MLS - EDUCATIONAL RESEARCH (MLSER)
http://mlsjournals.com/Educational-Research-Journal
ISSN: 2603-5820

*(2025) MLS-Educational Research, 9*(1), 193-209. doi.org/10.29314/mlser.v9i1.2523.

# VALIDITY, RELIABILITY, AND THREATS TO VALIDITY IN VIRTUAL ENGLISH I CLASS ASSESSMENTS IN THE FOREIGN LANGUAGES DEPARTMENT OF THE NATIONAL AUTONOMOUS UNIVERSITY OF HONDURAS

## VALIDEZ, CONFIABILIDAD Y AMENAZAS A LA VALIDEZ EN LAS EVALUACIONES DE LA CLASE DE INGLÉS I VIRTUAL DEL DEPARTAMENTO DE LENGUAS EXTRANJERAS DE LA UNIVERSIDAD NACIONAL AUTÓNOMA DE HONDURAS

**Bessy Valeska Mendoza Navas**[1]
Universidad Nacional Autónoma de Honduras, Honduras
(d.lenguasextranjeras@unah.edu.hn) (https://orcid.org/0000-0003-1898-256x)
**Leonel Armando Madrid Argenal**
Universidad Nacional Autónoma de Honduras, Honduras
(d.lenguasextranjeras@unah.edu.hn) (https://orcid.org/0000-0002-8354-7858)

**Manuscript information:**

**Recibido/Received:**06/12/2023
**Revisado/Reviewed:** 12/02/2024
**Aceptado/Accepted:** 09/04/2024

**Keywords:**
representations, validity, reliability, threats to validity, virtual English

### ABSTRACT

This article corresponds to a study carried out within the framework of the Master's Thesis called: "Students' representations on threats to validity, and reliability of assessments of the Virtual English I class of the Department of Foreign Languages of the National Autonomous University of Honduras" with the purpose to have references that serve to eliminate threats and correct security gaps to assessments of the virtual English I class and at the same time give out the background to develop a valid and as secure as possible evaluation system in virtual language classes of the UNAH. The study adopts a mixed methodology, the instrument used to collect data were two questionnaires applied to students of virtual English I of the III PAC 2022. The instruments were filled out online in a survey form using Microsoft Forms. The quantitative analysis was performed with IBM Spss software the quantitative analysis with Atlas/Ti. The results of this research have allowed validation of the Null Hypothesis about the evaluations of the virtual English I class do not have the validity or security that reflects the acquisition of language skills achieved by students according to the CEFR corresponding to 56 hours. Although the validity manages to comply with 4 of 5 validity shreds of evidence suggested by (Downing, 2003; Messick, 1989), some inferences put the safety of the evaluations at risk due to the open possibility of cheating and plagiarism in the evaluations.

[1] Corresponding author.

## RESUMEN

**Palabras clave:**
representaciones, validez, fiabilidad, amenazas a la validez, Inglés Virtual

El presente artículo corresponde a un estudio realizado en el marco de la tesis de maestría denominada: "Representaciones de estudiantes sobre las amenazas a la validez, y seguridad de las evaluaciones de la clase de Inglés I Virtual del Departamento de Lenguas Extranjeras (DLE) de la Universidad Nacional Autónoma de Honduras (UNAH)."con la finalidad de tener referencias que sirvan para eliminar amenazas y corregir brechas de seguridad a las evaluaciones de la clase de Inglés I virtual, y que al mismo tiempo sirva de antecedentes para desarrollar un sistema válido y lo más seguro posible de evaluación en clases virtuales de lengua del DLE de la UNAH. El estudio adopta una metodología mixta: los datos analizados fueron tomados de 2 instrumentos que se aplicaron a estudiantes de inglés I virtual del III periodo académico (PAC) 2022. Los instrumentos se llenaron en línea en forma de encuesta usando Microsoft Forms. El análisis cuantitativo se realizó con el software de IBM SPSS, y el análisis cuantitativo con Atlas/Ti. Los resultados de esta investigación han permitido validar la Hipótesis Nula: las evaluaciones de la clase de inglés I virtual no cuentan con la validez ni la seguridad que reflejen la adquisición de competencias lingüísticas alcanzadas por los estudiantes según el MCER correspondientes a 56 horas. Si bien la validez logra cumplir con 4 de 5 evidencias de validez sugeridas por Downing, 2003 y Messick, 1989; hay inferencias que ponen en riesgo la seguridad de las evaluaciones debido a la posibilidad abierta de trampas y plagio en las evaluaciones.

## Introduction

This research has its genesis in the doubts that language teachers, specifically teachers who teach the virtual English I class at UNAH, have about the validity and security of online assessments. Doubts about not knowing if it is indeed the students who make the evaluations or other people, if students copy or plagiarize, or if they modify previous work. This would be a false indicator of the validity of the results in the acquisition of language skills virtually. Other questions arise as to whether oral production and interaction can be developed and assessed online.

