

Examining the Context and Consequential Validity of a High-Stakes Writing Test in a Colombian Bachelor's in Foreign Languages

Bryamth Cárdenas Marín

Linda Cruz Lenis

Supervisor

Dr María Eugenia Guapacha Chamorro

A dissertation submitted in partial fulfillment of the requirements for the degree of Bachelor's
in Foreign Languages in English and French

Universidad del Valle

2024

Dedication and Acknowledgments

From Linda

First and foremost, I want to express my deepest gratitude to God for blessing me with the opportunity to be part of this prestigious university. His strength has been my anchor, enabling me to overcome the challenges and embrace the experiences that have enriched both my academic and personal growth.

I dedicate this achievement to my incredible father, who has been my constant pillar of support. His belief in me has never hesitated, even during my many detours. He has stood by my side, loyal and strong, especially during times when I have doubted myself.

My heartfelt thanks also go to my partner, whose devoted companionship, encouragement, and willingness to share the burdens along the way have been invaluable. Your support has been a constant source of strength and motivation.

I am profoundly grateful to my thesis director, María Eugenia Guapacha, and my colleague, Bryamth Cárdenas, for giving me the opportunity to be part of this remarkable project. Your guidance and collaboration have been instrumental in this journey.

Finally, I extend my sincere thanks to the bachelor's degree Program in Foreign Languages and its director, Hugo Areiza, for his exceptional support and guidance throughout this demanding yet rewarding journey. Your mentorship has been invaluable, and I am deeply appreciative of your dedication to my success.

Linda Lucía Cruz Lenis

From Bryamth

Reaching the final stage of this great and challenging project has been an arduous yet rewarding experience. Many nights were spent working tirelessly, but as the saying goes, nothing worth having comes easy. The true reward lies in the knowledge and growth I gained through this process, navigating and mastering a field that was initially unfamiliar to me.

I extend my heartfelt gratitude to those who were patient with me during this project. To my family and Vicky, your support during moments of anger and frustration was invaluable.

I am deeply appreciative of the guidance and dedication of our director, Professor María Eugenia Guapacha, whose unwavering advice and support were essential in the completion of this project. Thank you for your commitment and for sharing your expertise so generously. I would also like to thank my colleague Linda who offered her assistance and encouragement throughout this journey. Your knowledge contributed to the success of this project.

Finally, I am grateful to all the instructors and mentors from the BA in Foreign Languages who provided me with the foundational knowledge and skills necessary to undertake this project. Your dedication to teaching and mentorship has had a lasting impact on my academic and professional development.

Bryamth Cárdenas Marín

Table of Contents

Dedication and Acknowledgments	2
Table of Contents	4
List of Figures	7
List of Tables	8
Abstract	9
1 Introduction	15
1.1 Chapter overview	
1.2 Writing assessments' validity and fairness	
1.3 Research questions	
1.3.1 Main research question	18
1.3.2 Specific research questions	18
1.4 Significance and contribution of the study	
1.4.1 Theoretical significance	18
1.4.2 Research significance	18
1.4.3 Practical significance	19
1.5 Thesis outline	
2 Literature Review	20
2.1 Chapter overview	
2.2 Theoretical framework	
2.2.1 Validity and fairness in language assessment	20
2.2.2 Test validation	21
2.2.3 Shaw and Weir's (2007) sociocognitive framework for writing test validation	23
2.3 Empirical research on EFL writing test validation	
2.3.1 International studies on writing assessment validity	30
2.3.2 Studies on writing assessment validity in Colombia	32
2.3.3 Studies on writing assessment at Universidad del Valle	33
2.4 Concluding remarks and gaps to fill	
3 Research Methodology	36
3.1 Chapter overview	
3.2 Research approach and design	
3.3 Research site and participants	
3.3.1 Research site	37
3.3.2 Participants	38
3.4 Data collection methods, procedures and analysis	
3.4.1 The writing proficiency test	39
3.4.2 Students' writing performance data (test scores, time and length)	42
3.4.3 Students' perceptions data from a questionnaire	43

3.5 Data analyses	
3.5.1 Qualitative data analysis	44
3.5.2 Quantitative data analysis	45
3.6 Validity and reliability	
3.7 Ethical considerations	
3.8 Concluding remarks	
4 Results	50
4.1 Chapter overview	
4.2 RQ1. What is the performance of the students in the writing test concerning text length, time, and scores?	
4.2.1 Text length	50
4.2.2 Writing time	51
4.2.3 Writing test scores	52
4.3 RQ2. What are the students' perceptions of the writing test regarding context validity?	
4.3.1 Test-takers' negative perceptions regarding the test task setting	55
4.3.2 Test-takers' negative perceptions regarding the test administration setting	56
4.3.3 Test-takers' negative perceptions regarding linguistic demands	58
4.3.4 Test-takers' negative perceptions and experiences regarding psychological and experiential characteristics	58
4.3.5 Test-takers' positive perceptions of the writing test and themselves	59
4.4 RQ3. What are the students' perceptions of performance before and after the test?	
4.5 RQ4. What is the relationship between students' perceptions and performance on the writing test?	
4.6 Concluding remarks	
5 Discussion	67
5.1 Chapter overview	
5.2 Students' writing performance regarding scores, time and text length	
5.3 Students' perceptions of the test - context and consequential validity	
5.3.1 Students' perceptions of context validity: Task setting	69
5.3.2 Students' perceptions of context validity: Administration setting	71
5.3.3 Students' perceptions of context validity: linguistic demands	72
5.4 Students' perceptions related to test-takers' psychological and experiential characteristics	
5.5 Students' perceptions of performance before and after the test	
5.6 Relationships between perceptions and performance	
5.7 Examining the context and consequential validity evidence of the writing test	
5.7.1 Valid and fair aspects of the test regarding context validity and consequences	76
5.7.2 Aspects of the test that need improvement to enhance its validity and fairness	77
5.7.3 What are the positive and negative consequences of the test on students' performance and the academic program?	79

6	Conclusions	80
6.1	Chapter overview	
6.2	Summary of findings and conclusions	
6.3	Implications	
6.3.1	Implications for the writing Test: validity and fairness	82
6.3.2	Implications for the curriculum of the program	83
6.3.3	Implications for students as test-takers	84
6.4	Limitations and further research	
	References	87
	Appendices	90
	Appendix A	
	Appendix B	
	Appendix C	

List of Figures

Figure 2.1	Shaw and Weir's (2007) sociocognitive framework for conceptualizing writing test performance	23
Figure 2.2	Categories of context validity for writing assessment.	26
Figure 3.1	Ex post facto research design	36
Figure 3.2	Instructions for the first writing task (a sample prompt form of the trial test)	42
Figure 3.3	Instructions for the second writing task (a sample prompt form of the trial test)	42
Figure 3.4	Graph of text length for task 1 using JASP showing the normality of data assumption	44
Figure 3.5	Values for performing correlations	46
Figure 4.1	Students' perceptions of the writing test regarding context validity across cohorts	53
Figure 4.2	Number of students across proficiency levels: actual scores and perceptions	63

List of Tables

Table 3.1	Overview of the research questions and data analysis	39
Table 3.2	Writing proficiency test construct and context features	41
Table 3.3	Test impact classification by category (data analysis sample from Excel)	45
Table 3.4	Sample of alignment of students' perception and performance in Excel	45
Table 3.5	Pearson's correlation range	47
Table 4.1	Descriptive statistics for text length (word counts) across cohorts (2021 and 2022)	50
Table 4.2	Descriptive statistics for writing time (minutes) across cohorts (2021 and 2022)	51
Table 4.3	Equivalence between language levels and ratings	52
Table 4.4	Descriptive statistics for writing scores across tasks and cohorts	52
Table 4.5	Frequencies of students' positive and negative perceptions of the test by categories	54
Table 4.6	Students' perceptions of self-performance before and after the test.	61
Table 4.7	Descriptive statistics of actual scores and perceptions using the values 0 to 4 to represent the language levels across cohorts.	64
Table 4.8	Correlations between students' initial and final performance perceptions and test scores	64

Abstract

Writing assessments must be valid and fair for test-takers, as assessments evaluate students' performance and can impact their lives. To determine the validity and fairness of a writing test, it is essential to examine its context and consequential validity, aspects that can be examined through test-takers' perceptions. Test-takers' perceptions of a test provide valuable information to enhance assessment practices. While international studies on test validation have largely focused on the impact of EFL proficiency tests on test-takers to improve language assessments, little attention has been given to the validity and fairness of writing tests in Colombia, particularly by collecting context and consequential validity evidence from test-takers. Context validity refers to the appropriateness of the test's features for test-takers, whereas consequential validity pertains to the test's positive or negative impacts on test-takers. This study aims to shed light on the validity and fairness of writing assessments by examining the context and consequential validity of a high-stakes writing test in the Bachelor in Foreign Languages at Universidad del Valle.

The study drew on Shaw and Weir's (2007) sociocognitive framework for writing test validation, which includes context and consequential validity theory. An *ex post facto* research design was employed to analyze qualitative and quantitative data previously collected on test-takers' performance and perceptions. The data belonged to senior students from the 2021 and 2022 cohorts ($N = 103$). The study analyzed students' performance using scores, time length, and text length data. A survey was conducted to gather students' perceptions of the test's validity and fairness and its impact on them. Additionally, students' self-perceptions of performance were collected before and after the test to compare their views with their actual scores, thereby assessing the test's impact on their performance and perceptions.

Performance data results revealed that most students met the minimum text length requirement of 250 words within the 90-minute timeframe, suggesting that the text length and time allocation are generally valid and fair. However, some students struggled to reach 250 words, indicating potential issues related to their writing skills or familiarity with the test format. The average proficiency level was B2, which is below the expected C1 level, suggesting difficulties that could stem from various factors such as writing skills, test familiarity, curriculum issues, or pandemic-related disruptions.

Regarding students' perceptions of the writing test, it was found that time constraints and response format (context validity – task setting) and the physical conditions of the test (context validity – administration setting) were perceived as negative for students. Other

aspects of the test were not seen as problematic. Additionally, psychological and experiential factors, such as stress, nervousness, typing skills, and lack of preparation and familiarity negatively affected students. These findings indicate issues with students' time management skills and test preparation for data analysis and explanatory text writing under assessment conditions.

When comparing test-takers' perceptions of performance before and after the test, there were statistically significant positive correlations between perceptions and actual performance. Students' perceptions generally aligned with their performance, indicating a realistic understanding of their abilities and the test's demands. However, perceptions often decreased after the test, with students feeling less confident in their writing skills. This highlights the need for targeted training and support in test preparation to boost confidence and performance.

Based on these findings, the study concludes that while most aspects of the test are valid and fair regarding context and consequential validity, there are areas needing revision to better accommodate all test-takers. Revisions to the test should focus on improving students' familiarity with the test tasks and conditions and creating a supportive testing environment. Gathering validity evidence directly from test-takers is crucial for enhancing the fairness and validity of writing assessments. Including test-takers' perspectives in the design, administration, and evaluation processes can significantly improve the quality and credibility of high-stakes language assessments. The adoption of Shaw and Weir's (2007) sociocognitive framework for writing test validation was beneficial in this study, offering a structured approach to validating writing tests from the test-takers' viewpoint. This study provides valuable and innovative contributions and insights into the validity and fairness of writing assessments in our educational context, serving as a model for future research and practice in educational assessment.

Keywords: context validity, consequential validity, ex post facto research design, fairness, high-stakes writing assessment, pre-service language teachers, sociocognitive framework for test validation, validity.

Resumen

Las pruebas de escritura deben ser válidas y justas para los examinados, ya que éstas evalúan su desempeño y pueden influir significativamente en sus vidas. Para determinar la validez y justicia de una prueba de escritura, es esencial examinar su contexto y validez consecuencial, aspectos que se pueden evaluar a través de las percepciones de los examinados. Estas percepciones proporcionan información valiosa para mejorar las prácticas evaluativas. Aunque los estudios internacionales sobre la validación de pruebas se han centrado principalmente en el impacto de las pruebas de competencia en inglés como lengua extranjera (EFL) en los examinados para mejorar las evaluaciones de idiomas, se ha prestado poca atención a la validez y justicia de las pruebas de escritura en Colombia, especialmente recolectando evidencia de validez contextual y consecuencial de la prueba a través de los examinados. La validez contextual se refiere a la adecuación de las características de la prueba para los examinados, mientras que la validez consecuencial se refiere a los impactos positivos o negativos de la prueba sobre ellos.

Este estudio tiene como objetivo examinar la validez contextual y consecuencial de la prueba escrita en el examen de proficiencia del programa de Licenciatura en Lenguas Extranjeras de la Universidad del Valle. Para ello, se utilizó el marco sociocognitivo de validación de pruebas de escritura de Shaw y Weir (2007) que se enfoca en estos aspectos de validez. Se empleó un diseño de investigación ex post facto para analizar datos cualitativos y cuantitativos previamente recolectados sobre el desempeño y las percepciones de los examinados. Los datos comprendían estudiantes de último año de las cohortes de 2021 y 2022 ($N = 103$). El estudio analizó el desempeño de los estudiantes utilizando los puntajes de desempeño, tiempo empleado y longitud de los textos. Se realizó una encuesta para recopilar las percepciones de los estudiantes sobre la validez y justicia de la prueba y su impacto en ellos. Además, se recolectaron percepciones de desempeño por parte de los estudiantes antes y después de la prueba para comparar sus opiniones con sus puntuaciones, evaluando así el impacto de la prueba en su desempeño y percepciones.

Los resultados de desempeño revelaron que la mayoría de los estudiantes cumplió con el requisito mínimo de 250 palabras por texto dentro del tiempo límite de 90 minutos, lo que sugiere que la longitud del texto y el tiempo asignado son generalmente válidos. Sin embargo, algunos estudiantes tuvieron dificultades para alcanzar las 250 palabras, indicando posibles problemas relacionados con sus habilidades de escritura o familiaridad con el formato de la prueba. El nivel promedio de competencia fue B2, por debajo del nivel C1 esperado, lo que sugiere dificultades que podrían deberse a varios factores, como

habilidades de escritura, familiaridad con la prueba, problemas curriculares o aspectos relacionados con la pandemia.

Con respecto a las percepciones de los estudiantes sobre la prueba de escritura, las restricciones de tiempo y el formato de respuesta (validez contextual – parámetros de la tarea) y las condiciones físicas (validez contextual – parámetros de la administrativa) fueron percibidos como negativos. Otros aspectos de la prueba no fueron vistos como problemáticos. Además, factores psicológicos y experienciales como el estrés, los nervios, el uso del teclado y la falta de preparación y familiaridad afectaron negativamente a los estudiantes. Estos hallazgos indican problemas con las habilidades de gestión del tiempo, la preparación en análisis de datos y la escritura de textos explicativos en condiciones de evaluación.

Al comparar las percepciones del desempeño de los examinados antes y después de la prueba, se encontraron correlaciones positivas estadísticamente significativas entre las percepciones y el desempeño real. Las percepciones de los estudiantes generalmente se alinearon con su desempeño, indicando una comprensión realista de sus habilidades y las demandas de la prueba. Sin embargo, las percepciones a menudo disminuyeron después de la prueba, indicando una falta de confianza de los estudiantes en sus habilidades de escritura. Esto resalta la necesidad de apoyo específico durante la preparación para la prueba, para aumentar la confianza y el desempeño.

Basado en estos hallazgos, el estudio concluye que, aunque la mayoría de los aspectos de la prueba son válidos e imparciales en cuanto a la validez contextual y consecucional, hay áreas que necesitan revisión. Estas revisiones deben centrarse en mejorar la familiaridad de los estudiantes con las tareas y condiciones de la prueba, así como crear escenarios que sirvan de apoyo. Recoger evidencia de validez directamente de los examinados es crucial para mejorar la imparcialidad y validez de las evaluaciones de escritura. Incluir las perspectivas de los examinados en el diseño, administración y procesos de evaluación puede mejorar significativamente la calidad y credibilidad de las evaluaciones de idiomas de alta importancia.

La adopción del marco sociocognitivo de validación de pruebas escritas de Shaw y Weir (2007) fue beneficiosa en este estudio, ofreciendo un enfoque estructurado para validar las pruebas de escritura desde el punto de vista de los examinados. Este estudio proporciona información valiosa sobre la validez y justicia de las evaluaciones de escritura en nuestro contexto educativo, sirviendo como modelo para futuras investigaciones y prácticas en la evaluación educativa.

Palabras clave: validez contextual, validez consecuencial, diseño de investigación ex post facto, evaluación de escritura de alta importancia, justicia, marco sociocognitivo de validación de pruebas, profesores de idiomas en formación, validez.

Résumé

Les évaluations en écriture doivent être valides et équitables pour les candidats, car elles évaluent la performance des étudiants et peuvent influencer considérablement leur vie. Pour déterminer la validité et la justice d'un test d'écriture, il est essentiel d'examiner son contexte et sa validité conséquente, des aspects qui peuvent être évalués par les perceptions des candidats. Ces perceptions fournissent des informations précieuses pour améliorer les pratiques d'évaluation. Bien que les études internationales sur la validation des tests se soient largement concentrées sur l'impact des tests de compétence en anglais langue étrangère (EFL) sur les candidats pour améliorer les évaluations de langue, peu d'attention a été accordée à la validité et la justice des tests de rédaction en Colombie, notamment en recueillant des preuves de validité contextuelle et conséquente auprès des candidats. La validité contextuelle se réfère à la pertinence des caractéristiques du test pour les candidats, tandis que la validité conséquente concerne les impacts positifs ou négatifs du test sur eux.

Cette étude vise à examiner la validité contextuelle et conséquente du test d'écriture de l'examen de compétences dans le programme de Licence en Langues Étrangères de l'Universidad del Valle. Elle a utilisé le cadre sociocognitif de Shaw et Weir (2007), qui se concentre sur ces aspects de validité des preuves. Une recherche ex post facto a été employée pour analyser des données qualitatives et quantitatives précédemment recueillies sur la performance en écriture académique et les perceptions des candidats. Les données comprenaient des étudiants en dernière année des cohortes 2021 et 2022 (N = 103). L'étude a analysé la performance des étudiants en utilisant des scores, la durée et la longueur des textes. Une enquête a été réalisée pour recueillir les perceptions des étudiants sur la validité et la justice du test et son impact sur eux. De plus, les perceptions de la performance des étudiants ont été recueillies avant et après le test pour comparer leurs points de vue avec leurs scores réels, évaluant ainsi l'impact du test sur leur performance et leurs perceptions.

L'analyse des données de performance a révélé que la plupart des étudiants ont respecté l'exigence minimale de longueur de texte de 250 mots dans le délai de 90 minutes, ce qui suggère que la longueur du texte et l'allocation de temps sont généralement valides et équitables. Cependant, certains étudiants ont eu du mal à atteindre les 250 mots, indiquant des problèmes potentiels liés à leurs compétences en rédaction ou à leur familiarité avec le format du test. Le niveau moyen de compétence était B2, ce qui est en dessous du niveau C1 attendu, suggérant des difficultés qui pourraient découler de divers facteurs tels que les

compétences en rédaction, la familiarité avec le test, les problèmes curriculaires ou les perturbations liées à la pandémie.

Concernant les perceptions des étudiants sur le test de rédaction, les contraintes de temps et le format de réponse (validité contextuelle - cadre de la tâche) et les conditions physiques (validité contextuelle - cadre de la gestion) étaient des facteurs négatifs significatifs. D'autres aspects du test n'étaient pas considérés comme problématiques. De plus, des facteurs psychologiques et expérientiels comme le stress, la nervosité, les compétences en dactylographie et le manque de préparation et de familiarité ont affecté négativement les étudiants. Ces résultats indiquent des problèmes avec les compétences de gestion du temps des étudiants et la préparation à l'analyse de données et à la rédaction de textes explicatifs dans des conditions d'évaluation.

En comparant les perceptions de la performance des candidats avant et après le test, des corrélations positives statistiquement significatives ont été trouvées entre les perceptions et la performance réelle. Les perceptions des étudiants correspondaient généralement à leur performance, indiquant une compréhension réaliste de leurs capacités et des exigences du test. Cependant, les perceptions diminuaient souvent après le test, les étudiants se sentant moins confiants dans leurs compétences en rédaction. Cela souligne le besoin de formation et de soutien ciblés dans la préparation au test pour augmenter la confiance et la performance.

Sur la base de ces résultats, l'étude conclut que, bien que la plupart des aspects du test soient valides et équitables en ce qui concerne la validité contextuelle et conséquente, il y a des domaines nécessitant une révision pour mieux accommoder tous les candidats. Les révisions devraient se concentrer sur l'amélioration de la familiarité des étudiants avec les tâches et les conditions du test et la création d'un environnement de test favorable. Recueillir des preuves de validité directement auprès des candidats est crucial pour améliorer la justice et la validité des évaluations de rédaction. Inclure les perspectives des candidats dans le processus de conception, d'administration et d'évaluation peut améliorer considérablement la qualité et la crédibilité des évaluations linguistiques à enjeux élevés. L'adoption du cadre sociocognitif de Shaw et Weir (2007) a été bénéfique dans cette étude, offrant une approche structurée pour valider les tests de rédaction du point de vue des candidats. Cette étude fournit des informations précieuses sur la validité et la justice des évaluations de rédaction dans notre contexte éducatif, servant de modèle pour les futures recherches et pratiques en évaluation éducative.

Mots-clés: validité contextuelle, validité conséquentielle, recherche ex post facto, évaluation d'écriture, cadre de validation des tests sociocognitifs, enseignants de langues en formation initiale, justice validité.

1 Introduction

1.1 Chapter overview

In this chapter, we introduce and justify the importance of investigating the validity, fairness, and impact of writing assessments on test-takers, particularly the writing test used in the Bachelor of Foreign Languages at Universidad del Valle, to assess senior students' writing proficiency. The chapter discusses the relevance of writing assessment in language testing and the significance of collecting evidence from test-takers' perspectives to examine the validity and fairness of a test. It also addresses the research gap regarding the validity and fairness of writing tests in Colombia and explains the aim of this study. Also, it represents an innovative and significant contribution to the area of writing assessments validation in Colombia and to the assessment practices of the academic program regarding test validation processes.

1.2 The validity and fairness of writing assessments

From a sociocognitive perspective of writing assessment, adopted in the present study, writing is a social and cultural act as well as a cognitive activity with context, purpose, and audience as key parameters (Shaw & Weir, 2007). Writing is part of the language abilities EFL students and test-takers must master to show their language proficiency. Hence, writing assessments should be valid and fair for test-takers (Bachman & Palmer, 2010; Messick, 1996; Shaw & Weir, 2007).

In the research context, writing is one of the abilities measured in the English language proficiency test administered to senior students (as test-takers) in the academic program. In this context, a language proficiency test is defined as a test that enables students to transcend the context of their study program subjects and assess their language competence in a holistic manner (Universidad del Valle, 2021). This test allows undergraduate students to familiarize themselves with this type of test they will encounter in their future academic and professional lives.

This writing test assesses students' C1 level, meaning that students should be able to produce clear, well-structured, detailed texts on complex subjects, showing controlled use of organizational patterns, connectors, and cohesive devices (Council of Europe, 2020). This writing test should be well designed so as not to affect students' performance (Bachman & Palmer, 2010; Messick, 1996; Shaw & Weir, 2007). Although the writing test was revised by experts and validated with a small sample of students of the program, its validity and fairness have not been examined with a wide sample of participants. As the test characteristics are

part of the context of writing assessment, such features might affect students' performance, perceptions and experiences. Thus, it is relevant to examine its context and consequential validity evidence from test-takers' perceptions.

In language assessment and test validation, a current view of test validity refers to the evidence and theory that support the interpretations and uses of test scores, whereas test validation refers to process of collecting evidence to examine if a test is valid and fair (AERA, APA, & NCME, 2014). For a test to be valid and fair, different ways of validity evidence should be collected (Bachman & Palmer, 2010; Messick, 1996; Shaw & Weir, 2007). Several frameworks to test validation have been employed to explore the validity and fairness of language tests (as mentioned in Chapter 2 - Literature review). However, for the purpose of the present study, we drew on Shaw and Weir's (2007) writing test validation framework, which entails collecting evidence of different validity categories to evaluate a writing test's validity and fairness. For the present study, we examined the context and consequential validity evidence of the writing test of the academic program.

Context validity is the degree to which an assessment accurately reflects the real-life language use and communicative situations for which it is designed and ensures that the test reflects the writing of test-takers (Shaw & Weir, 2007). Consequential validity refers to the impact of test scores and testing procedures on test-takers and other stakeholders, as well as on the larger social context in which the test is implemented (Bachman & Palmer, 2010; Messick, 1996; Shaw & Weir, 2007). Consequential validity is important because it ensures that test scores are meaningful and relevant to the test-takers and other stakeholders, in this case, lecturers and administrators of the program.

