“La evaluación ha cambiado”: A Case Study on Construct and Face Validity in the Transition from Face-to-face to Computer-Assisted L2 assessment

Josefina Cataldo, Sofía Herrera, Álvaro Hidalgo, Paulina Oyanedel, Katherine Saavedra, Juan Torres & Emma Santibáñez

Universidad de Chile

Facultad de Filosofía y Humanidades

Departamento de Lingüística

Informe Final de Seminario de Grado para optar al grado de Licenciado/a en Lingüística y Literatura Inglesas

Profesor guía: Daniel Muñoz Acevedo

Santiago de Chile

Diciembre 2021

Acknowledgements

Entrar a estudiar a esta universidad fue un sueño para mí, estar culminando esta etapa es sentir que lo estoy cumpliendo por completo. Fue desde el comienzo, sin duda, un desafío, mucho trabajo y mucho que aprender, debo decir que lo pasé bien estando cada semana llena de cosas que hacer y a penas tener mis fin de semanas libres. Suena extraño pero aprender es una experiencia demasiado enriquecedora como para sentir otra cosa. Claramente implicó un esfuerzo y también un cansancio que pude sostener gracias a la ayuda de mis seres queridos.

Por sobre todo quiero agradecer a mi madre que además de tenerme diez meses dentro de ella es mi gran maestra, te amo infinitamente. Su apoyo incondicional, inteligencia y fortaleza han hecho de mí la mujer que soy hoy.

No menos importante es mi abuela quien también es mi mamá, gracias a su amor y comprensión pude seguir cumpliendo con cada desafío académico.

Gracias a mi tata que está en el cielo y que me acompaña y protege siempre, gracias por enseñarme a ser buena persona e introducirme al mundo del inglés con tus clases particulares, el mejor profesor de inglés siempre serás tú, te amo.

En especial quiero agradecer a mi pololo Matías quién llegó a mi vida a mejorarla sustancialmente sin previo aviso. Sin imaginarlo construimos nuestra propia familia y hoy somos tres junto a nuestra hija Lisa que nace en un par de meses. Gracias a él y nuestra hija mi vida es mejor día a día y camino acompañada por cualquier camino que la vida decida llevarme. Gracias por ser mi mejor amigo, mi partner incondicional y por complementarme con tu inteligencia, te amo muchísimo.

Gracias a ti mi Lisa por acompañarme la mayor parte de este año con tus movimientos y pataditas que me dieron ánimo cuando estaba muy cansada, gracias por enseñarme día a día lo que significa ser mamá.

Gracias a cada amiga que estuvo a mi lado en todo este trayecto, haciéndome reír, escuchándome y apoyándome cuando fue necesario, las quiero mucho.

Quiero agradecer a mis profesores que son unos secos y nos compartieron tantos conocimientos, los admiro profundamente a cada uno. En especial quisiera agradecer a nuestro profesor guía Daniel que sin presionarnos ni estresarnos logró sacar lo mejor de nosotros y nos apoyó hasta el final en este arduo proceso.

Muchas gracias a mis compañeros de tesis por lograr entre todos generar un buen ambiente de trabajo, compañerismo y apañe. Los admiro a cada uno por sacar adelante este trabajo de forma remota, tarea que no es nada fácil.

Por último, agradecer a la Facultad de Filosofía y Humanidades de la Universidad de Chile por ser un lugar en el que realmente se siente humanidad entre las personas, donde los problemas de uno son los problemas de todos y donde los derechos humanos importan antes que cualquier cosa. Gracias por hacerme sentir en casa, segura y que el mundo tiene personas buenas en él.

Sofía Herrera Vásquez.

A mi familia, seres queridos, equipo de tesis, profesor guía y profesores durante la carrera, estaré eternamente agradecido.

Juan Torres Millar.

En vista de lo caótico y cambiante que ha estado el planeta durante estos últimos años, jamás me imaginé terminando mi carrera. Mucho menos, me imaginé redactando mis agradecimientos en una tesis de grado, la cual terminé junto a un grupo de compañeres y un profesor guía bacanes. Es por ello, que quiero partir agradeciendo a elles, que siempre estuvieron dispuestos a sacar esto adelante, trabajando siempre en equipo en contra de todas las adversidades que pudo acarrear otro año de clases online. Sin elles, por supuesto, hubiera sido imposible terminar esta tesis, que se hizo con cada granito de arena que mis compañeres, el profe Daniel, y yo, pusimos a lo largo del año.

Quiero agradecer igualmente a mis papás y a mi hermana, que siempre estuvieron apoyándome y dándome ánimos para sobrellevar el estudio y la batalla contra el monstruo de la ansiedad. Con respecto a esto último, también quiero agradecerle a mi psicólogo Juan, que siempre estuvo presente, ayudándome y entregándome herramientas para sobrellevar las distintas emociones negativas que conlleva la carga académica, y, por qué no decirlo, el desarrollar un trabajo tan grande como lo es esta tesis.

También agradecer a mis amigues, por aguantarme cuando me ponía pesada, estresada y llorona. Gracias por ser tan geniales y por siempre estar disponibles cuando necesitaba unas buenas risas y unos buenos abrazos con palmadita en la espalda. Siempre serán pilares en los que sé que me puedo apoyar.

Finalmente, quiero agradecerle a mi perrita Freya, a mi gato regalón Gusanito y a mi pololo Raúl, por ser siempre respiros de aire fresco, y obviamente, por entregarme todo el amor, la comprensión y la felicidad que pude necesitar durante un año tan importante para mí. Sin mi

compañero y estos dos seres de naricitas mojadas, jamás hubiera tenido las fuerzas ni las ganas para llegar a donde estoy ahora.

Y ya, chao a este año y a esta etapa. Gracias, muchas gracias a todes quienes fueron parte de mi vida durante este proceso, y espero de todo corazón que a todos nos espere un futuro bueno, tranquilo y bonito <3 <3 <3.

Josefina Antonia Cataldo Rojas.

Para empezar, quiero agradecer a nuestro grupo de seminario porque fueron parte fundamental en este proceso. Chiques, gracias por el apañe, por el esfuerzo y dedicación, que a pesar de haber trabajado desde la virtualidad, pudimos llevar a cabo la tesis de la mejor manera posible. Son personas bacanes y super inteligentes, no dudo que llegarán muy lejos. Agradecimiento especial a las dos personitas que hicieron más llevadera la universidad (La Z & la L jaja) gracias por su amistad, por las risas, las conversas y por todo en general, les quiero y les deseo lo mejor. También agradecer al profe Daniel por su apoyo en este proceso, por su buena onda siempre, por tener la disposición de ayudarnos cada vez que lo necesitábamos, por sus enseñanzas y por guiarnos y motivarnos a terminar esta tesis. Agradecida del equipo que me tocó y también de les profes de la carrera.

También quiero darle las gracias a mi familia, a mi mami, a mi papi y a mi hermana. Son la familia más bacán que el universo pudo haberme entregado. Gracias por su apoyo incondicional y por seguir igual de unidos. Los amo mucho, estoy muy orgullosa de ustedes y espero que ustedes también estén orgullosos de mí. A mi weli Cristi, que gracias a ella pude estudiar sin preocuparme por pagar la universidad y por ser tan chistosa jaja. Agradecer también a mi pololo, por apoyarme, subirme el ánimo y darme mucho amor cuando lo necesitaba. Son personas muy importantes para mi y agradezco que sean parte de mi vida.

Ya que soy de una ciudad chiquita llamada Cabildo, creo que es importante decir que Santiago me entregó algo que no voy a olvidar nunca y que sigue siendo parte de mi y de lo que soy. El baile. Ir a bailar dancehall, ir a bailar twerk fue un escape del estrés de la universidad, fue una terapia, un reencuentro conmigo misma y una inyección de amor propio. Por eso y mucho más, gracias bailecito.

Finalmente quiero agradecer a mi minina y a mis cachorras (que llegaron justo en pandemia) porque cuando no estaba en mis mejores momentos, me quitaban todo lo malo con sus besitos y amor. Espero que en el futuro tratemos a todos los animales como tratamos a nuestras propias mascotas.

Gracias vida, gracias universo por toda la gente que conocí en este proceso, me llevo los mejores recuerdos.

Paulina Belén Oyanedel Pérez.

Hay tanto que agradecer que simples palabras no resultan suficientes para decir gracias a quienes acompañaron este proceso, esta escuela que fue la universidad. No podría empezar sin agradecer a todas las mujeres que abrieron este camino hace más de un siglo para que yo pueda estar aquí y ahora. Infinitas gracias por haber construido un camino que aún no terminamos de pavimentar pero que cada uno de los pasos que damos lo vuelve más seguro.

La universidad la viví en un proceso de profundos cambios sociales que me hicieron encontrar un rumbo el cual quiero seguir y desde dónde quiero y espero poder ser un aporte. Estoy agradecida de todas las personas que me encontré en el camino, luchando, conversando, reflexionando, sin duda fue un privilegio y algo que me permitirá ahora construir mi futuro con profunda consciencia de clase y género, sabiendo que todo lo que aquí viví fue un privilegio.

Y si bien soy una mujer profundamente privilegiada por estar aquí y aprender lo que aprendí, nada hubiera sido posible de no ser por mis mapadres. Gracias a mi mamá por darme la oportunidad de volver a empezar cuando lo necesite, por impulsarme a seguir cada vez que se volvía casi imposible, por ser el apoyo que siempre estuvo, ha estado y estará. Te admiro, sigo y amo con todas mis fuerzas, gracias, infinitas gracias por ser madre, amiga y compañera. A mi padre que hizo posible tantas cosas que él no tuvo el gusto de hacer ni tener, gracias por darme todo lo que necesite, por volver tangibles mis sueños y deseos. Te amo papá.

A Alexandra y Nicolás, mis hermanes, que siempre han creído en mis capacidades y han estado ahí para volver cualquier proceso más llevadero. Las risas, conversaciones y ridiculeces, son sólo nuestras y nada sería igual sin ustedes a diario.

Cinco años duró esté proceso, en donde me acompañé de hermosas personas, que aunque algunas ya no están fueron también parte del aprendizaje que me llevo. Gracias a mis amigas con

quienes nunca cesaré de tener recuerdos, es que la pasamos tan bien, fuimos niñas jugando a una temprana adultez, disfrutando y riendo cada viernes en Calama. Gracias Florencia y Niamh por siempre estar, el amor es infinito para ustedes, amigas del alma.

A mis bebés de cuatro patas. Charlito que me acompañó y comprendió incondicionalmente todo este camino. Al gato Salvador por hacerme entender que el amor no es justo muchas veces y que no se debe entregar por esperar recibir. A mi Fidelito a quien se supone salvamos, pero estoy segura que fue al revés, gracias por enseñarme de sencillez y humildad. Gracias mis wawitas por ser mi hakuna matata, les amo por la eternidad.

No puedo terminar de agradecer sin decirle infinitas (y de verdad infinitas) gracias a mi compañero, Nicolás. Me has acompañado incondicionalmente estos dos últimos años, me has impulsado y desafiado tanto como te lo he permitido. Me has hecho crecer, creer y sentir que soy capaz de todo lo que me proponga. De seguro, nada sería igual en este momento sin ti, gracias por todo el amor, las risas, los viajes, momentos y el maravilloso aprendizaje de poder construir con respeto, que es tanto y más importante que todo lo que pude aprender en la universidad. Gracias, gracias, gracias, te amo sin límites!