The teaching, acquisition, and evaluation of language skills in virtual environments can be affected by various factors such as aspects, previous technological and linguistic training, availability of technological resources, study habits and disciplines, time available, academic load and pressures, formation of values such as honesty, relevance of programs and content, appropriate methodology and practice of virtual teaching, design and application of assessments, feedback, among others. These circumstances make it complex to visualize the effects on learning and its results in the development of expected skills and competency achievements. It is therefore necessary to define, on the basis of precise diagnoses and specific research, the criteria, methodologies and instruments required for a solid and pertinent evaluation of the different components of a virtual environment. The present research work is justified by the need to seek solutions to the validity and security of the evaluations in the virtual English I class of the Foreign Languages Department of the National Autonomous University of Honduras (UNAH) with the objective of reducing the threats to the online evaluation processes.

Evaluation generates multiple reactions in students, not all of them pleasant. We usually say that organizational activities must be evaluated, that in order to improve we must evaluate, that without evaluation it is impossible to know exactly where we are, among other statements, but when it comes to personal evaluation the situation changes, especially when it is summative with high impact, and the news that we will be evaluated often generates a feeling of discomfort, anguish or even fear, and if we can defer or exempt it on occasions, we do so. (Sanchez, 2021).

Evaluating requires a reflective and mature attitude, resources to carry it out, personnel with training and experience in its methodological and technical nuances, time to plan, carry it out and analyze it, as well as infrastructure, all to document the different stages of the process. Effective evaluation processes require participatory organizational structures, not so vertical or hierarchical, that are willing to accept the results with enthusiasm and transparency, to act accordingly and to improve the structure, processes and results of the system. Systemic thinking and long-term vision are required for the evaluation process to be properly integrated into the system, as well as the active participation of the people who make up the different elements of the system. In summary, evaluation is not an easy or simple task; it requires individual and collective effort, as well as support from the various levels of the organizational structure. (Sanchez, 2021).

## Literature Review

The previous introduction leads to a bibliographic review of the central aspects addressed in this work, with the aim of providing reference arguments that can support the results of the study presented.

There are many definitions of the term "assessment" in education, Miller defines it as: "an umbrella term that includes a range of procedures for acquiring information about student learning and the formation of value judgments..." (Miller, 2012). This implies a systematic process of gathering information through the application of various instruments, such as written or oral examinations, to be analyzed with methodological rigor and thus provide the basis for decision making. The most recent edition of the AERA-APA-NCME Standards for Educational and Psychological Testing defines "assessment" as: "systematic method of obtaining information, used to formulate inferences about the characteristics of people, objects, or programs; systematic process for measuring or evaluating the characteristics or performance of individuals, programs, or other entities for the purpose of making inferences; sometimes used as a synonym for testing" (AERA, APA, and NCME, 2014).

Regardless of the technical definitions we use of assessment and its proximate concepts, teachers who have interactions with students should internalize assessment from a deeper view, as suggested in 1977 by Rowntree, when he says that a person consciously obtains and interprets information about another person's knowledge and understanding, skills and attitudes when that person interacts with another person directly or indirectly.

In recent years, the concept of "assessment of-for-as learning" ("assessment of-for-as learning" in English) has gained momentum, which aims to modify the emphasis that has existed on summative assessment, tests and grades, towards a broader and more integrated picture that leads us to anchor the entire assessment process with learning, the fundamental goal of the educational process (Ashford-Rowe et al, 2014; Bennett, 2015; Harapnuik, 2021; Maki, 2010; NFETLHE, 2017a). Teaching, learning and assessment are inextricably linked concepts and activities, and the alignment of these elements with curriculum planning, design and implementation is indispensable and becomes a key element for educational success.

- Learning assessment. According to several authors this type of evaluation is equivalent to summative evaluation, to document that learning occurred and the level of learning. Its nature is to evaluate activities that have already occurred, after or at the end of a learning period, and it emphasizes quantitative and numerical aspects, being associated with grades. When this assessment has significant consequences on the student, it is referred to as a "high impact assessment". In this type of assessment the main actor is the teacher or the organization that applies the assessment, who are the main decision makers, and the student is a passive participant who receives or to whom the exam or test is applied, in contrast to assessment for learning.
- Assessment FOR learning. As previously commented, the main goal of assessment should be to improve learning, not only to measure it, so when we talk about assessment for learning we refer to assessment traditionally called formative, linked to feedback (Maki, 2010; Man Sze Lau, 2016; Martínez Rizo, 2009; Wiliam, 2011). This assessment occurs throughout the teaching and learning process, is more longitudinal and represents a dialogue that occurs between teachers and students throughout their multiple interactions. It is

focused on helping the student, identifying their areas of opportunity and achievements, to guide them to progress in a better way in the educational process, without generating stress or wear, treating them as a person. It aims to move from an action that is done *to the* student, to a process that is done *with* the student. This assessment is inseparable from teaching and strongly supports learning, if carried out with professionalism and responsibility.