Validating tests includes examining and collecting evidence of factors, such as the motivation and engagement of test-takers, the fairness and equity of testing procedures, and the social and educational consequences of test scores (Bachman & Palmer, 2010; Messick, 1996; Shaw & Weir, 2007). For our study, it was relevant to identify students' perceptions of the context (i.e., content and administration) and impact of the test on them, as this proficiency test holds a pivotal position as a mandatory examination for graduation. In other words, students are required to pass the test to obtain their degree. The consequences of this test, whether positive or negative, are substantial. Failing the proficiency test can lead to an undesirable delay in students' graduation, necessitating additional preparation and testing. On the other hand, passing this high-stakes examination promises graduation and enhanced employability. If test-takers perceive a test as unfair or irrelevant, they may not be

motivated to perform their best, leading to inaccurate or unreliable test scores (Bachman & Palmer, 2010).

Another important aspect of context and consequential validity is the consideration of test-takers' perceptions and experiences regarding the clarity and fairness of testing instructions, the relevance and importance of the test content, and the perceived value of the test results (Bachman & Palmer, 2010; Brown & Abeywickrama, 2019; Cheng & Curtis, 2012; Im et al., 2019; Messick, 1996; Shaw & Weir, 2007). It is important to consider test-takers' perceptions of a test because they can impact their motivation, engagement, and performance on the test and their attitudes towards testing and education more broadly. Moreover, their perceptions can inform teaching and assessment practices. For instance, test-takers can provide insights into test preparation, adjustments or improvements in test design, content, and scoring criteria. This information is instrumental in refining the test itself and in shaping more targeted and effective test preparation strategies.

At the international level, the scope of studies on validity and reliability (i.e., consistency of scorings) is considerably broad, with most studies examining the validity and reliability of high-stakes standardized tests, such as TOEFL and IELTS (e.g., Bakri, 2023). However, despite the importance of test validity in educational assessments, this aspect is still largely underexplored in Colombia, even more so in tertiary education, where university students take high-stakes tests to show their English proficiency (Guapacha & Chaves, 2023; Guapacha, 2020, 2022). There have been a few studies into the validity of writing assessments in Colombia (Muñoz et al., 2006; Muñoz & Álvarez, 2008); however, the scarcity of studies in this context leads to limited knowledge about the current state of writing assessments (Guapacha & Chaves, 2023). Given the lack of studies on the validity and fairness of writing tests in the research context, it seems that EFL tests used in educational contexts might be designed without considering the potential consequences of their use, either for test-takers or the broader social context.

At Universidad del Valle, specifically at the School of Language Sciences, there have been studies on proficiency level, assessment practices, item analysis, and students' preferences for test mode (e.g., Gómez, 2016; Guapacha, 2020; Kostina, 2012; Kostina et al., 2013; Ramirez, 2020). However, there are no studies on test validation. As pre-service language teachers in this context take an English language proficiency test as a graduation requirement; hence, it is relevant to examine the validity and fairness of writing tests by collecting context and consequential validity evidence from test-takers.

The motivation for carrying out this study relies on the fact that there is no published research on the context and consequential validity of this in-house writing proficiency test. Previous studies in this context have addressed student performance and assessment practices in the program and curriculum; however, they have not explored the relationship between test-takers' perceptions and performances. We focus on the writing test because recent results and lecturers' views report that students have low performance in this skill. Hence, it is important to investigate the validity of the writing test from the test-taker's perspective to improve the test and to provide feedback to teachers and students in this context. Investigating this population is also relevant because, as future educators, pre-service language teachers' writing is a fundamental skill that they need to master, as it plays a central role in language instruction, lesson planning, creating instructional materials, providing feedback on assignments, and assessing student work. Additionally, language teachers serve as models for their students, and their proficiency in writing sets a standard that students strive to achieve. They must also be aware of the importance of valid and fair writing assessments.

1.3 Research questions

The purpose of this study is to examine the validity and fairness of the writing proficiency test of the bachelor's in foreign languages at the Universidad del Valle, regarding its context and consequential validity, from students' voices and perspectives.

1.3.1 Main research question

How valid and fair is the writing test regarding its context (content and administration) and consequential validity (positive or negative test impacts) from student's perspective?

1.3.2 Specific research questions

- RQ1. What is the performance of the students in the writing test regarding writing time, text length, and scores?
- RQ2. What are the students' perceptions of the test regarding context validity?
- RQ3. What are the students' perceptions of performance before and after the test (consequential validity)?
- RQ4. What is the relationship between their perceptions and performance in the writing test (consequential validity)?

1.4 Significance and contribution of the study

1.4.1 Theoretical significance

This research holds substantial theoretical significance within the domain of language assessment. It advances theory and practice in language test validation in our research context by exploring the concepts of test validity, fairness, context validity and consequential validity. The study investigates the relationship between assessment tools and their contextual relevance, thereby contributing to the theoretical understanding of how language assessments should be designed to align effectively with the objectives and educational contexts of a program. Furthermore, this study makes a significant contribution to test validation theory by exploring the implications of writing proficiency tests for EFL learners from an underexplored geographical context. By examining these implications, the research highlights how language assessments can transform teachers, learners and administrators' educational paths and practices. This exploration enriches our understanding, as students and future teachers, of the importance of language assessment and its impact on all stakeholders.

1.4.2 Research significance

This study is significant and innovative in its purpose and design because it provides empirical evidence to evaluate the validity and fairness of a writing proficiency test following underexplored theoretical and research procedures. The sociocognitive framework for writing test conceptualization is a comprehensive framework that guides teachers and researchers in examining the validity and fairness of a writing test. The incorporation of an ex post facto research design is a noteworthy and innovative aspect of this study for both the researches and the context because it analyzes genuine data already collected by the program about its students' performance and perceptions. This empirical methodological choice allows for an investigation into the natural outcomes of the test, making it especially valuable in our context. All this data helps to provide feedback to lecturers and students and revise our own practices.

1.4.3 Practical significance

This research project is highly pertinent in terms of its practical implications for language assessment and education, particularly in our academic program. It has the potential to elevate the quality of the writing proficiency test by making it contextually relevant, valid, reliable and fair for the students in this program. Consequently, the findings of this study directly inform language educators, program administrators, and assessment experts to enhance language assessment practices and better align them with the

pedagogical objectives of the program. Moreover, the findings of this study can significantly influence the broader educational policy landscape and serve as a model for further studies on test validation. By illuminating issues related to writing assessment, the research has the potential to shape language assessment policies at the institutional level, such as incorporating student feedback into test design and administration, implementing clearer guidelines for test timing, providing adequate resources for preparation and practice, and fostering a culture of continuous improvement in language assessment methodologies. These outcomes can contribute to enhancing the overall validity and fairness of language assessments, ultimately benefiting both students and educational institutions.

1.5 Thesis outline

This thesis is organized into six chapters. **Chapter 1: Introduction** introduces the research topic, its significance, and the context at Universidad del Valle. **Chapter 2: Literature Review** comprises two parts. The first one, the theoretical framework, delves into key language assessment theories, the evolution of the concept of validity, and Shaw and Weir's (2007) framework for writing proficiency tests validation. The second part, empirical research, summarizes relevant local, national and international studies regarding writing assessment, context validity and consequential validity. **Chapter 3: Research Methodology** elucidates the research design, data collection methods, and ethical considerations while also considering potential limitations. **Chapter 4: Findings** reveals the results regarding context validity and consequential validity, interpreting them within the context of the research objectives and discussing theoretical, practical, and policy implications. **Chapter 5: Discussion** further interprets and discusses the study results. The significance and implications of the findings are explained, discussed and related to the context and existing research in the field. **Chapter 6: Conclusion** summarizes key findings and reflects on their significance, providing recommendations for practice and future research. The reference section and appendices complement the study's academic rigor.

2 Literature Review

2.1 Chapter overview

This section covers the theoretical framework and empirical research on the context and consequential validity of writing assessments. The first part provides the theoretical definitions of test validity, fairness, test validation frameworks and context and consequential validity. The second part highlights some relevant previous studies related to writing test validation that informed our study.

2.2 Theoretical framework

2.2.1 *Validity and fairness in language assessment*

Validity and fairness are crucial considerations in evaluating language assessments, in this case, writing assessments. Historically, the term *validity* referred primarily to the extent of content coverage by a task. However, it has since evolved to incorporate a broader range of evidence. A traditional and narrowed view of validity refers to the extent to which an assessment accurately measures what it intends to measure; however, currently, *validity* entails gathering various sorts of data to determine whether a test is valid and fair in assessing a test-taker's performance (Akbari, 2012; Green, 2014; Im et al., 2019). The most current view of validity is "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 2014, p. 11). This view was influenced by Messick's (1989, 1998) view of validity. Messick (1989, 1998) stated that:

All validity is of one kind, namely, construct validity. Other so-called separate types of validity -whether labeled content validity, criterion-related validity, consequential validity, or whatever – cannot stand alone in validity arguments. Rather, these so-called validity types refer to complementary forms of evidence to be integrated into an overall judgment of construct validity (p. 37).

Messick's (1989,1998) views of validity became a framework that provides a comprehensive, multidimensional approach to assessing the validity of educational assessments. Messick's view recognizes that validity is not a fixed property of an evaluation, but rather an ongoing evaluation process. By considering multiple aspects of validity, this framework helps ensure that assessments are reliable, meaningful, and appropriate. Messick emphasized a broader perspective on validity, considering it as a multifaceted concept rather than a single test quality.

On the other hand, *fairness* ensures that the evaluation procedure and results are equal and unbiased for all test-takers (Bachman & Palmer, 2010; Phakiti & Isaacs, 2021;

Shaw & Weir, 2007; Weir, 2005). In his framework, Messick (1996) argued that a test must be fair because it can have unintended and intended positive and negative consequences or effects on test-takers, classrooms, and society. This effect type is also called consequential validity or test impact (Bachman & Palmer, 2010; Messick, 1996; Shaw & Weir, 2007). Assessments that are unfair to all students are not ethical (Phakiti & Isaacs, 2021). Assessments that are biased toward a specific group (e.g., using topics that favor one group over another for reasons extraneous to the ability being measured) are unfair. Another facet of fairness is ensuring equitable access and opportunity to undertake the assessment, including accommodations for students with special needs (Kunnan, 2018).

Shaw and Weir (2007) assert that a test of competences will be fair if it is reliable, open to administration, impartial in its assessment, has reasonable consequences and opportunities for student growth, and can be tailored to meet the needs of specific individuals. If any of these elements are weakened the assessment may be considered unfair. Fairness in exams is a crucial concern in education, particularly when talking about high stakes. In our context, it is relevant to identify whether the tasks, computer condition, and length, among other factors, are fair to these students from their perspectives.

All in all, our study adopts the current approach to validity because it is a dynamic concept that adapts to changing educational and social environments and encourages continuous, evidence-based assessment practice evaluation. By adopting this current view, we can collect different forms of validity evidence to understand that our evaluations are sensitive to people's needs and traits, resulting in more valid and fair results. For instance, we can collect evidence of the writing test's content, physical conditions, instructions, etc., to identify that these factors do not affect students' performance.

The current notions of validity and fairness, adopted in the present study, take on paramount importance when it comes to a public university language proficiency test, in this case, the writing test. Validity is crucial because it guarantees that the test measures the fundamental skills and knowledge that our students should be able to demonstrate, maintaining the academic standards of the institution and enabling well-informed decisions about future steps. Fairness is necessary to guarantee that the test does not establish unjustified obstacles or disadvantages for any group of test-takers, promoting equal opportunities for everyone, irrespective of their circumstances, socioeconomic background, or cultural differences. Equally important is understanding the context of test-takers, which recognizes that different students may bring different preparation levels and life experiences to the exam, requiring fair and focused assessments that take these differences into account.

To examine the validity and fairness of a language test, multiple sources of evidence are necessary to support test score interpretations and decisions on test use; that is, to support a test validity and fairness. To do so, several frameworks for test validation have been proposed, as discussed below.

2.2.2 Test validation

Test validation is the process of collecting evidence that allows test users to have better interpretations of the test-takers' performance and test use; in other words, it is about justifying the claims made about the validity of the test (Bachman & Palmer, 2010). The concept of test validation is directly related to this research because it collects information from test-takers to justify the validity and fairness of the writing test. More specifically, the present study is interested in examining the context validity (i.e., content, features, and administration) and consequential validity (or the effect of the test on students) evidence of the high-stakes writing test from students' voices and perceptions.

Based on the work of eminent academics (e.g., Bachman, 2005; Bachman & Palmer, 1996; Carroll, 1980; Chapelle & Voss, 2014; Kane, 1992, 2001, 2002, 2004, 2006, 2013; Kane et al., 1999; Messick, 1989), we can articulate the evolution of test validity and validation in language testing research over the past few decades. The evolution includes four validation approaches: (1) the one-question approach; (2) evidence-gathering; (3) test usefulness; and (4) argument-based approach.

Messick's (1986, 1996) view of validity was the first approach or framework to examining the validity and fairness of a test. However, his framework was criticized because it did not provide specific criteria to collect validity evidence (Im et al., 2019). Drawing on Messick's (1986, 1996) work, several test validation frameworks have been employed to explore the validity and fairness of language tests. These frameworks encompass the argument-based approach (Kane, 2013), Test Usefulness (Bachman & Palmer, 1996), Assessment Use Argument (Bachman & Palmer, 2010), Evidence-based design (Mislevy & Risconscente, 2005), and the Sociocognitive framework (Weir, 2005). These frameworks incorporate several types of validity evidence, such as content, face, scoring, construct, and consequential validity evidence (Akbari, 2012; Green, 2014; Im et al., 2019).

Although previous test validation frameworks can assess the validity and fairness of any language test, Shaw and Weir's (2007) framework is specifically designed for validating writing tests. These authors propose a sociocognitive approach to test validation that includes addressing the context and consequential validity evidence of a writing test as it

emphasizes the importance of collecting test-takers' views of the validity and fairness of a writing test.

According to recent test validation frameworks, all validity evidence serves to develop a case concerning the validity of a test in terms of test interpretation and use. In the present study, we wanted to know whether the writing test designed by the B.A is valid and fair in assessing pre-service language teachers' academic writing skills when producing computer-based academic essays. This study aims to determine whether the writing test designed by the B.A program is valid and fair in assessing pre-service language teachers' academic writing skills, specifically when producing computer-based academic essays. It seeks to evaluate the validity and fairness of the test's features and administration in measuring these students' writing performance, as well as to identify any potential positive or negative effects the test may have on them.

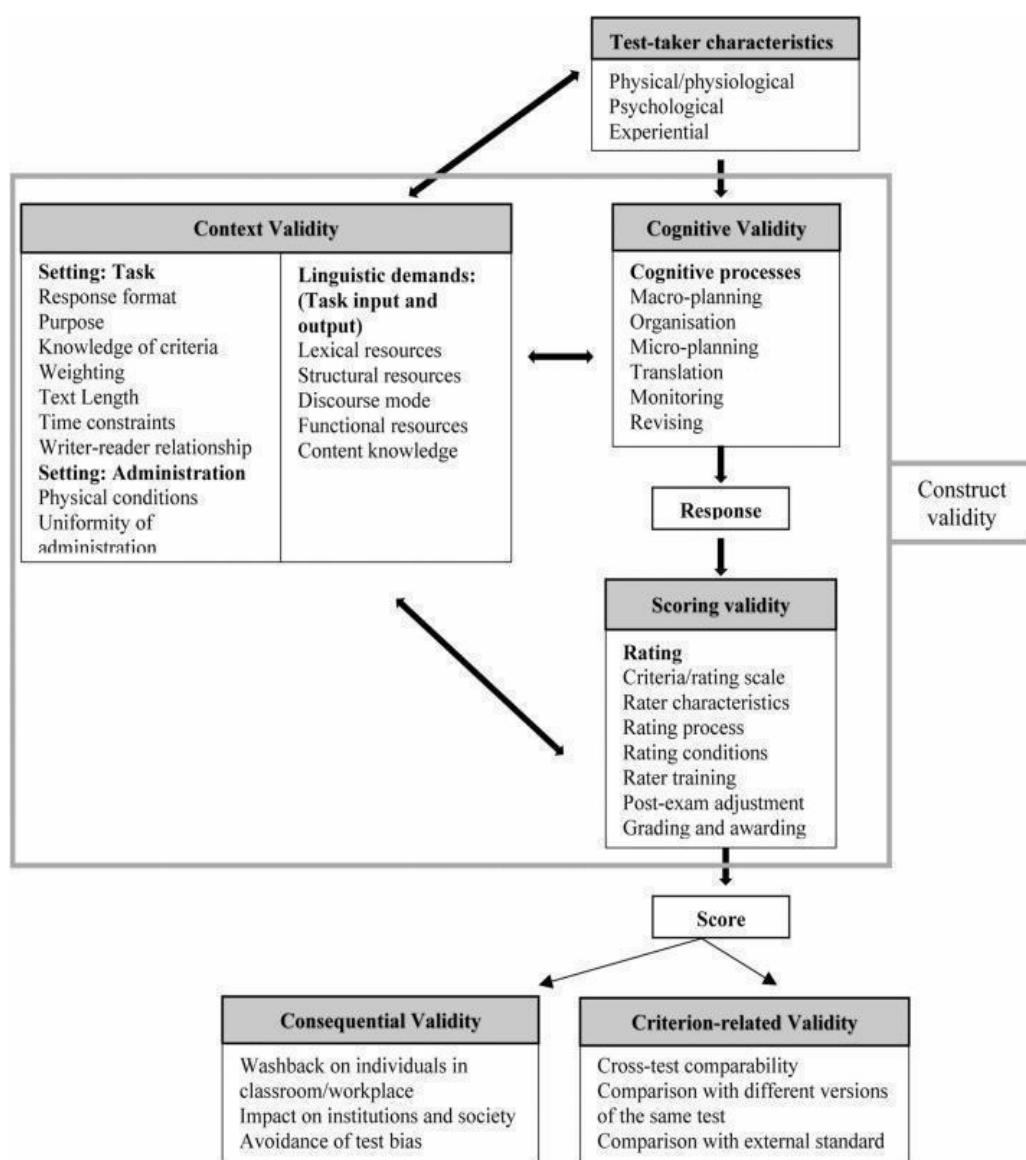
In this context, writing is one of the abilities measured in the English language proficiency test administered to senior students (as test-takers) in the academic program. In this context, a language proficiency test is defined as a test that enables students to transcend the context of their study program subjects and assess their language competence in a holistic manner (Universidad del Valle, 2021). This test allows undergraduate students to familiarize themselves with this type of test they will encounter in their future academic and professional lives. To examine the validity and fairness of this writing test, we adopted Shaw and Weir's (2007) sociocognitive framework for conceptualizing and validating second-language writing tests, as discussed below.

2.2.3 Shaw and Weir's (2007) sociocognitive framework for writing test validation

According to Shaw and Weir (2007), the sociocognitive framework "represents a unified approach to establishing the overall validity of the test" (p. 3). The various categories of evidence acquired are interconnected and strive to create a case for test validity and fairness (Shaw & Weir, 2007). Shaw and Weir's framework outlines six components that should be addressed during the test validation process: *test-taker characteristics*, *cognitive validity*, *context validity*, *scoring validity*, *consequential validity*, and *criterion-related validity*. All the validities together help to build up an argument about the validity of a test regarding the interpretations and uses of test scores. (Figure 2.1).

Figure 2.1

Shaw and Weir's (2007) sociocognitive framework for conceptualizing writing test performance



Shaw and Weir's (2007) *Sociocognitive Framework for Conceptualizing Writing Test Performance*. Figure taken from Guapacha (2020, p. 23)

Shaw and Weir contend that the *context*, *cognitive*, and *scoring validity*—collectively known as construct validity—have a symbiotic relationship and that evidence supporting the construct of writing skill must come from each of these three validities. The cognitive skills of test-takers (such as planning and revising), the task environment (such as academic writing and writing medium), and the scoring procedure (such as rater reliability) are related in a triangular pattern (refer to Figure 2.1). On the other hand, the effects and impact of test results, as well as how these results compare with other measures of the same construct in contexts outside of the test, are related to the external dimensions of a test, *consequential validity*, and *criterion-related validity*, respectively (Shaw & Weir, 2007).

Shaw and Weir's (2007) sociocognitive framework was chosen in our study because it provides a methodical and thorough way to design and validate writing tests. This framework can be used for examining the validity of high-stakes and classroom assessments (Guapacha, 2022). Furthermore, the writing test was designed based on this sociocognitive view of writing, which is also adopted in the program's curriculum. In the specific case of our study, this framework highlights the characteristics of the test-takers and other factors such as the environment, the tools, the devices, the test, and many others that directly and indirectly affect performance in a writing test. For the present study, the following components of the framework will be discussed as they relate to aspects of the test and test-taker that can affect their performance: *test-taker characteristics*, *context validity*, and *consequential validity*.

2.2.3.1 Test-takers' characteristics.

Bachman and Palmer (2010) believe that several variables related to test-takers can affect test results. The following are listed in order: 1) personal attributes (age, sex, nationality, resident status, length of residence, native language, level and type of general education, time, and amount of preparation or prior experience with a given time); 2) topical knowledge or content knowledge; 3) affective schemata (feelings associated with particular kinds of topical knowledge); 4) cognitive strategies (strategies that language users employ when executing plans); and 5) language ability. In a similar vein, Shaw and Weir's (2007) writing assessment framework offers insightful information about several other test-taker traits. These traits include physical, psychological and experiential characteristics, which offers a holistic perspective on the test-taker's abilities and competencies beyond mere

writing proficiency. For instance, students' handwriting might be a physical characteristic that affects students' performance on a computerized writing test. Similarly, nervousness, anxiety and lack of experience or of test preparation might produce a negative impact on students.

For our study, the taxonomy of test-taker characteristics by Shaw and Weir is useful to identify students' reported personal traits that affect their performance. Such information provides a washback effect (i.e., impact on classroom practices) and helps enhance the validity and fairness of the writing test.

1.1.1.1 Construct validity.

As already stated, context validity is part of the construct validity of a writing test. *Construct Validity* is relevant in language assessment because it is the extent to which a test adequately measures the underlying abilities or skills of interest (Bachman & Palmer, 2010). Shaw and Weir (2007) define construct validity as the interconnection between cognitive, context, and scoring validity. *Cognitive Validity* is determined by how well the test captures the cognitive processes (planning, revising, editing, etc.) required in writing situations other than the test, or in carrying out the task in real life (Shaw & Weir, 2007). *Scoring Validity* refers to the procedures that accurately reflect the test results. *Context Validity* (also referred to as content validity) is the degree to which the assessment accurately captures the information, skills, and abilities pertinent to the particular linguistic construct under assessment.

Although the construct of writing ability of the present study involves cognitive processing, the context of the test and the scoring process, the following section will discuss the context validity of the writing test, which is the focus of the study. It is aimed to determine whether the writing test context is valid and fair from the students' perspectives and if it has any positive or negative consequences for students' performance. We will examine the context validity of the writing test through its features and administration. Cognitive and scoring validity were not investigated because they involve collecting evidence of the test-taker's cognitive processes and raters' ratings, which is not the purpose of this study.

1.1.1.2 Context validity.

According to Shaw and Weir (2007), *context validity* is the degree to which an assessment accurately reflects the real-life language use and communicative settings for which it has been designed. It includes the ecological validity of an evaluation, guaranteeing that the activities and information correspond to the target language domain and the

particular situations in which the language is employed. The notion of physical context is suggested by Bachman and Palmer (2010) as one of the elements affecting context validity. The physical context corresponds to the external environment and circumstances in which language is utilized, including the communication tools, resources, and location. Therefore, context validity makes certain that the assessment accurately measures a test-taker's capacity to function in real-life communicative situations by considering the appropriateness and authenticity of the language tasks and content.

The context validity framework proposed by Shaw and Weir highlights the importance of a comprehensive and methodical approach to creating assessments that truly represent the material they are aimed at assessing. This will improve the overall standard and significance of language testing. In Shaw and Weir's framework, the following aspects need to be examined to assess the context validity of a writing assessment (See Figure 2.2).

Figure 2.2
Categories of context validity for writing assessment.

Context Validity	
Setting: Task Response format Purpose Knowledge of criteria Weighting Text Length Time constraints Writer-reader relationship Setting: Administration Physical conditions Uniformity of administration	Linguistic demands: (Task input and output) Lexical resources Structural resources Discourse mode Functional resources Content knowledge

Task Setting

- **Response format:** The response format used in a writing assessment determines how test-takers organize their written responses. The format could be a letter, an essay, or a brief response, for instance, and each needs a particular structure.
- **Purpose:** A writing assessment should have a particular objective that is in line with the abilities or competences being assessed. This objective should be achieved by the response format, which guarantees that the test assesses the necessary writing abilities.