A mis abuelas y tata, mi poroto, babita, equipo de tesis, profesor guía y todes quienes inspiraron e inspirarán este camino.

Katherine Saavedra Sánchez.

En primera instancia, quisiera agradecer a nuestro grupo de seminario. Nos tocó realizar la tesis en tiempos pandemicos bastante turbulentos para todos, en los cuales muchas veces nos enfrentamos a situaciones personales dificiles y cada uno supo sacar fuerza de donde no las había para seguir con este proceso. Un agradecimiento especial a mis amigos Paulina y Juan, chiquillos míos, su apoyo ha sido fundamental en todo el proceso. No habría sobrevivido la universidad sin ustedes, los tkm. A nuestro profesor guía Daniel Muñoz, por su indudable pasión en lo que respecta a la evaluación y por atender todas nuestras inquietudes no solo respecto a la tésis sino también de la vida en sí misma. El resultado final de esta tésis no habría visto la luz sin ninguno de ustedes, infinitamente agradecida.

Quisiera agradecer a mi familia, a mi hermana Lucia, que aburrida sería la vida sin tí; solo tu sabes convertir lágrimas de tristeza en lágrimas de risa y ese es tu superpoder, gracias por estar conmigo desde siempre y por ser mi mejor amiga, infinitas gracias por tu amor incondicional. A mi hermano Jorge, gracias por tus consejos y tu paciencia jefecito. A mis tías Paula y Alejandra, gracias por no haber soltado mi mano desde que mi papá falleció y por haberme dado todo el amor y la contención necesaria para ser quien soy el día de hoy.Por hacerme parte de sus familias y quererme como una hija más, les estaré siempre agradecida. A mi tío Rodrigo por sus consejos y su cariño, a mis primos Tomás y Rodrigo por enseñarme a tener paciencia y a recordarme que también fui niña. A mi tía Angélica y tío Marcelo, que me han tratado como a una hija, les agradezco infinitamente su preocupación y amor. A mi mamá, Selma, que me ha dado las herramientas para enfrentar cada momento difícil y me ha formado en carácter. A mis abuelos que ya no están conmigo, y aquí me extenderé un poquito. Querida abuelita Selma, gracias por haberme cuidado cuando era una cachorra de tan solo 5 años, me recibiste en tu hogar y me tomaste como una hija, me cobijaste y me diste todo aquello que hace felices a los niños: juegos, ilusiones, amor,

alegría. Querido tata Zenón, sin duda tu partida el año pasado por el covid fue uno de los eventos más duros de toda mi corta existencia. Dejaste un vacío en mi que aun estoy aprendiendo a sobrellevar. Fuiste abuelo y papá y te juro que nadie pudo haberlo hecho mejor. De ti aprendí que hay que seguir adelante no importa que tan dura sea la situación, siempre habrá una pequeña luz al final del camino. Querida familia, si hoy estoy de pie, presentando esta tésis, es gracias a ustedes. Son el cimiento de todos mis valores y los amaré incondicionalmente siempre.

Quisiera agradecer a la luz que apareció en mi vida a mitad de año. Mi compañero de vida, Alejandro De La Torre; gracias por el amor que me has otorgado y por tu paciencia sobre todo cuando estaba con los ánimos por el suelo. Por los tecitos y los cafecitos. Tu abrazo y tus cariños han sido parte crucial para culminar este proceso. Gracias por llenarme de ilusiones y nuevos sueños que sé construiremos día a día juntos. Te amo.

A mis amigos, que han estado conmigo todos estos años, en especial a mi querido Benjamin Osses. Benjita, cuando el mundo se me vino abajo y me quiso aplastar, tú, con tus palabras, salidas al río, abrazos y tu amor incondicional lo levantaste.

Finalmente, a los hermanos pequeños que me han acompañado; mis 13 gatos: flochi, mota, llamon, astri, tito, sr. chiquitito, kenai, chiki, ginny, duquesita, judith, lisa y la princesa guatona. Mis dos caninos bellos: la luna y la muñeca. Una mención honrosa al canino bobby y la felina bobba. Gracias a ustedes mi vida es multicolor; tienen un lugar especial en mi corazón cada uno de ustedes.

We finally did it y como diría mi tata “díganme licenciada” ¡Gracias!

Emma Andrea Santibáñez Cruz.

Estos cuatro años han sido sin duda alguna una experiencia inolvidable, los buenos recuerdos los guardo bajo la almohada y los malos los llevo conmigo para nunca repetirlos. Conocí nuevas personas que no solo me ayudaron en sobrellevar este nuevo proceso, sino que me apoyaron en momentos difíciles, su impacto en mi vida es irreversible, para bien o para mal.

Quisiera expresar mis más sinceras gratitudes a cada uno de los integrantes del equipo que realizó esta tesis, sin elles nada de esto hubiera sido posible. Agradezco su paciencia y diligencia, siempre dispuestos a ayudar y preguntar, logramos juntes armar este trabajo de manera remota, lo cual me parece simplemente increíble; sin les cabres y su increíble apañe nunca hubiera podido dar el siguiente paso, trabajar con ustedes fue una experiencia única, me enseñaron el valor de la colaboración y la comunicación. Dentro de ese equipo incluyo también agradecimientos al grandioso profesor guía que nos ayudó y lideró no solo durante la etapa de aprendizaje de las temáticas que luego nos permitirían desarrollar este estudio, sino que también su preocupación y motivación durante el proceso de llevar a cabo la investigación y reporte.

También reconocer el invaluable apoyo que recibí de parte de todos mis profesores, que desde mi etapa escolar hasta mi etapa universitaria me han brindado las herramientas adecuadas para lograr mis aspiraciones. Igualmente quiero agradecer al departamento de Filosofía y Humanidades de la Universidad de Chile por brindar los espacios adecuados no solo para llevar a cabo mis estudios, sino que también un ambiente en el cual desarrollarme como individuo.

Es increíblemente importante agradecer a mi familia, la cual me ha apoyado desde siempre, como sobrino y nieto mayor en la familia de mi madre he sido utilizado como ejemplo frente a mis primos, pero el ejemplo original fue el que ellos me dieron a mí. Sobre todo, agradezco desde lo más hondo de mi corazón a mi madre, Ivon Cuevas, la cual ha sido una montaña a mi lado, desde

la muerte de mi padre que ambos nos hemos apoyado uno del otro, siempre apoyándome y motivándome a buscar nuevos desafíos, preocupándose simultáneamente de que no me esté sobre esforzando, ella es la razón que me esfuerzo cada día y ella es la razón por la que cerrar exitosamente este proceso significa tanto para mi y mi familia.

No menos son mis amigos, a los cuales les agradezco las risas y momentos inolvidables que me ayudaron a mantener la sanidad durante estos cuatro años, incluso luego de estar lo más hondo que he estado en mi vida algunos de ellos no se rindieron en mi y me ayudaron a reconstruir mi vida. Seba, nos conocimos en mi peor momento y aun así siempre estuviste dispuesto a juntarse y discutir lo que nos afligía durante la semana o nuestras metas más recientes; por ello te agradezco infinitamente. Fuiste mi apoyo principal cuando la carrera, o la vida, me superaba y jugaba con la idea de rendirme siempre me tiraste para arriba y motivaste a seguir intentando, sin tu guía quien sabe que hubiera sido de mí. Miguel, Gum para los admins, es quien me entiende desde lo más dentro de mí, ambos vamos por esta vida fingiendo que sabemos lo que hacemos. Has sido una influencia increíblemente positiva, pertenecemos a grupos distintos de cierto modo, pero algo como ello nunca nos ha detenido en nada, y no empezaremos ahora, por ello estoy realmente agradecido.

Finalmente quiero agradecer a todos aquellos que fueron parte de mi vida en algún punto y me brindaron su apoyo y cariño, incluso si nos separamos en malos términos o simplemente no mantuvimos comunicación es innegable que fueron parte en la formación de quien soy hoy, me disculpo por los males que provoqué y espero que algún día me pueda volver a presentar, seguro de quien soy y activamente mejorando como persona.

Álvaro Esteban Hidalgo Cuevas.

Table of contents

List of Tables

Abstract

As a response to the global pandemic caused by COVID-19, Chile was forced into a transition from face-to-face format to online computer-based tests to measure L2 proficiency. In our programme, *Licenciatura en Lingüística y Literatura Inglesa* this was echoed in a new format for the L2 proficiency assessment called Oral Project. By using a matrix of dimensions describing Computer Assisted Language Testing (CALT) proposed by Suvorov & Hegelheimer (2013), we observed the changes in construct validity of an English L2 oral proficiency test produced by the transition from a face-to-face to an online test. We elaborated a survey based on the same dimensions to see if face validity had also been affected. The survey was applied to 24 undergraduate students of an English Language and Literature programme in Chile who had taken the two versions of the test. Results indicated that the change of format affected construct validity in several dimensions. However, face validity was not affected as students, even when identifying the changes of the test, still believed that it was measuring the same construct. These results suggest that the relationship between construct and face validity is complex and that the main construct and test purpose are strong factors when establishing face validity.

Key words: l2 assessment, online assessment, construct validity, face validity, oral project, language proficiency, face-to-face format, online format.

1. Introduction

Among changes observed in education within the pandemic context, the ones concerning teaching, learning and assessment have been both sudden and unforeseen. When COVID-19 pandemic hit our lives worldwide, schools and universities faced unexpected closures and most people started working and studying from their homes.

In the field of L2 learning, the last two years have allowed us to advance in the replication of teaching and learning practices that had been done physically via the utilization of a computer. The use of technology to this purpose includes language assessment and testing by using computer-based tests with streaming video, oral-response-recording options, automated (or human-generated) feedback mechanisms, and dynamic or enhanced input (such as hyperlinked text, answer choices, or interactive objects) in several fields (Winke & Isbell, 2016).

Indeed, language testing through technology has been a matter of concern among researchers in the field of teaching and learning process, especially today as technology is being intensively used worldwide. Literature on this topic clearly establishes that a change of a test to a computer-based format implies a change in the ways in which language is being taught and acquired (Gacs et al, 2020). When talking about Computer-Assisted Language Testing, researchers agree that it is fundamental to know if changes of test format cause a change in the construct that it is being measured. That is why many of the studies concerning online assessment seek to answer if evaluations are correlated with their face-to-face counterpart regarding what is being measured and if that is maintained in a computer-based test (Dooey, 2008).

At the same time, when tests change their format, the perspective of test-takers about the construct that the test is measuring becomes very important. This is because test-takers' perceptions should be considered to achieve high acceptance of the new test. This acceptance means that

stakeholders believe that the implementation of changes does not imply a loss of test quality (Sato & Ikeda, 2015).

In Chile, the Ministry of Education implemented a virtual system which made teachers and students use technologies such as computers, tablets, and cellphones. The implementation of distance, computer-based learning and teaching in Chile has opened several dimensions of concern: the provision of tools necessary for the schools, the devices necessary for the students, changes in the educational curriculum (as an appropriate response to the lack of face-to-face environments), mental health due to confinement, among others. For second language teaching, learning and assessment, the last two years have seen a profound transformation, as assessment was performed in a face-to-face format by default, especially because of the interactional nature of language performance.

