UNIVERSIDADE DE BRASÍLIA

INSTITUTO DE LETRAS

DEPARTAMENTO DE LÍNGUAS ESTRANGEIRAS E TRADUÇÃO

Face validity of the TOEFL-ITP according to UnB students

Student: Raul Vitor Sandrini da Rocha

Research field: Applied Linguistics

Advisor: Prof. Dra. Gladys Plens de Quevedo Pereira de Camargo

Final paper presented to the department of Foreign Languages and Translation of the University of Brasília, as requested for concluding the English Language course.

Brasília-DF, maio de 2021.

UNIVERSIDADE DE BRASÍLIA

INSTITUTO DE LETRAS

DEPARTAMENTO DE LÍNGUAS ESTRANGEIRAS E TRADUÇÃO

Face validity of the TOEFL-ITP according to UnB students

RAUL VITOR SANDRINI DA ROCHA

APPROVED BY:

_____

Prof. Dra. Gladys Plens de Quevedo Pereira de Camargo (LET-UnB)
(Presidente)

_____

Prof. Liberato Silva Santos (IFG)
(Avaliador externo)

_____

Prof. Avram Stanley Blum (LET-UnB)
(Avaliador interno)

_____

Profa. Vanessa Borges de Almeida (LET-UnB)
(Suplente)

BRASÍLIA - DF, May 16th.

SUMMARY

FIGURES, GRAPHS AND CHARTS

ABSTRACT

The TOEFL-ITP test is a language proficiency test made by ETS, and a very important one for its ease-of-use and application. It is crafted for institutions that want to provide *en masse* English proficiency testing, and here we analyze the context of said tests that were provided by the LwB (language without borders) program, at UnB. In language assessment, there are ways to determine whether a test can be trusted or not - its validity. It is split into four kinds of validity: content, criterion, construct and face validity. In this paper, we aimed at determining whether the test had face validity among UnB students, who had the opportunity to take the test free of charge, and for that we used a questionnaire based off of qualitative research and the Likert scale, since face validity is subjective rather than objective. The results were that the test does possess face validity, but some aspects of it may need more research, such as the time given for the realization of the test, since some students considered it unsatisfactory. In suma, the paper aims at being a stepping stone to assist further projects related to validity of proficiency tests.

Key-words: TOEFL-ITP; Face validity; Validity; Proficiency tests.

1 - INTRODUCTION

Evaluations and assessments have been present in our lives for a long period, such as the *Dokimasia*, where Greeks would assess the capacities of citizens to attain public positions, rights, and duties (QUEVEDO-CAMARGO, 2014). Yet, it has been a fairly recent preoccupation of countries and governments, especially in Brazil, where evaluation institutes such as INEP have been in charge for around 30 years only (GATTI, 2014). With that in mind, there is a crescent need for data and research on the topic, and this paper aims to be a stepping stone for a more detailed analysis of the English proficiency test (PT) made by ETS and applied at UnB (as part of the actions of the ISF (languages without borders) program and completely free of charge, being last applied in 2019), TOEFL ITP , while also instilling a better reading construction on the topic of assessments and validity. Throughout the article, the face validity of the aforementioned test will be put into light using a qualitative method, with questions based upon the Likert scale, altogether with definitions, explanations, and resources used for the analysis.

2 - MOTIVATION FOR THE STUDY

Becoming a teacher is a challenge that we, educators, face every day. The myriad of factors that must be taken into consideration when planning a class is ever-increasing, and within these factors, one rises higher than most: the results that students achieve in standardized tests. Sometimes these demands get so high that we enter a stage of role inversion: we begin "*teaching for the test*", a term coined by Jonathan Kozol in his book *The Shame of the Nation (2005)*, to name the way the Bush presidency handled education in its administration after the *No Child Left Behind* act, which drove many schools to adopt a policy of accountability, where the teachers were blamed for their students' (poor) results. The question that led to this research was to exactly what point do students of the Letras institute trust and rely upon a test like TOEFL-ITP, and whether there were cases of students being taught for the test.

Hence, the main guide behind this paper is providing data about the TOEFL - ITP in order to conduct deeper and more complex analyses, and not only about this test, but to compare its aspects and nuances with other proficiency tests as well in the near future. Furthermore, there is a crescent need for research on this topic according to a personal interview with Meire Souza (UnB)[1], and also because of a personal curiosity to know what was thought among peers about this aspect of the test.

[1] Meire Nadja Meira de Souza is a Master of Education by the University of Brasília, and a personal acquaintance of the author. Her master's thesis defense was what brought the main idea of this paper into light.

3 - LITERATURE REVIEW

Since language testing and assessment are connected directly to language learners, teachers, test designers, and others who have a great importance when it comes to teaching and learning English, we can safely put this topic into applied linguistics (BACHMAN, 1990). To (2001 apud HUONG, 2020) classifies language assessments as a most useful tool to create a powerful (positive) washback effect, via creating a feeling of competition, all based off of Davies (1978 apud HUONG, 2020, p. 2) who first stated that "[...] qualified English language tests can help students learn the language by asking them to study hard, emphasizing students' promotion as well as teachers' evaluation. ".

Validity is not the only principle when it comes to a language assessment: practicality and reliability (QUEVEDO-CAMARGO; SANTOS, 2020) are also main principles for a test.

Face validity is, on its own, regarded as one of the less relevant forms of validity (LAERD, 2021, p.1), but it is a concept present in any kind of questionnaire, be it a research, evaluation, or even an interview. DAVIES et al. (1999, p. 59) define face validity as "the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer". Thus, a test's face validity is merely a result of its most superficial aspect being on par with what people think *should* be being evaluated and also *how* people think the best way to evaluate a topic: if a research about people's stamina is proposed and the way to assess that is an aerobic exercise such as running or biking, it would have a greater face validity (LAERD, 2021).

Throughout the researching period, there was plenty of material to be used in regards to the TOEFL-ITP test, such as *Critical Analysis on TOEFL ITP as A Language Assessment*, by Taufiq *et al*. (2018) . In this text, the TOEFL-ITP test is analyzed critically and curiously reaches a similar level of face validity when compared to the result of this paper, converging to classifying the test as facially valid, while also stating that "[...] More students want to get the lessons to raise the score to be higher" (p.1), thus creating room for more research on its backwash effect and providing a more solid conclusion.