- **Knowledge of criteria:** Test-takers should be well aware of the standards by which their answers will be judged. This information enables them to respond in the designated answer format while meeting the necessary specifications.
- **Weighting:** Different sections or aspects of a writing exam may have varying weights. For example, longer essay responses might be given more weight. Weighting also applies to different writing standards, such as structure, language use, mechanics, and content.
- **Text length:** The expected text length is integral to the response format in writing assessments. The expected word count, paragraphs, and sections. The length of the text can impact the assessment, especially when there are specified length requirements. If a response is too short or excessively long, it may affect the overall evaluation.
- **Time constraints:** Time constraints must be considered because they may impact the writer's capacity to organize, develop, and revise their response. It might also have an impact on the writing's general quality.
- **Writer-reader relationship:** This leads to a correlation with how successfully the author interacts with the reader and builds an accurate connection through communication. Coherence, tone, and cohesion can all affect this relationship and, consequently, the assessment itself.

Administration setting

- **Physical conditions:** The assessment process requires maintaining uniform physical circumstances for each test-taker. It includes the influence of location, comfort, technology and tools, distractions, resources, ergonomics, etc. Those conditions can change, which could lead to prejudice and disadvantage certain test-takers while benefiting others. Ensuring equitable conditions for all participants is crucial.
- **Uniformity of administration:** This concept ensures the consistent and standardized application of administrative processes and policies across an organization, emphasizing consistent procedures, standardized policies, fairness, predictability, efficiency, and accountability. This approach ensures that all members are treated equally, processes are understandable, operations are streamlined, and compliance with standards is easily monitored, fostering a cohesive, transparent, and efficient organizational environment.

Linguistic demands (Task input and output)

- **Lexical resources:** This section focuses on the writer's language and vocabulary selections. Examiners determine whether the writer successfully communicates their argument or point of view using acceptable language and vocabulary.
- **Structural resources:** The response's structural composition is essential. This entails elements like headings, paragraphing, and the coherence of ideas.
- **Discourse mode:** The purpose and genre of the writing must be supported by the discourse mode (e.g., narrative, expository, persuasive) that is selected. An improper speech mode may have an unfavorable impact on the assessment.
- **Functional resources:** It is known as the ability of the writer to utilize language to perform specific assignments, including describing, persuading, or explaining.
- **Content knowledge:** This component evaluates the writer's subject-matter knowledge. According to Shaw and Weir (2007), the subject matter of a writing assessment is usually the subject or topic that test-takers are required to write about in their responses. It includes the particular topic or issue that the writing assignment is meant to assess.

According to Shaw and Weir's conceptualization, the categories of context validity are a key factor in determining how well test-takers perform in language tests. Context validity helps to guarantee that the goals and the subject matter of the test reflects pragmatic use of language instances, which can improve test-takers' confidence, motivation, and level of engagement. Test-takers are more likely to demonstrate their genuine linguistic ability and perform at their best when they believe that the assessment is pertinent and important to their language learning objectives and daily communication demands. Furthermore, because context-valid assessments provide test-takers with a clear idea of what is expected of them and how they will be evaluated, they have a greater opportunity to approach the tasks strategically. A positive testing environment is fostered when the assessment context and test-takers' experiences are compatible. This can have a favorable effect on the test-takers' assessment experience and performance.

Overall, the framework of context validity was useful in the present study as the baseline to identify students' positive or negative perceptions of the test, which informed context and consequential validity. The next section discusses the relevance of consequential validity evidence of a writing test.

2.2.3.2 Consequential validity.

The terms *consequential validity* (Messick, 1996; Shaw & Weir, 2007), *test impact* (Bachman & Palmer, 1996), and *test effect* (Phakiti & Isaacs, 2021) all allude to the impact that testing has on educational environments and society. These terms also address the ethical and societal ramifications of how tests are interpreted and used. A test should bring positive rather than negative consequences to students' lives and outcomes. Shaw and Weir (2007) propose three viewpoints of the consequential validity of a test: test bias avoidance, impact on institutions and society, and washback on individuals in classrooms. We will focus on avoidance of test bias and washback.

In this study, we will focus on the avoidance of test bias and washback, as these directly impact the individual test-takers and the classroom environment, which are the primary concerns of this research. While the impact on institutions and society is equally important, it falls outside the scope of this study, which is centered on the immediate effects on students and their educational experiences. A broader exploration of societal and institutional impacts could be considered for future research.

Avoidance of biases. Avoidance of bias means checking that no unfair bias occurred for individuals due to decisions taken earlier about contextual features of the test (Shaw & Weir, 2007). For example, in a group with diverse cultural backgrounds, writing topics must be familiar and meaningful to all test takers to avoid. This is a crucial procedure that ensures that none of the participants is unfairly given an advantage or disadvantage. It begins with an extensive review to remove any prejudice that can result from parameters such language variety, social status, or cultural background, among others. Shaw and Weir (2007) stress the significance of this aspect in guaranteeing the impartiality and equity of evaluations, emphasizing the necessity of addressing prejudices that may impact test-takers' performance results.

Washback is the effect of assessment on teaching and learning as a result of a testing system; it informs the intended and unintended consequences of test use (Bachman & Palmer, 1996; Messick, 1996; Shaw & Weir, 2007). The idea of washback acknowledges that assessments can influence how teachers teach and how students learn, going beyond simple evaluations by asking students about their needs and perceptions of performance. Teachers should not necessarily focused on teaching to the test but on providing students with skills to be demonstrated in real-life situations. It draws attention to how closely instruction, assessment, and learning outcomes are related. Clear standards and criteria for writing assignments can be provided by effective assessments with positive washback, which include designing the test according to the construct, test-takers' characteristics and needs,

etc. Negative washback, such as inappropriate prompts or students' low performance on a test provide insights into the issues of the test and teaching, which could be improved in the classroom. These assessments can also inspire students to practice writing in a meaningful way by promoting writing methods and skills. This can then result in enhanced linguistic ability and writing performance. For instance, if students report that they are unfamiliar with the type of task or that the text length and time are difficult, unfair or inappropriate, it is necessary to help students familiarize themselves with the task features or make revisions to the test. For example, they will be able to realize if the design of the test on the virtual campus is suitable for the correct development of the test or if, on the contrary, they should make changes and improvements for an optimal enhancement of the test. They will also be able to know if the teachers in charge are sufficiently prepared to guide an exam of such magnitude, taking into account aspects such as virtuality and what it entails, like the use of digital devices.

Face validity. Face validity, as described by Brown and Abeywickrama (2019), is a form of consequential validity that focuses on how test users perceive the appropriateness of the test (i.e., its design, preparation, subject matter, and administration). This type of evidence is relevant to our study, as it allows us to explore students' perceptions of whether the writing test is suitable for its intended purpose. By examining face validity, we can gain insights into how students view the test's overall appropriateness.

To sum up this section, we explored the notions of validity, context validity, consequential validity, and fairness to support our study by drawing on Shaw and Weir's (2007) framework. It is relevant to examine the context and consequential validity of language assessments, particularly that of the writing test for the BA in Foreign Languages, as standardized tests often benchmark language proficiency for purposes such as hiring, promotion, overseas assignments, graduation requirements, and admissions. For instance, university admissions decisions are made using the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS) Academic, while hiring and promotion decisions are made using the Test of English for International Communication (TOEIC). Standardized language assessments carry significant consequences and carry high stakes for test-takers.

In the context of the present study, the language proficiency test is high-stakes because it determines whether students have achieved the advanced level and met the graduation requirement. The writing component has the same weight as the other components, hence failing this component will affect students' overall score considerably.

Such a failure can bring negative consequences to students to obtain their degree. Therefore, it is important to find out whether the features and administration of the writing test are indicators of test-takers' language ability or proficiency to make valid decisions about test use and identify the consequences of using such a test. Identifying the relationship between students' perceptions of performance and their actual performance is also relevant to knowing how accurate their views of writing ability are and to inform teachers regarding classroom practice and test preparation. Collecting evidence of the appropriateness and fairness of the test provides a positive or negative washback effect to the academic program, mainly for teachers and students. The next section reviews selected empirical studies on EFL writing assessment validity and fairness.

2.3 Empirical research on EFL writing test validation

This section reviews previous research on L2 writing test validation. The scope of these studies addresses face validity, consequential validity, construct validity, comparative validity, washback effect, assessment design, and test-takers' performance. As discussed in the literature review, while most of the studies on consequential validity and face validity adopted a qualitative approach, those on performance and the washback effect employed a quantitative approach. Some other mixed-method studies explored the relationship between performance and stakeholders' perceptions; however, this relationship was examined through the lenses of content validity and comparative validity. The relationship between test-takers' performance and their perceptions regarding context validity represents a gap in the research presented in this review.

2.3.1 *International studies on writing assessment validity*

At an international level, the scope of studies on face validity and reliability (i.e., consistency of scorings) is considerably broad, regarding approach, methodology, participants, and purposes. For the present study, relevant empirical research was analyzed to obtain the state of the art regarding the viewpoints and experiences of test-takers, comprehensive evidence supporting the validity, the diverse contexts in which language assessments are used, and recommendations for enhancing the fairness, validity, and reliability of assessments, addressing test-taker needs and concerns, and improving overall assessment processes.

These studies include the work of Lee (2006), which examined writing assessment from a process-oriented approach. This qualitative research examined the promises and challenges of an ESL writing assessment course at the University of Illinois, including learner

engagement, feedback provision, and its impact on writing development. Through surveys, observation, and field diaries, the researcher recorded 100 students' experiences with the test. The study found that the anxiety level of the test-takers dropped significantly when they knew the nature of the questions. The study concluded that writing as a process reduced the negative impact of English proficiency tests.

Another study at the University of Illinois also considered face validity as a key element in writing assessment. Lee (2008) addressed the relationship between writers' perceptions and their performance on a field-specific writing test. Through two surveys, this mixed-methods study collected the perceptions of 124 graduate students from 20 different L1 backgrounds and with half of the participants coming from Asian countries. The study found that the theme of the exam was not related to the performance of the students. However, the researcher concluded that the students' familiarity with the instructions of the written test did seem to contribute positively to their performance. As a limitation, the researchers recognized the need to explore this relationship with other variables such as revision time or peer response.

From a quantitative approach, Cheng and DeLuca (2011) addressed face validity and test-takers' voices in order to gather further evidence for language assessment validation and its use. Through interviews and surveys, their study collected the insights from 59 test-takers at a large English-medium university in Asia. This approach allowed them to explore participants' experiences, perceptions, and suggestions regarding language assessment, validity and test-implementation. The results showed that unfamiliar test topics, inconsistency between test preparation and test, as well as physiological factors (e.g., age, gender, short- and long-term disabilities) related to test performance and its consequences undermined test-takers' performance and confidence in their own proficiency. The researchers concluded that there is an intertwined relationship between the aspects of the test design and the consequences for the test-takers.

From a qualitative approach, Allen (2016) investigated the washback effect of the IELTS test on the Japanese tertiary context. The study collected information from surveys, interviews and observations to explore 300 learners' experiences, motivations, and perceptions related to the test. It was found that most students showed little familiarity with the test, which resulted in a limited level of confidence. However, on the second attempt, students reported increased fluency, particularly in writing tasks, leading to a positive washback effect. The researchers concluded that the IELTS test generated a positive washback effect on productive skills when students had previous experience. Since the

findings were derived from participants' self-reports, the researchers suggested triangulating the information in future interventions for reliability.

A quantitative research study by Jin and Yan (2017) examined the construct validity of a high-stakes computer-based writing assessment. Their study involved 125 EFL learners as participants and explored the relationship between computer literacy and test performance. It was found that regardless of the digital literacy of participants, most of them perceived themselves as 'good writers' before the test. The authors concluded that digital literacy could become an even more common element in education if it were integrated into elementary school education. As a limitation, the researchers reported that the results could significantly differ among students from other universities in the country, so it was recommended to consider the socioeconomic background of each group.

Overall, previous studies examined the validity and fairness of language tests in evaluating individuals' language proficiency, while addressing the needs of diverse test-taker populations and providing valuable insights and recommendations for improving assessment practices. These studies explored the validation of test content to ensure it aligned with the intended language constructs of the assessment; the analysis of testing procedures to identify and reduce potential biases; and the exploration of test-taker characteristics such as cultural background, linguistic diversity, motivation, and test-taking strategies. However, there is a need to analyze the impact of a high-stakes writing assessment, specifically examining context validity with EFL pre-service language teachers.

2.3.2 Studies on writing assessment validity in Colombia

In Colombia, although previous studies have focused on writing instruction (e.g., Colmenares, 2013; Diaz, 2014; Vergara, 2016), few of them have addressed writing assessment (Guapacha & Chaves, 2023). For instance, Muñoz et al. (2006) conducted a qualitative study that focused on the design and validation of a writing assessment system in the Universidad EAFIT, in Medellín. The authors analyzed the validation of the construct, and the writing test prompts, as well as the design and validation of scoring instruments. The construct validation involved defining and aligning the construct, standards, and the nature of the assessment tasks. The validation of tasks was determined by analyzing task content against a set of criteria to design appropriate writing tasks. Additionally, a survey was conducted to gauge students' understanding of the prompts, with 137 adults and 126 adolescents participating. The results indicated that certain tasks in both the adult and adolescent programs required redesigning in terms of their alignment with course standards, organizational plan or presentation format, genre, and intended audience. The researchers

concluded that the lack of specification in these areas could lead to students misunderstanding the prompts.

As a continuation of the previous study, Muñoz et al. (2008) conducted a qualitative study at Universidad EAFIT to evaluate the impact of a writing assessment system on 27 adult students enrolled in an English program. The researchers measured the participants' improvement over time by analyzing the syntactic complexity of their texts using the average number of words per T-unit (e.g., sentences). The results of the study indicated some improvement in the students' writing skills as they received meaningful feedback. Furthermore, the study highlighted the importance of teachers providing better feedback to students by effectively utilizing the required assessment tools.

Mojica (2009) studied the perception of stakeholders in a language extension program in order to assess the relevance of the placement test. This quantitative study implemented surveys, interviews, field notes and retrospection to collect the responses of 115 students from the Universidad Nacional de Colombia aspiring to the English Philology Program. The study found that the effects of the test-taking experience, the decisions made based on the results obtained, the test implementation process, and the test itself had a negative impact on students. The study concluded that the test cannot be said to be valid when there is a large percentage of dissatisfied students. The researchers suggested adopting the practice of permanent evaluation of the different aspects of the curriculum present in these exams.

Tejada (2010) investigated the backwash effect of the ECAES high-stakes exam to know the legal status, the theoretical foundations, the objectives, the contents and the type of questions used in the test. This quantitative study used the results of the 2009 ECAES tests to carry out a critical analysis of the theoretical framework of the test. The results of this research mainly showed that the test reduced communicative competence to organizational competence and its three components: grammatical competence, textual competence and textual coherence. The study concluded that the English test generated negative effects in the participants regarding processes, products, teaching and learning of the foreign language. They acknowledged the need of a larger sample to perform more detailed analyses.

Overall, the aforementioned studies primarily focused on score and construct validity of proficiency tests, as well as the significance of placement tests for university admission. Consequential validity was also addressed to analyze the test's impact on students' performance and learning. Generally, these research studies underscored the importance of incorporating students' feedback into test design and curriculum enhancements, along with

providing process-oriented feedback to improve students' writing proficiency. However, the studies on test-takers' perceptions primarily aimed to analyze students' performance and learning based on the test itself, often neglecting contextual validity aspects such as administration settings. One motivation behind the present study is to ascertain the relationship between students' perceptions of contextual validity and their performance.

2.3.3 *Studies on writing assessment at Universidad del Valle*

At Universidad del Valle, initiatives on proficiency tests and washback effect were found. For instance, Kostina (2012) studied the English proficiency level of future language teachers to identify English proficiency levels. This quantitative descriptive study used the database of the results of the English proficiency exam that used to be applied every year to ninth semester students at the Universidad del Valle, from 2007 to 2012. The results showed that approximately half of the future teachers achieved a B2 level, the minimum required for the bachelor's degree, while the other half did not. It was concluded that it was necessary to implement actions aimed at strengthening English not only in communicative and evaluative practices in the classroom, but also in the implementation of curricular changes in the program. As a limitation, the author recognized that variables such as the perception of the students and how this affected their performance in the test were not analyzed.

In another study, Kostina et al. (2013) addressed the assessment practices in the English and French components of the BA in Foreign Languages. In order to analyze the assessment practices carried out by teachers in English and French courses, this qualitative study utilized a corpus of official documents, discussion sessions with professors and focus groups held with 73 students. The results showed that assessment practices were heterogeneous which was due not only to factors identified in literature but also due to important factors depending on situations of a public higher education institution in Colombia. The researchers concluded that it was important to take advantage of a combination of two main functions of assessment, formative and summative, and to know how summative assessment information may be used for formative purposes. The researchers suggested that the results were exploratory and proposed future studies to expand the conclusions.

Gómez (2016) addressed the proficiency exams in English and French of the morning and afternoon shifts at Universidad del Valle. This quantitative study used the results of the 2011 English proficiency exam applied to eighth-semester students, to make a comparison and contribute to the evaluation of the academic processes of the program. The results of the analysis of the collected information showed that the decisive factors in the performance of the students in the tests were the autonomous preparation and the constant work during the

academic training. The study concluded that it was necessary not only to reorient the role of teachers and students in terms of preparation for tests, but also to consider some modifications in the curriculum of the BA in Foreign Languages that respond to current needs in the optimal preparation of the students. The researcher recognized that this research was synchronous, that is, it only analyzed the results of the 2011 exams. In this sense, the changes that the results have undergone over the years were not considered.

Ramirez (2020) addressed the results of an English placement test to report the measurements of the difficulty index and the discrimination index in an English placement test designed by and for the Universidad del Valle. This quantitative study implemented the Moodle platform to analyze the responses of 815 students admitted to the first semester of various programs at the University. The results of the study suggested that the items found to be incorrect corresponded to questions whose instructions were ambiguous or whose answer options contained non-functional distractors. It was concluded that the test was functional and fulfilled its purpose of establishing the knowledge level of ESL and the placement of first semester students in the different levels offered by the University. The researcher acknowledged that the efficiency index on the distractors was not measured and that future studies should be proposed around the validity and reliability of the test.

Guapacha (2020) addressed the comparative validity of CB (Computer-based) and PB (Paper-based) writing tests and their effects on test-takers and raters to demonstrate whether the writing medium would make a difference in students' cognitive processes, performance and preferences for test mode, as well as in EFL raters' scores, perceptions of texts and preferences for test mode. This mixed-methods study implemented questionnaires, scoring and interviews to analyze the perception of 38 second-year EFL students of Universidad del Valle. The findings suggested that both writing media appeared to have positive and negative impacts on students' micro-planning, at the individual level, and that students interacted differently with the writing media. The researcher concluded that the writing medium needed to be considered in direct assessment of L2 writing and the writing medium affected EFL stakeholders in different ways. The researcher suggested that sample size was small so it did not represent the majority of the EFL university students in Colombia and other research contexts and could differ from other EFL learners, regarding background, level of education and age.

Previous studies on language assessment at Universidad del Valle have generally focused on assessment design, evaluation practices, test-takers' performance, and the integration of stakeholders' perspectives into the assessment process. Furthermore, some of

these studies (e.g., Gómez, 2016; Guapacha, 2020; Kostina, 2012) have specifically focused their analysis on the proficiency test within the BA in Foreign Languages, with test-takers being English pre-service teachers. This research background evidences the importance of test-takers' voices regarding assessment practices and design as well as curriculum enhancement. This type of research is relevant to language teacher education because they are in charge of training professionals who are able to design and perform language tests.

Nevertheless, there is a need for a study that focuses uniquely on writing assessment for pre-service teachers and that investigates the relationship between their performance and their perceptions of context and consequential validity through a framework intended for writing test validation. Nowadays, academic writing is assessed for different jobs and academic positions. In an attempt to offer a more comprehensive panorama of the intricate relationship between students' characteristics, test administration, curriculum, and performance, this study aims to fill this research gap.

2.4 Concluding remarks and gaps to fill

The studies previously reviewed contribute to our understanding of language assessment and highlight different aspects of the validity and consequences of assessing writing. Most of the analyzed studies employed either qualitative or quantitative methods exclusively, while a lower number if them employed a mixed-methods approach to examine the association between consequential validity and test-takers' performance (Muñoz et al., 2008; Guapacha, 2020; Lee, 2008). Furthermore, a significant portion of these studies focused on exploring the washback effect by analyzing participants' perceptions through test validation frameworks such as Messick's and Shaw and Weir's. Nevertheless, they tend to overlook the proposition of alternative approaches to improve current assessment methods with the aim of minimizing negative consequential validity. This research gap presents an opportunity to explore the educational consequences and impacts of the EFL proficiency test in a context as diverse as that of Universidad del Valle, thus furthering the understanding of the test's effectiveness and potential improvements.

Analyzing the writing component of the English language proficiency test at the Universidad del Valle promotes the enhancement of English writing skills in the academic program and the validity and fairness of the test. The test's design, parameters, and criteria can be updated through feedback provided to evaluators, teachers, and students. The next chapter describes the methodology employed in this study.

3 Research Methodology

3.1 Chapter overview

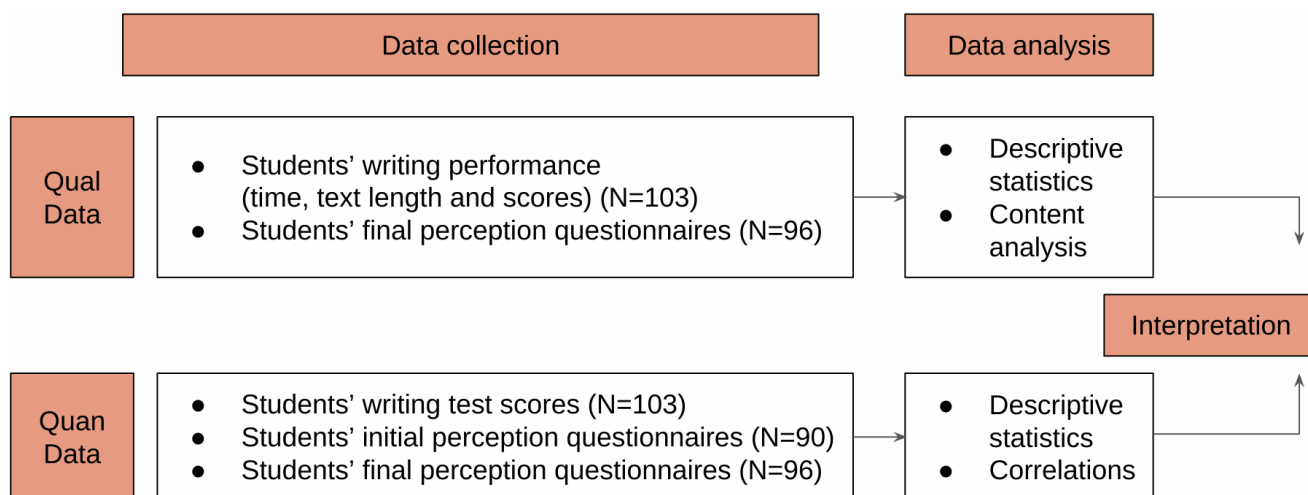
As seen in the literature review, previous studies used qualitative approaches to investigate students' perceptions of English proficiency tests, whereas other studies focused on students' performance using a quantitative approach where researchers influenced the study's design (e.g., design the test or questionnaire). Only a few have implemented a mixed-methods approach to investigating the impact of writing assessments on stakeholders, and none have implemented ex post facto research designs. The present study aims to fill this gap using an ex post facto research design to examine the context and consequential validity of the B.A writing test through the relationship between test-takers' perceptions and performance from the data already collected.

3.2 Research approach and design

An ex post facto research design is used for determining potential causes of events that have already occurred and cannot be planned or influenced by the researcher (Tavakoli, 2012). In an ex post facto research design, the researcher does not manipulate any variables because the event has already occurred or has been naturally determined by external circumstances. In the present study, we analyzed the data already collected by the English Department about students' perceptions of the writing test and performance, as well as their scores. We were granted access to the data through the English Department Coordinator (See Consent Form in [Appendix A](#)). In the qualitative part of this study, students' perceptions of the writing proficiency test regarding features and administration were analyzed, whereas the quantitative part identified relationships between students' perceptions of performance and their actual performance. Since the current study aims to examine previously collected data on test-takers' perceptions of a writing test in a real-world context and to establish links between these perceptions and performance, an ex post facto research design was suitable (See Figure 3.1).

Figure 3.1

Ex post facto research design



This research design is suitable and innovative in the context of the study because it used data and evidence collected from students in this program, which was analyzed to provide feedback to the program and teachers. This feedback might help improve the program and teachers' practices regarding writing teaching, learning and assessment. It might help to design valid and fair writing tests and prepare their students for future examinations. This design might also be helpful for other language teachers to examine the data collected from their students' perceptions and performance.

