

# Meaningfulness, Generalizability and Consistency of a Spanish Placement Test: Insights from Assessment Use Argument Framework



## Authors:

Fati Osuman, Ph.D.  
*University of Ottawa*

## ABSTRACT

This study validates the Spanish Language Placement Test (SLPT) using the Assessment Use Argument (AUA) framework to evaluate the consistency, meaningfulness, and generalizability of score interpretations. Based on analyzed data from 105 students, including internal consistency metrics, item analysis, conditional standard error of measurement (CSEM), alignment with instructional tasks, and administrative observations, results showed the SLPT is internally consistent ( $\alpha = .94$ ) and procedurally stable but lacks precision at advanced levels. The narrow item difficulty range and increased standard error reduced reliability for higher-proficiency placements. A construct analysis grounded in grammatical knowledge revealed a limited focus on morphosyntactic and lexical skills, excluding key instructional components such as listening, writing, and sociocultural competence. These rebuttals weaken the meaningfulness and generalizability of scores. The study recommended expanding the test construct to include multiple language skills and offers implications for AUA research in non-English contexts. Future research should integrate stakeholder input and track longitudinal placement outcomes.

**Correspondence E-mail:** fatymaosuman@gmail.com

**Keywords:** Reliability, Assessment records, Score-based interpretations, Grammatical knowledge, Language test.

Language placement tests are designed to assess learners' knowledge and assign them to instructional levels where they can thrive. They must be carefully examined to justify their meaningfulness, generalisability, and reliability to ensure valid score-based interpretations. Accurate placement of students into language courses is essential for providing effective instruction, supporting learner engagement, and avoiding misalignment that can lead to frustration or disengagement (Lam, 2010; Hille & Cho, 2020). Misplacement affects students' motivation and learning outcomes and can result in additional costs for learners and institutions (Johnson, 2011; Johnson & Riazi, 2015).

The Spanish Language Placement Test (SLPT) is used to assign incoming university students to beginner, intermediate, or advanced Common European Framework of Reference (CEFR)-aligned Spanish courses. However, it has been in use for several years and has not been systematically evaluated and updated to reflect the changes to the courses' content, goals and shift in teaching methodology. Therefore, the SLPT stakeholders (including the coordinator and administrators) have raised questions regarding the extent to which the SLPT effectively reflects and assesses the communicative competencies targeted in the curriculum, whether it predicts the types of tasks students will encounter in their courses, and whether its results are consistent enough to support placement decisions.

Moreover, course instructors—often graduate students with limited training in placement testing—are tasked with identifying potentially misplaced students during the initial weeks of classes and recommending reassignments based on students' performance. However, their lack of expertise in assessment may lead to errors in these recommendations. To address this, SLPT stakeholders emphasise the need for a valid and reliable placement test to minimise measurement errors and ensure accurate course placements. In this context, an Assessment Use Argument (AUA) serves as a framework to evaluate whether the SLPT's placement decisions are valid, reliable, and fit for their intended purpose—assessing students' language abilities to assign them to appropriate courses. Using a combination of task analysis, alignment comparisons, reliability statistics, and placement outcome data, the AUA approach is structured around claims, warrants, backings, and rebuttals to present an argument that addresses the SLPT's strengths and limitations.

The study's significance lies in its theoretical contribution to the AUA framework as applied to a Spanish placement test—a relatively underrepresented context in the examples provided by Bachman & Damböck (2018), which has focused mainly on English-language classroom-based assessments. Beyond that, it provides empirical evidence to SLPT stakeholders regarding its qualities to support score-based decisions. The findings contribute to discussions about the best placement test design and use practices, particularly in communicative language programs.

## **Literature Review**

This section provides some background for the study, emphasizing how it has been informed by Bachman and Colleagues' works (2010, 2018). It then discusses the nature of language placement tests, the grammatical knowledge model (Purpura, 2004) and the Communicative Language Competence (CLC) as they relate to this study.

### *Assessment Use Argument Framework*

The Assessment Use Argument (AUA) framework was first introduced by Bachman (2003, 2005) and further refined in Bachman and Palmer (2010). Bachman and Damböck (2018) later adapted the AUA framework for classroom contexts, making it more accessible to educators and researchers engaged in small-scale or program-specific assessments. It seeks to ensure that interpretations and uses of test scores are justified, appropriate, and supported by evidence. Recently, there has been a shift from traditional validation methods to an argument-based approach based on Toulmin's (2003) methodology. The AUA framework closely overlaps with Kane's interpretation–use argument (IUA) approach to validation, as both use Toulmin's argument structure to justify score interpretations and uses (Kane, 2013). Thus, drawing on Toulmin's argumentation structure and Messick's (1989) unifying model, the AUA framework breaks down validation into claims, warrants, backings, and potential rebuttals, providing transparency and coherence in test justification. Thus, the AUA focuses on collecting different types of evidence to justify the use of a test for a specific purpose. This includes evidence supporting the consistency of assessment records (Claim 4), which refers to how well a test produces stable and reliable results when measuring the same construct across different items, tasks, or occasions. It also involves gathering evidence to support the meaningfulness and generalizability of interpretations (elements of Claim 3). Meaningfulness concerns whether score interpretations accurately reflect the ability being assessed and provide stakeholders with useful, relevant information. Generalizability refers to the extent to which the test tasks correspond to real-world Target Language Use (TLU) tasks in their characteristics, ensuring that interpretations can be applied beyond the testing context (Bachman & Damböck, 2018; Bachman & Palmer, 2010).

Concerns about the SLPT reflect broader issues in unvalidated assessments: unclear constructs and inconsistent interpretations can lead to misplacements and undermine stakeholder trust. In line with current demands, where modern second language education prioritises communicative competence, transparency, and accountability, validating the SLPT through the AUA framework will provide insights into aspects needing improvement. Consequently, it will ensure improved alignment with instructional goals, reliable scoring, precise placement decisions, and increased stakeholder confidence in the assessment process.