All of these reasons motivated us, as L2 learners and students of linguistics, to examine the effects of the transition from face-to-face to online, computer-based testing of L2 proficiency in our program. More specifically, we wondered about the specific changes that have happened in the way we are being tested, how these have affected the validity of the tests, whether those changes have been noticed by test-takers, and what their perceptions are regarding the benefits of those changes.

To this purpose, we examined the case of a test of English L2 oral proficiency called Oral Project. Overall, the face-to-face version of this test was characterized by a situation where test-takers presented a topic in front of a teacher and an audience (classmates). After the presentation, students could go through a question-answer conversation with the examiners. The corresponding

oral performance was in general characterized by spontaneity and an important degree of improvisation.

From our experience as test-takers, we knew that these conditions were normally associated with test anxiety and its corresponding effects in our performance. However, the new online version of the Oral project Online, mediated by technological devices from home, made us think that perhaps now students could relax more when dealing with the presentation. If so, we assumed that the new format could allow the improvement of both proficiency in English L2 and the corresponding grades.

In view of the significant change in the test procedures and tools that is implied in the transition from face-to-face to online, computer-based testing, it seems relevant to observe potential changes in the construct validity of the test, i.e. what the new format of the test is evaluating regarding oral skills and proficiency in an L2. At the same time, it also seems relevant to observe possible changes in the test face validity, i.e. how the changes are being perceived by test-takers regarding the capacity of the test to measure what it measured in the original face-to-face format and the benefits of the new version.

To this purpose, we conducted a study consisting of an analysis of both face-to-face and online Oral Project tests using a set of dimensions proposed by Suvorov & Hegelheimer (2013) for the description of Computer-Assisted Language Testing. On the other hand, a survey was administered to a group of undergraduate students of the *Licenciatura en Lingüística y Literatura Inglesas* programme at *Universidad de Chile*, to observe the way they perceived the changes implemented regarding the skills that were evaluated in both versions of the test as well as their perceptions of their potential benefits.

We decided to carry out this case study to test our assumptions against empirical evidence. Furthermore, we believe that observing, identifying, and analyzing the differences between the face-to-face and computer-based tests can help to inform test developers and users to maintain the quality and precision of their tests when implementing the transition to online, computer-based language testing.

In the following chapter we will present the theoretical framework within which the study was conducted. Firstly, it includes the reasons for using technology in the language field as a response to the COVID-19 pandemic (in this particular test). Secondly, the topic of Validity in L2 language assessment and its characteristics. Thirdly, the construct and face validity of language tests. Fourthly, the differences between the synchronous and asynchronous assessments. Fifthly, the main features of L2 computer-based assessment. And finally, the CALT tools are explained alongside its dimensions.

2. Theoretical Framework

The Coronavirus (COVID-19) pandemic originated in 2019 in Hubei, province of the Republic of China, and has become a global crisis during the past three years. In early 2020, the World Health Organisation (WHO) Emergency Committee announced a global health emergency because of the increasing cases of the virus in many countries over the world (McAleer, 2020; Velavan & Meyer, 2020, in Priyo & Nugroho, 2020). As a consequence of the COVID-19 pandemic, there have been worldwide class suspensions, resulting in the need for online learning (Moorhouse, 2020, in Priyo & Nugroho, 2020).

The COVID-19 pandemic in Chile has been a major issue since the first case appeared in March 2020. As a response to the emergency, the Ministries of Education and Health decided to suspend face-to-face classes in schools and universities. The purpose of the suspension was to ensure that education continued while protecting students, teachers, and staff from contagion (Ministerio de Educación, 2020). Despite the plans of fully coming back to face-to-face classes in 2022 (Palma, 2021), the current situation for most Chilean students is that classes are carried out exclusively via online technologies. This has meant a big challenge not only for the learners but also for the teachers, who needed to implement new ways of teaching and evaluating.

Online education has been an available option since before the pandemic started, but COVID-19 prompted a sudden interest in this medium all over the world (Alghammas, 2020). The result is that many schools and universities started to offer classes based on full-time online contexts. According to Priyo & Nugroho (2020), online teaching means “a set of learning activities in a subject delivered through a network giving access and exchange of knowledge” (p. 53). The term was coined to indicate the approach of teaching and learning that involves the use of the internet and technologies related to it (Priyo & Nugroho, 2020). As a crucial process in teaching

and learning, second language instructors also need to evaluate their students' skills. Assessment enables students to guide their learning process and teachers to effectively implement their teaching methods (Alghammas, 2020).

For this purpose, a number of available platforms have been implemented in many Chilean educational establishments during the COVID-19 pandemic, such as Zoom, Gmail, Google Meets and Google Forms. Moreover, some institutions also decided to develop and implement their own platforms in order to facilitate the connection between students and teachers. For assessment purposes, there are platforms with assessment functions, such as online test software programs, sections to upload material, email clients, forums, and alerts to connect to class via Zoom or other video chat programs.

In the present chapter, we present the theoretical framework required to observe the effects of the transition from face-to-face to online, computer-based assessment of oral L2 proficiency on the validity of the tests used for that assessment. Accordingly, the following sections explain validity in L2 assessment and L2 asynchronous online and computer-based assessments.

2.1 Validity in L2 assessment

The property of validity in language assessment can be described from different perspectives. The most traditional one is that, if a test measures what it says it does, then the test is valid (Lado, 1961). Chapelle (1999) suggests that this might be a somewhat oversimplified definition, as it considers validity from an all-or-nothing point of view when validity refers to a very complex phenomenon. Accordingly, the idea has been developed over time that validity in L2 assessment is achieved by the contribution of many sources, such as the construct that is to be assessed, the reliability of the process, test consequences and perception of test users. However,

defined, the validity of a test can always be observed in the way test results can be interpreted (Akbari, 2012).

Assessment validation processes are therefore crucial in every instance of assessment, as they show how well a test is measuring what it is supposed to measure. In second language assessment, validity is crucial as it "affects all language test users because accepted practices of test validation are critical to decisions about what constitutes a good language test for a particular situation." (Chapelle, 1999) This means that validation processes are important not only for establishing if the test is appropriate for its purposes, but also to determine what should be changed for the test to be useful for any particular situation.

In the case of L2 assessment, validity is established in relation to the capacity of a language test to determine the proficiency, achievement, or aptitude of an L2 learner on its diverse linguistic skills. Whenever an assessment process changes, therefore, its validity may also change. This is why Dooey (2008) indicates that it is necessary to reevaluate validity whenever a change into computer-based assessment happens (2008). Thus, methods through which validity is measured should be applied to check that a test is still applicable and efficient to assess what it has to assess despite the changes implemented.

2.2 Characteristics of validity

Since validity affects all language test users (Chapelle, 1999), the accepted practices to make a test valid are fundamental to decide what is an appropriate language test in a determined situation.

According to Chapelle (1999), test validity is achieved when test reliability is confirmed, that is to say, when the test yields the same results under the same conditions. The validity of a

test requires also establishing its correlation with other tests that are already valid and reliable. This source of validity is generally called criterion-related validity. Expert judgment about test content and construct validity can also be used to observe the degree to which a test or other measure assesses the theoretical construct it is supposed to measure (in L2 assessment, a language skill). These methods involve the observation of results from empirical research to inform theory-based expectations. This source of validity is called content-related validity.

Messick, in Chapelle (1999), identifies different types of evidence that can come into play in validation. The first one is content analysis, which consists in judgements about the content relevance, representativeness, and technical quality of the test material. This gives evidence to know if there is a relation between test ítems and the construct they are meant to measure. The second one is empirical ítem or task analysis. This provides evidence to know if the hypothesized knowledge and processes are the reason for the learner´s performance. The third one is dimensionality analysis, which observes the internal structure of the test. This is carried out by evaluating the test response data to a psychometric model which must be in correspondence with the construct theory.

The fourth source of evidence of test validity is the investigation of relationships of test scores with other tests and behaviors. This type of evidence is obtained by using multitrait-multimethod (MTMM) to observe different constructs in a test. In this process, the evidence to know if a test is valid or not are found in the correlations among the tests of the same construct that are stronger than the ones that have different constructs.

A fifth source of test validity is the results from research on differences in test performance. It begins by establishing a theory which explains that the test should behave differently through

different groups of test-takers and other factors of the test such as test task characteristics. Then those characteristics that affect performance are generalized.

The last source of test validity evidence comes from testing consequences, including the social consequences, implications and interpretations made from particular test scores. This continues as an important issue to be investigated as has to do with the ways in which the test impacts the people involved in it. Testing consequences present a different dimension for a validity argument than the other forms because they involve hypotheses and research directed beyond the test inferences to the ways in which the test impacts people involved with it.

It follows, therefore, that when we want to observe the validity of a test when it changes from face-to-face to an online format, we should take into account the sources of test validity. The main source of validity is the verification of the construct that is evaluated, as explained in the following section.

2.3 Construct Validity

Construct validity refers to the degree to which a test or other measure assesses the theoretical construct it is supposed to measure. In the case of L2 assessment, this means we need to make sure a test is accurately measuring the language skills or knowledge that test developers and language instructors are trying to evaluate. For example, if a test evaluates the construct of L2 oral proficiency and includes a task of writing a script for a spoken presentation, then the test construct may be compromised, as the task requires the implementation of writing skills as well as oral skills. In order to demonstrate construct validity, we need evidence that the test is measuring only oral proficiency and also evidence that it is not measuring other skills, such as written production skills.

The construct of a measuring tool has to be extensively revised so that the test assesses everything it says is assessing and not anything else. Along these lines, Zuskin (1993) explains that a construct can include relevant sociolinguistic parameters such as intonation, gestures and facial expressions mixed with communicative skills usually assessed in L2 tests.

Changes in the construct validity of a test can be observed in many dimensions, such as the delivery format and the media density, and the target skill of the test. Therefore, construct validity has to be a priority for test developers involved in the transition of one type of test to another.

Zuskin (1993) continues about the importance that construct validity has by saying that test developers should focus on 'the grander scheme of things' referring to be able to ensure the reliability a test has and assessing what it should assess before "they can assess whether or not second language (L2) students know when to speak, when not, and as to what to talk about, with whom, where and in what manner".

In the case that construct does change in an evaluation, this will affect other areas tied to its validity. This means that a change in the construct could be rapidly noted by the students as different effects, ranging from less trust in the assessment to worse grades. The skills and abilities evaluated could be different from the original assessment, creating a situation where the changes in the construct are perceived by the test-takers. When this happens, test-takers views become essential evidence for validation regarding the face validity a test has. Even when the construct changes, if the students accept the new test as equally valid as the previous version, it is evidence that indicates the test did not lose any validity through the change of construct.

2.4 Face Validity

Holden (2010) defines face validity as "the degree to which test respondents view the

content of a test and its items as relevant to the context in which the test is being administered" (p. 1). This type of validity is verified by the judgments, perceptions, perspectives, and views that people have regarding a test construct. From this perspective, a test should not only be valid, but it should also appear valid (Mosier, 1947). This means that tests not only need to measure what they are supposed to measure, but they also need to be seen by stakeholders and users as measuring the construct for what it was made. Therefore, if an L2 proficiency test is measuring vocabulary adequately, but for some reason the test-takers perceive that the test is measuring another ability unrelated to the construct, then the general validity of the test is being affected by losing face validity.