In *ex post facto* research, scholars gather information on the variables from surveys, questionnaires, and other sources and perform correlations or associations between variables (Tavakoli, 2012). Therefore, this study lied in the exploration and categorization of student perceptions through two questionnaires. For this, we analyzed the test-takers' initial and final perceptions of the English writing test and their performance. In this way, it could be found, for instance, if students' positive/high and negative/low perceptions of performance related to their actual performance. Investigating the correlation between perception and performance provides insights into the students' accurate view of their self-performance. This, in turn, informs the academic program to make decisions on test preparation. Thus, the triangulation of both methods and data sources (perceptions and scores) provides better interpretations of the results and complement each other by clarifying their findings (Creswell & Creswell, 2018; Ivankova & Greer, 2015).

3.3 Research site and participants

3.3.1 Research site

The present study was conducted in the BA in Foreign Languages at Universidad del Valle and examined the writing test, as part of the language proficiency test and as one of the graduation requirements of pre-service language teachers. The BA in Foreign Languages offers disciplinary knowledge in English and French, ensuring students develop a strong understanding of the languages themselves and knowledge of pedagogy, didactics and research to foster critical thinking and inquiry skills (Universidad del Valle, 2024).

According to Resolution 062, dated April 8, 2021, of the Academic Council of the Universidad del Valle, it is established that students in the BA in Foreign Languages (currently 3267) must demonstrate a C1 level of proficiency in both English and French, in accordance with the Common European Framework of Reference (CEFR), as specified in Resolution 18583, dated September 25, 2017, by the Colombian Ministry of Education (Ministerio de Educación Nacional (MEN), in Spanish). This resolution also states that the School of Language Sciences at the Universidad del Valle will verify this proficiency level using the results of the Saber Pro Tests or standardized exams. On the other hand, it is also stipulated that the students enrolled under the validity of Resolution 054 of May 24, 2001 (formerly 3263) are required to pass the proficiency exam in English and French with percentages corresponding to level B2.

For this study, we took as a basis the results of the English language proficiency test, administered to fourth-year students to assess their language proficiency (B2 and C1 levels, explained later). The English proficiency test is a graduation requirement for students of the BA in Foreign Languages (English and French) who have passed the seven levels of English offered in the degree. The test had previously been administered to students in person by the lecturers in the building of the School of Language Sciences. However, due to the Covid-19 pandemic in 2020, the in-person test was postponed and administered synchronously online in 2021 and 2022, with a low number of students taking the test in-person. The test was administered through Campus Virtual, the Universidad del Valle's Moodle platform.

As part of their preparation for the exam, the students attend an informative talk about the test delivered by the English language Department Coordinator. In that talk, students are informed about the regulations, the protocol and characteristics of the exam. They also receive some recommendations and strategies for preparing and taking the test. The students are also provided with a trial test on Campus Virtual and are granted the opportunity

to use it up to five times prior to the actual exam so that students can familiarize themselves with all the components of the test, including the writing tasks, the platform and test administration. This allows them to practice and become acquainted with the exam's format and content, thereby enhancing their readiness for the final assessment as well as the validity and fairness of the writing test.

The program also offers a course called Advanced English, which is offered to students from semesters 8th to 10th and seeks to enhance the linguistic and communicative elements that represent the advanced level of a proficient user in a foreign language (level C1). The course emphasizes the development of skills specific to the academic and professional field and familiarizes students with advanced-level English tests such as CAE, IELTS, TOEFL and DELNA. Therefore, students can practice writing strategies under exam conditions.

Furthermore, students receive the results of the English proficiency test and attend a meeting to analyze the global results, along with their initial and final perceptions about the test administration and performance. Additionally, general recommendations about the test are given for further improvement and test taking experiences, which are useful for those who failed and passed the test. All the previous procedures attest to the validity and fairness of the language test in this research site; however, it is necessary to confirm these claims through students' perceptions.

3.3.2 Participants

For the present study, the data on performance and perceptions belonged to 103 ninth-year students who took the English proficiency test in 2021 ($n = 49$) and 2022 ($n = 54$) and consented to use their data for research and academic purposes, as seen in the perception questionnaire (See [Appendix B](#)). These participants belonged to two cohorts or reformed programs: 3263 ($n = 49$) and 3267 ($n = 54$). The students belonging to the 3263-program had to demonstrate a B2 level. In writing, a B2 level indicates students' ability to prepare clear and detailed texts on a variety of topics related to their area of interest, synthesizing, and evaluating information and arguments from various sources (Council of Europe, 2020). In contrast, the students who belonged to the 3267-cohort had to demonstrate a C1 level. In writing, the C1 level indicates students' ability to adeptly employ the structure and conventions of different textual genres, skillfully adapting the tone, style, and register of their writing based on factors such as the audience, the type of text, and the topic (Council of Europe, 2020).

3.4 Data collection methods, procedures and analysis

To examine the context and consequential validity evidence of the writing proficiency test for the BA in Foreign Languages at Universidad del Valle, the present study analyzed students' performance results (scores, writing time and text length), students' perceptions of context and performance, and the test features. These data were gathered by the English Department of the program using the test results and two surveys as instruments. Table 3.1 provides an overview of the data collection methods and analyses in relation to the research questions.

Table 3.1

Overview of the research questions and data analysis

Research Question	Data Method	Data Analysis approach
RO1: What is the students' performance regarding writing time, text length and scores?	Time (minutes) Text length (Word counts) Scores (mean, standard deviation)	Qualitative (Descriptive statistics) Total participants and cohort analysis
RO2: What are the students' perceptions of context validity?	Final survey on test setting administration	Qualitative Content analysis and descriptive statistics
RO3: What are the students' perceptions of performance?	Initial and final surveys on self-assessment	Qualitative Descriptive statistics
RO4: What is the relationship between students' perceptions and performance?	Writing test scores and both initial and final surveys on perceptions of performance	Quantitative Descriptive statistics and correlations (Spearman's and Pearson's ranks)

3.4.1 The writing proficiency test

This section first describes the characteristics of the in-house English language test to continue with the description of the writing test and the data on students' performance and perceptions.

3.4.1.1 The in-house English Language Proficiency Test: Features.

The English language proficiency test assesses if the students have developed the necessary skills in the foreign language. The exam takes around 3 hours and 40 minutes, divided into two sessions. The first session assesses listening, reading, and writing skills. After a break of 90 minutes, the second session evaluates oral production. The test had always been given to students in person by lecturers at the School of Language Sciences

building. However, as mentioned before, the in-person exam has been postponed due to the COVID-19 pandemic in 2020.

The Language Proficiency Test allows undergraduate students, on the one hand, to familiarize themselves with this type of test they will encounter in their future academic and professional life. On the other hand, proficiency exams enable students to transcend the context of their study program subjects and assess their language competence in a holistic manner (Universidad del Valle, 2021). The English language test, designed by the School of Language Sciences for the BA in Foreign Languages, is based on broadly recognized standardized language tests, such as TOEFL iBT®, CAE, and IELTS, among others. Following current trends in international exams, the English Department updates the test's construct on a regular basis. Via the Exam Protocol, the exam structure, assessment rubrics, and succeeding percentages are shared at least two months before the exam's administration.

3.4.1.2 The in-house English language test: registration and administration.

To register for the exams, students are required to have attended and passed the English Composition course (level 7). Once registration is complete, students are granted access to a trial test, rubrics, and a test score calculator to their respective emails on the university's Moodle platform, Campus Virtual. This trial test consists of a mock test. Test-takers have two months to work on this practice alternative, providing ample time for preparation and practice. Additionally, a survey is conducted before registration to gather information about the students' physical contextual conditions regarding resources, time, and space. Furthermore, a Zoom meeting is held in which the coordinators of the English and French departments share the results of the survey, as well as the exam regulations, structure, requirements and important dates. All aspects are thoroughly explained. Here are the requirements to take the test online:

- Internet connection (carry-on data plans are not allowed)
- A computer with a camera, a microphone, and speakers (either in-built or external) (tablets, cell phones or other mobile devices are not allowed)
- A place with minimal interruptions during the test administration
- Remaining in front of the computer during the test administration.

Within a maximum period of two months after the test administration, the results are sent to the students' institutional emails. The results specify the test components, the

passing percentages, and an approval or non-approval status (Translated from Universidad del Valle, 2021).

3.4.1.3 Writing Test: Features and Administration.

Writing plays a significant role in students' performance, which might be influenced by contextual factors such as the mode of application (in-person or online). The writing test was designed and reviewed with expert lecturers and piloted and validated with 12 students. Nevertheless, it was relevant to collect the experiences of more individuals to identify if it is necessary to make adjustments to the test. This validation is relevant and will inform the context validity and consequential validity of the writing test in both modalities. Below, Table 3.2 specifies the writing test construct, structure and features of both writing tasks.

Table 3.2
Writing proficiency test construct and context features

	Writing Task 1	Writing Task 2
Construct of writing proficiency	To assess the candidate's ability to • demonstrate cognitive abilities (planning, monitoring and revising) and digital skills (typing skills, computer familiarity) to produce a text • produce non-specialized academic texts addressed to an academic audience within a limited time. • produce structured and coherent texts that reflect a good use and variety of vocabulary, register, punctuation, spelling, and grammatical structures. • produce an academic essay that states, explains, and develops the student's position on a familiar topic, providing specific details and examples. • Interpret, present clearly and correctly and comment on the most important points or trends contained in diagrams or other visual data on familiar or professional topics	
Response format	Opinion essay	Explanatory text
Topics and instruction	One of two prompts: 1. Expressing agreement or disagreement with the benefits of digital communication for communicating effectively. 2. Choosing the most suitable solution from three provided options to address a difficult problem in life. See Figure 3.2.	Test-takers were provided with a graph illustrating the shift in the frequency of daily activities from before to during the COVID-19 pandemic. See Figure 3.3.
Time	45 minutes suggested from a total of 90 minutes	
Text length	A minimum of 250 words	
Writing mode	Computer-based test	
Platform	Campus Virtual of the Universidad del Valle	
Additional resources	Trial test: up to 5 attempts until the day before the actual exam.	
Scores and levels	C2: 46-50 pts C1: 37-45 pts B2: 33-36 pts B1: 28-32 pts	

Figure 3.2

Instructions for the first writing task (a sample prompt form the trial test)

Personal essay – Stating an opinion

You have 45 minutes for this task. You should write at least 250 words (the size of the text box). Write an essay that states, explains, and supports your position on an issue. Support your position with details and examples. Choose **only one** of the two following topics or options, then write your text in the text box. **Copy and paste the option in your text:**

1) Do you agree or disagree with the following statement?
Email, text messages, and social media are improving our ability to communicate effectively.

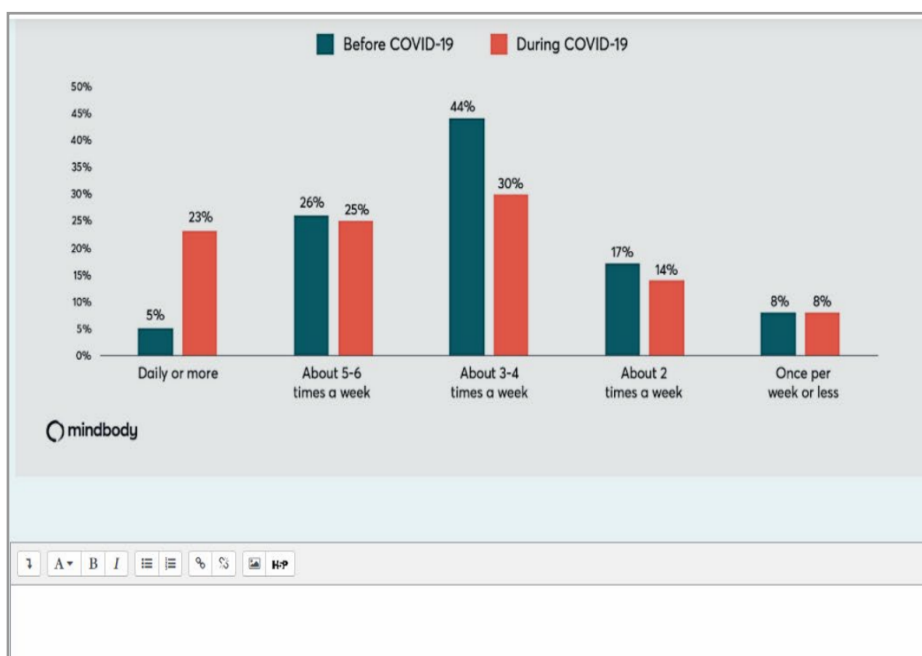
2) When you face a difficult problem in life, what do you feel is the best way to solve it?

- a) asking someone with more experience for advice about the problem
- b) finding information about the problem using the Internet
- c) taking a long time to think about the problem

Rich text editor toolbar: Bold, Italic, Underline, Bulleted List, Numbered List, Link, Unlink, Image, Print.

Figure 3.3

Instructions for the second writing task (a sample prompt form the trial test)



3.4.2 Students' writing performance data (test scores, time and length)

As detailed above, the writing test consists of two tasks with a total duration of 90 minutes, 45 minutes for each task, and a word count of a minimum of 250 words per text. Students' writing performance was assessed using an analytic scoring rubric which evaluates the components of content development, coherence, cohesion, style, and form (grammar, sentence structure, vocabulary, and spelling) (refer to [Appendix C](#)). The writing score was obtained from the average of all the component ratings awarded by two independent raters. All this information allowed us to determine whether the test-takers had enough time, produced enough words and passed the test. This information provides evidence of the context validity of the writing test regarding time, length and scores and helps us answer the research question: *What is the performance of the students in the writing test regarding writing time, text length, scores?* In this way, we can define if the proficiency test is fair and valid, or if, on the contrary, adjustments must be made to achieve it.

3.4.3 Students' perceptions data from a questionnaire

The students' perceptions data were collected through two surveys. The first survey was conducted before the exam, while the second was administered after the test. These surveys aimed to gather information regarding the students' perceptions of the test and their self-assessment of listening, speaking, reading, and writing skills before and after the test. For the purpose of this study, we focused on students' perceptions of the writing test. These perceptions provided insights into the validity and fairness of the writing test regarding context validity (features and administration) and how accurate their perceptions of performance were in regard to their actual performance.

The initial survey consisted of 10 questions distributed across seven sections, comprising five sections with closed-ended questions and two sections with open-ended questions (refer to [Appendix B](#), Table B1). Conversely, the subsequent survey included 15 questions divided into five sections, with three sections featuring closed-ended questions and two sections featuring open-ended questions (refer to [Appendix B](#), Table B2). As a survey is an instrument capable of offering both quantitative and qualitative descriptions of trends, attitudes, and opinions within the surveyed population (Creswell & Creswell, 2018), both surveys allowed us to collect test-takers' perceptions of context validity and performance and analyze the test impact on test-takers. Therefore, we could address our research questions: *What are the students' perceptions of the test regarding context validity?* and *What are the students' perceptions of performance before and after the test (consequential validity)?*

Additionally, by combining writing test scores with survey data, we analyzed the correlation between test-takers' perceptions of performance and their actual performance.

Consequently, we could answer the research question: *What is the relationship between their perceptions and performance in the writing test (consequential validity)?*

3.5 Data analyses

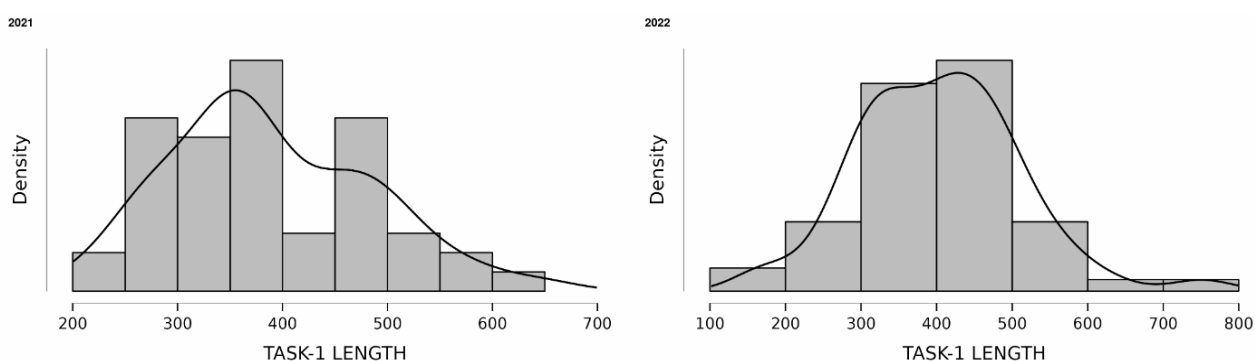
As the present ex post facto study is grounded in both qualitative and quantitative data, methods such as descriptive statistics, content analysis and correlations were employed for analyzing and interpreting the collected information. This section elucidates the procedures undertaken for both types of data.

3.5.1 Qualitative data analysis

RO1. In order to identify student's performance in the writing test regarding test scores, writing time and text length, the results of the writing test from the two cohorts were analyzed both separately and as a whole group of test-takers. The calculations of the maximum and minimum scores, measures of central tendency (mean and mode), as well as dispersion measures such as standard deviation were made using JASP¹ (a free software for quantitative analysis, <https://jasp-stats.org/>). Additionally, this software was used to graph the data. See a sample in Figure 3.4 for checking assumptions of normality.

Figure 3.4

Graph of text length for task 1 using JASP showing the normality of data assumption



RO2. To identify students' perceptions of context validity, the remarks provided by students in the second survey regarding the test administration were analyzed by using content analysis. It is a relevant method that extracts meaningful insights and patterns from the data, including techniques such as color coding (Krippendorff, 2009). We conducted inductive and deductive analysis. We first analyzed the data inductively through our lenses

¹ JASP (Jeffreys's Amazing Statistics Program) is a free and open-source statistical software package designed for both beginners and advanced users in the field of statistics. It provides a user-friendly interface with a wide range of statistical analyses and data visualization tools.

and then adjusted the categories to Shaw and Weir's (2007) framework of context validity. Consequently, through the use of labeling and color coding, it became possible to classify both the categories and subcategories, as well as the students' perceptions associated with each category (whether positive or negative). For instance, if a stakeholder mentioned that the proposed test time was insufficient, this could be interpreted as negative feedback for the timing category. The frequencies of each category were calculated for each cohort and perception (positive or negative) using Excel. Table 3.3 shows a sample of the data coding and analysis.

Table 3.3

Test impact classification by category (data analysis sample from Excel)

Students' comments	Perceptions	Categories
Ojalá poder presentar en dos días cada parte, porque siento que fue pesado la rutina en un solo día.	Negative	Time constraints
Se me borró toda la información de uno de los textos que realicé en el examen de escritura.	Negative	Physical conditions (Campus virtual)
Considero que se debe tener en cuenta este fallo en el campus virtual para la realización del siguiente examen de competencias [...]	Negative	Physical conditions (Campus virtual)
Referente a la escucha y la lectura creo que el examen pudo desempeñarse muy bien, todo fue muy ordenado y no se sintió tan diferente a la presencialidad [...]	Positive	Physical conditions (Place)
No tengo recomendaciones. Me pareció que el diseño del examen estuvo bien al igual que su aplicación. Todo estuvo muy bien organizado.	Positive	Uniformity of administration

RO3. To identify students' perceptions of performance before and after the test, the students' self-assessment remarks from both surveys were analyzed. The initial and final perceptions were compared with the final writing test scores. This comparison required assigning alignment values to each perception:

- Zero (0): when students' perception of performance is lower than the actual score
- One (1): when students' perception of performance is equal to the actual score
- Two (2): when students' perception of performance is higher than the actual score

The results were organized into categories based on low, aligned, and high perceptions, corresponding to alignment values of zero (0), one (1), and two (2), respectively.

Table 3.4 illustrates the comparison between students' perceptions of performance and final scores.

Table 3.4

Sample of alignment of students' perception and performance in Excel

Nº	Cohort	Final score	Initial perception	Alignment	Final perception	Alignment
1	2021	C2	C1	0	C1	0
2	2021	B1	C1	2	C1	2
3	2021	C1	C1	1	C1	1
4	2021	B1	B2	2	B2	2
5	2021	B2	C1	2	B2	1

3.5.2 Quantitative data analysis

RO4. In order to examine the relationship between test-takers' perceptions of performance and their actual performance, this study used correlations to identify whether the students' high or low perceptions of performance related to their high or lower levels of performance. Correlation analysis is used to describe the strength and direction of the linear relationship between two variables (Pallant 2016), in this case performance and perceptions. Scores were considered as quantifiable variables, whereas the cohorts and tasks were considered nominal variables.

A t-test for Pearson's r and Spearman's ρ was performed using JASP as both measures the strength and direction of association between two variables. They are important tools in statistical analysis, with Pearson's being more appropriate for linear relationships between continuous variables when data meets the assumptions. In the present study, assumptions for Pearson's correlations were checked and met. The p -value of Shapiro-Wilk was not significant in three out four datasets (Task 1 Cohort 2021, $p = 0.833$; Cohort 2022 = $p = 0.070$ and Task 2 Cohort 2021, $p = 0.096$; Cohort 2022, $p = 0.034$). Results were also confirmed using the non-parametric Spearman's Rho test.

For performing the correlations, data from scores and perceptions were used and related per cohort. We had to assign numerical values to the language proficiency levels across the test scores and perceptions. For instance, C2 was assigned the value 4; C1 was 3; B2 was 2; B1 was 1, and A2 and A1 were 0, as seen in Figure 3.5.

Figure 3.5

Values for performing correlations

 Cohort	 Final score Correl	 Initial perception Correl	 Final perception correlation
2021	4	3	3
2021	1	3	3
2021	3	3	3
2021	1	2	2
2021	2	3	2
2021	2	2	2
2021	3	4	3
2021	3	1	2
2021	3	4	2
2021	3	4	2
2021	1	1	1
2021	3	3	2
2021	3	4	2

It is important to know how to interpret the results of correlations. A p value higher or lower than alpha 0.5 indicates statistically significant or non-significant correlations, respectively. Then, we need to interpret the strength and direction of the correlation. Strong linear relationships are shown by high positive r values, and strong negative linear relationships by high negative r values. If the value is around 0, there are no linear relationships between the variables. After that, it is necessary to interpret the ranges of the correlations. This can range from -1 to 1 . This value will indicate the strength of the relationship between your two variables. A correlation of 0 indicates no relationship at all, a correlation of 1 indicates a perfect positive correlation, and a value of -1 indicates a perfect negative correlation (Pallant, 2016) (see Table 3.5).

Table 3.5
Pearson's correlation range

Correlation r value	
Interval correlation	Level of correlation
0.80-1.000	Very Strong
0.60-0.799	Strong
0.40-0.599	Strong enough
0.20-0.399	Weak
0.0-0.199	Very weak

Through these correlations it was possible to identify whether students' initial and final self-perceptions of performance had a strong/moderate/weak and negative/positive relationship with test scores. For instance, a strong positive correlation indicates that students' perceptions aligned with their actual performance.

3.6 Validity and reliability

Research validity, understood as the trustworthiness, credibility, and accuracy of findings, is concerned with ensuring that the data collected, and the interpretations made accurately reflect the participants' perspectives and the phenomenon being studied (Creswell & Creswell, 2018). In other words, research validity is not focused on statistical generalization but on establishing the quality and soundness of the research process and the resulting insights, avoiding any type of bias either from the researchers or from the data itself (Creswell & Creswell, 2018). Reliability has to do with the consistency in data coding and interpretations (Creswell & Creswell, 2018). To guarantee research validity and reliability in the interpretation of the data, this study implemented a set of strategies.

Adoption of appropriate research methods. Data were analyzed following the most appropriate methodology and design for the case. This means that the students' perceptions regarding context and consequential validity were interpreted and classified through a qualitative approach, while the correlation between students' perceptions and performance was addressed by using a quantitative approach.

Triangulation. Several sources of data were analyzed to inform this study. Data from performance and perceptions were suitable to answer the research questions.

Debriefing sessions between researchers. Each part of the analysis of the results was socialized by the researchers before proceeding to the next stage. This allowed for closer control of the study's progress and timely methodological adjustments if required.

Intercoding agreement. The nature of each piece of data and of each concept was analyzed and evaluated by the researchers before being incorporated into the study basis. This ensured that the theoretical framework and the categories for writing assessment were thoroughly analyzed and agreed upon by all researchers to avoid methodological gaps. Each qualitative analysis was first performed individually by each researcher before being discussed and, consequently, carried out systematically. The color coding for the interpretation of the categories, as well as the nature of the washback effect, was completed individually before being socialized. This approach avoided researcher bias while considering each researcher's perspective. Initially, both researchers used their own labeling codes to classify different samples of the data based on students' perceptions. Subsequently, both coding systems were compared in the context of the students' comments and the framework proposed by Shaw and Weir (2007). Similar categories were merged, derivative categories were removed and redefined, and pertinent new categories were incorporated into the

framework. For the students' performance and correlations, data were analyzed and interpreted simultaneously by researchers using the same units and labels from the beginning. The resulting categories and data from performance underwent revision by the research supervisor as a third reviewer.