- Evaluation AS learning. In this type of evaluation the student is empowered, has greater responsibility in the learning process and can be the key decision maker. Students need to acquire skills for the use of basic evaluation concepts in their personal development. Self-directed lifelong learning, autonomous learning, critical thinking, among others, require evaluating data and information on work and life situations, analyzing them, establishing value judgments, and making decisions on personal and professional issues. All this requires self-evaluation and the ability to make decisions based on the evaluation of complex contexts and realities. Although the teacher generally holds the hierarchical power in the formal educational process, assessment as learning moves this *locus* of external control to a more intrinsic control by the student body. However, the student requires support from teachers and peers to fully exercise the aforementioned skills. Assessment as learning helps students learn how to learn, fosters metacognition and self-regulated learning.

It is very important to review what is being taught and what students are expected to learn; the alignment of teaching, learning and assessment are essential in any field of education, because in this way the teacher ensures that what is being taught and learned is indeed what is being assessed. (Basabe et al. 2020).

It would be a wasted effort what is done in the evaluation if we do not have these objectives. Formal education is guided by processes clearly organized by a curriculum, syllabus, and subject matter programs. These plans and programs are the guide at the time of our evaluations and to make sure that indeed, what we teach and expect our students to have learned, is what we are actually going to evaluate.

It is therefore essential to have a thorough knowledge of these two documents in order to carefully review each of their elements and the function or raison d'être of each one. In the case of the study plan, it locates the subject and the connections it has with the rest of the subjects that integrate it. This can guide the evaluation of the objectives, not only of the subject, but also of the curriculum as a whole, thus contributing to the training of students. The curriculum is a system in which several gears are derived, and it is necessary that all of them work properly to achieve the established purposes. This is precisely what we mean when we talk about alignment of teaching, learning and assessment.

### *Validity, Reliability and Threats to Validity*

Throughout our lives as teachers we conduct many assessments to try to learn about the level of knowledge or performance of our students. This process involves the elaboration, application and interpretation of different types of tests: diagnostic, formative and summative. Regardless of its purpose, the goal of any assessment includes the identification of the level of some construct, such as written communication competence, oral communication competence or interaction in the case of foreign languages.

Assessment results should ideally reflect in an accurate and reproducible manner what is intended to be assessed, in order to be able to rationally interpret the assessment results and to be able to make inferences and decisions on a sound basis. When assessing students on a particular topic, you want to identify the process and learning outcomes that allow you to infer the level of performance on the constructs of interest. After applying the evaluations, we obtain results in the form of scores that help us to make decisions, which lead to the following questions: are we evaluating exactly what we want to evaluate, what do the results imply with respect to the student's academic progress, if it is a summative evaluation, what is the minimum grade to pass the course, how reproducible is the measurement, among many others. Evaluation in education is an increasingly sophisticated and research-based discipline that requires incorporating fundamental academic concepts to be carried out with professionalism and methodological soundness (Instituto Nacional para la Evaluación de la Educación, 2017).

The most important conceptual pillar of evaluation in education is validity. Today, the concept of validity has evolved from the traditional "measuring what it is intended to measure," to a broader and deeper model, in which it "refers to the degree to which evidence and theory support interpretations of a test's scores for proposed uses of the tests" (AERA, APA, & NCME, 2018). It is a set of actions that are placed throughout the evaluation process, to support the interpretation of the results and thus generate inferences. Validity analysis, or validation, is the process by which we evaluate the evidence presented to determine what the degree of validity is (Cook and Hatala, 2016). It can be performed for different types of examinations, diagnostic, formative and summative, although it is particularly relevant for high impact summative evaluations.

Traditionally, validity in education was classified as "the 3 Cs": content, criterion and construct validity (Cronbach and Meehl, 1955). In the current definition this distinction has disappeared, since the current model proposes different sources of evidence that shed light on different aspects of validity, not that they reflect different types of validity. Validity is a unitary concept, so all validity is considered to be construct validity.

Subsequently, in the late 20th century, a new validity framework was proposed and accepted by the major educational assessment and psychological testing organizations (American Educational Research Association et al., 2018), incorporating the holistic concept of construct validity. This model establishes that, in order to determine the degree of validity of the uses and interpretations of the results of an evaluation, several elements must be provided to demonstrate it (Downing, 2003). This scheme proposes the following elements as five sources of validity evidence (Downing, 2003; Messick, 1989):

- Evidence based on the content of the test.
- Evidence based on response processes.
- Evidence based on internal structure. The internal structure presents three basic characteristics: dimensionality, differential functioning and reliability (Rios and Wells, 2014). When designing the test, it must be determined which dimensions are to be assessed on the construct of interest, and this information is described in the test specification table.
- Evidence based on relationships with other variables.
- Evidence based on the consequences of the test. Test results.