Face validity is not based on the judgments of experts, but rather on the opinion of test-takers, depends on the obviousness of the test content, and is affected by the context in which a test is conducted (Holden, 2010). Face validity has to do with people's judgements toward the test and not with the empirical results of the test. In his review of the concept of face validity, Nevo (1985) highlighted that face validity should not be confused with other types of validity, such as content-related, content or construct validity. The main difference with those sources of validity is that face validity is "the most subjective member of the validity family" (Akbari, 2012, p. 31).

Because of its subjective nature, face validity is usually seen as the least important source of validity. For this reason, in some situations face validity can be overruled. For instance, a cloze test that measures proficiency may not appear to measure this skill to test respondents, but the theory behind the test supports its use for measuring language proficiency (Akbari, 2012). If this is the case, face validity can be dispensed as it is less relevant than construct validity. However, this should not be necessarily so. Indeed, Holden (2010) points out that, regardless of the accuracy of a test (i.e its construct validity and reliability), "the absence of face validity may result in test

respondents feeling dissatisfaction and anger, and of being cheated" (p. 1). In turn, low face validity in language testing might negatively influence the performance of test-takers if the test is perceived as irrelevant, therefore, scores may not reflect their ability (Sato & Ikeda, 2015). In other words, low face validity can result in low construct validity.

Sato and Ikeda (2015) highlight the importance of research on face validity and its benefits for language testing. They mention studies (such as Brown, 1993 and So, 2014) which show that "the test-taker perception of the test can be used to develop a fair and accessible test for all prospective test-takers" and "stakeholders' involvement in test development helps to improve test quality and leads to their acceptance of the test." (p. 3). From this point of view, test-taker perception in language testing is relevant to achieve high face validity and thus create the opportunity to measure the ability it is supposed to measure and thus contribute to its general validity.

In a case where a test that has been administered face-to-face changes to an online format, face validity may be affected if test-takers perceive that the construct that was measured in the face-to-face format changed when the test is taken via online. Therefore, if test-takers judge that the ability that was originally measured by a test changed or different constructs (skills or knowledge) have been included, there is a conceivable risk for the validity of the test to be affected.

2.5 From synchronous to asynchronous L2 proficiency assessment

The transition from face-to-face to online L2 proficiency assessment can be characterised as a change from synchronic to asynchronic teaching, learning, and assessment methods. According to Chen et al. (2007), synchronous media methods "are time-bound in which both the instructor and learners would need to be attending the session simultaneously". Such methods

include, for instance, videoconferencing, in which strong interactive capability is provided. On the other hand, an asynchronous system is one in which learners can work at their own places and preferred times" (Chen et al, 2007, p. 218). The most important feature of these methods is that learners are not present at the same time and can differ even in places. Such methods include the use of tools such as bulletin boards, e-mails, and pre-recorded videos.

When the pandemic threatened the educational system, videoconferencing was the preferred way of replacing the lack of body language interaction needed in L2 learning. In their study of the adoption of synchronous and asynchronous media in the teaching of a second language, Chen et al. (2007) found that participants have other types of problems when it comes to reaching media synchronous methods. In particular, participants' awareness of their exposure to a camera or video recorder seemed to cause them to feel "unnatural". This feeling, in turn, can cause a distraction from the general instructions. Even more, the extensiveness of technical functions of application packages could make both learners and instructors feel distracted and under stress.

In addition, the importance of having a bandwidth that is available and a network that must have exceptional traffic conditions become vital elements for making synchronicity work. This implies that learners' technological abilities and available resources affect the overall performance and the general output when learning a second language.

In conclusion "synchronous media give language learners more opportunities to practice authentic language, but before applying the media the instructors have to consider how learners adapt to such environments and their experience and ability to use the technology" (Chen et al, 2007).

Asynchronous systems can also be applied in L2 assessment. Evaluating pre-recorded videos and asynchronous tests have thus been the tendency along the pandemic period. When an asynchronous method is implemented, “participants are more independent and have more freedom in choosing when to participate in an instruction session” (Chen et al, 2007). Therefore, L2 learners have more chances to choose whether they feel comfortable enough to perform a task or not.

Nonetheless, one of the major problems with this system is the lack of interaction when instructions are given and the lack of interactivity among instructors and learners; even when this could be an advantage for those elementary learners who happen to have enough time to improve their own learning process at their own time. Also, when talking about skills and technical requirements, an asynchronous method is much more flexible than the synchronous one, since there is less pressure at the moment of recording a video than having to do a face-to-face oral presentation.

In conclusion, the transition from face-to-face to online L2 proficiency testing involves a movement from synchronous to asynchronous assessment. The changes involved in this transition affect the kind of performance that can be evaluated and the attitudes of test-takers and examiners towards the test.

2.5.1 L2 Computer-based assessment: Main features

Computer-based assessment takes advantage of technological tools for testing students such as computers, smartphones, tablets and other devices. Generally, the motivations to implement computer-based assessment include the desire to increase efficiency in assessment. However, in the context of the COVID-19 crisis, the transition to computer-based assessment has

responded to the need to keep the educational system working while protecting the health of the educational community.

Computer-based assessment (henceforth CBA) is defined as the assessment which is both delivered and scored by a computer (Nottingham Trent University, 2007). This kind of method can be used for formative and diagnostic assessment. Feedback that is essential in paper-based assessments (henceforth PBA), may or may not be automatically delivered to students. Tests are typically multiple-choice questions or objective question types.

According to Nottingham Trent University (2007) computer- based assessments has six main features and advantages as opposed to paper-based assessment:

- ➢ It can give detailed feedback in diagnostic, summative and formative assessments.
- ➢ Feedback can be delivered automatically.
- ➢ Assessments are scored automatically by a computer.
- ➢ It can be more flexible than PBA, since it can be available at any time and anywhere. Features such as the duration of the tests, the period in which the students are allowed to answer the test and the number of attempts permitted can fluctuate.
- ➢ Questions can be randomized in order to encourage students to work individually.
- ➢ Different questions can be reused or can be shared across different modules.

Although CBA holds these advantages against PBA, there are some cautions to be taken into account. For instance, novices in setting up CBA may take time to learn how to use this type of evaluative system. This means they will need to develop abilities and skills in order to create effective questions. Also, students may need some training to use the technological tools necessary to respond to the corresponding tests. This is so, especially in the current context, when the

transition from PBA to CBA has been very abrupt for educators and few people were really prepared to begin with these assessments immediately. Finally, the most important weakness of CBA is the impossibility of monitoring students at the moment of evaluation. For high-stakes summative assessments, additional processes are needed in order to prevent students from cheating and to have a back-up plan in case of system failure.

In conclusion, it seems that the adoption of computer-based assessment (CBA) is a reasonable way to overcome the need to teach and assess in a context where paper-based or face-to-face evaluations are not possible. However, the implementation of CBA brings about concerns regarding the recruitment of new abilities for the use of computer-based tools.

2.5.2 L2 Online assessment: Main features

Online assessment utilizes connection to the internet together with CBA tools. This allows online assessment to provide the same type of assessment as CBA with the exception of the possibility of being applied anytime.. This type of assessment utilizes the immediate transfer of data that the internet provides to test student's skills in real time, following a time limit for the assessment. This still allows for flexibility of test location and time for test-takers, as they can take the test as long as they have an internet connection and the means to use it.

As can be seen, his kind of assessment is heavily dependent on the circumstances of test-takers, as practical problems start to arise, such as laggy connections or no connection at all, downed links, and the possibility that a student could have access to the items that appear on a test and the answers to them..

As Dooey (2008) states "technology promises a range of capabilities, including speed, accuracy and efficiency" (p. 24), meaning that this type of assessment improves many aspects that

are reliant on human testers and replaces them with the accuracy and efficiency that coding and new technologies are capable of. Nonetheless, online-based assessment allows students to take tests without worrying about distance, making it possible for students that could never be present to be able to participate like all the others. Online-based assessment could also be cheaper and easier to code than a computer-based format, as Roever (2001) explains, it allows tests to be more accessible by teachers and students alike. All of this does not mean that any web-based assessment works for any case, but rather it allows testers to programme their tests to fit their necessities. In this sense, examiners can use self-scoring scripts or provide immediate feedback if the test is created in such a way that it can provide it as the student is doing the test.

In conclusion, online testing seems to allow for the necessary flexibility required by the need to assess test-takers remotely. However, the implementation of online testing implies the need to have access to adequate technological and internet equipment. Online based tests are a good tool to use during a pandemic that does not allow for face-to-face tests but at the risk of being more susceptible to changes in validity.

2.5.3 Computer Assisted Language Testing (CALT)

One of the new approaches regarding the assessment of language that appeared as an alternative to Computer-Based Assessment is called Computer Assisted Language Testing (henceforth CALT) or Computer Assisted Language Assessment (CALA), both terms being interchangeably (Mubarak, 2012). CALT focuses primarily on the assessment of language proficiency, as opposed to CBA, which encompasses different fields in the assessment through a computer.

Diverse definitions of CALT have been proposed, referring to it as an integrated procedure in which language performance is elicited and assessed with the help of a computer (Noijons, 1994, in Sulaiman and Khan, 2019). Computer-adaptive testing (CAT), the use of multimedia in language test tasks, and automatic response analysis are encompassed inside the CALT approach (Chapelle & Douglas, 2006, in Sulaiman & Khan, 2019).

According to Mubarak (2012) there are three major domains that permits the utilization of CALT: 1) generating tests automatically by using computers; 2) the online interaction between the computer and the test-takers; and finally, 3) the use of computers for the evaluation of the participant's responses. Because of these reasons, popular tests such as TOEFL, IELTS and DIALANG, among others, have shown an interest in using these new types of approaches (Mubarak, 2012).

This new computer assessment approach is different from CBA in the sense that even though assessment uses technological devices, assessment is fundamentally in the hands of teachers.

Advantages of CALT include, for example, what Meunier (1994, p. 23) explains as: ''… the opportunity for examinees to pace themselves; the ability to individualize tests; to make tests shorter; to promote a more positive attitude toward tests; to report test results immediately; to measure tests with precision; and to improve test security''. Furthermore, any type of assessment that uses computer technology creates other benefits, including greater authenticity, expedient delivery, digital record-keeping options, and automatic, objective scoring possibilities (Winke & Isbell, 2016).