3.1 - Theoretical Framework

In regards to theory, authors such as Brown (2000) and Akbari (2018) were the main basis for what is validity and its aspects. The works of Quevedo-Camargo (2018; 2020) were used as a touchstone to check related factors, such as other types of validity and the Brazilian scenario

on this kind of research. Moreover, Henning (1987, p.89) was also used to more clearly define what face validity is, “the test’s surface credibility or public acceptability”, while Arkoudis (2011, p. 33) was the basis to the main critique to this kind of test’s face validity, such as in "The [assessment] sector's blind faith in language testing inhibits the development of more robust ways of addressing English language outcomes for graduates".

Since this was a face validity assessment, the main guide for the research, as recommended by the advisor, was qualitative research. Face validity, being a more supplemental kind of validity by nature, relies on people’s *opinions* rather than solid facts, but that does not stop it from being accurate - far from that. The more experts consider a test or assessment facially valid, the better its results tend to be (LAERD, 2021).

Furthermore, this whole project idea came from Souza (2019), a master’s dissertation presented at UnB which revolved around applying formative yet ludic testing. According to a personal interview with the author, Meire Nadja Meire de Souza, there is a crescent need for a more liberating form of education, and therefore also a greater need for different ways of testing. Hence, in order to begin research on how to change the status quo, there must be research on the status quo itself in order to really know what can be changed and what should be changed, and if those changes can really be for the better.

3.2 - The TOEFL-ITP test

The test was created and still is managed by ETS, an American nonprofit private company, and, according to *www.fundinguniverse.com*, a website which hosts the history of educational testing service, it is the biggest company in this field. The company itself has employed several known names, such as Gary J. Ockey, author of *Assessing L2 Listening: Moving Towards Authenticity* (2018), one of the most quoted books on the subject according to *ResearchGate*, a website made almost exclusively for sharing articles and papers, and also an excellent reading.

According to its own website, *www.ets.org/toefl_itp/content*, and also a quick handbook to prepare test takers provided by ETS themselves, the test has three areas to assess students’ English levels, and the tests themselves also vary in level: Level 1 tests are designed to evaluate intermediate to advanced students, while level 2 is plotted to assess high beginning to intermediate students (source: the website help). The website does not mention what they consider specifically these levels to be, and the tests also differ in time and number of

questions: level 1 has 140 questions in total and a total time limit of 115 minutes, or approximately 2 hours, while level 2 has 95 questions and a time limit of 70 minutes, just above one hour.

The website also mentions the topics that can be mentioned throughout the tests; they can be of three major groups: Academic topics, such as Arts, Humanities, Life Sciences, Physical Sciences, and Social Sciences; Campus-life topics, such as classes, campus administration and activities; and finally General topics, such as business, environment, food, language and communication, media, objects, personal relations, planning and time management, purchases, recreation, transportation, and workplace.

The areas of assessment listed before are similar to the traditional (yet outdated since long (LOTHERINGTON, 2004) four language skills: Reading, writing, speaking and listening, but with a more functional approach rather than an objective assessment. Since the research is based only on the level 1 test, the level 2 test properties shall be left aside in order to provide a clearer picture. The areas of assessment are as follows:

3.2.1. Listening Comprehension

In this first section of the test, students are exposed to several audios which attempt to simulate a conversation or dialog, and are asked questions about each audio, totalling 50 questions, with a varying number of questions per audio, but usually one per short clip. All questions in this section of the test are multiple-choice questions, and the time limit for this part of the test is 35 minutes. This part of the test was considered by the interviewees the hardest part of the test in terms of time accounting and question clarity, as we shall observe later on.

3.2.2. Structure and Written Expression

The second and shortest part of the test, withholding only a maximum time frame of 25 minutes, and it is specifically designed so that the students' ability to recognize language that is appropriate for standard written English can be properly assessed. In this section, there are also two types of questions, one for structure assessment and another for written expression assessment. The first type consists of making students fill in blanks in the middle of sentences with one or more words, which are answering options (all questions are multiple-choice). The second type, though, revolves around

four underlined words in a sentence, but one of these words is ungrammatical; the student's job is to pick which word needs correction.

3.2.3. Reading Comprehension

The final and most time-consuming part of the test, totalling an allotted time frame of 55 minutes, is the reading part of the test. In this section, students are presented to a short text or paragraph themed around a specific scientific subject (in order to prevent some students which already know said subject from getting an advantage, the test contains enough context so that even the most unknowledgeable person can stand on even grounds when compared to someone that knows about the subject (ETS, 2017, p.14), which is followed by a number of questions, which can vary. The questions add up to 50 in total, and attempt to measure how well a student can infer information from a scientific paper, similar in topic and style to those found in universities and colleges. In the end, most of the hardships interviewees had throughout the test were concentrated on the listening part of the test, considering most of the structure and reading comprehension parts with high remarks.

4 - METHODOLOGY

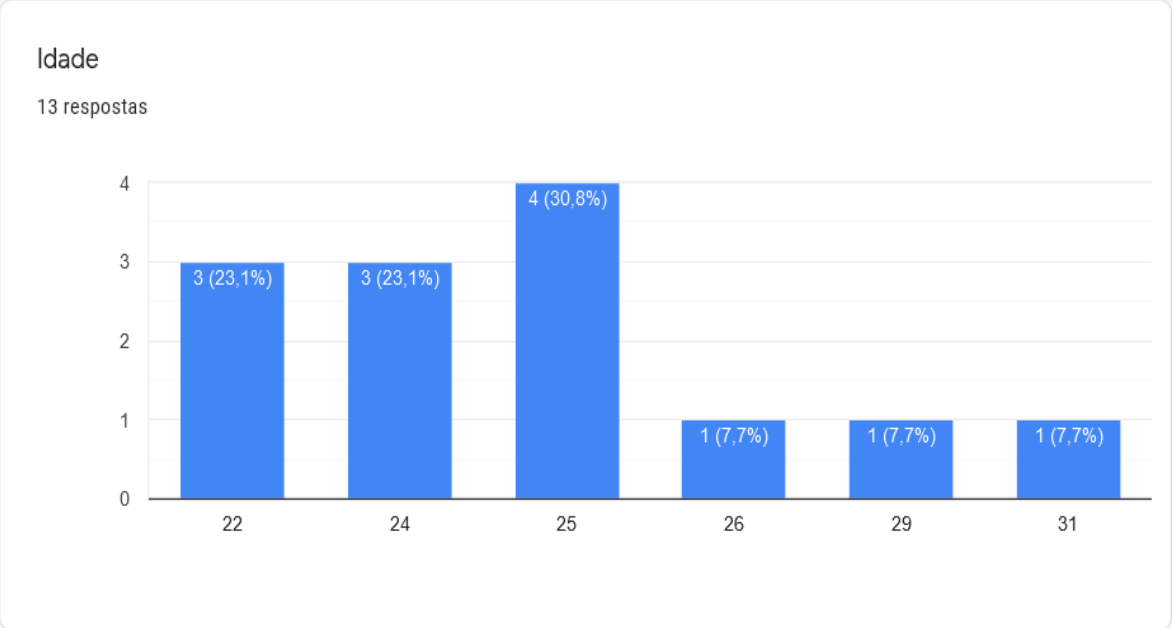
The main guide used to piece together the research was Prodanov e Freitas' *Metodologia do trabalho científico: Métodos e técnicas da pesquisa e do trabalho acadêmico*, which proved to be a fantastic guide on how to give adequate structure to this work.