### *Validation*

Validation is a process that should be planned at the same time as the test is designed, to ensure that the necessary sources of evidence are available to obtain the highest possible degree of validity in the interpretation of the test results. The following is a suggested way to carry out this process.

1. Specify the uses and interpretations of the scores

1a. Formulate the uses and interpretations. The uses and interpretations of the scores obtained in a test are different concepts and both should be clarified from the beginning of the test design.

1b. Establish the hypotheses. Hypotheses are questions we can ask ourselves about the evaluation being developed. They must be proven by means of the aforementioned sources of evidence.

2. Evaluate sources of evidence

2a. Create a plan to test the hypotheses. Based on the hypotheses selected, sources of evidence are sought and the corresponding information is gathered.

2b. Evaluate the evidence and formulate a judgment. In this last step, all the evidence is evaluated in order and the degree of validity of the interpretation of the test scores evaluated is established. This grade will depend on the quality of the evidence presented and also on the most important evidence, depending on the test.

### *Threats to Validity*

In addition to analyzing the sources of evidence of validity, it is suggested to identify elements that may affect the degree of validity of the evaluation results. This step is important because it gives strength to the decisions made based on the test results. Items that reduce the degree of validity are called threats to validity; they are so called because they interfere with the correct interpretation of the scores (Carrillo-Ávalos et al., 2020; Downing and Haladyna, 2004). These threats may be present in any type of assessment. In general, two types of threats to validity are recognized: construct underrepresentation (CS) and variance irrelevant to the construct (VIC) (Downing, 2003; Messick, 1989).

The first is the threat to validity due to underrepresentation of the construct. This refers to the fact that there is an inappropriate representation of the domains explored in the assessment of the content to be assessed by the tests. For example, when a test has too few items or too few questions that do not properly explore the area of knowledge to be reviewed. Another example is the distribution of reagents that do not faithfully follow the specification table. So some areas end up being over-explored and others under-explored. There are even times when there are areas that are not even explored in an exam. This obviously affects the validity of the use of the test. Another example is many items, many questions, that explore low-level cognitive processes, such as memory or factual data recognition, while the teaching objectives are ideally higher-level, such as application or problem solving. Another threat to validity, which has become increasingly important, is the phenomenon of teaching to the test. This means that the teacher overemphasizes in class what is to be included in the exam, thus distorting the curriculum, the educational goals and in general, the whole process; this has come to occur to such an extent that some teachers use test items in class to artificially increase their students' grades and thus improve the evaluations of their group or even their institution in this world of educational accountability.

The second major type of threat to validity is what we call construct-irrelevant variance. This refers to variables that systematically interfere with the ability to interpret the evaluation results in a meaningful way, and that cause, shall we say, noise in the measurement data. Examples of this type of threat to validity are reagents that have been developed with deficiencies and are flawed. Writing good test question questions is both an art and a science, and requires training and experience. It is not as easy as we often think. Another example is the problems that occur with test security and with information leakage or cheating on the test, cheating, using what we call accordions, so that the test result does not accurately reflect what the person really knows.

This obviously invalidates the test results, with complex ethical and resource implications, such as retesting, re-testing, or taking repressive measures with students. Most of the departments that offer online classes at UNAH have item banks that are not very large or do not have an item bank at all, so overexposing test questions becomes a major operational problem. On the other hand, creating a punitive or punitive culture around evaluations is not the message that teachers should ideally give to students, so these aspects should be taken into account when considering how to respond when these types of irregularities occur. There is also something called test-taking cunning; this occurs when students prepare with test-taking strategies and may get scores that do not necessarily reflect what they know, especially on tests that are not well done.

## Method

The research that we have carried out is inscribed within the methods of research in language learning as a non-experimental *ex post facto* field study, at a descriptive MIXED level, but with a correlational characteristic.

### *Participants*

On a self-selected sample of 163 subjects in single cross-section, the study has been conducted with students from 5 classes (Sections: 0702, 0800, 1005, 1401, 1802) who were taking General English I in the III academic period (PAC) 2022 in the Department of Foreign Languages of the National Autonomous University of Honduras.

### *Research Instrument*

A mixed self-administered questionnaire with both closed and semi-closed items was applied. The frequency of response for each item is presented in tabular and graphical form. Finally, open coding and selective coding have been implemented for textual citations.

### *Data Analysis*

For the analysis and interpretation of the data , different categories were chosen according to the main issues raised in the research: evaluation (E), security in evaluations (SE), use of technological tools (UHT), improvements (M) and level of satisfaction of the class (NS). In each category, a series of subcategories were identified in response to indicators provided by the various informants and directly linked to the main themes selected in advance, which made it possible to manage the accumulation of information gathered during the research and to present the results in accordance with the proposed objectives. Regarding the presentation of the results and interpretation of the open-ended questions used in the questionnaire that provide textual information, opinions, explanations, justifications, the analysis was implemented from the perspective of the different categories, entering into the respective subcategories defined.