The process of using CALT has been strictly related to the appearances of new technologies. Computers, multimedia, online gaming, tablets, smartphone, apps, and social media have presented a new environment for the entire world, changing the scenery of language learning over the past few years (Godwin-Jones 2011, in Winke & Isbell, 2016). Within this new landscape, three circumstances have been identified, as part of a process of ongoing normalization of tech-infused language assessment (Chambers and Bax, 2006, in Winke & Isbell, 2016):

> First, advances in technology are making it possible to have more sophisticated yet low-cost computerized testing programs that go beyond online multiple-choice testing. Second, commercial software packages for making online assessments are (or can be) embedded into popular learning management systems (LMS) With such tools, teachers can make short e-quizzes for students to take at home before they come into the classroom so that teachers know what to focus on most in their teachings (…) Third, the generation of languages learners typically found in the classroom has grown up with computers, tablets and smartphones. Earlier impediments that centered on computer access and familiarity are less concerning for mainstream test-takers (older adolescent and adult learners), and major testing operations have shifted almost exclusively to computer delivery. Thus, the field of computer-assisted language testing (CALT) has not just expanded, it has moved heavily toward normalization. (p. 2-3)

There are three main reasons to employ CALT: efficiency, equivalence, and innovation (Chapelle, 2010, in Sulaiman & Khan, 2019). *Efficiency* occurs when several technological tools are utilized regarding the interaction between a computer and language testing. Those tools are computer adaptive testing and analysis-based assessment, consisting of *automated writing evaluation* (AWE) or *automated speech evaluation* (ASE) *systems*. *Equivalence* happens when the

computerized tests are equivalent to paper and pencil tests, which are considered a *gold standard* in language testing (Suvorov and Hegelheimer, 2013). Finally, *innovation* refers to the idea in which technology becomes a real transformation in language testing, being CALT an instrumentalization of a new ability that selects and displays the necessary language, using necessary technologies, in particular contexts.

The use of CALT has gained tremendous popularity and acceptance in the field of *English as a Foreign Language* (EFL) teaching and *English for Specific Purposes* (ESP), both aiming for the particular linguistic training needs of students (writing, speaking, reading and listening) (Robinson, 2003, in Sulaiman & Khan, 2019). In order to fulfill those needs, gathering a big corpus of content, language, grammar, vocabulary, curriculum, and instructional materials is vital. This can help to test the communicative ability and efficiency of students in the classroom, testing the use of language in a certain context, for specific purposes, such as, business, medical, law, science, and technology (Sulaiman & Khan, 2019).

In order to characterize CALT resources and tests, a framework was proposed by Suvorov and Hegelheimer (2013). It consists of nine attributes, five of them exclusive to CALT, and the remaining four related also to traditional paper-based tests (Suvorov and Hegelheimer, 2013). Table 1 below summarizes the framework:

Table 1. Framework for the description of computer assisted-language tests

#	*Attribute*	*Categories*
1	Directionality	Linear, adaptive, and semi-adaptive testing
2	Delivery Format	Computer-based and Web-based testing
3	Media Density	Single medium and multimedia

4	Target Skill	Single language skills and integrated skills
5	Scoring Mechanism	Human-based, exact answer matching, and analysis-based scoring
6	Stakes	Low stakes, medium stakes, and high stakes
7	Purpose	Curriculum-related (achievement, admission, diagnosis, placement,progress) and non-curriculum-related (proficiency a and screening)
8	Response Type	Selected response and constructed response
9	Task Type	Selective (e.g., multiple choice), productive (e.g., short answer, cloze task, written and oral narratives), and interactive (e.g., matching, drag and drop)

Directionality

Suvorov and Hegelheimer (2013, p. 3) explain: ‘‘Computer assisted-language testing can be linear, adaptive, or semi adaptive’’. *Linear* means that the same number of test items are applied in the same order to the test-takers. In some of these tests, test-takers can review their answers, while in others, they can not. *Computer-adaptive testing* selects different items and displays them to the person being assessed, and then presents them in a sequence based on the test-taker's answer to each item. If the answer is correct, a more difficult item appears on the screen. On the contrary, if the response is incorrect, the level of difficulty of the following items decreases.

This type of test (CATs) is based on the item response theory (IRT) which follows two assumptions: a) unidimensionality and b) local independence. The former means that all test items measure the same construct, and the latter means the test-taker answers to each item independently

from the other responses (Suvorov and Hegelheimer, 2013). *Semi-adaptive tests* differ with the adaptive ones in that they are adaptive at the level of either group items called testlets or the test itself, which correspond to the proficiency of the test-taker on a pretest.

Delivery Format

Language tests taken with the help of computers are divided into *computer-based tests* (CBTs) and *web-based tests* (WBTs), both of them considered a part of computer-based assessment, specifically by engaging language. The first one is related to offline software, such as CDs, DVs and standalone software applications installed in a particular device; the second one includes performance assessed by online technologies (Suvorov and Hegelheimer, 2013).

Media Density

Computer assisted language testing has the advantage of the availability of different media formats and the possibility of their integration. CALT may be *single medium* (audio-only listening test or a text-based reading test) or *multimedia* (e.g., listening test with video or reading test with images and text) (Suvorov and Hegelheimer, 2013). Although the use of multimedia provides the opportunity to increase the authenticity of language tests, this also might pose a threat to test validity because this can also result in a more complex construct to measure (Douglas and Hegelheimer, 2007, in Suvorov and Hegelheimer, 2013).

Target Skill

Tests administered by computers or mobile devices can be designed to evaluate a *single language skill* -such as writing, reading, listening or speaking- or a set of *integrated skills*, such as listening and speaking. Integrated skills are said to better reflect the use of language in different complex contexts rather than a single ability in isolation. Moreover, it reflects the complexity of

language use in context (Chapelle, Grabe, & Berns, 2000, in Suvorov and Hegelheimer, 2013) and the improvement of the authenticity of language tests through integrated tasks (Ockey, 2009, in Suvorov and Hegelheimer, 2013).

Scoring Mechanism

Test-takers' performance on computer or mobile devices can be assessed either by human raters or by computers. There are two ways of scoring: *computerized scoring* and *analysis-based scoring*. In the first case, the input is obtained by analyzing test-takers responses, which are matched with the exact answers (correct preset responses) (Suvorov and Hegelheimer, 2013). This is commonly used for reading/listening, and sometimes for writing scoring in the form of one word or short phrases that are included in the pre-piloted list of acceptable answers and common spelling errors (Alderson, 2005, in Suvorov and Hegelheimer, 2013). In the second case, the analysis-based scoring permits performance based-testing, which means that the test-takers construct extended answers completing writing and speaking tasks. This type of scoring utilizes language processing methods, such as e-rater used in Criterion, an automated writing evaluation system (Attali & Burstein, 2006; Burstein & Chodorow, 2010, in Suvorov and Hegelheimer, 2013) and speech appraisals systems such as Ordinate in the Versant English Test (Downey, Farhady, Present-Thomas, Suzuki, & Van Moere, 2008, in Suvorov and Hegelheimer, 2013).

Stakes

Just like any other type of test existing for language assessments, CALT presents *low*, *medium* and *high stakes* for test participants (Suvorov and Hegelheimer, 2013). *Low stakes* have almost no consequences for test-takers, engaging tasks for practicing, self-studying, and personal purposes. *Medium stakes* represent some sort of impact on test-taker's lives. Finally, *high stakes* have life-changing consequences, which are mainly employed for granting citizenships,

professional certification, and admissions to educational programs (Roever, 2001, in Suvorov and Hegelheimer, 2013).

Purpose

The purpose represents the decisions which are made on the basis of the test output (Suvorov and Hegelheimer, 2013). The different purposes can be classified into two categories: *curriculum-related* and *non-curriculum related* (Carr, 2011, in Suvorov and Hegelheimer, 2013 in). On one hand, curriculum-related tests have to do with tests for admission, placement, and diagnosis purposes. On the other hand, non-curriculum-related tests are used to measure language proficiency of the students (Suvorov and Hegelheimer, 2013).

Response Type

Test-takers usually offer two core types during a computer-delivered language test, which corresponds to *constructed responses* and *selected responses* (Parshall, Davey, & Pashley, 2011, in Suvorov and Hegelheimer, 2013). Constructed responses addressed the idea of test-takers developing their own answers by providing different lengths of linguistic output (Suvorov and Hegelheimer, 2013). Meanwhile, selected responses only focus on test-takers choosing from a list with a certain range of options, like multiple choice questions (Suvorov and Hegelheimer, 2013). However, both of these response types can be noted at the same time in one answer, because some responses may possess both traits.

Task Type

In CALT, the task type can be divided into three categories. They can be *interactive*, by matching, dragging, and dropping items; *selective*, by answering yes/no questions and multiple-choice questions; and *productive*, by responding to short answer tasks, close tasks and written or

oral narratives (Suvorov and Hegelheimer, 2013). Nevertheless, there are some task types that are only possible through the utilization of computer or mobile devices.

All of these dimensions function as the core categories that allow for the description of tests within the CALT approach. When observing validity changes after the transition from face-to-face to online testing of L2 oral performance, these categories allow for the observation and comparison of attributes between one format and the other.

2.6 Conclusion to the theoretical framework

The framework just presented identifies construct and face validity as fundamental sources of validation to observe when a test is transformed or modified. At the same time, a descriptive framework has been explained that can be used to describe computer-assisted and online tests. Together, these two descriptive perspectives can be used to compare face-to-face and online versions of L2 proficiency tests in order to evaluate whether construct and face validity have been affected by the transition from one test format to the other.

In the next chapters, a study is reported which attempted to evaluate the effects on construct and face validity of a test of English L2 oral proficiency that made the transition from a face-to-face to an online format. In the following chapter, the methodology of the study is explained.

3. Methodology

This report describes a case study that aimed at elucidating if and how face and construct validity in an English L2 proficiency test changed. The test is called Oral Project, and it is applied as part of the English Language course in an English Language and Literature undergraduate programme in Chile. The study was guided by the following research questions:

- ➢ Research question 1: Did the construct validity of the Oral Project change from the face-to-face format to the online format? If so, how?

- ➢ Research question 2: Did the face validity of the Oral Project change from the face-to-face format to the online format? If so, how?

In order to answer Research Question 1 (henceforth RQ1), we described the changes observed between the face-to-face format of the Oral Project test according to the set of categories proposed in Suvorov and Hegelmeiher (2013) (see section 3.2.1),

In order to answer Research Question 2 (henceforth RQ2), a survey was designed having as a basis the same categories proposed by Suvorov and Hegelmeiher (2013). The survey was applied to 33 students of the Lengua y Literatura Inglesas programme that had taken the face-to-face (pre-pandemic) and the online version of the Oral Project test (post outbreak of the pandemic).

The questions in the survey aimed at knowing the perception of test-takers as to what changes in construct validity they had been able to observe and whether they thought those changes were positive or negative in general and why. Responses to the survey were organized in spreadsheets and analysed in terms of their content. The analysis consisted of identifying main themes and ideas presented by participants in their responses. Themes were grouped into ever more

general themes according to the relation they had among each other. Finally, we identified the ideas undergoing these themes and represented them as general statements regarding the perception of participants of the changes introduced to the test as part of its transition to an online format.

In this chapter, the procedures that we conducted to collect and analyse the data of the study are described and explained.

3.1 Data collection

For the present research, data was collected by different methods in order to answer each question. In the present section, the tools, and the procedures of how data was collected will be explained.

Data were collected through two different methods in the interest of answering each research question separately (RQ1 and RQ2). In order to collect data to answer our question about construct validity (RQ1), we gathered the rubrics and guidelines of both formats of the Oral Project (specifically the ones of the face-to-face format made in 2019 and the ones of the online format of 2021). These rubrics and guidelines were found in past emails sent by the OP evaluators and in the files uploaded in U-Cursos, a platform of the University of Chile that organizes course archives among other functions. Guidelines and rubrics can be consulted in Appendix 2.