3.7 Ethical considerations

The researchers had an ethical obligation to respect the rights, needs, values, and desires of the informants. This included ensuring informed consent, confidentiality, and voluntary participation (Creswell & Creswell, 2018). Respecting informants' rights involved obtaining their consent before involvement in the study, protecting their privacy, and considering their well-being throughout the research process. The values and desires of the informants were acknowledged and respected, ensuring their autonomy and dignity.

To safeguard personal information and maintain the integrity of the investigation, the following measures were implemented:

- Avoid sharing test and participant information with individuals not involved in the research.
- Refrain from contacting participants for purposes unrelated to this research.
- Do not publish study results that reveal participants' personal information.
- Use codes to anonymize test information before publication if necessary.
- Ensure that the analysis, interpretation, or conclusions drawn from test data do not include value judgments on participants' performance and remarks.
- Maintain the reliability of analyzed information by not modifying test results or arbitrarily altering participants' responses.

3.8 Concluding remarks

This chapter detailed the procedures used to examine the context and consequential validity of the writing test for pre-service language teachers in the BA in Foreign Languages at Universidad del Valle. The methodological contribution and innovation of this research are noteworthy for two reasons. Firstly, the *ex post facto* research design represents a new approach used in our research context. This approach allows for a comprehensive retrospective analysis of real-world data without manipulating variables, providing a more authentic reflection of the factors influencing students' performance and perceptions. Secondly, the analysis of students' perceptions and performance offers important insights.

By correlating how students view the writing test and how it might affect their actual performance, the study provides valuable feedback for enhancing teaching practices and assessment methods. This becomes a continuous improvement cycle that benefits current and future students, professors and evaluators.

Overall, this chapter outlines a strong methodological framework that addresses the research questions while contributing significantly to the field of language proficiency assessment. The innovative data analysis techniques and the suggested program improvement strategies, presented later in the thesis, have the potential to enhance the teaching and assessment of writing skills in foreign language programs, ultimately better preparing students for their academic and professional futures.

4 Results

4.1 Chapter overview

The present study aimed to investigate the context and consequential validity of the writing proficiency test of the B.A. in Foreign Languages at Universidad del Valle, through four research questions. In this chapter, we present the results per research question.

- **RQ1.** What is the performance of students in the writing test concerning factors such as time, text length, and scores?
- **RQ2.** What are the students' perceptions of the test regarding context validity?
- **RQ3.** What are the students' perceptions of performance before and after the test?
- **RQ4.** What is the relationship between their perceptions and performance in the writing test?

4.2 RQ1. What is the performance of the students in the writing test concerning text length, time, and scores?

4.2.1 Text length

The students' writing performance was assessed through two academic writing tasks. Task 1 was a personal opinion essay on a familiar topic, while Task 2 was an explanatory text based on statistical data from a graph. Regarding the length, the minimum suggested word count for each writing task was 250 words. Table 4.1 presents the results of both tasks word count across cohorts 2021 and 2022.

Table 4.1

Descriptive statistics for text length (word counts) across cohorts (2021 and 2022)

	TASK-1 TEXT LENGTH		TASK-2 TEXT LENGTH	
	2021	2022	2021	2022
Sample	49	54	49	54
Mean	388	403	355	348
Std. Deviation	97	107	113	160
Shapiro-Wilk	.973	.976	.969	.899
P-value of Shapiro-Wilk	.306	.349	.215	< .001
Minimum	204	174	119	52
Maximum	644	751	670	1055

As shown in Table 4.1, on average, both cohorts achieved the minimum word count required in each writing task. The students produced at least 100 words more than required. As seen in the maximum and minimum values, there were students who exceeded the text length limit by more than 100 words. However, some students could not meet the text length criteria. The standard deviation shows that Task 2 in 2022 had the highest dispersion ($SD = 160$), indicating variability in text length. As seen in the average, most students were able to produce more than 250 words in each writing task and the word limit requirement could be met, although some students could not achieve it. The fact that most students were able to meet the text length requirement provides positive insights into the test's validity and fairness regarding writing time. It could also be affirmed that the task topics are apparently appropriate for the students to produce lengthy texts. Difficult topics can affect students' writing fluency and cognitive process (planning for an answer thus taking time). Nevertheless, as seen in the minimum volume, there were students who struggled to produce lengthy texts probably due to factors such as the topic, typing ability and lack of time management. Hence, it is necessary to collect students' perceptions of the test to identify these potential factors affecting their text length performance.

4.2.2 Writing time

Regarding writing time (see Table 4.2), the students were given 90 minutes to perform two writing tasks. Each task had to be completed within 45 minutes. However, each test-taker could manage their own time; therefore, it was not possible to split the time between Task 1 and Task 2.

Table 4.2

Descriptive statistics for writing time (minutes) across cohorts (2021 and 2022)

	Writing time (mins)	
	2021	2022
Sample	49	54
Mean	87	88
Std. Deviation	5.7	3.1
Shapiro-Wilk	0.600	0.7
P-value of Shapiro-Wilk	< .001	< .001
Minimum	59	77
Maximum	90	90

According to the results in Table 4.2, both cohorts completed the writing test around two minutes below the maximum time given (90 minutes). Some students took less time to complete the test (around 59 and 77 minutes), whereas others employed all the time given. Nevertheless, it is interesting to note that students took longer than the time given by recognized standardized language tests (e.g., IELTS (60 minutes) and TOEFL iBT (30 minutes)). This indicates that the writing test is valid and fair regarding writing time because the time is appropriate and enough for the students to produce two academic lengthy texts. Also, the fact that these students take nearly 90 minutes to complete two writing tasks indicates that students require more training in time management and planning. It could also be that the complexity of the programs' writing test requires students to take longer time.

4.2.3 Writing test scores

Regarding students' scores, the ratings ranged between 1 (poor) to 5 (qualified), with 1 representing the lowest performance and 5 representing the highest. In order to interpret the scores, Table 4.3 shows the language levels and their corresponding scores, where students had to obtain the C1 level (3.7 - 4.5 points) or above to pass the writing test.

Table 4.3

Equivalence between language levels and ratings

Writing test scores and proficiency levels				
Less than B1	B1	B2	C1	C2
< 2.8 points	2.8 - 3.1 points	3.2 - 3.6 points	3.7 - 4.5 points	≥ 4.6 points

Table 4.4 shows the descriptive statistics for writing scores across tasks and cohorts.

Table 4.4

Descriptive statistics for writing scores across tasks and cohorts

	TASK-1 SCORE		TASK-2 SCORE	
	2021	2022	2021	2022
Sample	49	54	49	54
Mean	3.6	3.9	3.5	3.4
Std. Deviation	0.644	0.667	0.637	0.853
Shapiro-Wilk	0.986	0.960	0.960	0.953
P-value of Shapiro-Wilk	0.833	0.070	0.096	0.034
Minimum	2.2	1.9	2.0	1.2

Maximum	4.9	5.0	4.7	4.9
---------	-----	-----	-----	-----

As shown in Table 4.4, the students' overall performance was classified into B2 level (3.2 - 3.6 points), as seen in Task 1 (cohort 2021) ($n = 49$, $M = 3.6$, $SD = 0.64$) and Task 2 (cohort 2021) ($n = 49$, $M = 3.5$, $SD = 0.64$) and (cohort 2022) ($n = 54$, $M = 3.4$, $SD = 0.85$). Task 1 for cohort 2022 showed a C1 level ($n = 54$, $M = 3.9$, $SD = 0.67$). In general, students' scores in Task 1 were higher than their scores in Task 2 in both cohorts.

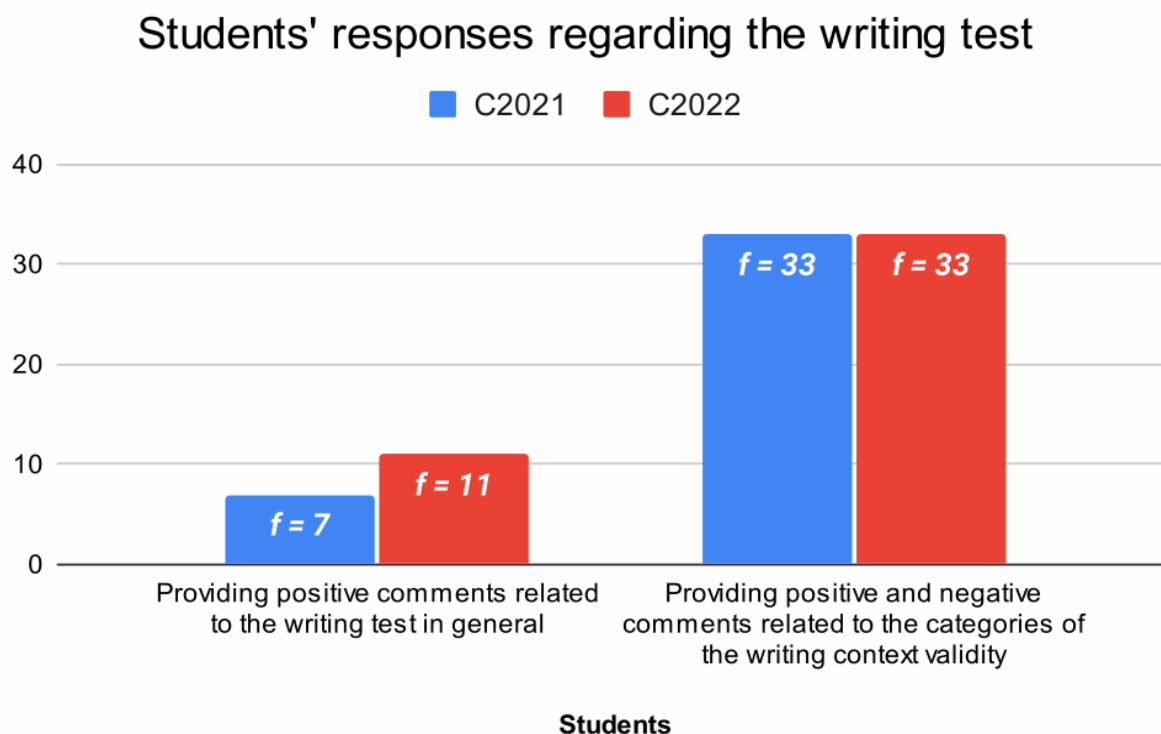
These results indicate that students' writing proficiency is below the expected C1 level but not far from the target, as scores of 3.5 and 3.6 can be considered upper B2 level. Additionally, it can be said that test-takers perform better in personal opinion essays (as required in Task 1) than in explanatory essays. This suggests that students need to improve their writing skills and get familiar with Task 2 about interpreting and reporting data from graphs. A first implication is that more training and familiarity with the test is needed to improve students' results. It is also possible that Task 2 is more challenging, and the sequence of the tasks may impact test-takers' performance.

4.3 RQ2. What are the students' perceptions of the writing test regarding context validity?

Students' perceptions of the test were collected through the final questionnaire. The students' answers were classified as positive and negative perceptions of the writing test. Of the total of test-takers ($N = 103$), 96 answered the questionnaire: 66 (64%) provided positive and negative responses related to specific categories of the context validity and test-takers' characteristics, and 18 (17%) provided positive perceptions of the writing test in general. Thus, with 81% of students reporting on their perceptions about the writing test, we can make claims about the validity and fairness of the test from students' perceptions. Figure 4.1 summarizes the students' responses from the final perception questionnaire across cohorts. It can be seen that both cohorts engaged in their responses in similar trends.

Figure 4.1

Students' perceptions of the writing test regarding context validity across cohorts



To examine the context and consequential validity evidence through students' perceptions of the test, we adopted Shaw and Weir's (2007) taxonomy of context validity evidence (refer to Figure 2.1 in Chapter 2). Table 4.5 outlines the students' perceptions of the writing test, organized according to the categories and subcategories proposed by the framework. Note that Table 4.5 presents the frequencies (*f*) of students' perceptions, rather than the number of students per category because a single student's comment may contribute to multiple categories, but each perception can be exclusively positive or negative.

Table 4.5

Frequencies of students' positive and negative perceptions of the test by categories

Category	Subcategory	Students' specific answers				Subtotals	
		2021 (n = 33)		2022 (n = 33)			
		Positive	Negative	Positive	Negative	Positive	Negative
		f	f	f	f	f	f

Test-takers' psychological characteristics	Stress	0	0	1	3	1	3
	Nervousness	0	0	0	3	0	3
	Subtotal	0	0	1	6	1	6
Test-takers' experiential characteristics	Typing skills	0	1	0	1	0	2
	Test preparation	0	6	0	1	0	7
	Subtotal	0	7	0	2	0	9
Context validity Setting: Task	Response format	0	1	2	5	2	6
	Purpose	0	1	0	0	0	1
	Knowledge of criteria	0	2	1	0	1	2
	Weighting	0	0	1	0	1	0
	Text length	0	0	0	0	0	0
	Time constraints	0	2	4	16	4	18
	Writer-reader relationship	0	0	0	0	0	0
	Subtotal	0	6	8	21	8	27
Context validity Setting: Administration	Physical conditions	3	14	4	1	7	15
	Uniformity of administration	3	3	6	1	9	4
	Subtotal	6	17	10	2	16	19
Context validity: Linguistic demands	Lexical resources	0	0	0	1	0	1
	Structural resources	0	0	0	0	0	0
	Discourse mode	0	0	0	0	0	0
	Functional resources	0	0	0	0	0	0
	Content knowledge (input and output)	0	3	0	3	0	6
	Subtotal	0	3	0	4	0	7
		Total frequencies				25	68

As shown in Table 4.5, the students reported more negative experiences or perceptions regarding the context validity – task setting ($f = 27$) and context validity – administration setting ($f = 19$). To a much lesser extent, context validity – linguistic demands ($f = 7$) were perceived as negative by fewer students. Noticeably, other perceptions were not related to the context validity of the test but with test-takers' characteristics, which are external to the test. There were more negative than positive perceptions of students associated with test-takers' experiential ($f = 9$) and psychological characteristics ($f = 6$).

Within the task setting category, most negative perceptions were associated with time constraints ($f = 18$), followed by response format ($f = 6$). Fewer students had negative perceptions regarding knowledge of criteria ($f = 2$) and purpose ($f = 1$). There were no perception answers regarding text length, weighting and writer-reader relationship.

Within the administration setting category, the difference between positive ($f = 16$) and negative ($f = 19$) perceptions was not so notable. While most of the positive perceptions were associated with the uniformity of administration ($f = 9$), most of the negative perceptions were related to physical conditions ($f = 15$).

Within the test-takers' characteristics category (factors external to the test), nervousness ($f = 3$) and stress ($f = 3$) appeared as psychological factors affecting a small

number of students' performance. However, concerning test-takers' experiential characteristics, most negative responses were associated with test preparation and familiarity ($f = 7$) and typing skills ($f = 2$).

The next sections provide some samples of students' negative and positive experiences and perceptions associated with the context validity of the test and test-takers' characteristics. These findings provide insights to discuss and conclude about the positive and negative consequences of the writing test on students' performance in the discussion and conclusion chapters.

4.3.1 Test-takers' negative perceptions regarding the test task setting

Regarding context validity – task setting, which relates to the requirements, instructions, and parameters within which the writing task should be performed, most negative perceptions were associated with time constraints ($f = 18$), followed to a lesser extent by Response format ($f = 6$). That is, some students had negative experiences with the time allotted (90 minutes) and with the type of writing task proposed, which were academic essays interpreting data.

- *Time constraints*

AC2022V: “Me parece que la prueba fue asertiva en la mayoría de los componentes, sin embargo, creo que para la parte escrita se debe tener en cuenta que todos tenemos procesos de escritura diferentes, que unos tomarán más tiempo que otros, por ejemplo. Escribir rápido no quiere decir más habilidad de escritura y tomar más tiempo tampoco indica falta de esta.”

JDP2022V: “El tiempo de las pruebas de escritura y lectura no fue suficiente.”

- *Response format*

AFS2022V: “en la organización de la parte escrita debió primero estar el análisis de datos, teniendo en cuenta que es más rápido de realizar [...]”

AH2022V: “En la parte de escritura se me hizo muy raro la parte de las gráficas, no sé si lo hice bien realmente, ya que se me pedía un ensayo académico, pero en las reuniones habían hablado de que era más de un informe interpretativo de las gráficas. No sé qué tan bien me fue en ese punto.”

These findings suggest a lack of students' familiarity and test preparation in both time management and data analysis skills. As shown in the findings regarding students' writing time (Section 4.2.2), the time was enough to complete both tasks; therefore, students' time-

related issues could be attributed to challenges in time management or curricular test preparation rather than an insufficient test duration. In fact, the in-house writing test allots extra time for students compared to other standardized writing tests. In the same way, these findings evidence a lack of students' preparation with explanatory texts, that could suggest a need for students to enhance their data interpretation skills, as required in Task 2, and to better prepare for discourse genres.

4.3.2 Test-takers' negative perceptions regarding the test administration setting

Regarding context validity – administration setting, which relates to the organization and management of the environment and the conditions under which the writing tests takes place, most negative experiences and perceptions were associated with the physical conditions ($f = 15$). Students reported negative experiences related to technological resources ($f = 9$), such as Internet connection and Campus Virtual, as well as to location ($f = 4$) and comfort ($f = 2$). In that regard, some test-takers mentioned that:

- *Physical conditions – Technological resources*

NVM2021V: “Creo que la única experiencia negativa que tuve fue con la conexión de internet, pero se puede solucionar rápidamente.”

CRG2021V: “En una ocasión mi conexión a internet fue inestable y el campus no me dejó continuar. Y el espacio en el que hice el examen no fue del todo cómodo, había mucho ruido y distracciones por lo que el tiempo no me fue suficiente.”

DCE2021V: “Se me borró toda la información de uno de los textos que realicé en el examen de escritura.”

2021: “Considero que se debe tener en cuenta este fallo en el campus virtual para la realización del siguiente examen de competencias, pues hubo varios estudiantes que se les borró la información de la parte escrita al igual que a mí.”

- *Physical conditions – Location*

YAR2021V: “[...] En cuanto a la escritura y el habla sí sentí que los resultados hubiesen sido mejores de haber estado en la presencialidad.”

- *Physical conditions – Comfort*

DXQ2021V: “Debido a ruidos externos en mi espacio físico, tuve inconveniente con la concentración.”

2021: “La sección de lectura y escritura virtualmente deja muy exhausto al estudiante.”

To a lesser extent, some students reported negative experiences regarding the uniformity of administration ($f = 4$), which relates to the consistency of the procedures, instructions, and guidelines when conducting the test. In that regard, some test-takers mentioned that:

- *Uniformity of administration*

GPS2021V: “Tal vez preparar hacer una capacitación previa para los profesores a cargo del examen (los que vigilaban) para saber manejar mejor el sistema.”

These findings suggest the need to revise issues related to technical resources such as the Internet connection to provide a smooth technical-related experience during the test. It is also necessary to revise issues related to the Moodle platform, Campus Virtual, as a technological resource for administering the test, and a suitable location for test administration, as some students preferred to take the test in-person and not online. However, it is important to note that most negative perceptions of these categories came from test-takers of 2021 cohort, which faced to a greater extent the challenges due to the COVID-19 pandemic.

Although context validity – administration setting arose as one of the categories perceived most negatively by students, it could be claimed that the validity and fairness of the writing test was not greatly impacted by the features of the test itself but to external factors, such as the pandemic. The negative perceptions were broadly distributed among the subcategories and were primarily due to the specific incidents related to the COVID-19 pandemic rather than systematic issues during the test administration. Although most students did not have negative experiences with the test features, it is important to consider the opinions of those who were affected in order to take action. For instance, test procedures and information about the demands and conditions of the test could improve and test preparation for students and staff should be provided.

4.3.3 Test-takers' negative perceptions regarding linguistic demands

To a much lesser extent, test-takers also reported negative experiences regarding context validity – linguistic demands, which relates to the ability to understand and respond to prompts appropriately. However, within this category, only the negative perceptions related to content knowledge ($f = 6$) were notable. Content knowledge evaluates the writer's

subject-matter or topic knowledge. This includes the particular topic or issue that the writing assignment is meant to assess. Some students mentioned that:

- *Content knowledge (task input/output)*

MAC2022V: “Desde el comienzo sabía que mi punto débil sería sobre todo la escritura, así que no me sorprende que haya sido lo peor para mí.”

BC2021V: “La parte de escritura fue diseñada para molestar más que para probar la suficiencia en la escritura bien. El segundo tipo de texto era muy confuso de escribir y pedían cosas muy diferentes dentro de este. Deberían hacer una sola consigna de escritura.”

These findings suggest that students' content knowledge issues may be related to the type of task and response format, as data interpretation posed a challenge for test-takers during the writing test. A notable aspect that we identified from the students' negative perceptions with this regard was the difficulty in reading the graphic associated with Task 2, which illustrated the shift in the frequency of daily activities from before to during the COVID-19 pandemic. This issue could be due to the lack of input students had about the graph topic which affected their output, as well as to insufficient students' preparation in interpreting data and writing explanatory texts. However, there is not enough evidence to imply the lack of validity of the task based on this aspect and a few students' perceptions.

4.3.4 Test-takers' negative perceptions and experiences regarding psychological and experiential characteristics

Apart from the features of the writing tasks, the students reported negative perceptions related to test-takers' psychological and experiential characteristics. Test-takers' psychological and experiential characteristics are defined as the mental, emotional, and personal background factors that influence how test-takers perceive, approach, and perform the writing test. These factors are not part of the test but can be triggered by it. Regarding psychological characteristics, students reported negative experiences associated with psychological factors, such as stress and nervousness during the writing test:

- *Stress*

DCE2022V: “El contexto del examen en general me provocó bastante estrés debido a la importancia que este tiene dentro de la carrera.”

- *Nervousness*

DHP2022V: “Al momento de hacer la parte de escritura sentí muchos nervios y me quedé en blanco en ciertos momentos haciendo que me tome más tiempo del debido.”

SM2022V: “Realizando este examen me bloqueé mucho, considero que estuve muy nerviosa y hubo momentos que leía y leía sin sentido.”

Similarly, students reported negative experiences associated with experiential characteristics during the writing test. These negative perceptions were associated with students’ test preparation and familiarity ($f = 7$) and typing skills ($f = 2$). Some negative experiences about test preparation and familiarity were related to students’ lack of experience with the trial test, the Campus Virtual platform, and the test itself. In this regard, some test-takers mentioned that:

- *Test preparation and familiarity*

LJH2022V: “Debí hacer el trial; sin embargo, de modo general siento que me esperaba que fuese más difícil [...]”

VT02021: “Que hagan ingresar al campus un día antes de el examen.”

ALM2021V: “Tener preparación previa desde al menos un semestre antes de la prueba ya que con los exámenes llevados a cabo en anteriores semestres se hacían varios simulacros que nos entrenaban y los resultados eran mucho mejores.”

- *Typing skills*

AC2022V: “Me parece que la prueba fue asertiva en la mayoría de los componentes, sin embargo, creo que para la parte escrita se debe tener en cuenta que todos tenemos procesos de escritura diferentes, que unos tomarán más tiempo que otros, por ejemplo. Escribir rápido no quiere decir más habilidad de escritura y tomar más tiempo tampoco indica falta de esta.”

These findings indicate a lack of test preparation and familiarity, as reported by some students. Since test-takers facing psychological (6%) and experiential (9%) challenges do not represent the majority of the sample, it can be implied that these negative factors are not directly associated with the test features, and they might stem from students’ lack of preparation or even challenges related to the pandemic. However, these findings also highlight the importance of continuing student training prior to the test by offering the Advanced English course, the trial test, and protocol meetings.

4.3.5 Test-takers' positive perceptions of the writing test and themselves

Although students' perceptions were mostly negative, as seen in the frequencies of negative reports ($f = 68$), compared to the positive reports ($f = 25$), it is worth highlighting that there were significant positive students' perceptions of the test, indicating that not all students had a negative experience or impact during the test. Interestingly, the positive aspects were associated with test-taker's context validity: setting of the task and setting of the administration, which were negatively perceived by other students. Students perceived most positively the uniformity of administration ($f = 9$), followed by Campus Virtual as a technological resource ($f = 6$), and time constraints ($f = 4$), as seen below in some quotes:

- *Uniformity of administration*

SC2022V: "Sentí que la prueba era acorde al nivel de exigencia."

LCL2021V: "No tengo recomendaciones. Me pareció que el diseño del examen estuvo bien al igual que su aplicación. Todo estuvo muy bien organizado."

- *Technological resources – Campus Virtual*

JMC2022V: "El Campus Virtual fue una muy acertada herramienta para la realización de esta prueba; funcionó óptimamente y, en los momentos de cierre por límite de tiempo, guardó acertadamente nuestras respuestas [...]".

- *Time constraints*

CF2022V: "En general, me parece que tuve una buena experiencia en el examen. El tiempo me parece adecuado, y el ambiente es adecuado para concentrarse."