The AUA proposes a unified validation framework encompassing four interrelated claims—Assessment Records, Interpretations, Decisions, and Consequences—and all four claims need to be considered in an overall evaluation. However, this study only focuses on aspects of Claim 3 and Claim 4, particularly the qualities of meaningfulness, generalizability, and consistency. These qualities were chosen based on the concerns raised by SLPT stakeholders and the availability of space and supporting evidence. In other words, the present study is a partial validation argument focused on key aspects of score-based interpretation in the SLPT. That said, it is worth noting that because the AUA claims are directly or indirectly related to one another, it can be considered, to some extent, that the study is addressing all claims in the AUA framework.

The AUA was applied to this study because it offers a clear, structured validation approach that fits a novice researcher's needs and experience level. Secondly, the AUA framework emphasises the importance of complementing quantitative test scores with qualitative evidence, such as coordinator reports, observed test-taker behaviour and content analysis, to support meaningful and responsible interpretation of scores.

Additionally, Bachman and colleagues' (Bachman & Damböck, 2018; Bachman & Palmer, 2010, 1996) Task Characteristics Framework (FTC), developed as part of their broader validation theory, offers a systematic way to analyse the extent to which test tasks reflect tasks in the instructional contexts. FTC is divided into five categories: characteristics of the setting (location, materials and participants), the rubric (task instructions, structure, time allotment, and scoring method), the input the expected response (type of input/response, language structure and the channel), and the relationship between the input and the response (interactiveness, scope and directness). In the present study, this framework was applied to examine the generalizability of SLPT tasks to subsequent classroom activities, making the AUA a suitable and comprehensive framework for evaluating both the interpretations and contextual use of the SLPT.

As part of the AUA argument process, SLPT scores, course curriculum procedures, and testing procedures were examined to evaluate the meaningfulness, generalizability, and consistency of SLPT scores to its intended purpose, providing evidence to accept or reject the claims made for the test. Each claim is supported by evidence (backing), including construct analysis based on Purpura (2004) and the components of communicative language competence (explained in sections 2.3 and 2.4); task characteristics comparisons; internal consistency measures; conditional standard error of measurement (CSEM); and coordinator feedback. Potential limitations and rebuttals, including construct under-representation, misalignment of goals, and low precision, are also acknowledged, making the AUA a comprehensive and transparent guiding framework. The claims, warrants, backing, and rebuttals developed for the SLPT were informed by the examples in Bachman & Damböck (2018) and presented in Table 1.

### *Placement Testing in Foreign Language Programs*

Placement tests have three main functions: placing students into suitable courses based on their language ability (Brown, 1989), assessing students' mastery of specific course contents (Green, 2012; Green & Weir, 2004), or as an exit test (Sood, 2017). A placement test may include various components assessing knowledge of lexis, morphological rules, speaking, pronunciation, writing, listening, and reading comprehension (Long et al., 2018; Pardo-Ballester, 2010). The assessment method may consist of objectively scored items (i.e. selected response), open-ended performance (e.g. speaking, writing), or a combination. Selected response items (e.g. multiple-choice, fill in the gap with options provided, true or false) require more time to construct but less time to score (Lee, 2000; Ruit & Carr, 2011). From practical considerations, the selected response formats are commonly used in placement tests due to administration and scoring efficiency (Downing, 2002; Green, 2018). However, a drawback is that it may promote "cueing," where learners guess the correct response from the options provided rather than deeply understand the concepts (Ruit & Carr, 2011). Therefore, placement tests must be designed and conducted responsibly to ensure their effectiveness in assessing students' language skills and placing them at the appropriate level. Language programs typically rely on either commercial placement instruments or develop their own in-house placement tests. Developing bespoke placement tests offers several advantages, such as better alignment with course objectives and cost-effectiveness. However, it may require time and expertise to develop and implement effectively (Jamieson et al., 2013; Lam, 2010; Mozgalina & Ryshina-Pankova, 2015).

**Table 1.** *Claims, Warrants, Rebuttals and Backing for the SLPT*

| Measures<br>(Assessment<br>Qualities) | AUA Approach                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|---------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Meaningfulness                        | <p>Claim 1: The SLPT construct meaningfully reflects the communicative competencies taught in the Spanish courses.</p> <p>Warrant 1.1: The SLPT assesses skills, knowledge, and abilities that align with the communicative components of the Spanish courses.</p> <p>Rebuttals 1.1:<br/>Construct under-representation.<br/>Misalignment with course objectives and expectations.</p> <p>Backing 1.1:<br/>Categorization of SLPT items by grammatical knowledge areas.<br/>Categorization of course content into CEFR communicative competencies (e.g., interaction, production, reception skills).<br/>Comparative analysis of the SLPT construct vs. course objectives and outcomes.<br/>Coordinator reports of SLPT and course construct alignment.</p>                                                                                                                                                                  |
| Generalizability                      | <p>Claim 2: SLPT tasks are generalizable to the kinds of tasks students will encounter in the instructional context.</p> <p>Warrant 2.1: The characteristics of SLPT tasks reflect those of tasks in the TLU domain (instructional context).</p> <p>Rebuttal 2.1: SLPT tasks differ in key aspects such as interactivity, language mode, and communicative intent, limiting generalizability to the TLU domain.</p> <p>Backing 2.1: Comparison of SLPT and five representative instructional tasks using the Framework of Task Characteristics (FTC), covering task setting, input/response, interactivity, and language use.</p>                                                                                                                                                                                                                                                                                            |
| Consistency                           | <p>Claim 3: SLPT assessment records are consistent and provide reliable scores for placement decisions.</p> <p>Warrant 3.1:<br/>The SLPT demonstrates acceptable internal consistency.<br/>CSEM values at the three placement cut-offs are acceptably low, minimizing classification errors.<br/>Test administration, scoring and formats support consistent assessment delivery.</p> <p>Rebuttals 3.1:<br/>Some items may disproportionately influence overall reliability.<br/>Student reassignments indicate measurement error.<br/>Device-related disparities (phone vs. laptop) might affect user interaction and response accuracy.<br/>Cases of test access or completion issues.</p> <p>Backing 3.1<br/>Cronbach's alpha value for the SLPT scores.<br/>CSEM values at beginner, intermediate, and advanced cut-offs.<br/>Documentation of testing procedures and coordinator reports..<br/>Item-level analysis.</p> |

### *Grammatical Knowledge Categorization*

The categorization of the SLPT items was based on the definition of grammatical knowledge (Purpura, 2004). Purpura (2004) proposed a model of grammatical knowledge that consists of two closely related components: grammatical form and semantic meaning. Grammatical form refers to the structure of the linguistic systems, whereas semantic meaning refers to the literal meanings underlying the grammatical forms. Form and meaning are analysed at both the sentence and discourse levels. The sentential level encompasses phonological/graphological, lexical, and morphosyntactic forms and meanings. In contrast, the discourse level includes cohesive devices, information management strategies, and interactional forms and meanings. The subcomponents measured by the SLPT are described in the following paragraphs.

