a draft essay in this test, then wrote it again neatly within the given time.	68%	49%

INTERNATIONAL JOURNAL OF
Practical and Pedagogical Issues in English Education
ISSN: 2980-9533
Arad Institute of Higher Education

37	After finishing the essay, I also thought for a while of those statements or thoughts I removed	56%	52%
38	I felt it was easy to review or revise the whole essay.	43%	38%

As indicated in Table 1, except for a couple of items that belong to different stages of writing, the findings indicated that the test takers went over different stages of writing, which are goal setting, topic/ genre modifying, generating, organizing, and translating to compose a text to perform their tasks. In other words, the tasks were successful in eliciting writing stages that are involved in writing in an authentic context.

Research question two: Do the contextual characteristics of the tasks and administration approve the suitability of the tasks for the participants?

As Weigle (2007) argues, the most significant part of a valid local writing test is the content of the test. The first part of the context validity has to do with the tasks. The response format of the tasks was the first issue; performance-based open-ended writing tasks were included in this administration; in both educational and non-educational contexts, the participants were required to write texts that resemble the present tasks; expert judgment results also admitted the similarity of these tasks to real-life tasks. The purpose of the task was stated clearly in simple words, so the participants were able to answer the tasks with no major difficulty. The purpose of the present tasks can be stated as tapping the participants' ability to write a report on a familiar topic and write a short descriptive or argumentative article within a limited time. The participants had been informed about the way their written products were scored a week before the examination. The scoring rubric (Jacobs, et al., 1981) was given to the participants to make sure about their cognizance of the criteria and the scores weighting. In regard to the time constraint, a period of 85 minutes was allocated to these two writing tasks.

The second section of context validity has to do with the administration issues. It was attempted to make the physical conditions in the examination hall the same for different participants. For sure, different factors such as air-conditioning, lights, and chairs were examined before the exam to ensure the participants' comfort. The participants were given a bottle of water, a pen, and a sketch paper. Their cell phones were collected before the examination. During the examination, 3 invigilators were sitting in non-intruding places but walked through the seats once in a while to make sure about the absence of any sign of cheating. Since all the participants were in the same examination hall, the uniformity of administration was to a large extent guaranteed. The same amount of time was given to the participants. Even the pen and the paper that were given to the participants were of the same brand to avoid any sort of, even negligible, unfair condition. In order to avoid any possible commotion, all students had to sit until the time was up. The security of the questions was ensured by writing the test on a computer which was not connected to the internet and the copies were made just an hour before the exam.

Research question three: Do the scoring, consequential, and criterion-related validity evidence approve the suitability of the tasks?

Scoring Validity

As mentioned earlier, the scoring rubric, a copy of which was also given to the student, was Jacobs et al. (1981), which is an analytic well-known rating scheme for assigning scores to the written products of second language writers. The papers were scored by two raters. Both raters were Ph.D. candidates who had 5 and 7 years of teaching experience. In a formal session, the researcher and the two raters gathered and reviewed the scheme to make sure about the uniformity of the scoring process; a couple of papers were rated there independently, and then the results were juxtaposed and discussed to reach a uniform decision upon the scoring procedure. In order to examine the extent to which the raters were uniform in assigning scores, around 40 % of the papers were scored by both raters, and inter-rater reliability computed by Pearson correlation was found to be .93 which is an acceptable value.

Criterion-related Validity

To check the extent to which the scores were in line with those of the external measures, predictive criterion-related validity evidence was collected; in other words, the scores of the present study were compared with the mid-term writing test which was taken from a well-established exam, i.e., Cambridge FCE. The results of the Pearson correlation indicated high levels of go-togetherness with the directed task (r=.93), and free writing task (r= .86). The topic of the mid-term exam was as follows:

Every country in the world has problems with pollution and damage to the environment. Do you think these problems can be solved?

Write about:

1. Transport

2. Rivers and seas

3…….(your own idea)

The source: https://s3-eu-west-1.amazonaws.com/cbpt/2015/fce-writing/index.xhtml

Consequential Validity

To ensure the fairness of the testing procedure, potential biases related to gender and prior attendance in intermediate-level courses at the institute were examined. Independent sample t-test was run two times to check possible sources of bias. The results indicated that no significant difference was found between the scores of old students and newcomers (t = .383, $p < .05$) and between males and females (t= .491, $p <. 05$).

Furthermore, cognitive validity evidence suggested that the tasks elicited cognitive processes closely aligned with those required for real-life writing tasks. In addition, the context validity

evidence showed that the features of the tasks and task administration and setting were similar to those of the real-life writing situations. These results suggest that the test tasks were not suffering from the two validity threats which are construct under-representation and construct irrelevance (Messick, 1989). When tests are reported to be contextually and cognitively valid, these threats are minimized and the adverse Washback effect is avoided. In addition, the positive nature of the tasks to a large extent guarantees a positive effect on the learning process (Bachman, 1990; Weir, 2005).