In terms of uniformity of administration, these findings suggest that for the test-takers that had a positive experience with the writing test, it was designed and administered according to the expected proficiency level for test-takers; the Campus Virtual platform served as an appropriate and functional tool, and that the allocated time for completing both writing tasks was sufficient. These findings underscore the validity of the writing test in terms of uniformity of administration, the utility of the Campus Virtual platform, and the adequacy of writing time for a group of test-takers.

To sum up, the overall findings on students' perceptions of the writing test regarding context validity suggest a higher frequency of negative experiences related to some features of the writing test. Although these frequencies do not represent the voices of all or the majority of students, it is important to take their views into consideration for test revision and students' preparation. The most negative aspects of the writing test were associated with the

task setting and administration setting categories. Specifically, time constraints and response format prompted negative experiences among some students. Similarly, a group of students had negative experiences related to physical conditions such as Internet connectivity, Campus Virtual and location.

The findings indicate that, although time constraints were predominantly viewed negatively, these difficulties were closely associated with students' lack of test familiarity and prior experiences. Similarly, students' difficulties regarding response format were linked to inadequate preparation in data interpretation and producing explanatory texts. Additionally, negative experiences reported by test-takers regarding technological resources and locations were mainly associated with the COVID-19 pandemic challenges and did not suggest general problems in the writing test administration.

To a lesser extent, students reported negative experiences regarding linguistic demands, content knowledge, and test-takers' psychological and experiential characteristics. The findings suggest that negative perceptions of content knowledge were associated with students' insufficient knowledge and preparation for data interpretation. Additionally, negative perceptions related to psychological and experiential factors were associated with students' lack of preparation and did not indicate general issues with the trial test, the informative meetings or the test itself.

4.4 RQ3. What are the students' perceptions of performance before and after the test?

Students' perceptions of performance before and after taking the test were investigated to understand how test-takers perceive the test as pertinent and impactful to their language learning objectives and daily communication needs. Students self-rated their performance using the numerical scale and criteria used to assess students and classify them into proficiency levels. We analyzed the data of the 79 students who answered both questionnaires to make more accurate comparisons and relationships. A relationship between students' perceptions of performance and their actual performance provides insights into the consequential validity of the test (positive or negative impact on students' self-perceptions and score alignment). These perceptions were gathered through two perceptions questionnaires; one questionnaire before taking the test, and one final questionnaire after taking it and without having received their test scores, so as not to bias their perceptions and responses. The findings about students' perceptions of their performance before and after taking the proficiency test are presented in Table 4.6.

Table 4.6*Students' perceptions of self-performance before and after the test.*

	ALL STUDENTS (N = 79)		2021 (n = 35)		2022 (n = 44)	
PERCEPTIONS OF PERFORMANCE	Before	After	Before	After	Before	After
Higher than the actual score	26	18	14	9	12	9
Lower than the actual score	22	32	12	16	10	15
Alignment between perception and score	31	29	9	10	22	20

In Table 4.6, we can interpret the results from two perspectives. First, we can see that by taking all students together and per cohort, the results about the alignment between perceptions and scores are higher than the results of the higher and lower perceptions trends. We can see this trend in the responses of all students before the test ($n = 31$) and after the test ($n = 29$) and in Cohort 2022 before ($n = 22$) and after ($n = 20$) the test. In Cohort 2021 the alignment numbers did not highly vary before ($n = 9$) and after ($n = 10$) the test. It is important to note that the students could have changed their views and responses in the questionnaires based on their test experience.

A second interpretation can be elaborated if we compare the alignment results with the total number of students reporting a higher and a lower perception of performance. In this case, the alignment results are lower. That is, higher perceptions before the test ($n = 26$) and lower perceptions before the test ($n = 22$) together ($n = 48$) are major trends than students' alignment before the test ($n = 31$). This indicates that there were more students with misalignments than with alignments between perception of performance and their actual scores. The fact that most students had a misalignment between their perceptions of performance with their actual scores indicates a lack of familiarity with the test and training to identify their accurate performance. Nonetheless, there were students who had an accurate view of their performance.

On the other hand, the results also reveal an interesting change in perceptions, showing that before the test most students were optimistic about their performance, but their views changed after taking it. Before the test, there was a high number of students whose self-perception of performance was higher ($n = 26$, 21.36%). However, this trend changed after the test; the percentage of students perceiving having a lower performance increased ($n = 32$, 40.51%). A similar trend was observed in each cohort before and after the test regarding a high view of performance. That is, a high number of test-takers from Cohort 2021 ($n = 14$) and Cohort 2022 ($n = 14$) had a high perception of performance before the test, then this positive view decreased after taking it (Cohort 2021 = $n = 9$; Cohort 2021 = $n = 9$).

It seems that students felt confident about their writing performance before taking the test, but then their experience with the test changed their perceptions. This shows that their views might have been more realistic after the test, that their writing level was not what they had expected, or that they felt unprepared or were unfamiliar with the writing tasks they faced.

The previous findings suggest that, overall, participants feel more confident and optimistic about their writing proficiency before taking the exam. However, after the test experience, their confidence noticeably decreases. This suggests that many participants feel uncertain about their English proficiency after they take the exam or are unfamiliar with the expectations and tasks of the test. Additionally, there was a minority of students whose perceptions of performance aligned with their actual performance. It could be that these students have done some practice or have confirmed their views with other activities.

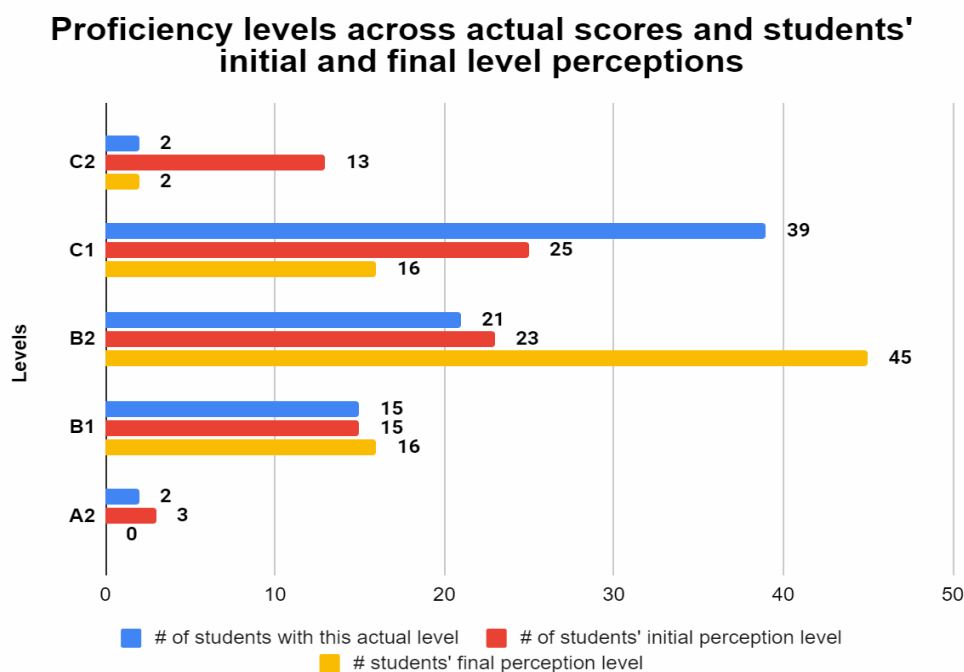
4.5 RQ4. What is the relationship between students' perceptions and performance on the writing test?

The previous section analyzed the perceptions data from a qualitative perspective, reporting trends. In this section, students' perceptions data were related to their scores through correlations to ascertain statistically if there was an alignment and relationship between perceptions and actual performance. These relationships provide feedback about students' views of performance in relation to the test. To identify whether students' perceptions are related to their performance, we implement two correlations. We first correlated students' writing test scores with their initial perceptions of performance, and then we correlated the writing test scores with their final perceptions of performance. The correlations were performed using the data from those participants who answered both initial and final perceptions questionnaires ($N = 79$). These analyses allowed us to understand, interpret and discuss whether students' views of self-performance might have any impact on their performance.

Before presenting the results of the correlations, it is important to remember the results of students' writing performance across the two cohorts. The average writing score for the 2021 cohort was 3.5 and for the 2022 cohort was 3.6, representing a B2 level (3.2–3.6 - 65%–74%). The expected level for the test was C1 (3.7–4.6 - 75%–94%). This B2 level was confirmed by a high number of students ($n = 45$) in their perceptions after the test (See Figure 4.2). This indicates that the average writing test score and students' final perceptions aligned.

Figure 4.2

Number of students across proficiency levels: actual scores and perceptions



In order to identify whether there were statistically significant correlations between actual scores and perceptions of performance, we implement Pearson's and Spearman's correlations. For performing the correlations, we used numerical values of 0 (A1, A2), 1 (B1), 2 (B2), 3 (C1), and 4 (C2) to represent each language level. Table 4.7 presents the descriptive statistics of test scores and perceptions per cohort, using the values that represent the language levels.

Table 4.7

Descriptive statistics of actual scores and perceptions using the values 0 to 4 to represent the language levels across cohorts.

	Test Scores		Initial perceptions		Final perceptions	
	2021	2022	2021	2022	2021	2022
Sample	35	44	35	44	35	44
Mean	2.2	2.3	2.3	2.3	2.0	2.1
Std. Deviation	1.0	0.8	1.0	1.1	0.5	0.7
Shapiro-Wilk	.836	.815	.871	.905	.756	.848
P-value of Shapiro-Wilk	< .001	< .001	< .001	.002	< .001	< .001
Minimum	0	1	1	0	1	1
Maximum	4	4	4	4	3	4

As seen in Table 4.7, the mean results show that students' test scores, initial perceptions and final perceptions classified into the numerical value 2, representing the B2 level. The results confirm the previous students' perceptions results. However, to confirm these descriptive statistics, correlations were performed (see Table 4.8).

Table 4.8

Correlations between students' initial and final performance perceptions and test scores

Variable		3. Test scores
1. Initial perception	<i>n</i>	79
	Pearson's r	0.424
	p-value	< .001
	Effect size (Fisher's z)	0.453
	SE Effect size	0.115
	Spearman's rho	0.421
	p-value	< .001
	Effect size (Fisher's z)	0.449
	SE Effect size	0.118
2. Final perception	<i>n</i>	79
	Pearson's r	0.354
	p-value	0.001
	Effect size (Fisher's z)	0.370
	SE Effect size	0.115
	Spearman's rho	0.359
	p-value	0.001
	Effect size (Fisher's z)	0.376
	SE Effect size	0.117

As shown in Table 4.8, there were statistically significant positive, moderate correlations between *students' writing test scores* and *their initial perceptions of performance* ($r(79) = .424$, $p = < .001$). Similarly, statistically significant and positive, but weak correlations were found between *students' writing test scores* and *their final perceptions of performance* ($r(79) = .354$, $p = .001$). These results suggest that students' perceptions of performance correlate with their actual performance. As the correlations were positive, we could say that both students' views of performance and scores go to the same direction, that is, if their perception of performance is high, the score will be high; and, if the perception of performance is low then the performance will be low, and the same way around, that is, if scores are high, the perceptions of performance will be high. We could say that students' views of writing are reflected in their performance and that perceptions might likely influence performance. Hence, it is important for test-takers to identify their performance and confirm their perceptions.

4.6 Concluding remarks

In this chapter, we presented the results of this study examining the context and consequential validity evidence of the high-stakes writing test of the **B.A in Foreign Language** at Universidad del Valle, through four research questions. First, we explored students' performance in the writing test regarding writing time, text length, and scores to have a view of their performance and understand the students' perceptions of these features of the test. Second, we explored students' perceptions of the test regarding context validity (task setting and administration setting) and consequential validity (positive and negative impacts of the test on students) to examine the validity and fairness of the test from students' voices. Third, we examined students' perceptions of performance before and after the test to identify their views of writing proficiency and how these views changed after taking the test as part of the test impact. Lastly, we analyzed the relationship between test-takers' perceptions of performance and their actual performance in the writing test to identify if students' views might influence their performance and vice versa, as part of the test impact too.

Regarding performance on text length, writing time and scores, the results revealed that, on average, students were able to meet the minimum text length (250 words) in each of the two writing tasks within the allocated time (90 minutes), indicating that the text length and time allotted are valid and fair aspects of the test' context validity. However, the minimum were not able to produce 250 words, which suggests issues probably related to their writing skills or their lack of familiarity with task setting and administration. These aspects might affect the validity and fairness of the test because some test-takers were affected by those features. Regarding test scores, on average, students exhibited a B2 proficiency level, which was below the C1 level they were expected to achieve. The existence of this performance gap suggests that difficulties could arise from factors related to students' writing skills, test familiarity, or curricular issues.

As to students' perceptions of the writing test, the results revealed that the most negatively perceived features were associated with the task setting, particularly *time constraints* and *response format*, as well as the features of the administration setting, notably technological resources, as part of the *physical conditions*. The findings suggest that issues with time constraints and response format were associated with students' time management skills and a lack of preparation in data analysis and writing explanatory texts under assessment conditions. Similarly, the findings reveal that difficulties due to technological resources were associated with the pandemic and technological expertise of participants and

do not indicate issues in the administration of the test. Other aspects of the test were not reported as negative. It was also found that for a great number of students, psychological and experiential characteristics were negative factors affecting students. These factors included stress, nervousness, typing skills and lack of test preparation and familiarity. Overall, the results about perceptions of the test (context and consequential validity) indicate that most aspects of the test are valid and fair; however, there are aspects that could be revised so as not to affect some test-takers.

Concerning test-takers' perceptions of performance before and after the test, the correlations identified statistically positive moderate and weak relationships between perceptions and performance. Students' perceptions of their performance changed after taking the test, with more students perceiving a lower performance. These results indicate that students' views generally aligned with their actual performance, and that they had a realistic understanding of their abilities and the test's demands. This alignment suggests that students' perceptions can be a reliable indicator of their performance and that their views might impact their performance too. This alignment highlights the importance of addressing students' perceptions in test design and preparation strategies. Also, the students' shifts indicated that they may have felt less confident about their writing skills after experiencing the test, highlighting potential areas for targeted training and support in test preparation.

In the Discussion chapter, we will explore the implications of these findings for writing assessment and teaching and for the validity and fairness of the writing test. It will also explore potential strategies to improve test design and administration based on these results.

5 Discussion

5.1 Chapter overview

In this section, we discuss the findings of the study regarding its four research questions, related to test-takers' writing performance, perceptions of the test, perceptions of performance before and after the test, and relationships between perceptions and performance. This discussion provides a deeper interpretation of the results in light of the test context, previous studies' findings, and the framework employed on context and consequential validity for writing assessment.

5.2 Students' writing performance regarding scores, time and text length

The first question in this study sought to determine test-takers' performance regarding score, writing time, and text length, in order to examine their performance and understand their perceptions of these features of the test. The findings showed that most students were classified into the B2 level of proficiency, demonstrating lower proficiency compared to the C1 level expected for this test. In fact, students' writing performance was the lowest of all the test components, as reported by the English language Coordinator and as perceived by the lecturers. This means that most students (achieving a B2 level) were able to perform both writing tasks, synthesizing and evaluating information from various sources. Additionally, they were expected to skillfully employ the structure and conventions of both tasks' genres (personal opinion essay and explanatory text) by adapting the writing tone, style, and register based on the audience, the type of text, and the topic, as determined by the C1 level (Council of Europe, 2020).

One of the reasons for this result could be the influence of virtual learning due to the COVID-19 pandemic, as the cohorts involved (2021 and 2022) took the test during this period. It could be that students lacked writing practice similar to writing test conditions (e.g., limited time, specific text length), which might have affected their performance. Unfortunately, we could not have access to writing test scores from other immediately previous cohorts to compare the results and students' performance, as there was not a writing test before. Another reason could be associated with the contextual features of the writing test, as discussed below in Section 5.3.

Findings on text length and writing time (see Tables 4.1 and 4.2, respectively) show that, on average, test-takers completed both writing tasks with the expected minimum word count (250 words) and within the required time (90 minutes). This is further evidenced by analyzing the minimum values for time and the maximum values for text length, which served

as proof of proficiency on an individual level. These findings show that students were able to produce lengthy texts on the computer for both writing tasks within that extended time given. Hence, writing time and text length were valid and fair aspects for most test-takers regarding the context validity of the task. However, as some other students could not meet the text length requirement and used all the time given for the task, it might be that these features affected them, students' writing skills were low or lack of test preparation. These aspects might affect the validity and fairness of the test.

Our findings also confirmed the low proficiency level found by Kostina (2012) with a different test used at the time. She pointed out that approximately half of the test-takers achieved a B2 level, while the other half did not. These results are concerning because 12 years ago, when there was less exposure to the English language and access to technology, some students were able to achieve the B2 level. In contrast, currently, when students have more exposure opportunities, their writing level remains lower than expected. This low proficiency in the test may indicate a general low proficiency level in the program, unfamiliarity with the test, a lack of test preparation throughout the curriculum, or a setback influenced by the pandemic.

The potential impact of COVID-19 on test-takers' performance might be a valid reason to interpret the low writing results. In fact, cohort 2022 showed, on average, better performance than cohort 2021, as the 2022 cohort returned to face-to-face interactions and might have had more writing practice in the classroom and autonomously. Also, the 2022 cohort had more training opportunities and familiarity with the writing test (see Table 4.4). Considering that cohort 2021 faced the initial repercussions of the pandemic and subsequently transitioned to virtual learning, this finding reinforces the idea of the influence of COVID-19 on test-takers.

Another hypothesis concerns the lack of opportunities to adequately prepare test-takers for assessment environments. Although the program curriculum includes an elective course (Advanced English) for eighth-semester students, not all have the opportunity to attend due to limited enrollment. In this regard, the BA in Foreign Languages might lack fairness in curricular preparation because it cannot be guaranteed that all students take the course prior to the test. Additionally, while the English Department allows students to take a trial test before the actual exam, this tool does not provide immediate feedback to students. However, the students have access to the rubric criteria to know what is expected from them. Students can also use the rubric to assess their performance on the writing test or ask an expert writer or peer to assess their texts based on the rubric criteria.

In this sense, students are given some tools to become familiar with the platform, test guidelines and the writing tasks but do not focus on explicit test preparation. This is coherent with what Cheng and DeLuca (2011) highlighted, that non meaningful feedback from test preparation undermined students' performance. Furthermore, although the program integrates writing tasks throughout the curriculum, it does not teach to the test. It seems that not all courses or teachers address the same writing tasks and conditions used in the exam; hence the writing done by students in class does not accurately reflect the conditions of the writing proficiency test.

A third hypothesis regarding performance is based on the lack of autonomous preparation. Although the BA in foreign languages program at Universidad del Valle prepares students to become language teachers, its focus is not solely on teaching the languages themselves. This means that part of the proficiency relies on students' autonomous work. It could be that a shift in assessment practices and views is affecting students' performance and test preparation. That is, the fact that most teachers might have quit implementing tests is affecting students' familiarity with test-taking experiences and strategies. As part of the fairness of the test, students need to be familiar with the test and receive test preparation (Bachman & Palmer, 2010). The implication is that tests should be part of the curriculum, as most students will take standardized tests under similar conditions in real life.

5.3 Students' perceptions of the test: context and consequential validity

The second question in this research aimed to analyze the students' perceptions of the test, which informs the context validity of the writing test and its impact on students' performance. Refer to Table 4.5 to see the frequencies of positive and negative remarks across different categories. The discussion of the results in this section is intertwined, meaning that the negative or positive perceptions and experiences of students with the writing test will reflect the impact of the contextual features of the test on students' performance. There is a close relationship between context and consequential validity because they inform each other.

5.3.1 Students' perceptions of context validity: Task setting

Context validity – task setting relates to the requirements, instructions, and parameters within which the writing task should be performed (Shaw & Weir, 2007). Here, we will discuss the specific features of time constraints and response format.

5.3.1.1 Time constraints.

Time constraints refer to the limited period allocated for test-takers to complete a writing task (Shaw & Weir, 2007). The validity of this test feature, in the context of the Writing Proficiency Test, implies that the allocated time is adequate for students to perform both writing tasks and demonstrate the expected skills. In this regard, one interesting finding was related to the negative experiences and impact that *time* triggered on a group of test-takers. Eighteen (out of 96) students who answered the survey explicitly mentioned that the allotted time was insufficient to complete both tasks. Negative perceptions of time constraints were associated with other aspects, such as test-takers' cognitive process of writing. For instance, planning and revising require time for some test-takers. Here is the comment of one of the test-takers:

FL2022V: “[...] en general los planteamientos de esta prueba estandarizada son de ejercicios de metacognición y pensamiento superior, se debe dar más tiempo en la planeación escrita. Me encanta escribir y tengo mi propia estrategia y en fin yo cuestiono la estructura de las pruebas estandarizadas en estos dos aspectos específicamente lectura y escritura en cuanto a tiempos.”

Standardized proficiency tests, such as IELTS and TOEFL iBT, allocate 60 and 30 minutes, respectively, as the time constraints for two writing tasks, and in that time, test-takers are expected to plan, write, and revise their texts. Consequently, as the writing proficiency test of the program allocates 90 minutes to perform both tasks, negative remarks of time evidence that students lack test preparation. Students need to learn planning and revising strategies under test conditions.

Other negative remarks regarding time are linked to low proficiency at the individual level, which supports the idea that, to some extent, the test-takers' performance is due to their dedication to autonomous work. This is also arguable considering other students' positive experiences regarding time. These findings suggest that students' commitment to becoming familiar with the test is fundamental to guaranteeing the effectiveness of the test preparation offered by the program.

Additionally, time and psychological factors also showed a frequent relationship. For instance, some students commented that:

DS2022V: “Creo que en parte, **la falta de tiempo y los nervios** hicieron que tuviera experiencias negativas durante el examen.”

SVV2022V: “A lo que la escritura respecta siento que hice un buen trabajo y fui capaz de seguir lo que se me pedía en las consignas. **El tiempo fue sin duda un factor de estrés** pero aún así considero que logré hacer un buen trabajo con lo que tenía [...]”

Therefore, time constraints, rather than being an invalid aspect of the Writing Proficiency Test, reflect underlying issues related to students’ lack of confidence and preparation. These comments support findings from previous studies on language assessment (e.g., Allen, 2016; Cheng & DeLuca, 2011), which concluded that high-stakes tests tend to be associated with negative psychological factors, such as nervousness, stress, and anxiety. Consequently, these factors affect test-takers’ proficiency and confidence. The suggestion here is that students’ nervousness, stress and anxiety can be mitigated with test familiarity.

5.3.1.2 Response format.

Response format refers to the specific format: genre, conventions and structure in which test-takers are required to produce their written responses (Shaw & Weir, 2007). For the Writing Proficiency Test, the validity of this feature implies that test-takers understand and are able to perform the type of text required in each writing task, exposing the expected writing skills. In this regard, although both tasks were academic essays, they had a different format and students experienced them differently. Task 1 involved writing a personal opinion essay, whereas Task 2 required writing an explanatory essay summarizing, interpreting and discussing data from a graph. Students commented that they encountered more difficulties with the second task, specifically with interpreting the graph.

On one hand, there is a sense of unfamiliarity with the format of the second task; some comments show that test-takers found the requirements of the second task ambiguous and meaningless. On the other hand, the difficulty in interpreting the graph could indicate that students do not have enough training in data analysis. While the first statement suggests a potential issue with test design, the second one suggests a deficiency in the curriculum in terms of test preparation. In fact, test-takers’ performance in the second task evidenced a more challenging experience, as the score for Task 2 was lower than that for Task 1 for both cohorts (see Table 4.4). Therefore, the hypothesis regarding unfamiliarity with Task 2 format might be the case.

Additionally, since Task 2 involved writing a text based on the analysis of a graph, it is also plausible that students overall struggle with data analysis. This could also indicate that students feel more comfortable with narrative texts, as those are the texts they work with the

most throughout the program. Although a group of students reported being unfamiliar with Task 2, this is a common task in real life and in other standardized tests, such as IELTS. University students often read research papers that include graphs and tables in different courses and semesters (e.g., Language Acquisition, Sociolinguistics, Classroom Research, and Research Methodology), when writing their theses, they should report, summarize, interpret and discuss the data from graphs and tables. In this sense, Task 2 is valid, but its fairness needs to be enhanced by familiarizing students with this type of task earlier in the curriculum.

5.3.2 Students' perceptions of context validity: Administration setting

Context validity – administration setting relates to the organization and management of the environment and the conditions under which the writing test takes place (Shaw & Weir, 2007). This comprises the space, resources, and supervisors. Here, we will discuss the features of physical conditions and the uniformity of administration.

5.3.2.1 Physical conditions.

Physical conditions refer to the tangible, environmental, and situational factors that can influence the test-takers' writing process (Shaw & Weir, 2007). These include aspects such as technological resources, location, and comfort. Ensuring the validity of these test features means that students' writing processes are not negatively impacted by these factors.