The study aimed to successfully assess the face validity of the TOEFL-ITP among UnB students, and use the provided data to further solidify future research on the topic.

4.1 - Participants

The target audience of the interview were students at UnB who had taken the TOEFL-ITP test. In total, there were 13 participants, but 2 had not yet taken the test, only heard of it. UnB students were the targeted audience because, as mentioned earlier, the test is provided freely every semester at the university.

Graph 1 - Interviewees' age. Source: The author.



4.2 - Instruments, procedures and analyses

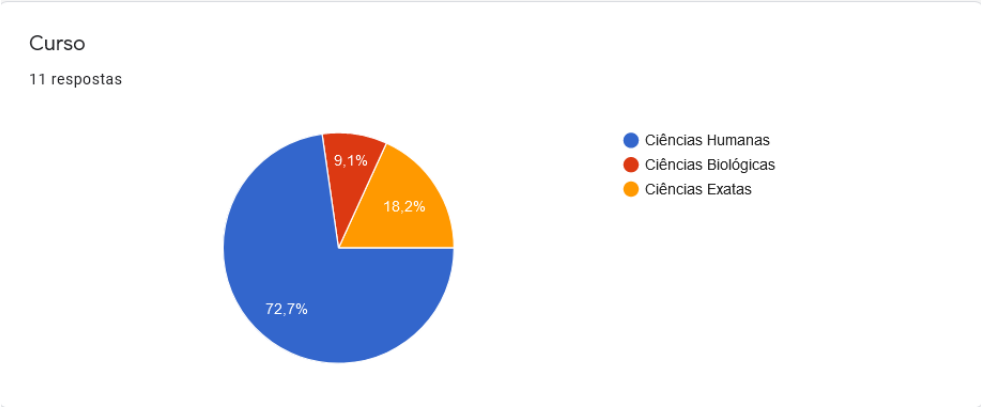
The data gathering tool used was a questionnaire, which is a quantitative way to gather data about a specific subject within a very large database due to its practicality, based off of a qualitative scale, more specifically the Likert scale, which involves asking interviewees their beliefs in a 1-5 range, leaving room for neutrality and smaller agreements/disagreements, since we are dealing with something more subjective than objective (LAERD, 2021).

The data were analyzed via use of simple arithmetic mean and percentages, and the graphics were either provided by Google Forms itself or made via Google Sheets.

5 - RESULTS

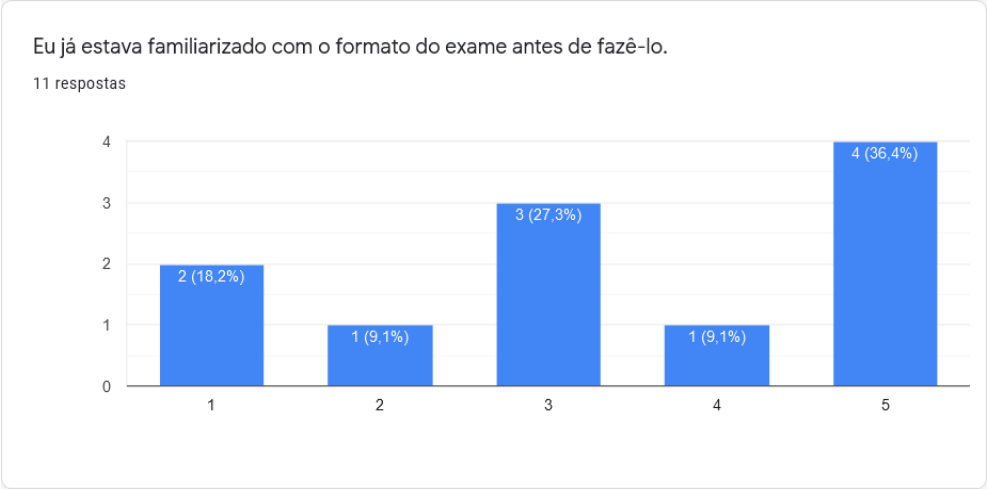
As mentioned earlier, two of the interviewees had not undergone the test, so their answers were discarded, leaving a total of 11 interviewed people. The majority of interviewees were female, with a total representation of approximately 72% (8).

Graph 2 - Interviewees' course. Source: The author.