For collecting the data concerning our RQ2 about face validity, we used the descriptive categories proposed by Suvorov and Hegelheimer (2013) (see section 2.5.3) to develop the survey that was responded to by the participants. The survey asked the test-takers about changes in construct validity of the online test when compared to the previous face-to-face version.

We sent a survey through a link that was sent to the Whatsapp chat group where the participants were present. The members of the chat group were students undergoing their last year studying the *Licenciatura en Lingüística y Literatura Inglesas* program at *Universidad de Chile*. The survey was completed online. After applying the survey, 30 responses were obtained. Out of these responses, 8 participants seemed to have confused the object of the questions (i.e., the Oral Project) for the Oral Test, which is a different instance of oral assessment of the program. We excluded their answers from the study and contacted these participants again through email in order to apply the survey on them again. Only 2 of them responded to the survey again. The final data set included, therefore, 24 responses.

3.1.1 Participants

The survey to answer RQ2 on face validity was applied to students undergoing their last year studying the *Licenciatura en Lingüística y Literatura Inglesas* program at *Universidad de Chile*. We chose final-year students to develop the study because we required participants who had taken both the Oral Project test in face-to-face format (before COVID-19 pandemics) as well as in the online format (during COVID-19 pandemics).

A first version of the survey was tested on 5 participants from senior year who were personally reached by the researchers. Feedback from this pilot was fed into a final version of the instrument. These responses were not included in the analysis of this study.

3.1.2 Data Collection Tool: Online Survey

The online survey used in this study was designed to gather data revealing opinions and perceptions of learners regarding the change of format made during the COVID-19 pandemic (from face-to-face to online) of the Oral Project assessment (OP). The purpose of the survey was to discover how face validity might have or might not have changed concerning the OP transition from a face-to-face to an online format.

The survey was made using Google Forms and started with a message of prior information for the students in order to answer the following questions. This information stated that they were able to answer either in Spanish (their first language) or English, that every answer was valid, and that it would take around 15 to 20 minutes to answer. Moreover, it also asked the participants to answer each question individually even though some of the questions might seem to lead to similar responses.

The survey consisted of 10 sets of questions. Each set consisted of a mix of closed and open-ended questions and was related to a specific dimension proposed by Suvorov & Hegelheimer (2013). The closed questions aimed at establishing the recognition of changes to posteriorly value them (positively vs negatively). The open-ended questions aimed at obtaining descriptions of the changes observed by participants in each dimension and explanations of the values they expressed in the closed question.

A copy of the survey can be found in Appendix 1.

3.2 Data analysis

Two sets of data were collected in order to compare the characteristics of the two versions of the Oral Project test. One set was constituted by the research team observations of the features of the test as indicated in guidelines, rubrics and the experience of the research team as test-takers. A comparison was carried out between the two versions of the test in order to observe changes of the construct validity of the test.

The other set was constituted by the responses of the participants to the online survey. The responses of the participants were analyzed in terms of the content of the corresponding responses. To answer RQ2, participants' perceptions of changes in the test and their views on their potential benefit were identified and characterized.

3.2.1 Construct validity analysis

The first set of data, which aimed to answer the question of construct validity (RQ1), constituted a matrix of analysis. This matrix was elaborated based on the categories proposed by Suvorov & Hegelheimer (2013) to describe Computer-Assisted Language Testing (see section 2.5.3). The matrix was designed as a grid of nine dimensions that allowed to compare the two formats according to each dimension, side by side. The purpose of the comparison was to discover what aspects of the construct validity of the evaluation had changed or remained the same. Table 2 below, shows the resulting matrix:

Table 2. Grid for the Matrix of Dimensions

Dimension	**Oral Project face-to-face**	**Oral Project online**	**Construct validity effect**	**Surface validity effect**
Directionality				
Delivery format				
Media density				
Target skill				
Scoring mechanism				
Stakes				
Purpose				
Response Type				
Task Type				

The matrix of analysis included the following components:

1. The name of the dimensions: The sets of attributes proposed by Suvorov & Hegelheimer (2013) for computer assisted tests. A further dimension, Delivery Format, was created by the researchers in order to account for the way in which the performance was presented to examiners in both tests.

2. The description of each test version: The descriptions of each test under each of the descriptive dimensions proposed by Suvorov & Hegelheimer (2013).

4. Construct Validity Effects: A brief explanation of how the changes identified between the two formats of the test seemed to have affected its Construct Validity.

5. Researchers' Face Validity Expectations: In this final column, we recorded our expectations as researchers of the effects that the changes in the different dimensions of the matrix on the face validity of the test. The expectations established here were later contrasted with the perceptions of participants (see results in 4.3).

The analysis was conducted in meetings where discussions took place amongst the researchers and the research supervisor. All descriptions of changes in each dimension were established by unanimous agreement.

3.2.2 Face Validity Analysis

As mentioned in 3.1.2 above, data for RQ2 consisted of the observations and opinions students had about changes (or the absence of changes) in the online version of the Oral Project test. Responses indicated changes observed by learners in each dimension and their opinion as to whether those changes were seen as positive, negative or neutral.

Responses were organized in a Google spreadsheet. Data was organized such that rows represented each participant and columns the responses they provided to each question of the questionnaire. The result was a grid of 24 rows and 10 columns to contain the responses to the 10 questions a participant had.

For each open question response, an analysis of what the main and secondary themes in each were. For this purpose, responses of four participants were assigned to each member of the research team for the first round of theme analysis. Results were shared later in short briefings by each team member. Two more rounds were conducted on the themes identified in the previous rounds in order to find broader concepts/ themes that could be identified as general responses.

Additionally, we added three new columns. Both the first and second columns were used to compare between the participants that thought that the change present was positive versus the ones who thought they were negative, while the third was used to establish if we thought that the answer given by the participants aligned with what they wrote in the survey; checking the surface validity of the test by putting 1 if they aligned and a 0 if they did not.

The data obtained in this section was discussed with the purpose of finding the themes that exists inside the ten dimensions we used, this allowed us to categorize the data, using a criterion of quantity, in groups that reflected the positive and negative perceptions of the participants in the closed questions, i.e., the number of participants that chose the positive, negative or neither alternative in the survey.

Then we grouped the themes according to a qualitative criterion that allowed us to group the open question responses into positive and negative attitudes, i.e., we read the open responses and found different themes that helped us to classify what the participants were feeling into two groups, positive attitudes, and negative attitudes. This data classification served the purpose of making it easier for us to study and discuss the changes in the face validity obtained through the open responses the participants gave us.

We observed initially which dimensions presented changes according to participants. To that purpose, we designed a simple grid in which the first column enumerated the questions from 1 to 10, the second indicated how many participants thought a dimension had changed, and in the third column were the participants that thought the dimensions did not change. The fourth column provided the ten different dimensions tied to each question of the survey; this can be seen in the grid that follows.

Table 3. Grid of changes on each dimension

Question	It changed	It did not change	Dimension
1			Directionality (Lack of the round of questions)
2			F-t-F to Online format
3			Target Skills (New-Lost Skills)
4			Scoring mechanism
5			Higher grades
6			Importance (stakes)
7			Purpose
8			Response type
9			Task type / Media Density (Addition of the Handout)
10			Delivery format

Being able to compare the number of participants that thought a dimension changed between assessments with the ones that thought it did not change, and then grouping them based on if they had a positive or negative perception of the change in a given dimension allowed us access to preliminary information regarding face validity, In particular, we could see the extent to which participants observed changes in the test as a whole and which aspects or dimensions of the test contributed to that perception of change.

The next five columns contained the data regarding the number of people that found these changes rather positive, negative or neither and the results this data provided. This information was interpreted as an aspect of face validity where participants stated whether the changes they had observed in the test made participants' experience different. Options here included the perception of positive changes that resulted in, for example, improving their performance in oral skills, allowing them to improve their grades, and improvement of the assessment. Perceptions could also be negative, as in causing more stress preparing the testorrecording videos, or by having less trust that the test was evaluating oral proficiency as adequately as its previous version.

Finally, we also interpreted results considering those perceptions which were shared by most of the participants, a group of participants or those held by one or a few participants. This criterion was applied broadly to establish which ideas seemed to reflect more generally the face validity of the test.

3.3 Conclusion to the methodology section

The procedures and tools used to collect and analyze the data in this study were designed with the purpose of providing evidence of the changes through which the Oral Project assessment has been through. Particularly we intended to elucidate if the Construct Validity had changed or had been affected due to the change of format. This was made through a comparison between the face-to-face format and online format of the test in each dimension of CALT proposed by Suvorov & Hegelheimer (2013) were compared between.

Simultaneously, we applied a survey to find out about the perceptions that participants have concerning the validity of the assessment. Responses consisted of perceptions that were interpreted in terms of possible changes in the face validity of the Oral Project test. In the following section, the results of the analysis are presented and explained.

4. Results

In this chapter, we account for the results of the analysis explained in the previous chapter. First, we present the results of the face-to-face and online assessment comparison in order to answer our RQ1, which seeked to find out whether the construct validity was affected by the change of format and how. Then, we introduce the results of the analysis of the responses to the survey, which seeked to answer our RQ2, regarding changes in the face validity of the test.

This chapter is divided into two sections. Firstly, the results of the comparison between the two versions of the test in each dimension of the matrix applied. Accordingly, the observed differences, similarities, and aspects that did not change from the Oral Project from the face-to-face version versus the online version are explained.

Afterward, we present and interpret the results of the analysis of the responses to the survey, which examined the perception that the students have of the change of format of the Oral Project, as an indication of changes in face validity (RQ2). This section is divided into two subsections. The first one displays the results concerning positive, negative, or neutral perception of the changes. The second presents the different themes and ideas participants expressed as changes they observed due to the application of the online version of the test.

4.1 Construct validity results: Matrix of dimensions

By using an analysis matrix, we were able to compare the two formats of the Oral Project Evaluation and to identify several differences and similarities according to the different dimensions of the matrix (see section 2.5.3 above). Table 4 below summarizes the results of this analysis:

Table 4. Matrix of Dimensions and its Results

Dimension	**Oral Project face-to-face**	**Oral Project online**	**Construct validity effect**	**Surface validity effect**
Directionality	**Linear-Adaptive** Time Target grammatical Structures Content Round of questions (Interview)	**Linear** Time Target grammatical Structures Content No round questions (No interview)	As a consequence of the change of Directionality, interview skills were lost in the Online format, reducing the construct validity. Only presentation skills remained. On the other hand, the same tool, requirements, and parameters were still applied to all test-takers. This contributes to reliability and construct validity.	The change in directionality raises the sensation that the new test lacks the opportunity to add new information (and thus, to demonstrate more of your knowledge) with the questions item. Therefore, the new format has reduced surface validity.
Delivery format	**Face-to-face** Live presentation to classmates and the teacher You can rehearse Memorization discouraged One-shot Bad timing affects the performance and the score Natural sound and visuals	**Computer-based** Recorded privately (Off-line). The recording can only be seen by the teacher You can rehearse Memorization cannot be controlled Several tries (you can repair and try again) Bad timing affects the performance and the score (But timing fails can be repaired) Microphone sound and camera visuals	A consequence of the new delivery format (recorded, more time to hand it in), the test now measures new abilities, such as the capacity to memorize and repeat a sketch/script (before it was asked in the rubric to Not to memorize the script). Thus, spontaneity is lost. Contrary, the new audience conditions (consequence of the new delivery format) contribute to validity: test-takers are less anxious and may perform better.	The perception is that the new task requires more workload because now it comprises new tasks, such as the handout and making a video. This last element requires some technological management that in the past was not required. Besides 3 new sources of information are required in the new task, which also has to be properly cited, which adds more demand to the student. Therefore, the surface validity is reduced.