Discussion and Conclusion

The findings demonstrated how researchers employed the well-regarded Weir & Shaw (2007) writing model to gather evidence supporting the intended interpretations and uses of a local writing test. The researchers could collect and present data to show that the test was a valid one. The findings indicated that the characteristics of test items were not in favor or against different religious, ethnical, or political groups or different experiential or biological groups (e.g., old students vs. newcomers; right-handed vs. left-handed test-takers). Furthermore, the results of the questionnaire administrations indicated that in order to compose the texts, test-takers went through different cognitive stages which were proposed in the employed writing theory.

With regard to context validity, scoring validity, criterion-related validity, and consequential validity, evidence was collected and presented to show the test tasks were contextually appropriate with regard to task setting, task demands, and administration. The findings also indicated that there was a significant consistency between the scores of the two scorers. The comparison of the scores obtained from the present tasks and a task from a well-known test (i.e., FCE) indicated a strong and significant relationship between the scores which provided the evidence related to criterion-related validity. In regard to consequential validity, two major threats that could affect the performance of test-takers were gender and their familiarity with the institute and its physical atmosphere; the results of the comparisons indicated no significant difference between different groups which shows the absence of bias with regard to these two factors.

As Bachman and Palmer (2010) and Huggins-Manley et al. (2024) state, one of the misconceptions about second language testing is teachers' inability to develop and conduct tests in an acceptable manner. This misconception can lead to two detrimental consequences which are the loss of teachers' confidence in their abilities and having a test which is developed by a test developer who is an outsider to the context. Enabling teachers to develop, administer, and collect a body of evidence to justify the use and interpretation of their tests is a feasible solution. Pre-service and in-service workshops can include a couple of sessions pertinent to practice test development and evidence generation (Hoomanfard et al., 2018; McNamara, 2001). Moreover, the review of low-stakes test validation reports such as the present study and Zainal (2012) can shed light on the process of test validation. The present study is a genuine example of the test developers' need to be accountable to different stakeholders. Upon the demand of the institute officials, the test developer who is a teacher of the institute, to demonstrate the appropriateness of the inferences drawn from the test scores, collected a body of evidence.

Allison (1999) and Ali and Kaivanpanah (2024) caution foreign language teachers that they should not be paralyzed by the complexity of terms used in language testing textbooks. In the

present study, since the test was a low-stakes local test, the researcher resorted to valid and manageable choices in the process of data collection. For instance, to collect cognitive validity evidence, the researcher selected the questionnaire rather than the think-aloud retrospective instrument due to feasibility issues. In regard to criterion-related validity, the researcher compared the scores with the mid-term exam, but not with the teachers' assessment or self-assessments which were other possible choices. For differential validity, the researcher selected a simple statistical measure (i.e., t-test), and did not employ more sophisticated tests such as Rasch Model. It seems logical to employ the easier choices since the test is not a high-stakes one and the decisions do not bear some weight in the testees' future.

All in all, it should be noted that the expectations from ESL teachers should be logical since they are already overwhelmed by an increasing workload (Yang, 2007); thus, school officials should take care of their teachers to avoid their professional burnout (Allright, 2005; Jafarigohar et al., 2018). At the same time, teachers should be well-equipped to be able to validate their own tests by the employment of simple and acceptable research methods to be taken as accountable testers to stakeholders. Local officials should take the issue of accountability seriously by shouldering responsibility for educating teachers and providing suitable incentives.

This study had a few limitations that can be avoided in future studies. This study employed a questionnaire to validate the developed test. Although questionnaires are regarded to be reliable sources of knowledge, the teacher could examine the students' cognitive processes using screencast or stimulated-recall interview data. The triangulation could ensure the precision of collecting and interpreting data. Although the simplicity of the validation process is emphasized in low-stakes tests, the teacher could use more sophisticated measures to assess a few students' cognitive processes. In addition, the present study focused on the micro-social consequential effects of the test. Given the low stakes of the developed test, while it seems reasonable to focus on micro-social effects, other studies can find easy-to-implement solutions to enable teachers to collect macro-social consequential effects as well. In addition to these suggestions for further research, other scholars can validate tests that examine L2 students' speaking, reading, and listening ability. Finally, the effects of collaborative validation projects conducted by teachers on individual teachers' language assessment knowledge.

References

Ali, A., & Kaivanpanah, S. (2024). Writing Assessment Literacy in an EFL Context: Insights from Iraqi Kurdish Teachers. *Applied Research on English Language*, *13*(1), 55-78.

Allan, D. (2004). *Oxford Placement Test*. Oxford University Press.

Allison, D. (1999). *Language testing and evaluation: An introductory course*. Singapore: Singapore University Press/ World Scientific.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly 19* (4), 453-476.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* Oxford University Press.