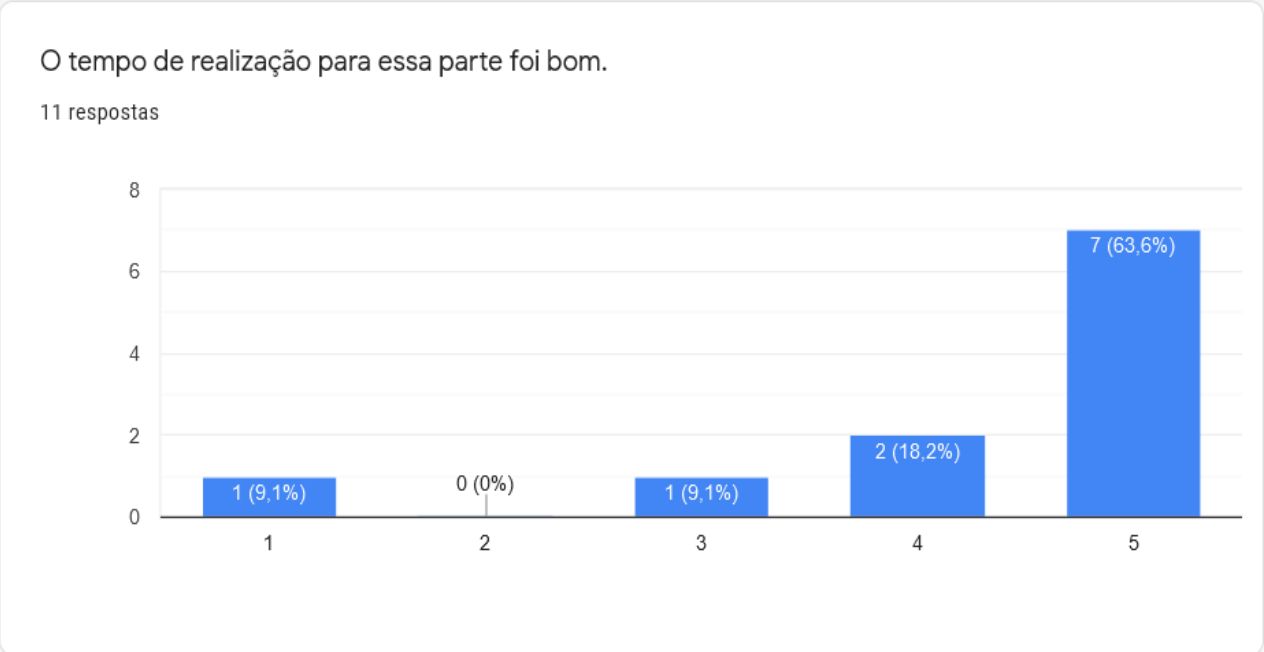
As what is being assessed here is face validity, we must be aware if the students have themselves already studied or undergone specific courses to prepare themselves for the test. The interview points to a null rate of preparation courses taken for the TOEFL-ITP, even though most of the interviewees had some previous knowledge about the test.

Graph 3 - interviewees' familiarity with the test. Source: The author.



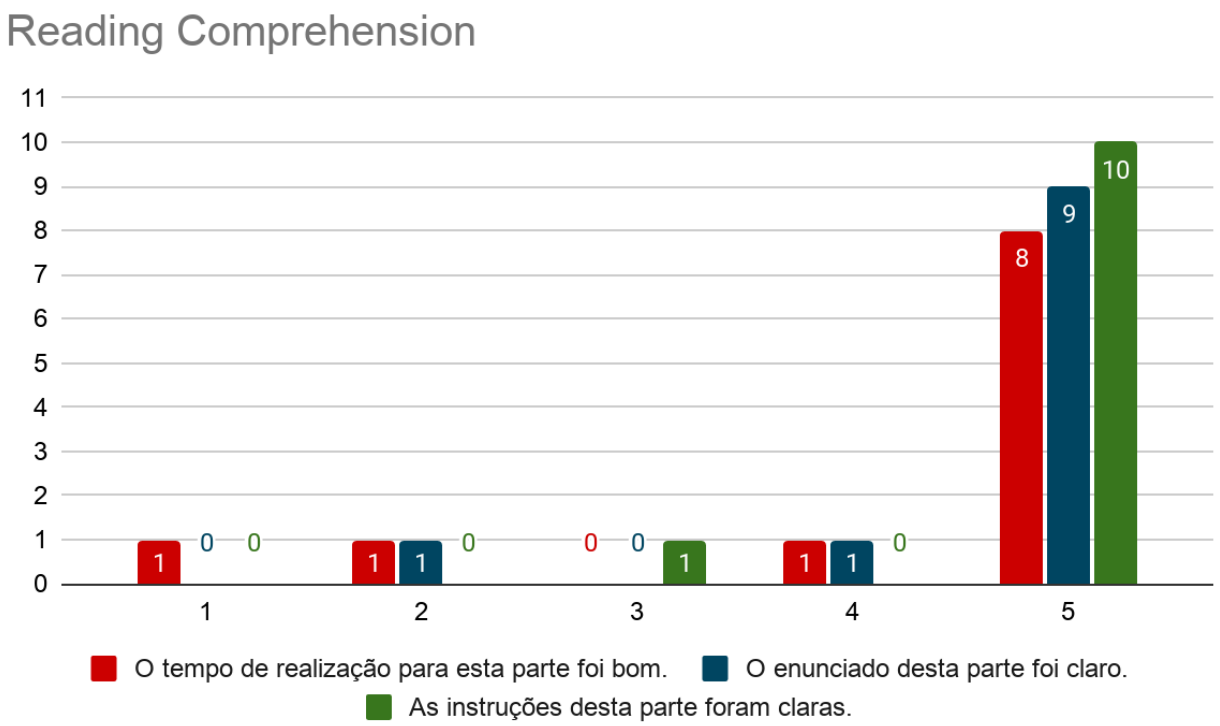
Because of those factors, it is safe to infer that the TOEFL-ITP test is not considered too daunting of an assessment by the students, although the listening part of the test was the one which fared the worst when compared to the other two parts of the test, as can be seen on Graph 4. Still, the overall opinion of the students always remained above the 3.0 mark, which can be considered as facially valid.

Graph 4 - Time allocation, reading part. Source: The author.