	Instantly delivered	(phone, tablet or computer) Delivered through Google Form. The new handout has to be filled in a word document and sent together with the recorded presentation.		
Media density	**Single-medium** Mainly oral Powerpoint support is encouraged in the guidelines but not evaluated in the rubric (not scored)	**Multimedia** Oral, written and visual Powerpoint support is not encouraged in the guidelines and not evaluated in the rubric (not scored). However, there is a written handout that is obligatory. Rubric instructs on technical qualities of the video/audio delivery	The fact that now the assignment is done in a multimedia environment, makes it possible for students to record it more than once or look more easily at their written notes or scripts, generating a decrease in the construct validity. For the same reasons, it also loses spontaneity.	The use of computers may create certain problems, as technological devices and wi-fi are likely to fail. Moreover, the students may perceive that the test is measuring other abilities such as technology skills or memorization as a consequence of the format change.
Target skill	**Integrated Skills** Use of time Non-verbal communication Oral expression Memory Interview (Round of questions)	**Integrated Skills** Use of time Non-verbal communication Oral expression Memory (Deployed differently) No interview (Round of questions)	As the round of questions was suppressed, there are less skills to evaluate, the ones used in interviews. On the other hand, other new skills were added, such as memorizing a script. Thus, construct validity is reduced, as now different skills are evaluated.	The sensation generated by the new evaluated integrated skills, is that the new test does not measure the same abilities as before or that it measures more abilities.
Scoring	**Human-based**	**Human-based**	Not applicable	Not applicable

mechanism				
Stakes	**Low/medium** 20%-30% of the final mark of the Practice module	**Medium/high** 60% of the final mark of the Practice module	-	The perception regarding the face validity changed negatively, as now the new task is perceived as more demanding, requires more tasks, and also it has considerably more percentage in the final course grade.
Purpose	**Non-curriculum related** Measures proficiency	**Non-curriculum related** Measures proficiency	Not applicable	Not applicable
Response Type	**Constructed response**	**Constructed response**	Not applicable	Not applicable
Task Type	**Productive:** Oral Optionally Written (if it was supported by presentation)	**Productive** Oral Obligatory Written (The handout was obligatory and the presentation optional)	Even though both formats possess productive task types, the new format of the test requires the student to choose and group certain grammatical structures in a document. As a consequence, the abilities to write and to remember these structures arise. Therefore, the new test has less construct validity, as now it requires the students to deal with new skills associated with the new tasks.	The evaluators do not point out what is the purpose or the use of the new requirement of the assignment (handout). This decreases face validity because it is not present in the rubric but the task is more time-consuming.

In what follows, we explain the results of the analysis for each dimension.

Directionality

The directionality dimension has to do with the way in which the tests adapt (or not) to the test-takers. It could be linear (the same number of test items are applied in the same order to the test-takers), adaptative (each task is selected in the base of the test-taker's performance on the previous task), or semi-adaptative (adaptive at the level of either group items called 'testlets' or the test itself which will correspond to the proficiency of the test-taker on a pretest) (Suvorov and Hegelheimer, 2013).

Regarding the face-to-face format of the Oral Project, the research team agreed that this had a linear-adaptive directionality, since it required all students to deliver their presentation at the same time, using the same grammatical structures and the same type of contents. However, the face-to-face format of the test also has one adaptive component: the round of questions. This round only happened in the case that the student did not use all the required structures on their presentation, i.e., this round was a result of the proficiency of the test-taker on the previous task, as judged by the examiners. In contrast, the online format of the OP was considered to be linear, as the round of questions was suppressed, but it kept requiring the test-takers to deliver their presentation under the same parameters of time and content. Thus, we agreed that the construct validity of the oral project partially changed, as some skills were lost during the change of format, particularly the skills required in an interview task. However, other skills and requirements stayed the same, such as presentation skills.

Finally, in the discussion of the face validity effects regarding directionality, the researchers perceived that with the absence of the round of questions, the online format lacks an

opportunity to add new information and knowledge. Thus, we expected the face validity of the test to be reduced.

Delivery format

The delivery format dimension refers to the way in which the tests are delivered. It could be computer-based (off-line), web-based (Online) (Suvorov and Hegelheimer, 2013), or face-to-face. In the Face-to-face format of the Oral Project, it was clear that the delivery format was face-to-face, as no computer or internet was involved in the tasks. In this format, the test-taker had to prepare a live presentation to the classmates and the teacher, where memorization was discouraged, and the students only had one try to deliver their presentation. On the other hand, the Online format of the OP was considered to be Computer-based, as the task required a privately recorded presentation that was delivered privately to the teacher through email or Google Forms.

Moreover, a handout that had to be filled using Microsoft Word was required to be sent together with the recording. This handout was a word document where the test-taker had to write all the important grammatical structures they would use in their recorded presentation. The addition of this new computer-based task, also requires a new skill, memorization, that was discouraged in the older version of the OP. In the online format OP, memorization cannot be controlled, the test-taker can learn the structures written on the handout and has several tries to record, where they could repair their failures of time or language.

Thus, as the delivery format dimension changed from one format of the OP, we agreed that the construct validity of the test changed for two main reasons. Firstly, the Online format measures new abilities, such as the capacity to memorize and repeat the structures on the handout with adequate intonation and pronunciation. This is the result of the fact that now the test-takers can

record several times their performance until they achieve the correct phonetic structures or target performance. Secondly, the spontaneity of the face-to-face format is lost, as now the presentation is not delivered in one shot and requires an amount of memorization.

Furthermore, the online format also implies a change in the construct validity of the OP test because of the devices required to produce the task. In this version of the test, the microphone sound and camera visuals of the device used by the test-taker (phone, tablet, or computer) might negatively affect the possibilities of the teacher to evaluate and the students to get a fair grade. The reason for this is that each student uses a different device to record the task (the one they have available). As we know, the current devices, depending on the range of prices, count with different qualities of camera and microphone, i.e, some test-takers record better images and sounds than others.

Last of all, in the discussion of the surface validity effects, we observed that the new test requires effort to perform new tasks, such as preparing the additional handout and making a video. This last task in particular requires skills for technology-related tasks, which in the past were not required for the test. The addition of new tasks and skills might affect the surface validity negatively. Nevertheless, we also agreed that the fact that the students can record themselves several times in their privacy can positively affect the construct validity, as this could mean less anxiety, less stress, and thus a better performance.

Media Density

As explained in section 2.5.3 above, CALT may use different media formats, such as multimedia (listening test with video or reading test with images and text) or single medium (audio-only listening test or a text-based reading test). These media density formats are considered

an advantage for the availability of the several integrations that the media could provide for these types of tests (Suvorov and Hegelheimer, 2013).

We encountered a clear difference regarding this dimension between the face-to-face Oral Project and the online Oral Project. The core difference is that in the former case, we see the participants engage the task by presenting the oral project to the professor and the audience (classmates) in the classroom. The main context in which the assessment occurred were mainly oral and face-to-face. Hence, media density was very low. In the online format of the Oral Project, in contrast, students are faced with several requirements to use multimedia. The use of technological devices is needed (desktop, laptop, cellphone, tablet, etc) and the media format is varied, including the audio of the presentation, the video of the presentation (both in the recorded file) and also images, videos, graphics, links, news (mainly in the handout that was asked to deliver alongside the video) reaching to the teachers by online platforms, such as, Gmail or Google drive.

Regarding construct validity, the online task now adds a dimension for the assessment that was not present before. Hence, we interpreted this addition as a decrease in the construct validity of the test because this new dimension is not related to the construct of English L2 oral proficiency.

In terms of face validity, and regardless of the change in construct validity just observed, we agreed that in general, the online version of the test is still mainly measuring oral proficiency of English L2.

Target skill

The different types of tests that are supported by computers or mobile devices may be utilized to evaluate a specific language skill (writing, reading, listening, or speaking), or an aggregate of integrated skills (listening and speaking). Integrated skills are assumed to display a better use of language in diverse complex linguistic situations rather than focusing on a single ability. Comparing the Oral Project face-to-face with the Oral Project online, both of them are aimed at evaluating integrated skills, as every student needs to display several languages skills simultaneously in their presentation. The skills of writing, reading, listening, or speaking are meant to be used together in the face-to-face format and the online format.

However, there are differences affecting the construct that can be seen. New abilities are required by test-takers in order to perform the task (the technological devices knowledge and its functions), but these are not explicitly considered in the rubric. Additionally, the elimination of the round of questions at the end of the face-to-face Oral Project implies that the interactional skills associated with that task are not evaluated in the new format of the test.

Finally, we formed the expectation that face validity could be affected significantly to the extent that participants realize the test is actually measuring different skills than in the previous version.

Scoring mechanism

The scoring mechanism is the medium by which a test is scored. The possible mechanisms to score a test under this dimension are human-based, exact answer matching, and analysis-based scoring, that can be done either by computers or by a human rater.

Regarding the two versions of the Oral Project, both face-to-face and online format use a human-based mechanism to score the task since the teachers are the ones who evaluate each Oral Project. The construct and face validity are not mainly affected by this dimension as the scoring mechanism remains the same.

Stakes

This dimension refers to the level of importance the different versions of the Oral Project have regarding what the results could mean for the test-taker. There are three levels to measure the stakes of the evaluation described by Roever (2001); slow-stakes, medium-stakes, and high-stakes; it is also possible to create a mix of two of them to create a medium level.

We classified the online version of the OP as a medium-high stakes assessment, as it represents more than 50% of the final grade obtained in the Medium/Advanced English course and it could potentially decide if a student passes the course or not. In the case a student failed the Medium / Advanced English course because of the grade obtained in the Oral Project it would mean that students would have to do the course again the next year and thus add a whole year to the duration of the course. This can be considered as to what Roever (2001) defines as close to a 'life changing event that does not have broad, life altering consequences.' (91)

The new stakes associated with the online format may also affect the face validity of the test, compared to the face-to-face format, which is considered as a low-medium stakes assessment. The importance of the test has increased, while at the same time some skills and abilities measured by the previous format have been removed from the test. This could mean that test-takers may consider the increase of the stakes of the assessment not consistent with what the new format evaluates and its increase of certain negative effects in the students that took the test.

Purpose

The research team agreed that the main purpose of the Oral Project has not changed between versions, as they are both *non-curriculum related*. None of the versions are used for admission, placement, or diagnosis purposes, but rather to measure the language proficiency of the students. Even if the different versions do include different skills to be evaluated, the main objective and associated decisions of the test remain the same: use oral discourse to evaluate the proficiency levels achieved by test-takers during the course and decide on which students pass or not the English Language course. We expected thus that participants recognised that the purpose remained the same and thus face validity would not be affected by this factor.