## Results

With the findings found throughout this research we were able to validate our null hypothesis ***Ho:*** The evaluations of the virtual English I class of the Department of Foreign Languages of the UNAH do not have the validity nor the security to reflect the acquisition of linguistic competencies achieved by the students according to the CEFR corresponding to 56 hours.

The findings revealed that:

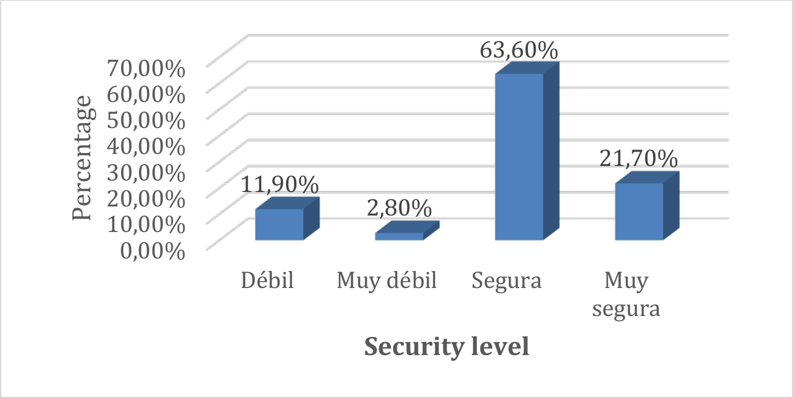
- The overall course evaluations are in the range of .854 of Alpha Cronbach's.
- Threats to the validity and security of English I class assessments are around 78.2%, which demands urgent updating of the class.
- While the assessments show a safe percentage in the development, at the time of response, there are several threats that need to be addressed urgently.

The analyses of inferences made to the validity of the evaluations, in relation to their security variable, gave the following results that show latent threats to the validity of the evaluations:

- 14.7% of the respondents stated that the level of security for evaluations and tasks was weak or very weak.

**Figure 1**

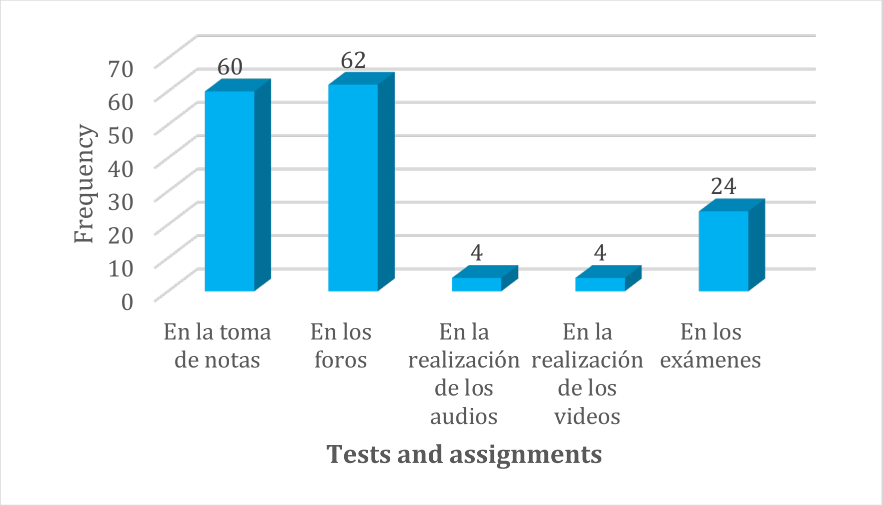
*Item C1. P12 security level*



The activities in which it is easiest to cheat or copy according to respondents, and which present threats to validity are:

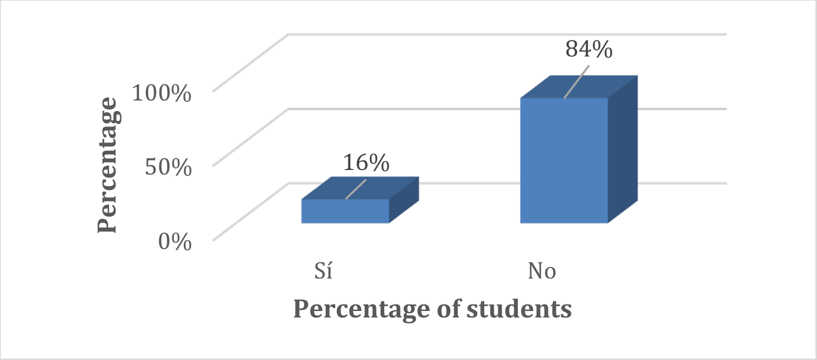
- Forums (40%)
- Note taking (39%)
- Exams (16%)
- Audio production (3%)
- Video production (3%)

**Figure 2**
*Item C1. P16 easiest test or task to cheat on*



Respondents reported hearing in 16% that their peers had copied or cheated on assessments (C1.p18).