Knowledge of lexical forms (LFORM) “enables us to understand and produce those features of words that encode grammar rather than those that reveal meaning” (Purpura, 2004, p. 93). This includes morphological irregularity, parts of speech (e.g. rude - Adjective, rudeness- Noun), compounding (e.g. makeup, water tank), and co-occurrence restrictions with prepositions and formulaic expressions. Consider the following example of LFORM:

“Juan siempre piensa \_\_\_\_\_ sus padres (John always thinks \_\_\_\_\_ her parents)”

- a. en (about) \*
- b. con (with)
- c. a (to)

In the example above, the word *piensa* (*thinks*) is followed by the preposition *en* (which contextually translates into English as *about*). This represents a co-occurrence restriction with prepositions and is considered the grammatical dimension of lexis. Lexical meaning (LMEAN) is closely related to lexical forms. Whereas LFORM deals with the grammatical structure of a word, LMEAN focuses on the literal meaning of a word (i.e. vocabulary in context). An example of LMEAN follows:

“El opuesto de barato es \_\_\_\_\_. (The opposite of cheap is \_\_\_\_\_.)”

- a. hermoso (beautiful)
- b. caro (expensive)\*
- c. gratis (free)

All four answer options are adjectives. So, this item does not measure the form of the word; rather, it measures students’ understanding of the meaning of the word *caro* (*expensive*). In this context, it means ‘costing a lot of money’ and is the most appropriate option.

Another component tested is the morphological form (MFORM). MFORM deals with the morphological and syntactic forms of the language. Some of its features are articles, prepositions, pronouns, affixes, simple, compound, complex sentences, mood,

and voice. On the other hand, morphosyntactic meanings (MMEAN) are derived from morphosyntactic forms, such as the meaning of time and aspects that parallel inflectional forms. See an example of MFORM below:

“El año pasado, \_\_\_\_\_ mucho (Last year, it  
\_\_\_\_\_ a lot)”

- a. llovió (rained) \*
- b. llueve (rains)
- c. llover (to rain)
- d. lloviendo (raining)

The preterit form *llovió (rained)* is the most appropriate choice in the example above. The same word is presented in alternative forms to measure students’ ability to recognise and understand the appropriate MFORM.

The last subcomponent is the cohesive form (CFORM) and meaning (CMEAN). Knowledge of CFORM “enables us to use the phonological, lexical and morphosyntactic features of the language to interpret and express cohesion on both sentence and discourse levels” (Purpura, 2004, p. 95). CFORM is closely linked to cohesive meaning (CMEAN). Using cohesive devices, CMEAN creates connections between forms and their referential meanings within the linguistic environment. They include cohesive devices such as ellipsis (e.g. When? [is Marta arriving?]), pronoun referents, comparative referents and logical connectors (e.g. though, however). Consider the example below:

“Lisa condujo a la escuela en vez de caminar \_\_\_\_\_ estaba demasiado cansada para hacerlo a pie. (Lisa drove to school instead \_\_\_\_\_ she was too tired to walk)”

- a. a pesar de (despite)
- b. porque (because) \*
- c. sin embargo (however)

Option b, *because*, is the most appropriate choice in the context, as it connects the reason (*he was too tired*) to the action (*drove to school*). Option a, *however*, is not suitable for the context since it introduces a contradiction. Option c, *despite* being a preposition, does not provide an accurate CMEAN in this context. This item assesses test-takers’ understanding of the use of cohesive meaning at the suprasentential level.

### *Communicative Language Competence (CLC)*

The description of the objectives and content of the instructional domain was based on the definition of communicative language competence as outlined in the CEFR. CEFR is recognised worldwide for describing language ability. The SLPT places students into Spanish language courses that follow CEFR guidelines and its definition of communicative language competence. Under CEFR, communicative language competence comprises three components: linguistic, sociolinguistic and pragmatic competencies. Linguistic competence encompasses grammatical (e.g. lexical,

phonological, semantic, cohesive) and language features as a system (Council of Europe, 2002, 2005). Sociolinguistic competence refers to the sociocultural dimensions of language use. It includes the skills required to have sensitivity to social conventions, such as rules of politeness and register differences, as well as dialects and accents. Pragmatic competence is concerned with the functional use of language to interact. It is also concerned with the mastery of discourse, appropriate use of cohesion, coherence, and the ability to identify text types and forms. CLC emphasises four primary language skills: receptive (listening and reading), oral and written production, interaction, and mediation.

### **The Present Study**

This article examines how well the SLPT supports its intended placement purposes within the current program context. Thus, the objective of the study is to evaluate the SLPT across three core qualities: meaningfulness, generalizability, and consistency, using the Assessment Use Argument (AUA) validation approach. Meanwhile, the three main guiding research questions of the study are:

1. To what extent does the SLPT meaningfully reflect the communicative competencies emphasised in the Spanish curriculum?
2. To what extent are the tasks in the SLPT representative of the tasks in the instructional context?
3. How consistent and reliable are SLPT scores for supporting accurate and fair placement decisions?

## **Method**

### **Participants and Procedures**

The study recruited 105 students enrolled in Spanish language courses at a large Canadian research-intensive university. While participation was voluntary, the sample reflected the ability range typically observed in routine SLPT administration, with representation across beginner, intermediate, and advanced score levels. Routine administrative data were not accessible for research during the study period, so recruited participants were used to support program evaluation and validation efforts. No demographic data were collected, but the Spanish program usually serves a diverse student population where learners enrol to fulfil degree or language requirements, complete a minor, take an elective, or strengthen personal and family linguistic connections. Typical course enrolments include students from a wide range of faculties such as Arts, Social Sciences, Business, and Engineering, with varying prior exposure to Spanish. The Spanish program coordinator informed the study by providing contextual information about the SLPT's history, testing procedures, and usage concerns, as well as the need for a structured evaluation of the test. This information was obtained through an informal phone conversation and follow-up email correspondence and was used solely to contextualise the findings rather than as a formal source of empirical data.