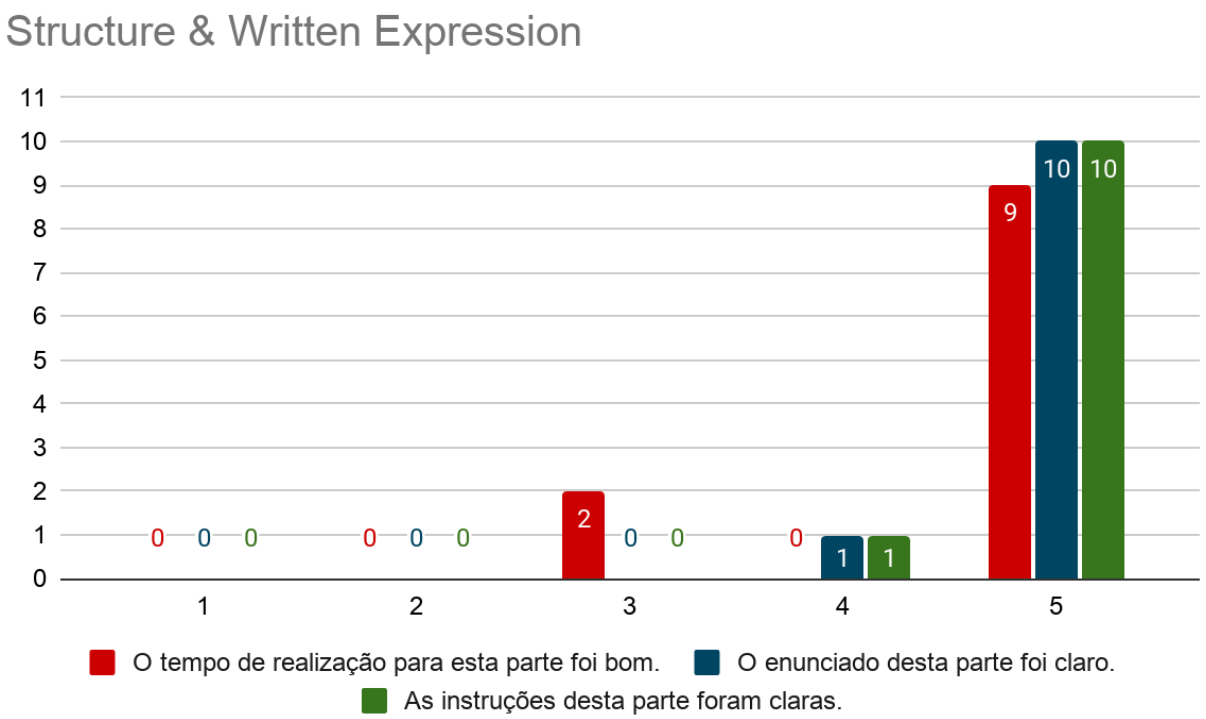
Finally, the latter parts of the text were almost unanimously clear and valid according to the interviewees, as can be seen in images 5 and 6.

Graph 5 - Reading comprehension statistics. Source: The author.



Again, it is noticeable that the aspect students have the most trouble with is time and its management. This discrepancy can be assessed in a separate study with a wider array of subjects in order to determine if the test is really lacking in terms of time displacement with more accuracy. The same pattern can be observed in the next chart.

Graph 6 - Structure and written expression statistics. Source: The author.



6 - FINAL CONSIDERATIONS

In a nutshell and in a similar tone to research made by Taufiq (2018), the TOEFL-ITP can be considered facially valid amongst UnB students of varied courses, but mostly from human sciences courses, which is quite significant for the test, since face validity is considered more relevant when it comes from opinions that have some level of expertise on the subject (LAERD, 2021); Also, there is room for research on the matter of time allocation for the test, since many students reported not being as satisfied with the time limit they had for the test's realization throughout all three aspects.

This paper proved to be a stepping stone for further research on the topic of proficiency tests' validity and washback effect, providing interesting data about what students consider about the test and can be made use of by the college itself in order to help determine whether the test should be recontinued after the pandemic has passed and *in persona* classes and assessments have made a return.

7 - REFERENCES

AKBARI, R. Validity. **The TESOL encyclopedia of English language teaching**, 2018, p. 1-6.

ARKOUDIS, S. English language development faces some testing challenges. **The Australian Higher Education, University of Melbourne**, 2011, October, p. 33. Available at: http://www.theaustralian.com.au/higher-education/english-language-development-faces-some-testing-challenges/story-e6frgcjx-1226164268886. Retrieved on April 20, 2021.

BACHMAN, L.. **Fundamental considerations in language testing**. Oxford: Oxford University Press, 1990.

BROWN, H. D. **Principles of language learning and teaching**. New York: Longman, 2000.

DAVIES, Alan et al. **Dictionary of language testing**. Cambridge University Press, 1999.

ETS. **Test taker handbook.** Available at: https://www.ets.org/s/toefl_itp/pdf/toefl_itp_test_taker_handbook.pdf. Retrieved on May 15, 2021.

OCKEY, G.J.; WAGNER, E. (Ed.). **Assessing L2 listening**: moving towards authenticity. Amsterdam: John Benjamins Publishing Company, 2018.

GATTI, B.A. Avaliação: contexto, história e perspectivas. **Olhares**: Revista do Departamento de Educação da Unifesp, v. 2, n. 1, p. 08-26, 2014.

HENNING, G. **A guide to language testing**: development, evaluation, research. Rowley, Massachusetts: Newbury House, 1987.

HUONG, N. T. H. . Face validity of the institutional English based on the common European framework of reference at a public university in Vietnam. **VNU Journal of Foreign Studies**, v. 36, n. 1, p. 81-102, 2020.

KOZOL, J. **The shame of the nation**: the restoration of apartheid schooling in America. Crown, 2005.

LAERD Dissertation. (2021). Non-probability sampling. *Dissertations and theses: An online textbook*. Retrieved from https://dissertation.laerd.com/

LOTHERINGTON, H. What four skills? Redefining language and literacy standards for ELT in the digital era. **TESL Canada Journal**, v. 22, n. 1, p. 64-77, 2004.

PRODANOV, C.C.; FREITAS, E.C. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico, 2ª Edição. Novo Hamburgo, RS: Editora Feevale, 2013.

QUEVEDO-CAMARGO, G. Formação de professores de línguas adicionais e letramento em avaliação: breve panorama e desafios para os cursos de licenciatura em LEM no Brasil. **Calidoscópio**, v. 18, n. 2, p. 435-459, 2020.

QUEVEDO-CAMARGO, G.; SANTOS, G.M. Validade de conteúdo das provas de um curso de compreensão oral para fins acadêmicos: relato de uma experiência. **Olhares & Trilhas**, v. 22, n. 2, p. 289-308, 2020.

QUEVEDO-CAMARGO, G.; SCARAMUCCI, M.V.R. O conceito de letramento em avaliação de línguas: origem de relevância para o contexto brasileiro. **Linguagem**: Estudos e Pesquisas, v. 22, n. 1, p. 225-245, 2018.