**Figure 3**
*Item C1. P18 English I students copy or cheat*



The answers to question C1.p18 show the ways in which students cheat in which they highlight:

- Using an assignment already presented in previous classes c1.p18-c,
- Asking someone else to do homework for him/her, a family member, or paying someone bilingual c1.p18-a,
- Having the internet, books, or notes open when taking the exam, committing plagiarism over the internet c1.p18-b,

**Table 1**

*Item C1. P19 how do they cheat in the virtual English I class?*

| **Ways to cheat on virtual English I assessments, according to respondents.** | | | |
|---|---|---|---|
| **Code** | **Form** | **Frequency** | **%** |
| c1.p18-a | Asking someone else to do homework for you, a family member, or paying someone bilingual. | 45 | 25% |
| c1.p18-b | Having open internet, books, or notes when taking the exam, Committing plagiarism over the internet | 34 | 18.90 % |
| c1.p18-c | Using a task already presented in previous classes | 55 | 30.60 % |
| c1.p18-d | Working with another on an individual task | 16 | 8.90% |
| c1.p18-e | Using digital tools to modify texts, changing the name of the lessons | 9 | 5% |
| c1.p18-f | Using digital tools to modify audios | 2 | 1.10% |
| c1.p18-g | Using digital tools to modify videos | 2 | 1.10% |
| c1.p18-g | Using audio tools to fake the voice | 1 | 0.60% |
| c1.p18-i | Others (suggested by students) | | |
| c1.p18-i-1 | Sharing screenshots of exams, having someone else do the exam for you or doing it with someone else, sharing reviews with others, using phones when taking exams | 11 | 6.10% |
| c1.p18-i-2 | Using translator or looking for someone to translate for you | 2 | 1.10% |
| c1.p18-i-3 | Presenting previous class assignments when repeating the class | 2 | 1.10% |
| c1.p18-i-4 | Copying answers from forums | 1 | 0.60% |
| | **Total, frequency and percentages** | **180** | **100%** |

Therefore, the Null Hypothesis is tested, although the validity manages to comply with 4 out of 5 evidences of validity, there are inferences that put the security of the evaluations at risk due to the open possibility of cheating and plagiarism in the evaluations.

On the other hand, the results revealed that despite threats to validity in the assessments, the virtual English I class helped students to:

- Written production, the class contains 12 video lessons of about 1 hour each (12hrs in total). Students should watch the lessons, and take notes from a note-taking guide for each lesson. In addition, students were required to complete written assignments in forums.
- Improve oral and written comprehension by watching videos. The video lessons, in addition to helping improve writing, also helped improve listening and reading comprehension.
- Perform oral productions through audio and video. Oral production or *speaking* was developed and evaluated through the production of audios through a simple tool (Vocaroo.com), and through the production of videos to ensure that the person in the video is the student who performs the oral production.
- Desire to learn more of the language. The data is relevant because a desire to learn more of the language was sown through the class.

Other contributions made by the class were to improve reading, engage in simple conversation, gain confidence in speaking, achieve autonomy in learning, develop technological skills and be more organized in study time. Less than 1% (0.5%) said that the class had not helped them at all, and 0.5% said that it had helped them in other aspects, but did not mention what those aspects were. Despite the threats to validity and security in the evaluations, the class contributed to the students' linguistic formation, and to the development of values and motivation to learn more of the language.

## Discussion and Conclusions

This research had as its main objective to analyze the representations of English I - III PAC 2022 students of the Department of Foreign Languages of the UNAH about the threats to the validity and security of the online assessments of the virtual English I class. The research was conducted in order to have data to eliminate threats and correct security gaps in the evaluations of the virtual English I class, and at the same time serve as a reference to develop a valid and as secure as possible evaluation system in virtual language classes of the DLE of the UNAH. The researcher started by testing the following 2 hypotheses:

1. Null Hypothesis $\boldsymbol{Ho}$: The evaluations of the virtual English I class of the DLE of the UNAH have neither the validity nor the security to reflect the acquisition of language skills acquired by students according to the CEFR for 56 hours.
2. Alternative Hypothesis $\boldsymbol{Ha}$: The evaluations of the virtual English I class of the UNAH DLE DO have the validity and security to reflect the acquisition of language skills acquired by students according to the CEFR corresponding to 56 hours.