After obtaining ethics approval, the SLPT data were collected in Winter 2021. The placement test was hosted on Google Forms, and the course instructors shared the test links with students through email. The data collection procedure mimicked the process of administering the SLPT test— routine administration of the SLPT for placement purposes is online, not timed, and not proctored. However, students were advised to complete the

test within two hours, not to consult any materials and to answer items to the best of their knowledge. The resulting quantitative data were analysed through SPSS, whereas the qualitative data were analysed manually.

## **Quantitative Data Sources and Analysis Process**

### *Spanish Language Placement Test (SLPT)*

The SLPT is an in-house web-based Spanish placement test used to sort students into one of three course levels (Beginner, Intermediate and Advanced levels I & II) at a Canadian university. The test is divided into two sections, consisting of 100 items in total. The items measure knowledge of lexical, morphosyntactic, and cohesive forms and meanings. As the SLPT is online, students take the test at a time and place of their convenience. Participants completed the test during a period of two weeks, and reminders were sent at appropriate times. Responses were scored dichotomously—a correct answer was awarded one point, and an incorrect answer received no points. A total score is the sum of the correct responses to all test items, and the maximum possible score is 100 points. In the program, students receive their score and placement decision within 1 to 2 days and then register for the course they have been placed into. The test has a multiple-choice format with four response options.

### *Cronbach's Alpha*

Internal consistency evaluates the degree to which items on the same test are correlated and measure the same construct (Livingston, 2018). Cronbach's Alpha is the most widely used statistical measure of the internal consistency of a test (Cronbach, 1951). It was used to calculate the internal consistency of the SLPT for the test scores of the 105 participants. To analyse in SPSS, responses to the SLPT were coded as 1 (correct) and 0 (incorrect). Also, the alpha coefficient of each SLPT test item was examined to determine whether there were items significantly influencing the overall alpha value obtained for the test. To do this, Cronbach's alpha if an item was dropped was inspected for each SLPT item.

### *The Conditional Standard Error of Measurement (CSEM)*

Cronbach's alpha offers an overall estimate of test consistency, but it does not reflect score precision across ability levels. CSEM addresses this by quantifying error at specific points on the score scale (Emons, 2023). For the SLPT, CSEM was calculated at the cut-off scores—Beginner (0–55), Intermediate (56–79), and Advanced (80–100)—to assess the precision of placement decisions. There are several methods for estimating CSEM, including Classical Test Theory (CTT), Generalizability Theory (G-Theory), the Rasch model, and Item Response Theory (IRT). IRT was selected—specifically the 2-Parameter Logistic (2PL) model—because it offers ability-specific CSEM estimates that are particularly useful for tests like the SLPT, which consists of four-option multiple-choice items scored dichotomously. Unlike CTT, which assumes uniform error across scores, IRT accounts for variation in measurement precision at different ability levels. G-Theory is better suited for performance-based assessments with multiple facets (e.g. raters, tasks); hence, it is not applicable.

Although 2PL models generally benefit from larger calibration samples (Guo et al., 2024; Schroeders & Gnambs, 2025), smaller samples can still yield interpretable results

when items are dichotomous, and the sample spans the full ability range, both of which characterize the present dataset. The sample of 105 students reflected the typical distribution of SLPT performance, allowing the estimation of item parameters and the examination of measurement precision across placement levels. The implications of the modest sample size for parameter stability and precision, particularly at the advanced cut-score, are addressed in the limitations section. The analysis was conducted using Python to estimate item difficulty and discrimination parameters, and then the Test Information Function (TIF) and CSEM were derived from cut-off scores.

## **Qualitative Data Sources and Analysis Process**

### *Observation and Documentation of Testing Procedures*

Policy documents provided information about the procedures for administering and scoring the SLPT. The researcher also took the test and obtained a score and a placement decision. They helped to gather evidence about the processes for administering and scoring the SLPT to assess the extent of consistency and accuracy of the testing procedures.

### *SLPT Item Categorisation, Syllabi and Course Content*

To provide evidence for meaningfulness, SLPT and course content and objectives were compared to determine the extent of representativeness. First, SLPT items were categorised based on what component of grammatical knowledge each was tapping into (Purpura, 2004). Then, the course content and objectives were described and summarised based on the CLC components. Descriptions were provided for the CEFR levels: Beginner (Level A), Intermediate (Level B), and Advanced (Level C) (See Appendix B). Finally, the grammatical knowledge assessed by the SLPT was compared with the courses' Spanish knowledge, skills, and objectives. The principal researcher did the categorisation and the research supervisor reviewed it.

### *Course Textbooks*

The characteristics of the SLPT and five TLU tasks sampled from course textbooks were described and compared to provide evidence for generalizability. Reviewing course syllabi and textbooks determined the five most frequent TLU tasks. We systematically examined a sample of syllabi from the three course levels and the corresponding textbooks to identify the most common tasks. The tasks that were consistently featured across multiple syllabi and textbooks were selected. They included: (i) a multiple-choice grammar exercise, (ii) a fill-in-the-blank lexical task, (iii) an audio task, (iv) a reading comprehension task, and (v) an essay writing task.

## **Results**

The results are organised according to the three qualities and corresponding warrants investigated, starting with meaningfulness, generalizability, and consistency.

### **Claim 1: Meaningfulness**

#### *Warrant 1.1: Alignment with Communicative Competencies and Course Objectives*

As stated, SLPT items were examined and categorised based on the aspects of the Spanish language knowledge that each assessed. Based on this analysis, the construct of

the SLPT was defined as “assessment of lexical, morphological and cohesive knowledge”. The results of the categorisation of SLPT items indicated that 34% of the items measured lexical knowledge, the majority (i.e. 61%) primarily targeted morphosyntactic knowledge, while cohesive knowledge accounted for the remaining 5% (See Appendix A for the complete categorisation).