SOUZA, M.N.M. de. **Avaliação formativa em matemática no contexto de jogos**: a interação entre pares, a autorregulação das aprendizagens e a construção de conceitos. 2019. 195 f. Dissertação (Mestrado em Educação)—Universidade de Brasília, Brasília, 2019.

TAUFIQ, W.; SANTOSO, D.R.; FEDIYANTO, N. Critical Analysis on TOEFL ITP as a language assessment. **Advances in Social Science, Education and Humanities Research (ASSEHR)**, volume 125, **1st International Conference on Intellectuals' Global Responsibility (ICIGR 2017)**. Atlantis Press, 2018. Available at https://www.researchgate.net/publication/323179784_Critical_Analysis_on_TOEFL_ITP_as_A_Language_Assessment. Retrieved on May 15, 2021.

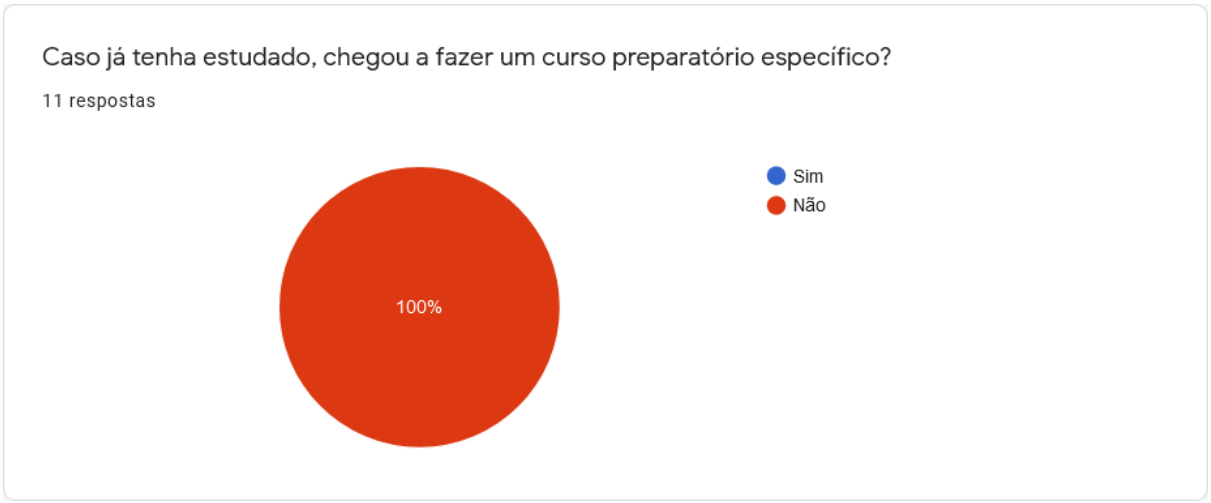
7.1 - Websites

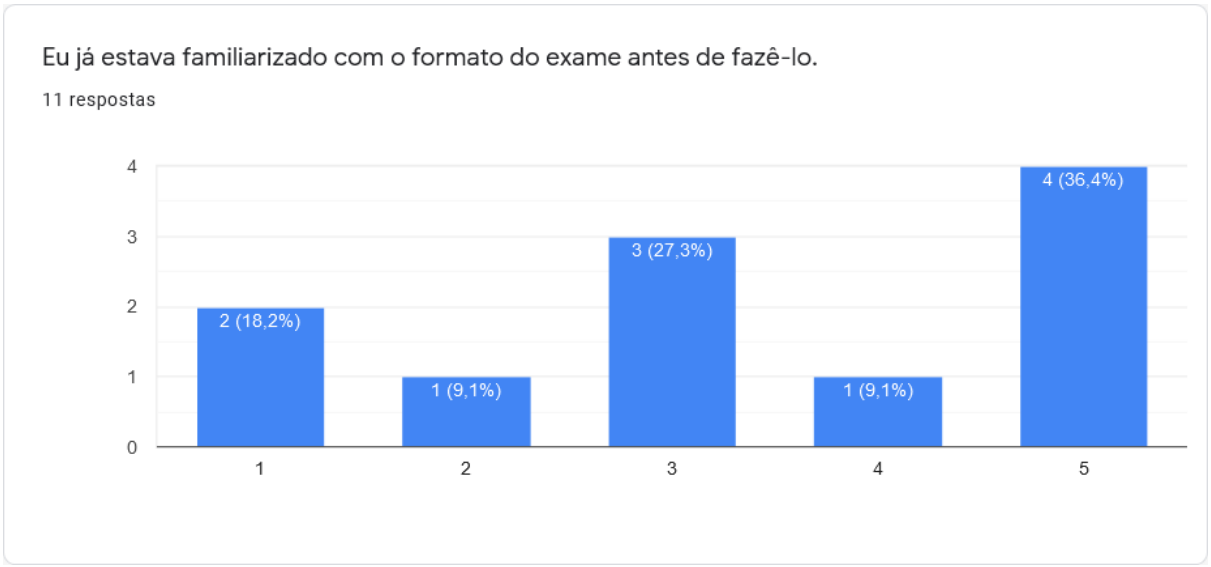
www.ets.org/toefl_itp/content . Last accessed on May 16th, 2021.

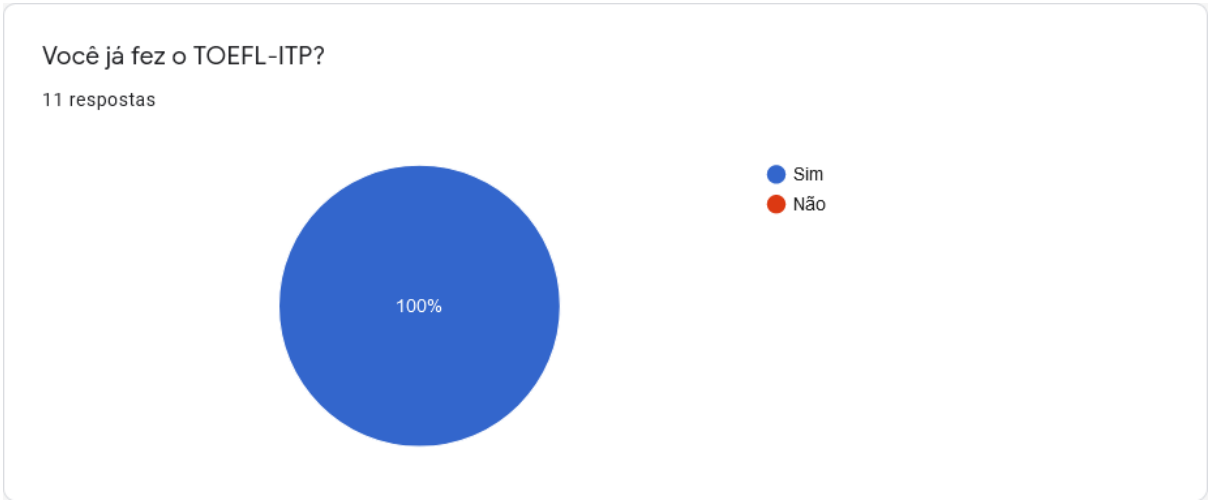
www.fundinguniverse.com . Last accessed on May 16th, 2021.