Brunfaut, T. (2023). Future challenges and opportunities in language testing and assessment: Basic questions and principles at the forefront. *Language Testing*, *40*(1), 15-23.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52, 281–302.

Fattah, Z. (2024). The Effect of Portfolio Assessment on Iranian EFL Learners' Writing Ability. *International Journal of Practical and Pedagogical Issues in English Education, 2(2), 13-34.*

Field, J. (2005). *Second language writing: a language problem or a writing problem?* Paper presented at IATEFL 'Writing Revisited' Conference, Cambridge, 25-27 February 2005.

Ginther, A., & Grant, L. (1997). The influence of proficiency, language background, and topic on the production of grammatical form and error on the Test of Written English. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 385–97). University of Jyväskyla.

Granena, G. (2023). Cognitive individual differences in the process and product of L2 writing. *Studies in Second Language Acquisition*, *45*(3), 765-785.

Hoomanfard, M. H., & Meshkat, M. (2015). Writing on a computer and using paper and pencil: Is there any difference in the internal cognitive processes? *GEMA Online Journal of Language Studies*, *15*(2), 17-31.

Hoomanfard, M. H., Jafarigohar, M., Jalilifar, A., & Masum, S. M. H. (2018). Comparative study of graduate students' self-perceived needs for written feedback and supervisors' perceptions. *Journal of Research in Applied Linguistics*, *9*(2), 24-46.

Huggins-Manley, A. C., Huang, J., Danso, J. A., Li, W., & Leite, W. L. (2024). Classroom assessment and instructional modes: An exploration of school-level contextualized psychometric challenges. *The Journal of Experimental Education*, *92*(3), 513-530.

Izadpanah, M. A., Rakhshandehroo, F., Hoomanfard, H. M., & Mahmoudikia, M. (2014). On the consensus between holistic rating system and analytic rating system: A comparison between TOEFL IBT and Jacobs et. al. composition profile. *International Journal of Language Learning and Applied Linguistic World*, *6*(1), 170-187.

Jacobs, H. L., Zinkgraf, D. R., Wormuth, V. F., Hartfiel, V. F. & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Jafarigohar, M., Hoomanfard, M. H., & Jalilifar, A. (2018). A Typology of Supervisor Written Feedback on L2 Students' Theses/Dissertations. *Iranian Journal of Applied Linguistics*, *21*(2), 43-87.

Jensen, C. & Hansen, C. (1995). The effect of prior knowledge on EAP listening-test performance. *Language Testing, 12*(1), 99-119.

Jia, W., & Zhang, P. (2023). Rater cognitive processes in integrated writing tasks: From the perspective of problem-solving. *Language Testing in Asia*, *13*(1), 50.

Kane, M. T. (2006). Validation. In Brennan, R. L. (Ed.), *Educational measurement* (pp.18–64). American Council on Education/Praeger.

Kim, M., Tian, Y., & Crossley, S. A. (2021). Exploring the relationships among cognitive and linguistic resources, writing processes, and written products in second language writing. *Journal of Second Language Writing*, *53*, 100824.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing, 18*(4), 333-350.

Mehrpour, S., Hoomanfard, M. H., & Vazin, E. (2023). Peer Feedback Accuracy in Synchronous and Asynchronous Computer-Mediated Conditions in an EFL Context. *Iranian Journal of Language Teaching Research*, *11*(1), 97-116.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds.). *Test Validity*. Lawrence Erlbaum Associates. 33-45.

Messick, S. (1989). 'Validity.' In Linn, R. L. (ed.) *Educational measurement* (pp. 13-103). Macmillan/American Council on Education.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of assessment arguments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Norton, B. (1997). 'Accountability in language assessment.' In Clapham, C. and Corson, D.(eds) *Encyclopedia of Language and Education*, vol. 7: *Language Testing and Assessment* (313–322). Kluwer Academic Publishers.

Shaw, S., & Weir, C. J. (2007). *Examining writing in a second language, studies in language testing 26*. Cambridge University Press/Cambridge ESOL.

Sunderland, J. (2000). New understandings of gender and language classroom research: texts, teacher talk and student talk. *Language Teaching Research* 4 (2), 149-173.

Van Dijk, T. A. (2011). *Discourse studies: A multidisciplinary introduction* (2nd ed.). SAGE Publications.

Wang, T., & Lajoie, S. P. (2023). How does cognitive load interact with self-regulated learning? A dynamic and integrative model. *Educational Psychology Review*, *35*(3), 69.

Weigle, S.C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing, 16*(3), 194-209.

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

Yakut, I., & Bada, E. (2022). Interlanguage development of Turkish speakers of English: Exploring oral and written communication strategies. *Journal of Teaching English for Specific and Academic Purposes*, 611-626.

Zainal, A. (2012). Validation of an ESL writing test in a Malaysian Secondary School Context. *Assessing Writing 17*(1), 1-17.