**Table 2.** *Construct Measurement of SLPT Items*

| Component of Grammatical Knowledge         | SLPT |
|--------------------------------------------|------|
| Lexical knowledge (form and meaning)       | 34%  |
| Morphological knowledge (form and meaning) | 61%  |
| Cohesive knowledge (form and meaning)      | 5%   |

A content analysis of the Spanish curriculum and syllabi according to the components of communicative language competence (see Appendix B) showed that the Spanish courses cover three competencies: linguistic (including grammatical knowledge), pragmatic, and sociolinguistic competencies. They also emphasise skills such as interaction, problem-solving, and production and reception skills development. The learning objectives and outcomes prioritise authentic language use.

The construct of grammatical knowledge assessed by the SLPT was compared with the knowledge, abilities and competences assessed and taught in the Spanish courses. The analysis revealed that the items related to morphosyntactic, lexical, and cohesive knowledge were representative of three subcomponents of grammatical knowledge on which students received instruction and assessment in the classroom. Concerning morphosyntactic knowledge, the structures examined in the SLPT reflected some of those taught in the Spanish classes, e.g. structures such as conjugation for the appropriate tense, mood, aspect, and contrast of SER and ESTAR, which the SLPT assesses receive instruction in the Beginner and Intermediate Spanish I and II classes. In the case of lexical knowledge, the type of words and the method of assessment of the SLPT were similar to how lexical knowledge is taught and evaluated in the Spanish classes, e.g. vocabulary related to adjectives, descriptions of places and things are similar to those taught in classes. Additionally, like in the classroom, the SLPT assesses the use of vocabulary in the appropriate context.

*Rebuttals: Construct under-representation and misalignment with course objectives*

The analysis revealed two potential rebuttals against the quality of meaningfulness of interpretations. The SLPT construct does not fully represent the broader Spanish language knowledge taught and evaluated in the Spanish courses. Spanish language knowledge is based on the construct of communicative competence and thus goes beyond grammatical knowledge. In other words, grammatical knowledge is a small part of what the courses assess; thus, the SLPT overlooks pragmatic and sociolinguistic competence, which are integral to the courses students are assigned. In addition, the classes teach and evaluate all six sub-components of grammatical knowledge. However, the SLPT only

measures three subcomponents of grammatical knowledge. It does not assess the subcomponents of phonological, information management, and interactional forms/meaning of the grammatical knowledge model. The narrow construct indicates that the resulting scores may not meaningfully indicate a student's readiness for the course level, thereby limiting the justification of its intended use. Furthermore, the coordinator's concerns about misalignment with the course objectives are a rebuttal against the SLPT's meaningfulness.

### **Claim 2: Generalizability**

#### *Warrant 2.1: SLPT Tasks Reflect TLU Domain Tasks*

The comparison of task characteristics between the SLPT and the TLU domain highlights key areas where the SLPT tasks reflect instructional task types. Regarding the setting, both SLPT and some classroom tasks are delivered online. The SLPT has no fixed time limits, whereas the instructional tasks are generally time-constrained. The assessment rubric in both contexts helps to define the criteria for correctness. SLPT tasks primarily rely on one criterion of right or wrong, while TLU tasks can involve varied criteria depending on the task type. The format of the SLPT is based on selected responses, with four options for each item, mirroring the straightforward nature of some classroom tasks.

#### *Rebuttal 3.1: Limited Generalizability of SLPT Tasks*

Despite the similarities, several notable differences between the SLPT and TLU tasks limit the generalizability of the SLPT tasks to actual language use in Spanish courses. First, the SLPT tasks primarily focus on isolated responses with a simple format, such as single words or selected responses. These tasks do not replicate the complexity or breadth of interaction in TLU tasks, often involving more varied, reciprocal interactions in authentic settings. The TLU domain typically incorporates audio, reading, and writing exercises involving interactive and dynamic use of language across a range of contexts. This requires learners to demonstrate both formal and informal registers and respond to more complex inputs that include cultural and non-cultural references. Sociocultural awareness is also a core component of the Spanish courses, where students are expected to demonstrate deeper understanding through tasks that require engagement with cultural references.

In contrast, SLPT tasks involve predominantly one-way communication and are confined to basic grammar tasks that lack the depth of language needed for advanced communicative or integrated tasks. Moreover, the SLPT's non-recursive, direct relationship between input and expected response is limited in its ability to capture the reciprocal and adaptive interactions in classroom tasks. As such, the SLPT may not effectively measure the higher-level, more interactive skills advanced learners possess, leading to concerns over its ability to generalise well to more complex language tasks in the TLU domain. A summary of the comparison of TLU and SLPT tasks is provided in Table 3 (See Appendices C and D for the detailed analysis).

**Table 3.** *Summary of Comparison of TLU and SLPT Task Characteristics*

| Characteristic                                                                                   | Instructional Tasks                                                                                                                                                                 | SLPT Tasks                                                                        |
|--------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|
| Setting: location, time                                                                          | Online and in class.                                                                                                                                                                | Online.                                                                           |
|                                                                                                  | Time of the test varies.                                                                                                                                                            | Time of the test varies.                                                          |
|                                                                                                  | Duration – Tasks are mostly timed or constrained by class time.                                                                                                                     | Test is not timed                                                                 |
| Assessment Rubric: duration, instructions, correction and recording.                             | Written and oral instruction. Teacher discussion to activate relevant schema for the task.                                                                                          | Written instructions only                                                         |
|                                                                                                  | The type of graded task dictates the criteria for correctness. One criterion for grammar exercise and vocabulary tasks, criterion varies for audio, writing and reading tasks.      | One criterion for correctness (right/wrong)                                       |
|                                                                                                  | Both visual and oral                                                                                                                                                                | Visual only                                                                       |
| Input: format, type, language, channel                                                           | Language is grammatically, textually, functionally, and sociologically more complex.                                                                                                | Language is grammatically and syntactically less complex.                         |
|                                                                                                  | Length and type of response varies depending on the type of task.                                                                                                                   | The length is short (single word), and the type of response is selected response. |
|                                                                                                  | Language varies from simple to complex grammatical, functional, and sociolinguistic elements. Includes cultural and non-cultural referents and register is both formal and informal | Language is grammatically less complex. Use of non-cultural referents.            |
| Relationship between input and expected response: type of interactivensess, scope and directness | Reciprocal and non-reciprocal                                                                                                                                                       | Non-reciprocal                                                                    |
|                                                                                                  | Broad and narrow scope                                                                                                                                                              | Narrow                                                                            |
|                                                                                                  | Consists of both direct and indirect.                                                                                                                                               | Direct                                                                            |