8 - ATTACHMENTS

8.1 - Graphs from the 11 TOEFL-taking students

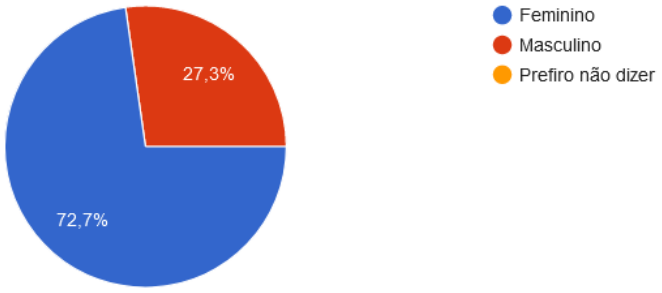






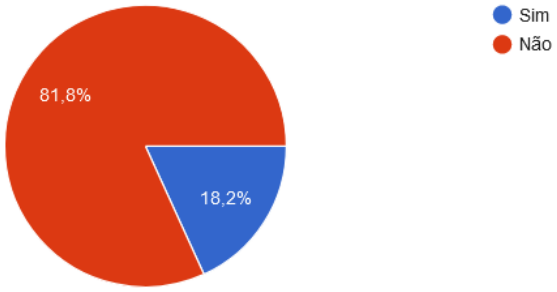
Gênero

11 respostas



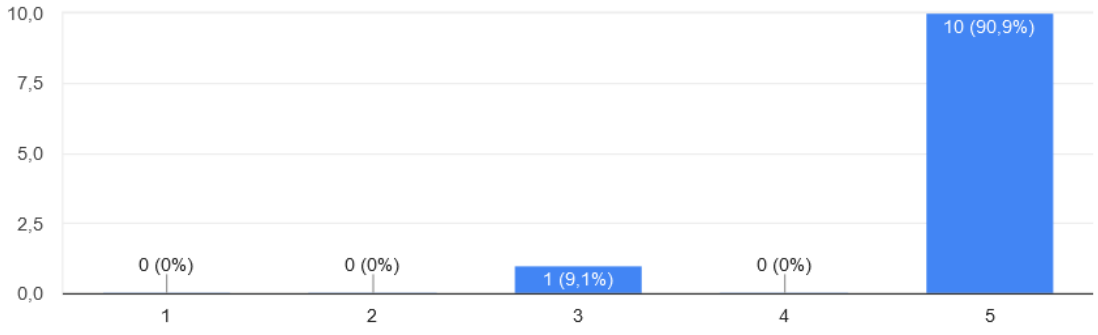
Você já estudou para o TOEFL-ITP?

11 respostas



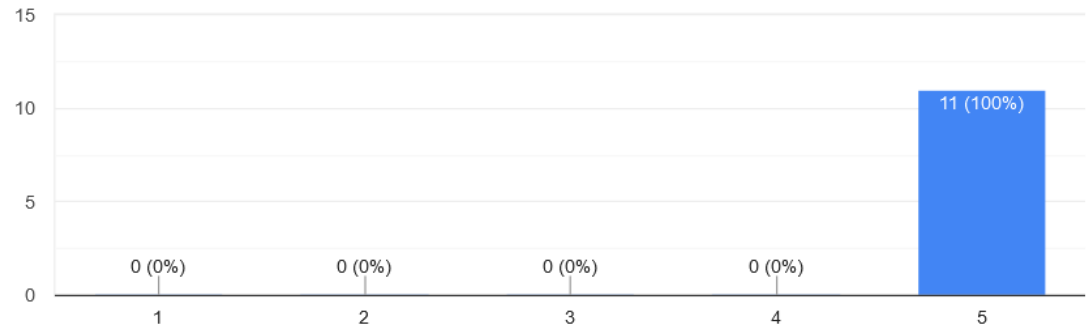
Os enunciados e alternativas desta parte foram claros.

11 respostas



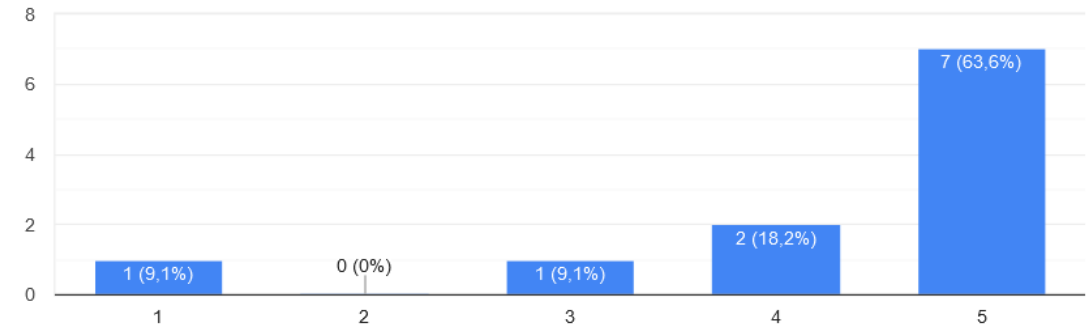
As instruções para a realização desta parte foram claras.