The results showed that a group of test-takers (15 out of 96) reported negative experiences regarding technological resources, which included internet outages, malfunctions of the Campus Virtual platform, and disturbance in the location because of noise and several distractions. However, it is important to note that the cohorts involved in the writing proficiency test in 2021 and 2022 faced, to a greater extent, the incidence of virtuality due to COVID-19. This event not only challenged students in taking the test but also evaluators and monitors in administering it. Therefore, issues related to technological equipment, location, and comfort might primarily stem from the unique challenges posed by the pandemic and the inadequate preparation of both staff and test-takers to manage the writing test during this period.

Since Campus Virtual is the current platform used to administer the test, it is essential to ensure that it works correctly in assessment environments. For instance, a group of students reported a malfunction in this tool when saving their texts, which clearly affected their writing process. As these students represent a small portion of the sample (4 out of 96,

4%), we cannot assume this issue is widespread across the platform. However, it is recommended to review the answer storage functionality of the Campus Virtual platform to prevent such issues from affecting students' performance in future tests. Nevertheless, we acknowledge that there are external factors that might affect such as power outages that are beyond the test and the campus virtual administrators' reach.

5.3.2.2 Uniformity of administration.

Uniformity of administration refers to the consistent and standardized procedures used during the administration of a writing assessment (Shaw & Weir, 2007). In this way, the validity of this test feature ensures that all test-takers are given the same instructions, time limits, conditions, and scoring criteria. The difficulties regarding uniformity of administration were also closely related to virtual challenges and the lack of preparation of the participants, test-takers and monitors. Some students reported being distracted by other test-takers who had trouble starting the test, as well as by monitors attempting to assist them, which interrupted their writing process. Although test-takers' technological expertise relates to their experiential characteristics, students who experienced these issues (4 out of 96, 4%) perceived them as unfair conditions during the test because of the exceptional circumstances under which it was conducted, i.e., the pandemic. Evidently, this group does not represent the majority of the sample, so these difficulties do not suggest unfairness during the test administration; instead, they might stem from students' lack of preparation. Therefore, it is important to familiarize test-takers with computer-based assessment to avoid these factors conditioning their writing test process.

5.3.3 Students' perceptions of context validity: linguistic demands

Context validity – linguistic demands relate to the ability to respond appropriately and effectively to writing prompts. This category encompasses key features such as lexical, structural and functional resources, discourse mode, and content knowledge (Shaw & Weir, 2007). We will discuss the feature of content knowledge (task input/output), as it was the only notable aspect test-takers reported negatively.

Content knowledge refers to the expertise that test-takers have regarding the subject matter they need to write about (Shaw & Weir, 2007). Hence, the validity of this test feature ensures that test-takers can skillfully interpret and report information on both writing tasks. In this regard, a small group of students (4 out of 96, 4%) reported challenges in interpreting the graph and writing the explanatory text required in Task 2. These difficulties in understanding and making inferences from the graph (content knowledge–input) and explaining its

implications (content knowledge–output) show a gap between the critical analysis expected of students and the analytical skills they demonstrate. Although students are required, throughout the curriculum, to take courses where they learn to interpret qualitative and quantitative data, these courses typically do not involve writing about the data and rarely assess students on this skill. It might also be that different methodologies among teachers in the preparatory courses prevent some students from training their analytical skills.

As these negative reports do not comprise a significant part of the sample, we cannot determine if they were related to the students' preparation in analytical skills or the clarity of the graph associated with Task 2. However, these findings highlight the importance of ensuring that the curriculum provides students with consistent academic training and the development of the analytical skills they are expected to demonstrate in the writing test.

5.4 Students' perceptions related to test-takers' psychological and experiential characteristics

Test-takers' characteristics are factors that can affect test-takers' performance (Bachman & Palmer, 2010; Shaw & Weir, 2007). These characteristics are external to the test, but they can be triggered by it either positively or negatively, depending on students' preparation. Students' positive and negative experiences associated with psychological factors stem from the circumstances of the test itself and issues associated with the test-taker. On one hand, the repercussions of failing the test, as it is a high-stakes exam, undermine students' confidence, causing nervousness and stress and consequently affecting their performance (Cheng & DeLuca, 2011). Thus, the mere act of taking the test without prior experience in similar conditions negatively impacts test-takers' writing performance. On the other hand, test-takers' negative experiences with other aspects of the test, such as time management, test preparation, and familiarity with the Campus Virtual platform, have an impact on test-takers' perceptions of the writing test.

These findings reinforce the idea that students should be familiar with the test autonomously and by receiving formal training throughout the curriculum, focusing not only on content and language skills but also on test-taking skills and strategies. This is consistent with previous research that concluded that when students are familiar with the nature of the questions, receive meaningful feedback, and have prior experience with similar tests, negative psychological factors decrease and the test generates a positive washback effect (Allen, 2016; Cheng & De Luca, 2011; Lee, 2006).

5.5 Students' perceptions of performance before and after the test

The third question of the study sought to examine the students' perceptions of performance before and after the test, to identify their views of writing proficiency and how these views changed after taking the test. Based on the results, test-takers' perceptions of their writing skills before and after the proficiency test changed, as seen by the adverse shift in perception after the test (see Table 4.6). This shift may result from several context validity aspects, including the test's difficulty, the participants' performance in comparison to expectations, psychological elements, or preparation level. It is important to create a positive testing environment. A test that aligns well with test-takers' experiences fosters a positive testing environment, reducing anxiety and stress. This positive environment can enhance overall performance and provide a more accurate measure of their abilities.

Their view changing following the test, as seen in students' perceptions, raises concerns about how test-taking experiences affect students' self-perception and confidence levels. This study may have implications for developing interventions or support mechanisms to assist test-takers in managing their perceptions and emotions during and after testing. This aligns with Shaw & Weir's (2007) framework, emphasizing the role of self-assessment and the impact of test experiences on individuals' perceptions of their performance. While it is clear that success in language learning and performance assessment depends primarily on an individual's ability in the intended construct, there are of course other variables that are likely to impact performance and relate to the personal characteristics of the individual test-taker; these include factors such as age, interests, experience, knowledge, and motivation (Shaw & Weir's 2007).

Moreover, the apparent change in their perception after the test shows that participants experienced test-related stress or uncertainty. This is consistent with sociocognitive theories, which highlight the impact of cognitive and emotional aspects on performance assessment. These elements should be considered by educators and test designers when developing assessments in order to establish a supportive testing atmosphere that reduces anxiety and promotes accurate self-assessment.

These results open avenues for further research into factors influencing test-takers' perceptions, strategies to enhance confidence and performance alignment, and the long-term impacts of testing experiences on individuals' self-assessments and motivation. Understanding how test-takers perceive the test helps ensure that it meets the expectations of all stakeholders, including educators, institutions, and the test-takers themselves (Bachman & Palmer, 2010; Shaw & Weir, 2007; Slomp, 2016). This alignment with stakeholder expectations is key for the test's acceptance and success.

In summary, the study's findings highlight the complex relationship between students' perceptions, actual performance, and test experience. Educators can use insights from Shaw and Weir's (2007) sociocognitive framework to create tailored interventions that improve students' self-assessment abilities, minimize test-related anxiety, and encourage a more accurate design and assessment of language competency. These activities are critical for promoting meaningful learning outcomes and maintaining the validity of language assessments.

5.6 Relationships between perceptions and performance

The fourth question in this study addressed the relationship between test-takers' perceptions of performance and their actual performance on the writing test. The findings revealed that there was a noticeable decrease in the number of test-takers who had a positive perception of their abilities after taking the exam. This indicates that students were very optimistic about their abilities before the test, but then their views changed to a more negative perception of performance after taking the test, probably due to test demands or the test-taker's psychological factors. As Shaw and Weir (2007) emphasize, test-takers' perceptions and their subsequent test performance are influenced by a range of cognitive and emotional factors that need to be carefully considered in test design and administration.

The analysis of the correlations of the results indicated a statistically significant relationship between students' perceptions (before and after the test) and their performance on the writing test. Most students who perceived their performance positively achieved higher scores, while those with negative perceptions tended to score lower. This finding suggests that students' perceptions can be a reliable indicator of their actual performance and highlights the importance of addressing these perceptions in test design and preparation strategies. Additionally, the alignment between perceptions and performance reinforces the test's validity from the examinees' perspective, indicating that students have a realistic understanding of their abilities and the test's demands. Shaw and Weir (2007) support the notion that perceptions of test-takers can serve as valuable feedback for validating the test's effectiveness and fairness.

In conclusion, while the writing test is mostly valid and fair, some aspects need revision to avoid being unfair to certain groups of students. It is essential to make adjustments in how students are prepared and familiarized with the test tasks and conditions and to create a more supportive evaluation environment. Incorporating examinees' perspectives into test design and administration can significantly enhance the quality and credibility of language assessments. This study emphasizes the usefulness of Shaw and

Weir's sociocognitive framework in writing test validation, providing a structured approach to understanding and validating these tests from the examinees' viewpoint and highlighting the need to consider their experiences and perceptions to improve assessment practices.

5.7 Examining the context and consequential validity evidence of the writing test

This last section of this chapter discusses the results in relation to the framework on context and consequential validity.

5.7.1 *Valid and fair aspects of the test regarding context validity and consequences*

According to the evidence, the following aspects of the writing test were identified as valid and fair from the students' voices, as there were not any reported negative experiences about them.

- **Purpose.** The reason for assessing writing is well defined and students are aware of the abilities they must demonstrate. This brings positive consequences to students in their performance as they know the purpose of the assessment.
- **Knowledge of criteria.** Students are explained the standards and guidelines by which the writing test is conducted. This brings positive consequences to students because they are informed about the level and performance expected.
- **Weighting.** The value given to the different aspects of both tasks was perceived as fair and valid by test-takers. This is positive because students know that both tasks mean similar demands.
- **Text length.** It can be stated that 250 words are valid and fair for the writing proficiency test because most students met this requirement, which is usually asked in class activities and other standardized exams. This requirement is positive for test-takers to demonstrate their performance at length, without word limits constraints.
- **Writer-reader relationship.** It can be said that students understand the expected writer-reader interaction for both tasks. This is positive because they know they have to be prepared for writing to a proficient academic audience.
- **Lexical resources.** Test-takers did not report having difficulties with the lexical requirements. This is positive because the test topics are appropriate for their level.
- **Structural resources.** Test-takers did not report problems organizing ideas coherently and logically. This means that the task purpose and instructions provide test-takers with opportunities to structure and organize their ideas.

- **Functional resources.** It can be stated that students did not report having problems using the language to achieve specific communicative goals. This is positive because it means that students were aware of the language and functions implicated in their responses.

All in all, most aspects of the writing test are valid and fair and could continue to be implemented in its administration.

5.7.2 Aspects of the test that need improvement to enhance its validity and fairness

- **Response format.** It was found that test-takers felt more comfortable writing the academic essay in Task 1 than with the explanatory essay in Task 2. According to some students, the instructions for Task 2 should be clearer and more detailed. They suggested receiving more information about the task requirements from the English Department during the introductory meetings. If some test-takers have issues with a task, this will bring negative consequences on their performance. Hence, providing more detailed information about the tasks' purpose and format for the following writing tests could not only reduce confusion and ensure that test-takers fully understand the requirements, but also guarantee the enhancement of the overall test experience.
- **Time constraints.** It was identified that 90 minutes for both writing tasks was sufficient according to the quantitative analysis, as on average, students completed the test before the allocated time. Furthermore, the time given is sufficient for producing a minimum of 250 words and is longer than the time given by standardized language tests. However, some students felt that this time was not enough. According to some comments, they experienced stress and nervousness and had difficulties planning and revising their texts due to the insufficient amount of time. The fact that some test-takers have issues with the 90 minutes means that their performance can be negatively impacted. This finding suggests providing more time or better preparing students to perform under time constraints.
- **Physical conditions – technological resources.** The Campus Virtual platform experienced malfunctions during the writing test. According to some students, the resource did not save their responses when they moved to other sections of the test within the same platform. If test-takers encounter issues with the storage functionality of Campus Virtual, it will disrupt their writing process and negatively impact their performance. Therefore, it is recommended to organize test simulations or test the

platform under conditions similar to those of the actual exam. These measures will help prevent potential errors during future test administrations.

- **Physical conditions – location and comfort.** As the cohorts involved in this test were assessed mainly virtually, students experienced distractions from external noise and interruptions from other participants struggling to start the test. If participants do not feel comfortable and focused while taking the test, their performance can be negatively affected. Therefore, it is important to ensure all test-takers are in a suitable space that minimizes external distractions, and to guarantee they all understand the technical aspects, such as logging into Campus Virtual and starting the test, to avoid interruptions during its administration.
- **Uniformity of administration.** During the test administration, test-takers had the same conditions regarding time constraints, text length, task instructions, knowledge of criteria and rubric scoring. However, since the test was conducted mainly virtually, they did not share the same physical conditions. Students faced challenges such as external noise, internet outages, and issues with their equipment. To prevent these factors from negatively affecting them in future tests, it is important to ensure that all test-takers have a suitable space and access to reliable technological resources.
- **Content knowledge (input/output).** The writing tasks are valid tasks, as they are commonly used in other standardized tests and reflect real-life academic experiences of university students. However, a group of test-takers struggled to interpret the graph associated with Task 2. They had difficulties making inferences from the data and, consequently, writing the explanatory text. If test-takers do not understand the provided information from the graphs, their compositions will not meet the criteria, negatively affecting their performance. Hence, it is recommended to revise how students are being trained in analytical skills prior to the test to reinforce common weaknesses.
- **Test preparation.** Although students were given access to a trial test, materials and informative meetings about the test, there were students who experienced issues with time management, unfamiliarity with the Campus Virtual platform and test structure, difficulties with typing, and psychological challenges. Although the trial test was intended to introduce test features to students, it was optional (not mandatory), and not all of them took it before the actual test. These findings suggest that students should be trained under conditions similar to those of the exam and throughout the curriculum in order to receive feedback and become familiar with the writing

assessment environment. Students should also be autonomous by developing test taking strategies by themselves to minimize negative experiences with the test.

- **Discourse mode.** Although test-takers did not mention this aspect explicitly in their answers, Task 2 was perceived as more difficult than Task 1 because they reported being unfamiliar with the genre.

Other aspects that could be improved, although not mentioned by the students but identified by the researchers, include test accommodation. For instance, the test could consider giving two modes of assessment, i.e., computer and paper-based, as some test-takers felt more comfortable with traditional paper-based tests. This is particularly relevant because, at the time this test was administered, many test-takers were more familiar with paper-based assessments. The sudden shift to a computer-based format during the pandemic presented challenges, including the need to ensure that students had functional resources at home, such as reliable internet connections and working devices. Additionally, considerations for students with specific needs, such as those with color blindness, should be addressed to ensure that the test is accessible to all. Task 2, for example, includes a graph using colors such as green, yellow, or red, which can be challenging for students with color blindness. Addressing these issues by collecting prior information from test takers' characteristics, providing alternative assessment modes and additional support can help create a more equitable testing environment.

5.7.3 What are the positive and negative consequences of the test on students' performance and the academic program?

5.7.3.1 Positive consequences of the test.

Apart from the above mentioned features of the test related to its context validity, we identified the following positive impacts of the test on students and the academic program (washback effect).

- **Performance.** The test shows that students are able to perform academic writing tasks, providing feedback to test-takers and the program. Students' performance aligned with their perceptions meaning that students should experience more tests and have pre- and post-test surveys and reflections to identify their performance and align their views with their scores.
- **Preparation for high-stakes tests.** The test serves as preparation for future high-stakes tests and certifications. This indicates that students will be able to perform

other tests with less stress and will be familiar with the demands and conditions of standardized tests. Hence, the teachers and the program could do mock tests more frequently.

- **Testing of curriculum.** The test enables researchers to identify strengths and weaknesses in the curriculum and assessment practices, particularly in terms of the alignment between course content and test demands. Additionally, the test evaluates the curriculum's effectiveness in fostering students' academic writing skills and preparing them for high-stakes assessments.

5.7.3.2 Negative consequences of the test.

- **Psychological aspects.** The test creates nervousness and stress among some students. This pressure can have negative effects on confidence, potentially leading to burnout or disengagement from sitting similar tests. Exposing students to actual tests and mock tests and promoting the use of the trial test and preparation in other courses could minimize these negative effects.
- **No teaching for the test.** The fact that the curriculum develops students' writing skills that will be assessed but lacks a strong focus on test-taking strategies and preparations for standardized tests might create a gap between what is taught and how it is assessed. For instance, the academic program fosters students' writing proficiency and develops skills that can be demonstrated in any language test (e.g., placement test, achievement test, proficiency test); however, test preparation might help students perform better. Although time and emotional management skills are not the focus of the test, they significantly influence how students approach it. If teachers employ holistic approaches that better integrate teaching and test-taking strategies, this could prevent students from feeling overwhelmed during the actual test.

6 Conclusions

6.1 Chapter overview

This chapter summarizes the results and provides the main conclusions regarding the context and consequential validity of the writing test administered to senior students in the BA in Foreign Languages at Universidad del Valle. The chapter also provides implications for the validity and fairness of the writing test, the program curriculum, and students as test-takers. The limitations of the study are acknowledged, and further research is suggested.

6.2 Summary of findings and conclusions

This study examined test-takers' writing performance, their perceptions of the test and of their performance, and the relationship between perceptions and performance as evidence to make claims about the validity and fairness of the writing test.

The results about performance on text length, writing and scores showed that most students were able to meet the minimum text length (250 words) in each of the two writing tasks within the allocated time (90 minutes), indicating that the text length and time allotted are valid and fair for most students. However, some students were not able to produce 250 words, which suggests issues probably related to their writing skills or the lack of familiarity with the task setting and the administration setting. Regarding test scores, on average, students exhibited a B2 proficiency level, which was below the C1 level they were expected to achieve. The existence of this performance gap suggests that difficulties could stem from factors associated to students' writing skills, test familiarity, curricular issues or the pandemic.

As to students' perceptions of the writing test, the results revealed that *time constraints* and *response format*, particularly in Task 2, (context validity – task setting), and *physical conditions* (context validity – administration setting) were features of the test that affected most students. Other aspects of the test were not reported as negative. It was also found that some test-takers' psychological and experiential characteristics, such as *stress*, *nervousness*, *typing skills* and *the lack of test preparation and familiarity* were negative factors affecting students. These findings suggest that there are issues with students' time management skills and a lack of preparation in data analysis and writing explanatory texts under assessment conditions.

Concerning test-takers' perceptions of performance before and after taking the test, the correlations identified statistically positive relationships between perceptions and

performance. Students' perceptions changed after taking the test, with more students perceiving a lower performance. These results indicate that students' views generally aligned with their actual performance, and that they have a realistic understanding of their abilities and the test's demands. This alignment suggests that students' perceptions can be a reliable indicator of their performance and that their views might impact their performance, too. This alignment highlights the importance of addressing students' perceptions in test design and preparation strategies. The shift in students' perceptions after the test, where they felt less confident about their writing skills, suggests areas that require targeted training and support. However, for claiming a general alignment, both pre- and post-test perceptions should be considered. Focusing solely on pre-test perceptions may overlook the impact the test had on students' confidence and experiences. Post-test perceptions provide valuable insights into how the actual test conditions affect students, which is crucial for implementing meaningful changes.

Based on these findings, the study concludes that most aspects of the test are valid and fair regarding context validity and consequential validity. However, there are aspects that could be revised so as not to affect some test-takers. In fact, the test is valid and fair regarding *purpose, knowledge of criteria, weighting, text length, writer-reader relationship, lexical resources, structural resources, discourse mode* and *functional resources*. Conversely, improvements are recommended in features such as *response format, time constraints, physical conditions, uniformity of administration, content knowledge, and test preparation*. Additionally, this study found that the Writing Proficiency Test has both positive and negative consequences for students and the Bachelor's degree. Positive impacts include the evidence of students' performance, preparation for other high-stakes tests, and alignment with the curriculum. On the other hand, negative impacts involve psychological factors associated with the test and methodologies that fail to integrate teaching with assessment.

We also conclude that investigating test-takers' perceptions of the test provides context and consequential validity evidence, which demonstrates how the test impacts those who take it. This evidence is crucial for showing that the test is effective, fair, and beneficial for the test-takers. Moreover, feedback given by test-takers can highlight areas where the test might be misaligned with their expectations or experiences. This information is highly important for revising the test to better meet its intended goals.

Shaw and Weir's (2007) sociocognitive framework and the ex post facto research approach were adequate theoretical and methodological choices for this study. Firstly, Shaw and Weir's framework provides a comprehensive approach to validating writing tests by

considering various dimensions of validity. By focusing on context validity, the framework allowed us to examine how well the test tasks reflect real-world writing situations, ensuring that the test is relevant and meaningful for students. It also considers the test-takers' characteristics and the testing environment, which are crucial for understanding how different factors might influence students' performance.

Additionally, the framework's emphasis on consequential validity was particularly useful for our study, as it helped us explore the positive and negative impacts of the test on students. This includes examining how the test affects students' anxiety, and overall perceptions of their writing abilities. By using Shaw and Weir's framework, we could examine comprehensive evidence to support the test's validity and fairness.

The ex post facto research approach was an appropriate methodological choice for this study. This approach involves analyzing data that has already been collected to investigate potential causes and relationships between variables. It was particularly useful for our study because it allowed us to examine students' performance and perceptions without manipulating any variables, the data was recorded years before, thereby maintaining the natural context of the test-taking experience.

Using this approach, we analyzed existing data on students' test scores, writing performance, and perceptions before and after the test. This retrospective analysis provided valuable insights into how various factors, such as test difficulty, psychological elements, and preparation levels, influenced students' performance and perceptions. Moreover, the ex post facto design enabled us to identify correlations between students' self-assessments and their actual performance, helping us understand the complex relationship between their perceptions and outcomes. Overall, this approach was instrumental in providing a deep evaluation of the writing test's validity and its impact on students.

Finally, this study contributes valuable insights, to us as future teachers and to the BA program, into the validity and fairness of writing assessments, highlighting areas for improvement in test design, administration, and preparation. The implications of the research extend to curriculum development, test-taker support, and the overall enhancement of writing assessment practices at Universidad del Valle. It aims to promote meaningful learning outcomes and maintain the validity of language assessments, ultimately benefiting both students and educational institutions. Addressing the concerns and perceptions of test-takers is essential for developing fair and valid assessments. Continuous feedback and adaptation will ensure that the writing tests not only measure proficiency accurately but also support the educational and professional growth of future language teachers.

6.3 Implications

6.3.1 Implications for the writing Test: validity and fairness

The study confirmed the validity and fairness of the Writing Proficiency Test regarding the features of *task setting* and *linguistic demands* categories. The majority of students met the minimum text length and completed both tasks within the allocated time. Additionally, most test-takers felt that the linguistic demands aligned with the expected proficiency level and the linguistic skills they were taught throughout the curriculum. We consider that these test features must continue to be implemented in future tests to maintain validity and fairness.

Task 1 and Task 2 are valid tasks as they are commonly used in other standardized tests and reflect the real-life academic experiences of university students. However, this study confirmed that test-takers struggled with Task 2 due to difficulties in reading graphs, making inferences, and writing explanatory texts. Therefore, we recommend implementing more data analysis exercises throughout preparatory courses, such as Classroom Research I and II, and requiring more explanatory texts as assessment activities. These measures could enhance students' analytical skills prior to the test and address common weaknesses.

Improvements in physical conditions and uniformity of administration should be considered to ensure the validity and fairness of *task administration* features. For instance, it is important that all students take the test in a suitable location that minimizes external distractions. The English Department should ensure that test-takers have access to quality technological resources, such as computers and the Campus Virtual platform, and guarantee that these resources will work optimally during the test administration. We recommend testing the Campus Virtual platform and the necessary equipment under conditions similar to the actual exam to avoid malfunctions during the administration of the test.

The surveys on students' initial and final perceptions provided valuable and consistent feedback about students' views on writing performance, their notion of the test context validity, and the test impact. Test-takers' perceptions offer necessary insights to confirm which test elements are working correctly and which need improvement. Therefore, it is important to continue integrating students' perception surveys into the Writing Proficiency Test to ensure its continuous improvement. Furthermore, we recommend the implementation of mock tests along with corresponding students' perception surveys to identify students' needs and mitigate negative consequences in the actual exam.

There are aspects of test-takers' characteristics that are not being considered when measuring the validity and fairness of the writing test such as visual disabilities. For instance, the design of the graph associated with Task 2 uses colors as the only elements to distinguish between variables and trend lines. Consequently, a test-taker with color blindness will not be able to distinguish between these variables because there are no labels or other signs apart from colors to differentiate them. It is recommended to employ several types of markers in the graph design to make the writing tasks more inclusive.