### **Claim 3: Consistency**

#### *Warrant and Rebuttals 3.1: Internal Consistency of SLPT*

The internal consistency of the SLPT was measured using Cronbach's alpha ( $K = 100$ ,  $N = 105$ ), yielding a value of  $\alpha = .94$ . Reliability guidelines suggest that higher coefficients, often at or above .90, are desirable when decisions carry meaningful individual consequences (Nunnally & Bernstein, 1994), and values in the .80–.90 range are commonly viewed as adequate for medium-stakes individual-level classification (DeVellis, 2017). The obtained alpha suggests that the test items exhibit high internal consistency and are strongly correlated with each other, indicating that they are measuring a similar underlying construct.

A potential rebuttal to the internal consistency warrant is that problematic items on the test can adversely influence the alpha coefficient of the test. To determine whether specific test items significantly affected the internal consistency of the SLPT, Cronbach's Alpha-if-item-dropped was inspected. The analysis revealed that no item influenced the overall internal consistency of the test by more than 0.001 if deleted.

#### *Warrant and Rebuttal 3.1: Conditional Standard Error of Measurement*

In Table 4, the theta values ( $\theta = 0, 1, 2$ ) were selected as reference points corresponding to the program's Beginner, Intermediate, and Advanced cut-score ranges, rather than representing population norms. In this context,  $\theta = 0$  does not indicate that the average student is a Beginner but serves as an anchor point for evaluating measurement precision at that placement level. The Test Information Function (TIF) describes how much measurement precision SLPT provides at the three placement levels, while CSEM values reflect the expected measurement error at the three cut-off score ranges. The results revealed that, at the Beginner level ( $\theta = 0$ ), the TIF is 113.92, and the CSEM is 0.094, indicating very high measurement precision and minimal error in estimating beginner-level proficiency. At the Intermediate level ( $\theta = 1$ ), the TIF remains strong at 101.03, with a slightly higher but still low CSEM of 0.100, suggesting the SLPT is also highly reliable for intermediate learners. However, at the Advanced level ( $\theta = 2$ ), the TIF drops significantly to 13.75, and the CSEM increases to 0.370, revealing a notable decline in measurement precision. This suggests that while the SLPT is well-calibrated to distinguish among beginner and intermediate students, it may be less effective at accurately assessing advanced learners, which has implications for placement decisions at the higher end of the language ability scale. Combined with the limited item difficulty range (average  $b = 0.37$ ), this suggests the test is not sufficiently challenging for advanced learners, reducing both score precision and the meaningfulness of interpretations.

Moreover, coordinator reports indicate that about 3-12 out of 200 students are reassigned each term, suggesting possible placement errors. However, since reassignments are not formally tracked and are made by instructors without training in placement testing, the true extent of misclassification is uncertain. This lack of oversight undermines the consistency and accuracy of SLPT placement decisions.

**Table 4.** *Summary of TIF and CSEM Values at the Designated Cut-off Scores*

| Theta (Ability Level) | Test Information (TIF) | CSEM  |
|-----------------------|------------------------|-------|
| 0 (0-55)              | 113.920                | 0.094 |
| 1 (56-79)             | 101.032                | 0.100 |
| 2 (80-100)            | 13.754                 | 0.370 |

### *Warrant and Rebuttal 3.1: Consistency of Testing Procedures*

Observation of the test administration and scoring procedures of the SLPT revealed that the same procedures are adhered to across different occasions and for all test takers. The SLPT is a web-based test, and students can access it via a computer or smartphone using their student numbers or email. Although the testing medium (e.g. phone vs. computer) could affect user experience, such as navigation or completion time, the researcher completed the test on both devices and observed no differences in test scores. Furthermore, previous research (Huff, 2015) found differences in usability ( $p < .001$ ,  $d = 1.059$ ) and completion time ( $p < .001$ ,  $d = -1.638$ ) between smartphones and computers on a personality assessment. However, no significant differences were found in the personality scores ( $p = 0.95$ ,  $d = 1.708$ ), indicating that the testing medium did not affect the test results. Therefore, it is unlikely that the medium of access in the SLPT would impact the consistency of the assessment records produced. Also, the test is automatically scored, suggesting that the same scoring procedures are always adhered to for all students and on all occasions. The test was not timed, but the completion time for participants was an average of 45 minutes. No test preparation information is provided to test takers. Moreover, students are informed that no preparation is required to take the test since it assesses their current Spanish level to place them into a course that matches their abilities. The coordinator reported that no technical issues related to students accessing the test online have been reported in the past few years.

## Discussion

### Research Question 1

The first research question examined the extent to which the SLPT construct provides meaningful score-based interpretations that reflect the competencies required in Spanish courses.

The results underscore a partial alignment between the SLPT construct and the instructional content of the Spanish courses. While the test appropriately targets core grammatical components—morphosyntactic, lexical, and cohesive knowledge—explicitly taught and assessed in class, it falls short in capturing the broader communicative competencies emphasised in the curriculum. This construct under-representation weakens the meaningfulness of score interpretations, as the SLPT does not reflect key aspects of communicative language ability such as interaction, mediation, and real-world application, now central to the CEFR-informed instructional objectives.

The coordinator's concerns about the test's outdated construct and limited scope are substantiated by this analysis. The evolution of the Spanish program to include more

functional and discourse-based skills implies that students are expected to do more than manipulate grammar in isolated contexts. If the SLPT continues to assess only a narrow slice of these competencies, its scores risk misrepresenting students' preparedness for course demands, particularly in spoken and interactive skills. Thus, while the SLPT may still offer some value for assessing grammatical knowledge, its interpretive power as a placement tool aligned with current course expectations is significantly diminished.