To test the hypotheses, we followed the null hypothesis form suggested by Sheaham (2006), and five sources of validity evidence suggested by Downing, 2003; Messick and 1989:

1. Evidence based on the content of the test. (The evidence rests on the proposals of Sireci and Faulkner-Bond, 2014):

a. Domain definition. The detailed description of the content areas and cognitive skills to be assessed from the construct defined in the curriculum and the learning activity outcomes were analyzed. It was found that the content areas are framed in the descriptors of the CEFR with the cognitive linguistic skills typical of an A1 level of the same Framework.

b. Domain representation. We analyzed in the tests whether the questions were set according to the learning objectives or goals and found that they were.

c. Domain relevance. The items were found to be important with respect to the aspect of the construct being measured in the class.

d**.** Appropriate test design procedures. The test items were tested prior to the start of the class in pilot projects. The review of test items is done in each period by content experts to ensure their technical accuracy. They verify that they are well elaborated.

e. Credentials of test developers, item developers, and content experts. The evaluations were prepared by experts in language teaching and experts in content design and management of virtual platforms from the Directorate for Educational Innovation (DIE).

2. Evidence based on response processes. Although in the evaluations of the English I class, there are exams with multiple choice questions, the evaluations demand

oral and written productions to verify if the student applies the acquired knowledge to real life (introduce himself, talk about his activities, his family, among others). the validation of the correct answer sheet, the quality control of the report of the results, among others, as suggested by (Downing, 2003), was carried out by experts from the DLE and the DIE.

3. Evidence based on internal structure. Three basic characteristics were analyzed as suggested by Ríos and Wells, 2014:

a. Dimensionality. (oral and written comprehension, oral and written production, interaction).

b. Differential functioning (Leenen,2014; Rios and Wells, 2014). We analyzed 8,505 results of 45 tests applied to 189 students from 5 different sections of English I- III PAC 2022. The same test was applied to both men and women of different ages.

c. Reliability. A *Cronbach's Alpha reliability scale analysis was performed with a high score of .854.*

4. Evidence based on relationships with other variables. Another test with international standards such as IELTS, TOEFL, TOEIC, Cambridge, or others, could not be taken due to financial issues.

5. Evidence based on the consequences of the test. To this effect, respondents were asked about what had helped them most in the class c1.p25 providing evidence of the consequences of the tests. For future research work on the same problem, it is recommended, as Lane, 2014 does, to conduct interviews, focus groups, questionnaires, to find out what are the most important components of academic programs and their points of greatest impact in the area of language knowledge.

After the analysis, it is concluded that the validity complied with 4 of the 5 evidences, it cannot be said if it complies or not with evidence # 4 based on the relationships with other variables since it could not be analyzed with other variables because there is no similar test that is economically accessible to all respondents.

Although it was found that the evaluations do comply with the variable of validity of the evaluations according to 4 of the 5 evidences suggested by Downing, 2003 and Messick, 1989, the variable of security in the evaluations should also be analyzed so that the validity has "a holistic and integrative evaluative judgment that requires multiple sources of evidence for its interpretation", and "that attempts to answer the question: ¿what inferences can be made about the person based on the test results?" (Downing, 2003).

The analyses of inferences made to the validity of the evaluations, in relation to their security variable, gave the following results that show latent threats to the validity of the evaluations:

- 14.7% of the respondents stated that the level of security for evaluations and tasks was weak or very weak.
- Forums, note taking, were the activities in which it is easiest to cheat or copy according to the respondents, and which present threats to validity.
- Respondents reported hearing in 16% that their peers had copied or cheated on assessments (C1.p18).
- The answers to question C1.p19 show the ways in which students cheat, in which they excel:
    - using an assignment already presented in previous classes c1.p18-c,
    - asking someone else to do homework, a family member, or paying someone bilingual c1.p18-a,

- having the internet, books, or notes open when taking the exam, committing plagiarism over the internet c1.p18-b,
- other forms of cheating are described in the descriptive table in question C1.p18 which summarizes the respondents' answers.

Therefore, the Null Hypothesis is tested, although the validity manages to comply with 4 out of 5 evidences of validity, there are inferences that put the security of the evaluations at risk due to the open possibility of cheating and plagiarism in the evaluations.

## Recommendations

To ensure validity and safety in the evaluations, it is recommended:

1. Update the programs, methodology, and evaluation of the virtual English I class.
2. Create different versions of exams with a large bank of questions and answers so that different types of exams can be applied.
3. The review of note-taking assignments should be face-to-face to reduce the frequency of digital files being passed and/or modified. Note-taking should stay with the teacher. In the forums, the answer must be submitted before viewing the contributions of the other participants. Audio assignments should be transferred to video to avoid editing. If the teacher has doubts about the authenticity of the work, he/she should compare the voice and the student with the video presentation, which should be mandatory in order to start doing the homework. That the platform be configured in such a way that the student cannot advance if he/she does not have a grade in the presentation forum.
4. Conduct exams at class time to avoid passing questions and give feedback at a reasonable time afterwards to avoid displaying answers or give overall feedback via Zoom of corrections, without leaving answers open-ended.
5. A regulation should be created, socialized, and applied to sanction fraud or plagiarism in accordance with the Academic Norms and the Student Regulations of the UNAH.
6. Shortening the duration of video lessons, segmenting them, for note-taking.
7. To carry out a single face-to-face evaluation with a value of 60%, either oral and/or written, and that the value of the evaluations and assignments be converted to 40%. This will make the student worry about being more linguistically prepared for the final exam.
8. Evaluations should be done at class time with the camera on, and with the teacher's supervision, if it is on a platform specialized in exams, the better. To this end, a budget for licenses must be included in the Annual Operating Plan.
9. Create a values training program on honesty, ethics, and responsibility to be included in the course in order to reduce or eliminate fraud.
10. Create a plan for meetings, attendance, test proctoring, feedback to students from advisors.
11. Maintain ongoing meetings with classroom staff to identify, report, and correct threats to the validity and security of assessments, creation of resources to replace those that need to be updated.