11 respostas



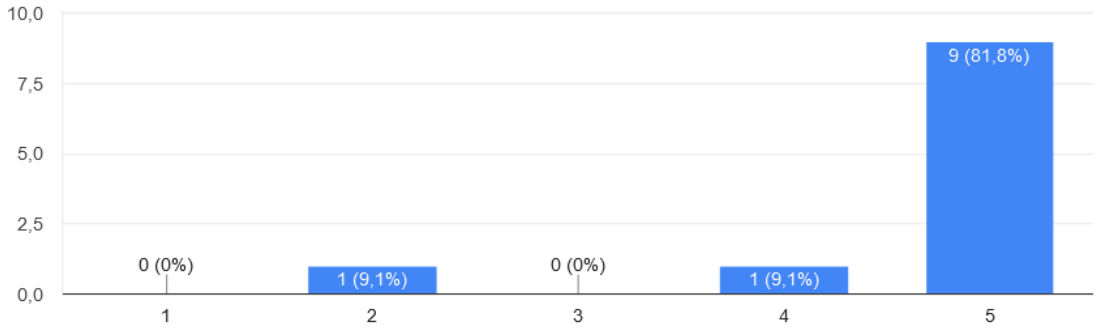
O tempo de realização para essa parte foi bom.

11 respostas



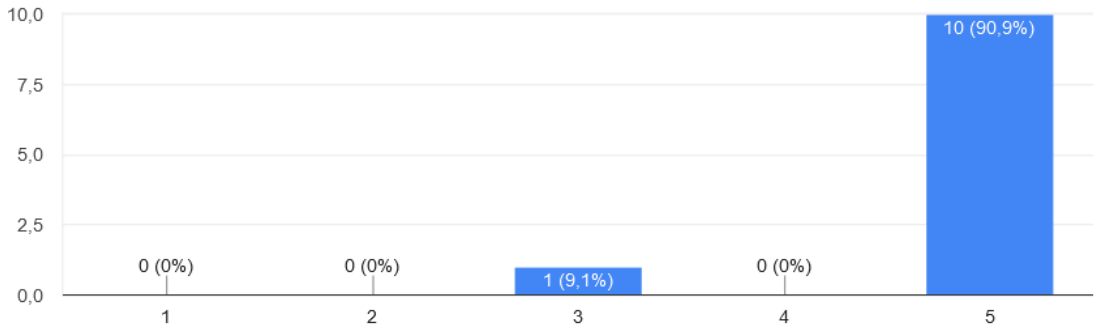
Os enunciados e alternativas desta parte foram claros.

11 respostas



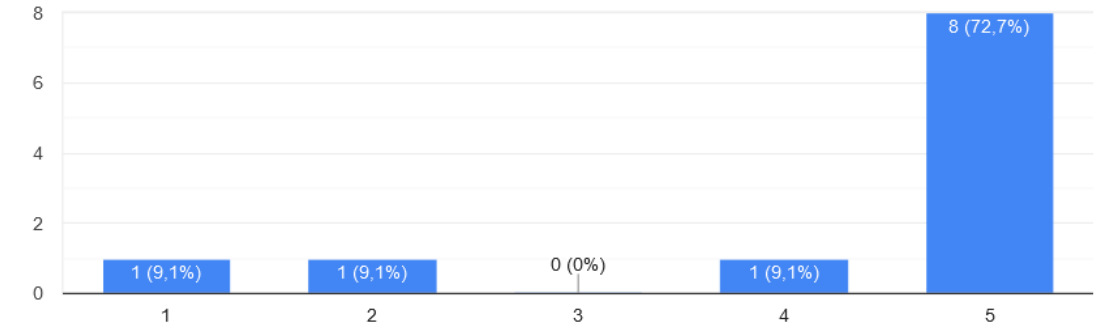
As instruções para a realização desta parte foram claras.

11 respostas



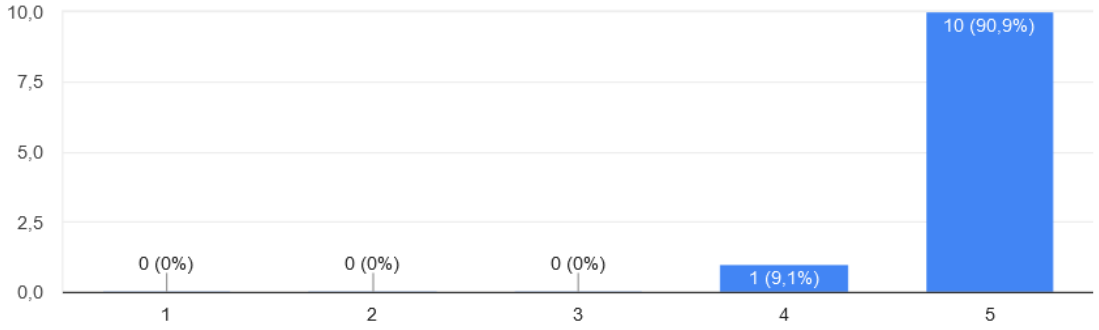
O tempo de realização para essa parte foi bom.

11 respostas



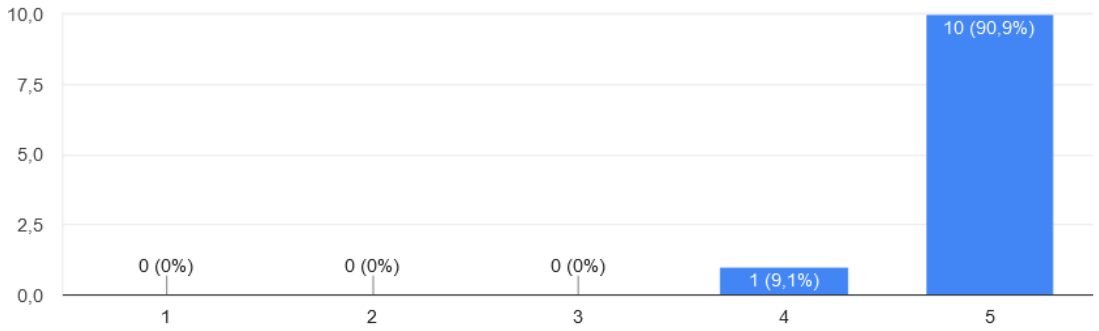
Os enunciados e alternativas desta parte foram claros.

11 respostas



As instruções para a realização desta parte foram claras.

11 respostas



O tempo de realização para essa parte foi bom.

11 respostas