### **Research Question 2**

Research question two investigated how well SLPT tasks generalise to the tasks in the instructional domain. Although the SLPT tasks show alignment with specific instructional tasks in their focus on morphosyntactic, lexical, and cohesive knowledge, they fall short in reflecting the broader range of classroom activities. The test lacks task variety and omits essential components such as pronunciation, reading, listening comprehension, writing, and sociocultural content, elements central to Spanish instruction. As a result, while the SLPT may provide a basic snapshot of grammatical knowledge, its narrow task design undermines its generalizability to the more authentic, reciprocal, and multimodal tasks encountered in the Spanish courses. This discrepancy weakens the claim that SLPT performance can meaningfully inform placement decisions across all course levels.

### **Research Question 3**

The analysis for Research Question Three examined the internal consistency and reliability of the SLPT assessment records. The test demonstrated strong internal consistency, and no item improved the reliability coefficient by more than 0.001 when removed. These findings indicate that the SLPT reliably differentiates among students. The reliability estimate is also consistent with expectations for medium-stakes placement contexts, where higher levels of internal consistency are desirable for individual-level classification.

However, the CSEM results raised concerns about measurement precision for advanced placements. The drop in information at the highest cut-score suggests that placement decisions for advanced learners may be more error-prone, which weakens the accuracy of score interpretations at the upper end of the scale. Although the SLPT provides placement recommendations across all three levels, relatively few students qualify for placement into the Advanced courses in this program. Typical enrollment patterns in university language programs indicate that most incoming students enter beginner or intermediate levels. Given this distribution, the reduced measurement precision observed at the Advanced cut-score affects a smaller proportion of all placement decisions, though it remains important for the students who fall near this threshold.

The consistency of assessment records is further influenced by program-level practices. Although test administration procedures were consistent across sessions and examinees, the coordinator reports of 3 to 12 student reassignments each term (made without formal tracking or placement-testing training) point to additional concerns regarding the stability of placement decisions. The program does not record whether reassigned students were moved to higher or lower levels, hence, the direction and pattern of misclassifications cannot be determined. These reassignments cannot be attributed

solely to SLPT score imprecision, but they highlight the need for clearer guidelines and closer alignment between test score interpretations and instructional expectations.

Participation in the study was voluntary, which may introduce some variation in students' test-taking effort, in turn, consistency estimates (see the Limitations section). However, the reliability patterns observed in this study suggest that effort levels were generally adequate for the purposes of the analysis.

### **Limitations of the Study**

A limitation of this study is that it broadly defined grammatical knowledge as including lexical, morphosyntactic, and cohesive knowledge. However, future studies could benefit from examining the specific types of knowledge within each subcomponent, such as referential forms (e.g., pronouns) versus logical connectors (e.g., *therefore*), to better understand how these subcomponents manifest in test items and their impact on test performance.

Secondly, the study did not incorporate feedback from students, test administrators, and course instructors. Their perspectives are crucial in evaluating the impact of the SLPT on placement decisions. Finally, the study did not include a longitudinal comparison of student scores on SLPT with their performance in courses, which limits claims about the test's ability to predict academic success in Spanish courses.

Participants were recruited for the study, which may introduce self-selection effects. Voluntary participants may differ from the wider placement population, either because they are more motivated or because they invest less effort when the activity has no academic consequences (Wise & DeMars, 2005). Variation in test-taking effort can influence internal consistency and IRT estimates. In this study, however, the high internal consistency, stable item–total correlations, and strong test information at the lower and mid cut-scores suggest that motivation did not substantially affect the reliability or measurement precision estimates. Nonetheless, the voluntary nature of participation should be taken into account when interpreting the findings.

Finally, although IRT was used to estimate CSEM, the sample size of 105 is modest for a 2PL model. Larger samples are typically recommended for optimal parameter precision, particularly for discrimination estimates (Guo et al., 2024; Schroeders & Gnams, 2025). The present sample did span the full ability range typically observed in SLPT administration, which supported model estimation, but the small size likely contributed to reduced precision at the advanced cut-score. These factors should be considered when interpreting the IRT-based findings.

### **Conclusions, Implications, and Future Directions**

This study found that while the SLPT demonstrates strong internal consistency and alignment with some classroom content, its limited construct coverage and reduced measurement precision at higher proficiency levels undermine the meaningfulness, generalizability, and consistency of score-based placement decisions. This misalignment poses risks for accurate placement, especially at the intermediate and advanced levels, where learners often demonstrate broader language competencies that the SLPT does not measure. Neglecting these concerns may result in continued placement inaccuracies, which could hinder student engagement and achievement while putting pressure on instructors to manage misplacements informally.

The analysis identified key areas for improvement that will be addressed in the next phase of test revision, including expanding the construct coverage, increasing item difficulty for advanced learners, and integrating more authentic, communicative task types to better align with instructional goals. First, to enhance the meaningfulness and generalizability of test interpretations, the SLPT construct should be expanded beyond grammatical knowledge to include integrated assessment of listening, reading, writing, and speaking skills. While resource limitations may preclude full-scale performance assessments, solutions such as scenario-based items and adaptive testing platforms offer feasible ways to capture a broader range of language ability. Incorporating AI-driven tools for scoring written or oral responses, particularly for flagged cases, could also provide a scalable and valid measurement without overburdening faculty. Furthermore, to address the issues with consistency and meaningfulness, the item pool needs to be revised to better target a broader range of abilities.

Moreover, this study contributes methodologically by applying the Assessment Use Argument (AUA) framework to a Spanish-language placement test. Additionally, the application of Purpura's (2004) model of grammatical knowledge to categorise SLPT items grounded the construct analysis in a theoretically informed categorisation. Thus, it offers a replicable method for evaluating language test content in relation to instructional goals.

Future research should explore the implementation and impact of integrated or adaptive SLPT formats, investigate stakeholder perceptions of revised testing procedures. The use of technology to support task variety and scoring efficiency, as well as further exploration into the training and decision-making processes of instructors, would also enhance the robustness of placement practices.