## References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, y Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. AERA. https://www.testingstandards.net/open-access-files.html

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psico lógicas.* American Educational Research Association.

Ashford-Rowe, K., Herrington, J., & Brown, C. (2014). Establishing the critical elements that determine authentic assessment. *Assessment & Evaluation in Higher Education, 39*(2), 205-222. https://doi.org/10.1080/02602938.2013.819566

Basabe, L., Leal Falduti, B., & Tornese, D. (2020). *Diseño de exámenes con ítems de respuesta abier ta. Citep. Centro de Innovación en Tecnología y Pedagogía*. http://citep.rec.uba.ar/wp-content/uploads/2020/05/AcaDocs_D09_Dise%C3%B1o-de-ex%C3%A1menes-escritos-con-%C3%ADtems-de-respuesta-abierta-1.pdf

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education, 39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Carrillo-Avalos, B. A., Sánchez-Mendiola, M., & Leenen, I. (2020). Amenazas a la validez en evaluación: implicaciones en educación médica. *Investigación en Educación Médica, 9*(34), 100–107. https://doi.org/10.22201/facmed.20075057e.2020.34.221

Cook, D. A., & Hatala, R. (2016). Validation of educational assessments: a primer for simulation and beyond. *Advances in Simulation, 1*(1), 1–12. https://doi.org/10.1186/s41077-016-0033-y

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Downing, S. M. & Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327–333. https://doi.org/10.1046/j.1365-2923.2004.01777.x

Harapnuik, D. K. (2021). Assessment OF/FOR/AS Learning. [Blog] It's About Learning. Creating Significant Learning Environments. https://www.harapnuik.org/?page_id=8900

Instituto Nacional para la Evaluación de la Educación. (2017). Criterios técnicos para el desarrollo, uso y mantenimiento de instrumentos de evaluación. *Diario Oficial de la Federación.* https://www.inee.edu.mx/wp-content/uploads/2019/04/P1E104.pdf

Maki, P. L. (2010). *Assessing for Learning: Building a Sustainable Commitment across the Institution*. Stylus Publishing, LLC.

Man Sze Lau, A. (2016). "Formative good, summative bad?" – A review of the dichotomy in assessment literature. *Journal of Further and Higher Education, 40*(4), 509-525. https://doi.org/10.1080/0309877X.2014.984600

Martínez Rizo, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un Sistema más equilibrado. *Revista Electrónica de Investigación Educativa, 11*(2). http://redie.uabc.mx/redie/article/view/231

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13–103).

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2012). *Measurement and Assessment in Teaching* (11ª Ed.). Pearson.

McKean M. & Aitken E.N. (2016) Educational Renovations: Nailing Down Terminology in Assessment. In: Scott S., Scott D., Webber C. (eds) *Leadership of Assessment, Inclusion, and Learning. The Enabling Power of Assessment.* Springer, Cham. https://doi.org/10.1007/978-3-319-23347-5_2

National Forum for the Enhancement of Teaching and Learning in Higher Education (2017a). *Assessment OF/FOR/AS Learning*. https://www.teachingandlearning.ie/our-priorities/stu-dent-success/assessment-of-for-as-learning/

Ríos, J. & Wells, C. (2014). Evidencia de validez basada en la estructura interna. *Psicothema, 26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260

Rowntree, D. (1977). *Assessing students: How shall we know them?* Kogan Page Ltd.

Sánchez, M. & Martínez, A. (Eds.) (2022). *Evaluación y aprendizaje en educación universitaria: estrategias e instrumentos.* UNAM.

Sánchez, M. et al., (2021). *Evaluación del y para el aprendizaje a distancia: Recomendaciones para docentes de educación media superior y superior.* CUAIEED, UNAM. https://cuaieed.unam.mx/descargas/investigacion/evaluacion-del-y-para-el-aprendiza-je-V02.pdf

Sireci, S. & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100–107. https://doi.org/10.7334/psicothema2013.256

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3-14, https://doi.org/10.1016/j.stueduc.2011.03.001.