6.3.2 *Implications for the curriculum of the program*

This study showed that the average students' proficiency level was B2, below the expected C1 level, indicating a gap in performance. The findings suggest that by the eighth semester, students may lack both the necessary knowledge for the evaluation and sufficient test preparation for standardized language proficiency tests. We advocate incorporating more writing practice and assessment tasks in preparatory courses such as English Integrated Skills and Writing Composition in English. By aligning these tasks with the test format and conditions, students will become more familiar and comfortable with the test requirements.

The BA in Foreign Languages supports students' test preparation by providing various resources integrated throughout the curriculum. For instance, the preparatory (mandatory) courses address linguistic demands and provide training for task setting; the Advanced English elective course trains students in assessment environments; and the Campus Virtual trial test allows students to become familiar with this platform. These resources are consistent with the writing test demands, so it is important to continue employing them. Additionally, we suggest extending the capacity of the Advanced English course to ensure all students can take it prior to the actual exam. Also, as the trial test does not offer feedback on performance, we recommend incorporating this resource into the same elective course to take advantage of it and allow students to receive feedback from teachers.

The study found that students' post-test perceptions of their performance more closely align with their actual performance. After taking the test, most students adjusted their views to more realistic assessments of their abilities and weaknesses. Therefore, implementing mock tests accompanied by post-test perception surveys could help the BA program identify students' needs, target improvements, and allow students to develop a more accurate awareness of their skills and areas for growth. We recommend incorporating mock tests and post-test perception surveys across the preparatory courses to identify weaknesses and

enable teachers to provide feedback. This approach can not only mitigate negative psychological and experiential factors but also enhance overall student performance.

6.3.3 Implications for students as test-takers

The study found that issues with test preparation and familiarity were also associated with students' lack of autonomous work. Although the BA in Foreign Languages offers consistent and timely test preparation resources, their effectiveness also relies on the students' commitment to utilizing them. It is important for students to review the test structure and evaluation criteria shared in the informative meeting to become familiar with the writing tasks and linguistic demands. We recommend that students use the test evaluation criteria to self-assess similar writing tasks and reflect on their performance, time management skills, and other related issues. This will help students anticipate potential difficulties, allowing them to improve before the actual test.

Students' familiarity with the Campus Virtual platform is instrumental, as it is the current platform for administering the test. The English Department offers a trial test on this platform to help students become familiar with the environment and the writing test setting. Although this measure is optional, students are responsible for ensuring they can log into the Campus Virtual platform and exploring its functionalities, identifying the writing test on the dashboard, and completing the attempts. Therefore, it is essential that test-takers utilize the trial test to understand this platform and all its technical requirements to avoid wasting time during the actual test.

Test-takers are not limited to only using the resources offered by the English Department and the mandatory courses throughout the curriculum. Since students of the BA in Foreign Languages are constantly engaged in collaborative work to develop group-work skills, they are familiar with addressing common academic difficulties and searching for solutions and strategies. Hence, the creation of study groups might help students feel more confident in recognizing common gaps and organizing peer-assessment activities. Collaborative learning allows students to share ideas and opinions, discuss different perspectives, and support each other in understanding the material. By working together, students can develop critical thinking skills and a deeper comprehension of the test content. Additionally, study groups provide a platform for peer support, where students can motivate and encourage each other, reducing the feelings of isolation and stress often associated with high-stakes testing. This collaborative environment fosters a sense of community and shared goals, which can be particularly beneficial in preparation for the writing test. Furthermore,

students can attend tutoring sessions with teachers or dedicated academic support groups when they have more specific questions.

Students have several self-paced resources to practice their writing skills. For instance, they can use online dictionaries or thesauri to increase their vocabulary, YouTube videos to learn strategies for high-stakes tests, and free language learning platforms like the British Council or Cambridge English to reinforce their writing skills. AI-based writing tools provide substantial support for students in reviewing and improving their texts. For instance, Grammarly and Hemingway offer instant feedback on grammar, vocabulary, style, and coherence, helping students polish their writing and improve overall readability. Additionally, students have access to dedicated materials from the School of Language Sciences and the Mario Carvajal Library at Universidad del Valle. All in all, these tools, resources, and materials can be invaluable in helping students enhance the quality of their writing and perform better on their assessments.

Finally, a crucial recommendation is that students should view assessments as an opportunity to measure their current language performance and identify areas for improvement, rather than seeing them as the ultimate level of proficiency they can achieve. Maintaining the right mindset during the evaluation process is essential for managing the psychological factors involved, interpreting the results objectively, and addressing areas of improvement, regardless of whether the performance meets expectations.

6.4 Limitations and further research

One of the primary limitations of this study was the sample size, as not all test-takers completed both the initial and final perception questionnaires. Students who did not complete both surveys were excluded from the study. However, the sample size was enough for performing correlations and was representative of this population. Additionally, the study could not make comparisons with previous cohorts because there was no writing exam in place, and the test applied earlier was different. The sample was also limited to the context of the pandemic, which means these results cannot be generalized to other contexts. This limitation restricts the ability to draw broader conclusions about the writing test's validity and fairness beyond this specific scenario.

Future research should consider a larger and more diverse sample size to ensure more comprehensive and generalizable results. It would also be beneficial to include a comparison of different cohorts, including those before and after the pandemic, to evaluate any changes in performance and perceptions over time. Additionally, research should

examine the long-term effects of the transition back to campus and whether it leads to improvements or declines in both performance and perception. It will be crucial to investigate whether students feel more comfortable with computer-based tests after increased exposure and determine if the technical issues experienced during remote testing persist. Test designers and administrators could also be included in future studies of this kind at the ECL. These considerations will provide valuable insights for enhancing the design and administration of future writing assessments.

References

- Akbari, R. (2012). Validity in language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyhoff (Eds.), *The Cambridge guide to second language assessment* (pp. 30-36). Cambridge University Press.
- Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia*, 6(1). <https://doi.org/10.1186/s40468-016-0030-z>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.
- Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
- Bakri, H. (2023). Evaluating and testing English language skills: Benchmarking the TOEFL and IELTS tests. *Bulletin of The Faculty of Languages & Translation*, 25(2), 243-266. <https://doi.org/10.21608/bflt.2023.319543>
- Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices*. Pearson Longman.
- Cheng, L., & Curtis, A. (2012). Test impact and washback: Implications for teaching and learning. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyhoff (Eds.), *The Cambridge guide to second language assessment* (pp. 89-95). Cambridge University Press.
- Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment*, 16(2), 104–122. <https://doi.org/10.1080/10627197.2011.584042>
- Colmenares Rodríguez, S. (2013). Prácticas de escritura académica en una universidad pública colombiana: Autoría, audiencia e interacción con otras voces. *Lenguaje*, 41(1), 201–227. <https://doi.org/10.25100/lenguaje.v41i1.4967>
- Council of Europe. (2020). *Common European Framework of Reference for Languages (CEFR) Descriptors*. <https://www.coe.int/en/web/common-european-framework-reference-languages>
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications, Inc.
- Díaz Ramírez, M. (2014). The impact of Project Work and the writing process method on writing production. *HOW*, 21(2), 31–53. <https://doi.org/10.19183/how.21.2.3>
- Educational Testing Service. (2024). *TOEFL iBT*. Retrieved from <https://www.ets.org/toefl/test-takers/ibt/about/content/writing.html>
- Gómez, G. (2016). Análisis descriptivo-comparativo de los resultados de los exámenes de competencia en inglés y francés de la jornada diurna y vespertina de la Universidad

- del Valle año 2011 (Bachelor's dissertation).
<https://bibliotecadigital.univalle.edu.co/handle/10893/9477>
- Green, A. (2014). *Exploring language assessment and testing*. Routledge.
- Guapacha Chamorro, M. E. (2020). Investigating the comparative validity of computer- and paper-based writing tests and differences in impact on EFL test-takers and raters (Doctoral dissertation). Retrieved from
<https://researchspace.auckland.ac.nz/handle/2292/53273>
- Guapacha Chamorro, M. E. (2022). Cognitive validity evidence of computer- and paper-based writing tests and differences in the impact on EFL test-takers in classroom assessment. *Assessing Writing*, 51. <https://doi.org/10.1016/j.asw.2021.100594>
- Guapacha Chamorro, M. E., & Chaves Varón, O. (2023). EFL writing studies in Colombia between 1990 and 2020: A qualitative research synthesis. *Profile: Issues in Teachers' Professional Development*, 25(1), 247–267.
<https://doi.org/10.15446/profile.v25n1.94798>
- Im, G. H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia*, 9(14), 1-26. <https://doi.org/10.1186/s40468-019-0089-4>
- International English Language Testing System. (2024). *IELTS*. Retrieved from
<https://ielts.org/take-a-test/preparation-resources/sample-test-questions/academic-test>
- Ivankova, N. V., & Greer, J. L. (2015). Mixed methods research and analysis. In B. Paltridge & A. Phakiti (Eds.), *Research methods in applied linguistics* (2nd ed., pp. 60–74). Bloomsbury
- Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly*, 14(2), 101–119. <https://doi.org/10.1080/15434303.2016.1261293>
- Kostina, I. (2012). El nivel de dominio de inglés de los futuros profesores de lenguas: Examen de competencias en inglés. *Lenguaje*, 40(2), 383–413.
<https://doi.org/10.25100/lenguaje.v40i2.4955>
- Kostina, I., Faustino Ruiz, C. C., & Vergara Luján, O. (2013). Assessment practices in the English and French component of a foreign languages teacher education program. *Lenguaje*, 41(2), 353–382. <https://doi.org/10.25100/lenguaje.v41i2.4973>
- Krippendorff, K. (2009). Conceptual foundation. In *Content analysis: An introduction to its methodology* (pp. 18-39). Sage Publications.
- Kunnan, A. J. (2018). *Test fairness and TESOL: Theoretical and practical issues*. Routledge.
- Lee, H. (2008). The relationship between writers' perceptions and their performance on a field-specific writing test. *Assessing Writing*, 13(2), 93–110.
<https://doi.org/10.1016/j.asw.2008.08.002>
- Lee, Y.-J. (2006). The process-oriented ESL writing assessment: Promises and challenges. *Journal of Second Language Writing*, 15(4), 307–330.
<https://doi.org/10.1016/j.jslw.2006.09.003>

- Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42(4), 448–457. <https://doi.org/10.1080/02796015.2013.12087465>
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241–256. <https://doi.org/10.1177/026553229601300302>
- Ministerio de Educación Nacional. (2017). *Resolución No. 18583*.
- Ministerio de Educación Nacional. (2019). *Programa Nacional de Bilingüismo*. https://www.mineducacion.gov.co/1621/articles-132560_recurso_pdf_programa_nacional_bilinguismo.pdf
- Mojica, C. P., Velásquez Díaz, Y. Y., & Casas, N. C. (2009). Percepciones de los agentes involucrados en la aplicación de una prueba de clasificación para alumnos de la extensión del Departamento de Lenguas de La Universidad Nacional de Colombia. *Matices en Lenguas Extranjeras*, (3). <https://revistas.unal.edu.co/index.php/male/article/view/13838>
- Muñoz, A., & Álvarez, M. (2008). Preliminary evaluation of the impact of a writing assessment system on teaching and learning. *Profile: Issues in Teachers' Professional Development*, 9(1), 89–110. Retrieved from <https://revistas.unal.edu.co/index.php/profile/article/view/10706/11177>
- Muñoz, A., Mueller, J., Álvarez, M., & Gaviria, S. (2006). Developing a coherent system for the assessment of writing abilities: Tasks and tools. *Íkala, Revista de Lenguaje y Cultura*, 11(1), 265–307. Retrieved from <http://aprendeenlinea.udea.edu.co/revistas/index.php/ikala/article/view/2791/2245>
- Phakiti, K., & Isaacs, T. (2021). *Classroom assessment and validity: Psychometric and educational measurement approaches*. Routledge.
- Pallant, J. (2016). *SPSS Survival Manual: A step by step guide to data analysis using IBM SPSS* (6th ed.). Open University Press. Chapter 11: [Correlations].
- Ramírez Espinoza, A. (2020). Análisis de ítems para prueba de clasificación en inglés en una universidad colombiana. *Cultura, Educación y Sociedad*, 11(2), 177–190. <https://doi.org/10.17981/cultedusoc.11.2.2020.11>
- Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.
- Slomp, D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English*, 48(3), 276–302.
- Tavakoli, H. (2012). *A dictionary of research methodology and statistics in applied linguistics*. Routledge.
- Tejada, H., & Castillo, N. (2010). El backwash effect o los “efectos colaterales” del examen ECAES, Prueba de inglés 2009. *Un análisis crítico. Lenguaje*, 38(2), 449–80. <https://doi.org/http://hdl.handle.net/10893/3516>
- Universidad del Valle, Consejo Académico. (2021). *Resolución No. 062*. Retrieved from https://drive.google.com/file/d/1qD2MA99zRI_idwG3yJ_gUnJaP77nw9Cn/view

Appendices

Appendix A

Consentimiento Informado Coordinadora de la Sección de inglés

Santiago de Cali, 21 de febrero de 2024

Profesora
María Eugenia Guapacha
Coordinadora Sección de Inglés
Licenciatura en Lenguas Extranjeras Inglés-Francés

Informed Consent

You are invited to participate in the research study entitled "Examining the Context and Consequential Validity of a High-Stakes Writing Test in a Colombian Bachelor in Foreign Languages Program". This research project is conducted by Bryamth Cárdenas and Linda Cruz as a requirement for the degree of Bachelor in Foreign Languages (English - French) from the School of Language Sciences at Universidad del Valle.

This study aims to examine how valid and fair the writing proficiency test of the Bachelor's degree in Foreign Language at the Universidad del Valle regarding its context validity is, by examining its consequential validity evidence (or effect of the test on students') from students' voices and perspectives. Additionally, the study aims to explore the students' perceptions of context validity and performance.

We are requesting access to the results of the English proficiency test and the perception surveys from students in the 2021 and 2022 cohorts to explore the context and consequential validity evidence. The access will be limited to the students' performance and perceptions only and will not include any personally identifying information about them. The data will be used solely for research purposes and will be anonymized to protect the privacy of the students. There are no potential risks associated with granting access to the data.

The benefits of participating in this study include exploring the implications of writing assessment on students' educational experiences and outcomes, providing empirical evidence to evaluate the validity and fairness of the writing proficiency test, and contributing to enhancing the overall validity and fairness of language assessments, ultimately benefiting both students and educational institutions.

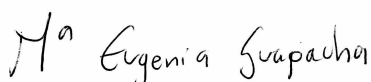
Granting us access to the essays is entirely voluntary, and all information provided will be kept confidential. If you have any questions or concerns about the study, you may contact Bryamth Cárdenas and Linda Cruz, at bryamth.cardenas@correounivalle.edu.co and cruz.linda@correounivalle.edu.co. Additionally, if you have any questions regarding your rights as a research participant, you may contact the supervisor Maria Eugenia Guapacha at maria.guapacha@correounivalle.edu.co

Consent:

I have read and understood the information provided in this form, and I voluntarily agree to grant access to the essays written during the Examen de Competencias for the purpose of the study described above.

English Coordination's Name: María Eugenia Guapacha

English Coordination's Signature:
Date: Febrero 24 de 2024



Appendix B

Questions Responses 47 Settings

Facultad de Humanidades
Escuela de Ciencias del Lenguaje



Section 1 of 3

Percepciones de proficiencia en inglés

Por favor responda esta breve encuesta sobre sus percepciones de proficiencia en el idioma inglés. Se le compartirán los resultados comparativos entre sus percepciones y desempeño real.

This form is automatically collecting emails from all respondents. [Change settings](#)

Autorizo que la información brindada en este cuestionario y en el examen se utilice anónimamente y para propósitos académicos. *

☐ Sí
☐ No

Question format for initial perceptions

Question
1 From 1 to 5 rate your performance in Listening Skill
2 From 1 to 5 rate your performance in Reading Skill
3 From 1 to 5 rate your performance in Speaking Skill
4 From 1 to 5 rate your performance in Writing Skill
5 I have better skill or performance in...
6 Why do you think you have better skills or performance in that skill? If you answered that your performance is the same in all skills or that you are not sure, please justify.
7 I have much more difficulty in...
8 Why do you think you have difficulties in that skill? If you answered that you have difficulties in all skills or that you are not sure, please justify.
9 Receptive skills: Listening and Reading Which of the following propositions applies to you? • I'm very good at both • My skill with -- is better than my skill with -- • I'm not sure
10 Productive skills: Writing and Speaking Which of the following propositions applies to you? • I'm very good at both • My skill with -- is better than my skill with -- • I'm not sure

Note. Questions 6 and 8 are open-ended.

Question format for final perceptions

Question	
1	How would you describe your experience of taking the proficiency test through the Campus Virtual?
2	How would you describe your experience in Listening?
3	How would you describe your experience in Reading?
4	How would you describe your experience in Writing?
5	How would you describe your experience in Speaking?
6	How would you describe your experience regarding the Campus Virtual?
7	How would you describe your experience regarding internet connection?
8	How would you describe your experience regarding the computer?
9	How would you describe your experience regarding the space in which the test was took?
10	What comments do you have about it?
11	What would you recommend to make the test better?
12	What do you think your listening performance was?
13	What do you think your reading performance was?
14	What do you think your writing performance was?
15	What do you think your speaking performance was?

Note. Questions 10 and 11 are open-ended.

Appendix C

Rubric: writing task 1

TASK ACHIEVEMENT & CONTENT DEVELOPMENT	5	4.5	4	3.5	3	2.5	2	1.5	1	0
	Fully addresses the task. Presents a clear and fully developed position with relevant, extended and well supported ideas, details and examples.		Sufficiently addresses the task. Presents a clear position, and presents, extends and supports main ideas with some details and examples.		Addresses the task. Although the position is presented, it is not sufficiently supported; there may be irrelevant detail and not enough examples.		Addresses the task only partially/ May lack focus. The position is unclear: supporting ideas, details / examples are unclear / repetitive / irrelevant. Shortish script (200 - 240 words)		Responds to the task only in a minimal way/the answer is completely unrelated to the task. Does not express a position. Presents few ideas, which are largely undeveloped or irrelevant.	Did not take the test
COHERENCE AND STYLE										
Coherence: Organisation of whole text; logical paragraphing, topic sentences. Reader engagement	Very well organised: logical paragraphs, topic sentences, clear visual introduction and conclusion. Message followed effortlessly.		Well organised: logical paragraphs, topic sentences, relevant introductory statement, may be no concluding statement. Message clear.		Writing organised into paragraphs but some paragraphs are not well structured. Introduction or conclusion missing. Message generally clear.		Poor organisation: illogical/ splintered /misplaced / no paragraphing but ideas can be followed. Hard to follow (needs rereading) at some parts. Introduction and conclusion missing. Shortish script (200 - 240 words)		Lacks organisation: no paragraphing. Incomprehensible / hard to follow throughout. Insufficient length (less than 200 words) and substance.	Did not take the test
Cohesion inside text: Use of cohesive devices (linking expressions) and referencing (this, these, those, which, it)	Skilful non-formulaic use of a range of cohesion and referencing devices.		Appropriate use but slightly formulaic / lower range of cohesive devices. Good use of referencing.		Formulaic use /overuse of cohesive devices. Repetitious referencing.		Insufficient/simple or repetitive connectives (and,but,so). Some incorrect cohesive devices and unclear referencing.		Lacks cohesion: many cohesive devices incorrectly used. Disjointed script.	Did not take the test
Style: Formal, impersonal, academic tone or register; hedging (e.g., possibly, might, may, could be). Not oral, colloquial or informal. Informal language includes abbreviations, contractions, etc.	Clear academic style and good use of hedging		Most aspects of academic style with some hedging. May be some slight informality		Some evidence of academic style, some informal or colloquial language. No / repetitive hedging.		Little evidence of academic style; Frequent informal / colloquial language, wordiness. No hedging.		Overuse of personal and informal / colloquial language.	Did not take the test
FORM										
Sentence structure: accuracy, variety, and complexity of sentences; punctuation	Controlled and varied sentence structure.		Adequate variety: occasional errors in complex sentences.		Limited range: errors in complex sentences may be frequent.		Limited control; many incomplete / rambling / convoluted sentences. Lacks punct.		Lacks control of sentence structure throughout-short script.	Did not take the test
Grammar: parts of speech (verbs, articles, prepositions, agreement, plurals)	Virtually error free		A few minor errors		Minor/repeated errors		A range of significant errors affect expression of ideas.		Short script (less than 200 words) with a range of significant errors affecting expression of ideas.	Did not take the test
Vocabulary and spelling: Appropriate or accurate use of words. Range (richness or sophistication): Form (spelling and word form)	Wide range of vocabulary, appropriately used. Very few minor misspellings.		Vocabulary generally appropriate and good range. Some minor misspellings.		Vocabulary appropriate but lacks range. Few errors in word spelling errors.		Vocabulary simple / repetitive / frequently inaccurate or inappropriate. Many intrusive misspellings.		Range and use of vocabulary inadequate. Numerous errors in word use and spelling disproportionate to length and complexity of script.	Did not take the test

Rubric: writing task 2

CONTENT	5	4.5	4	3.5	3	2.5	2	1.5	1
Part 1 of question. Summary of data: it includes context, time, place, figures, trends, clarity, detail	Accurate, very clear summary of all data: context, time, place, figures and trends		Data summarised mostly accurately; includes some significant figures and trends		Data summary lacks clarity; under/over emphasis on figures / misreading of some figures or trends		Data partially summarised; may be one piece of information only, and inaccuracy/ confusion or repetition.		Data not summarised / incomprehensible
Part 2 of question. Interpretation of data - reasons explained	Interpretation logical; reasons given with sufficient elaboration.		Interpretation generally adequate; at least two relevant reasons but more elaboration needed.		Interpretation brief. May include relevant and irrelevant reasons.		Interpretation very brief. May include irrelevant/ inaccurate reasons.		Interpretation unclear
Part 3 of question. Development of discussion/ extension of ideas with supported proposals	Ideas relevant - supported appropriately and in detail.		Ideas generally relevant and supported appropriately.		Ideas relevant with some supporting evidence.		Ideas not always clear; brief / may be insufficient supporting evidence.		Lacks relevant ideas; re No ideas/ incomprehensible
COHERENCE and STYLE									
Coherence: Organisation of whole text; logical paragraphing, topic sentences. Reader engagement	Very well organised: logical paragraphs, topic sentences, clear visual introduction and conclusion. Message followed effortlessly.	4.5	Well organised: logical paragraphs, topic sentences, relevant introductory statement, may be no concluding statement. Message clear.	3.5	Writing organised into paragraphs but some paragraphs are not well structured. Introduction or conclusion missing. Message generally clear.	2.5	Poor organisation: illogical/ splintered / misplaced / no paragraphing but ideas can be followed. Hard to follow (needs rereading) at some parts. Introduction and conclusion missing. Shortish script (200 - 240 words)	1.5	Lacks organisation: no Incomprehensible / hard to follow throughout. Insufficient (less than 200 words) and so
Cohesion inside text: Use of cohesive devices (linking expressions) and referencing (this, these, those, which, it)	Skilful non- formulate use of a range of cohesion and referencing devices.		Appropriate use but slightly formulaic / lower range of cohesive devices. Good use of referencing.		Formulate use /overuse of cohesive devices. Repetitious referencing.		Insufficient/simple or repetitive connectives (and,but,so). Some incorrect cohesive devices and unclear referencing.		Lacks cohesion; many devices incorrectly used in script.
Style: Formal, impersonal, academic tone or register; hedging (e.g., possibly, might, may, could be). Not oral, colloquial or informal. Informal language includes abbreviations, contractions, etc.	Clear academic style and good use of hedging		Most aspects of academic style with some hedging. May be some slight informality		Some evidence of academic style, some informal or colloquial language. No / repetitive hedging.		Little evidence of academic style: Frequent informal / colloquial language, wordiness. No hedging.		Overuse of personal and colloquial language.
FORM									
Sentence structure: accuracy, variety, and complexity of sentences; punctuation	Controlled and varied sentence structure.	4.5	Satisfactory variety: occasional errors in complex sentences.	3.5	Limited range; errors in complex sentences may be frequent.	2.5	Limited control; many incomplete /rambling/convoluted sentences. Lacks punct.	1.5	Lacks control of sentence structure throughout-short script
Grammar: parts of speech (verbs, articles, prepositions, agreement, plurals)	Virtually error free		A few minor errors		Minor/ repeated errors		A range of significant errors affect expression of ideas.		Short script (less than 200 words) with a range of significant errors affecting expression of ideas
Vocabulary: Appropriate or accurate use of words; Range (richness or sophistication): Form (spelling and word form)	Wide range of vocabulary, appropriately used. Very few minor misspellings.		Vocabulary generally appropriate and good range. Some minor misspellings.		Vocabulary appropriate but lacks range. Few errors in word use. Some important, intrusive spelling errors.		Vocabulary simple / repetitive / frequently inaccurate or inappropriate. Many intrusive misspellings.		Range and use of vocabulary inadequate. Numerous errors in use and spelling disproportionate to length and complexity