An important direction for future validation work also involves examining the consequences of SLPT score use through a consequential validity lens. Such research could investigate how initial placement decisions relate to subsequent student outcomes, including course performance, persistence in the language sequence, and self-reported engagement or achievement. Analyses comparing correctly placed and misclassified students would offer insight into whether placement accuracy meaningfully affects instructional alignment and learner progress. Future studies should also document the direction of any reassignments, as information about whether students are moved to higher or lower levels is essential for understanding the nature and impact of misplacements. Additionally, research could explore whether certain groups of students, such as heritage speakers or students with uneven language backgrounds, experience differential placement outcomes. A more systematic examination of these consequences would provide stronger evidence for the impact of the SLPT on learners and programs and would help evaluate whether revised placement practices promote equitable and educationally useful placement decisions.

## Appendices

[Appendices A-D: Grammatical Knowledge Categorisation, Objectives and Content, Characteristics of Instructional Tasks, Analysis of SLPT Task Characteristics](#)

## References

- Bachman, L. F., & Damböck, B. (2018). *Language assessment for classroom teachers*. Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
- Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34.
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Brown, J. D. (1989). Improving ESL placement tests using two perspectives. *TESOL Quarterly*, 23(1), 65–83.
- Council of Europe (2002). *Marco común europeo de referencia para las lenguas: aprendizaje, enseñanza, evaluación*. Instituto Cervantes, Ministerio de Educación, Cultura y Deporte y Editorial Anaya.
- Council of Europe. (2005). *Common European framework of reference for languages: Learning, teaching, assessment*. Press Syndicate of the University of Cambridge.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334.
- DeVellis, R. F. (2017). *Scale development: Theory and applications* (4<sup>th</sup> ed.). Sage.
- Downing, S. M. (2002). Assessment of knowledge with written test forms. *Springer Netherlands*, 647–672. [https://doi.org/10.1007/978-94-010-0462-6\\_25](https://doi.org/10.1007/978-94-010-0462-6_25)
- Emons, W.H.M. (2023). Methods for Estimating Conditional Standard Errors of Measurement and Some Critical Reflections. In: van der Ark, L.A., Emons, W.H.M., Meijer, R.R. (Eds.), *Essays on Contemporary Psychometrics*. Methodology of Educational Measurement and Assessment. Springer, Cham. [https://doi.org/10.1007/978-3-031-10370-4\\_11](https://doi.org/10.1007/978-3-031-10370-4_11)
- Green, A. (2018). Placement Testing. In J. I. Lontas, M. DelliCarpini, & C. Coombe (Eds.), *The TESOL Encyclopedia of English Language Teaching* (pp. 1–6). John Wiley & Sons, Inc.
- Green, A. (2012). Placement testing. In C. Coombe, B. O’Sullivan, P. Davidson, & S. Stoyloff (Eds.), *The Cambridge guide to language assessment* (pp. 164–170). Cambridge University Press.
- Green, A. B., & Weir, C. J. (2004). Can placement tests inform instructional decisions? *Language Testing*, 21(4), 467–494.
- Guo, H., Johnson, M.S., McCaffrey, D.F. and Gu, L. (2024). Practical considerations in item calibration with small samples under multistage test design: A case study. *ETS Research Report Series*, 2024: 1–21. <https://doi.org/10.1002/ets2.12376>
- Hille, K., & Cho, Y. (2020). Placement testing: One test, two tests, three tests? How many tests are sufficient? *Language Testing*, 37(3), 453–471. <https://doi-org.proxy.bib.uottawa.ca/10.1177/0265532220912412>
- Jamieson, J., , L., & Church, J. (2013). In-house or commercial speaking tests: Evaluating strengths for EAP placement. *Journal of English for Academic Purposes*, 12(4), 288–298.
- Johnson, R. C. (2011). *Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context*. [Unpublished PhD thesis]. Macquarie University.
- Johnson, R. C., Riazi A. M. (2015). Accuplacer Companion in a foreign language context: An argument-based validation of both test score meaning and impact. *Papers in Language Testing and Assessment*, 4(1), 31–58. [http://www.altanz.org/uploads/5/9/0/8/5908292/johnson\\_riazi.pdf](http://www.altanz.org/uploads/5/9/0/8/5908292/johnson_riazi.pdf)
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>

- Lam, Y. (2010). Yes/No tests for foreign language placement at the post-secondary level. *Canadian Journal of Applied Linguistics/Revue canadienne de linguistique appliquee*, 13(2), 54–72.
- Lee, Y. (2000). Designing a Korean language placement test. *The Korean Language in America*, 4, 181–210.
- Livingston, S. A. (2018). *Test reliability—Basic concepts* (Research Memorandum No. RM-18-01). Princeton, NJ: Educational Testing Service.  
<https://www.ets.org/Media/Research/pdf/RM-18-01.pdf>
- Long, A. Y., Shin, S. Y., Geeslin, K., & Willis, E. W. (2018). Does the test work? Evaluating a web-based language placement test. *Language Learning & Technology*, 22(1), 137–156.
- Messick, S. J. (1989). Validity. In Linn, R. L. (Ed.), *Educational Measurement* (pp. 13–103). Macmillan Publishing.
- Mozgalina, A., & Ryshina-Pankova, M. (2015). Meeting the challenges of curriculum construction and change: revision and validity evaluation of a placement test. *The Modern Language Journal*, 99(2), 346–370.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3<sup>rd</sup> ed.). McGraw-Hill.
- Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: test usefulness evaluation. *Language Assessment Quarterly*, 7(2), 137–159.
- Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.
- Ruit, K., & Carr, P. (2011). Comparison of student performance on “selected-response” versus “constructed-response” question formats in a medical neuroscience laboratory practical examination. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, 25(S1).  
[https://doi.org/10.1096/fasebj.25.1\\_supplement.182.6](https://doi.org/10.1096/fasebj.25.1_supplement.182.6)
- Schroeders, U., & Gnambs, T. (2025). Sample-Size Planning in Item-Response Theory: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 8(1), 1–13.  
<https://doi.org/10.1177/25152459251314798>
- Sood, P. (2017). Language proficiency and assessing classroom achievement: a literature review. *Issues in Language Instruction*, 2(3), 9. <https://doi.org/10.17161/ili.v2i0.6933>
- Toulmin, S. E. (2003). *The uses of argument* (updated edition). Cambridge University Press.
- Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment*, 10(1), 1–17.  
[https://doi.org/10.1207/s15326977ea1001\\_1](https://doi.org/10.1207/s15326977ea1001_1)