Response type

The response type dimension makes reference to the kind of response expected from the test-takers. *Selection-responses* occur when questions have predetermined answers (multiple choice). For *constructed responses*, the test-takers have to develop their own answers. The oral project in both its face-to-face and online format has a constructed response type since test-takers are expected to develop a full argumentative presentation from one topic. Since the skills for this task are the same, construct and face validity are not particularly affected. Accordingly, we did not expect effects on the face validity of the test.

Task type

The task type dimension refers to the manner the task is to be developed. There could be three task types: *selective*, *productive*, and *interactive*. The oral project, in both its face-to-face and online format, consists of a *productive task type*. Oral production is required to succeed in both versions. Nonetheless, while in the face-to-face format the written part was only present if it

supported the presentation, in the online format it is mandatory to hand in a hand-out with specific grammatical structures and useful vocabulary which became part of the rubric. In regard to construct validity, the evaluation remains the same in the manner of developing the task but with the incorporation of mandatory metalinguistic knowledge in the online format.

4.2 Face validity results: survey analysis

In this section, we report the results we obtained from the survey which was administered to the participants of the study. The questions of the survey were formulated in a way that they addressed the same dimensions used for the construct validity analysis. All the questions in the survey had the possibility for participants to evaluate if the change was positive, negative, or neutral for them. Accordingly, in this section we report the results concerning participants' perceptions of the change and their evaluation of the benefits of observed changes.

4.2.1 The test changed

Responses to the questionnaire allowed us to categorize the perception of the changes occurring in the ten different dimensions of analysis in this study. These perceptions, in turn, allowed us to analyze changes in the face validity of the Oral Project.

Results indicate that there are several dimensions where changes were recognised by participants. Table 5 below summarizes the results of this analysis.

Table 5. Perception of the changes in each dimension

Question	It changed	It did not change	Dimension
1	17	7	Directionality (Lack of the round of questions)
2	22	2	F-t-F to Online format
3	11	13	Target Skills (New-Lost Skills)
4	13	11	Scoring mechanism
5	18	6	Higher grades
6	15	9	Importance (stakes)
7	7	17	Purpose
8	16	8	Response type
9	13	11	Task type / Media Density (Addition of the Handout)
10	17	7	Delivery format

Participants agreed on perceiving a change on six out of the ten dimensions, Directionality, Delivery Format, Scoring Mechanism, Stakes, Response Type and Media Density. Participants seemed aware of changes in the different dimensions such as the subtraction of spontaneity skills, the addition of metalinguistic skills and the introduction of technology-related skills in the new format.

A dimension was considered positive in general when the changes involved improved the performance of the students, improved their grades, and overall improved the assessment process.

Changes in two out of the six dimensions were considered to be perceived as completely positive after having more than two thirds of the participants agreed that the changes were perceived as such; these changes were part of the dimensions of Delivery Format and Stakes. Table 6 below summarizes the results of this analysis.

Table 6. Positive vs Negative perception results I

Dimension	Positive	Negative
Directionality (Lack of the round of questions)	9	6
F-t-F to Online format	14	7
Scoring mechanism	5	3
Higher grades	13	2
Response type	6	5
Task type / Media Density (Addition of the Handout)	9	2
Delivery format	11	5

The dimensions that were considered to have a negative perception were Target Skills and Stakes (or importance). The participants perceived the changes occured in these dimensions, such as new added grammar skills from the handout or the increment of the importance the new format has, as negative changes. Open responses from the participants back up the notion that these changes produce more stress, anxiety, and a general uncertainty about the task among the participants; these dimensions are shown below.

Table 7. Positive vs Negative perception results II

Dimension	Positive	Negative
Target Skills (New-Lost Skills)	1	5
Importance (stakes)	5	8

The dimension that has neither a positive nor negative perception is the one that objectively did not change from one format to the other: its purpose. This dimension was considered as such because participants did not show any inclination, as there was no change in the dimension, regardless of the participants who found one, that could be considered to have neither implicances.

Table 8. Positive vs Negative perception results III

Dimension	Positive	Negative	Neither
Purpose	2	2	4

4.2.2 What changed according to participants?

In this section, we present the results of the analysis of the responses regarding participants' perception of changes from one format to the other. Results of the content analysis of the responses to open questions can be presented in two different groups. Firstly, we could recognise five major themes regarding changes in the construct validity of the test . Secondly, we recognised several themes related to attitudinal effects observed by students caused by the change of format of the test.

Results will also be presented making a distinction between those ideas that were held across participants and those which are held by one or a few of them.

Results of the survey indicate that participants recognise changes in at least four ways described below.

Perceptions across participants

Media density: More trials

Fourteen participants mentioned that, since the presentation is evaluated as a pre-recorded video, there are more chances to improve performance in the online format of the test at the moment of being assessed. P16 declares her/his has increased “mainly because I have the opportunity to record again and again when I make a mistake". The face-to-face version of the oral project did not provide this possibility, since it was a face-to-face oral presentation which you could only take once. In this respect, P20 indicates "It is easier to record a video, as it can be re-recorded as many times as you like. In the face-to-face format, there was more pressure and only

one opportunity". Thus, the change into a pre-recorded evaluation was not only perceived as easier but also as producing less pressure for the test-taker, which was considered as a positive change.

However, P25 does not agree that the change is only positive. She/he says that "It has positive points such as the awareness of the tenses, but negative things like the extra work for both the students and the teachers". On this matter, being able to record presentations gives extra work for both students and teachers as explained in section 2.5.

Seven participants also mentioned that the Oral project had a better organization and some of them indicated that this was due to the introduction of a mandatory hand-out as part of the presentation. P17 states, for example, that “Maybe because the online format allows more time to prepare and it has a defined structure”. When approached if he/she believes that the change of face-to-face format to an online format has affected the evaluation of the Oral Project, he/she responded that “yes, because it allows me to have more time to prepare the evaluation …It gives me a better structure for My speech”. Besides, P17 also states that a clearer written format, as it is the hand-out, gives test-takers the possibility to develop a better speech structure.

Along the same line, P01 also states that “It seems to me that now the evaluation rubric is better structured and divided in relation to the points to be evaluated”, which also supports the idea that a better structure outline results in better outcomes.

In the face-to-face format of the oral project, it was more difficult to produce a specific grammatical or phonological structure or new vocabulary since these were not specified. By providing a predetermined manner of developing the presentation task, the online format allows students to know what is to be expected from them. This seems to explain why the observed change was considered positively.

Stakes: Higher ponderation

Eight participants mentioned that the online format had a higher contribution to the score in the English Language course than the previous version (see section 4.2 for those results). P21 states that "Because it has more impact on our final grade we have to take it more seriously than before. So we have to spend more time preparing for it". Even though it is not explicitly specified in test tools and resources why the test has now a higher ponderation, test-takers are motivated to take it more seriously in terms of preparation.

The online version of the test is also seen as less complex and easier, and grades had improved for most participants. With the increase of the value of the test for the final grade of the course, the new version of the test, it seems reasonable that changes in this dimension, when recognised, are seen as positive,

Less frequent perceptions

Directionality: Less tasks

Concerning the number of tasks that the online format has, in contrast with the face-to-face format, the first one lacks the final round of questions. This constitutes a crucial change regarding construct validity (see section 2.3 above). According to 5 participants, this change in the construction of the test has affected them in a negative way. Thus, P03 responded “With the questions one could expand a little and demonstrate greater command of the language". The round of questions gave test-takers more possibility of compensating the test performance since, if anything was missing in the oral presentation, the questions made by the examiner lead the test-takers to reach that grammatical structure or necessary vocabulary missing according to the test rubric. Along the same line, P25 states that “since there is no extra development of the idea by answering questions if something was not clear”. Therefore, test-takers have only one opportunity

to demonstrate the command of the language in the video, without the chance of answering any questions from the examiner.

These participants were also aware of the fact that, even though having the chance to pre-record a video multiple times and thus amend performance mistakes, there was now no chance to repair remaining mistakes.

Target skill: No improvisation

Finally, four participants indicated that with the online version of the oral project the spontaneity factor is missing. Since the test is now assessed as a video that can be recorded as many times as possible, language performance lacks spontaneity. P07 states that "In face-to-face, we had to improvise. Online oral project evaluates the same aspects but leaves out spontaneity because a recorded version may have lots of hours of practice and more than one attempt of recording".

P05 considers this as a negative change, “porque quizás se pierde el cara a cara y la improvisación que había, factores que se tomaban en cuenta en presentaciones". Thus, participants perceive a change in how test-takers are being evaluated. Without the spontaneity of the previous format, most of the participants indicated that they were in front of a script rather than a test that measures their proficiency in the L2. Thus, P20 explains that "I think it's important to create instances where students have to improvise and respond without a pre-prepared speech". From this perspective, there can be a negative perception of the test-takers when they consider that they are missing the opportunity to show their command of the language in normal interactional contexts.

4.2.4 Attitudinal effects

The effects described below were characterized as attitudinal effects, as they all have to do with the self-perception of participants in relation to the way they perform the assignment on the online format and how different it makes them feel. Results are here presented distinguishing between positive and negative attitudes and, as in the previous section, between some ideas present across participants and others held by one or a few of them.

Positive attitudes across participants

These effects were identified across 14 participants, who perceived these changes identified them as something positive.

Less pressure

This attitudinal effect was mentioned by 7 of the 14 participants throughout the survey. Most of them indicated that in this new version of the OP they felt more relaxed while developing and performing the task. Among the responses we received, participants shared similar reasons to state that the new design of the task made them feel less pressure. For instance, P21 states that "(...)the round of questions, from my perspective, was designed to pressure the student". We can argue here that having the questions erased lowered the pressure that students had when performing this task. Indeed, P21 considered that the final round of questions of the Oral Project had no other purpose than putting the students under a situation of pressure.

Also, P30 considers that “Because of myself. It clearly has been a change, bit because of my way of performance. I don't feel the same pressure like before, that took me back un a lot of ways.” Since pressure has decreased, participants (i.e., students) had perceived the test with less stress

P14 also stated that "There is less pressure to perform, as if you make a mistake you can simply record another take until it's perfect". P20 also states that, "In the face-to-face format, there was more pressure and only one opportunity". For both of the participants, the fact that now they have as many chances as they want to record the video and therefore to perform the task, makes them perceive the Oral Project as less pressuring, as they now have more chances to improve their performance and prove their skills.

Less anxiety

Having less anxiety while performing the OP was perceived and mentioned by 9 of the 24 participants who answered the survey. In this category diverse reasons were mentioned to affirm that the Oral Project produced less anxiety.

Some participants point to a relationship between less anxiety and better performance in the test. P19 pointed out that "I get to show what I know, with anxiety I usually do not get to show my knowledge". Here we can see that the participant sees that anxiety did not allow him/her to show her knowledge and, therefore, that affected his/her performance negatively.

Moreover, in relation to the changes in the format of the Oral Project itself, P12 considered that not having a final round of questions was good along with not having an audience looking at them:

> Gracias a que no tengo el nerviosismo de las preguntas al final ni de tener que presentar frente a una audiencia que va a estar observándome, tengo el control de todo lo que digo y cuanto tiempo me toma, debido a que puedo regrabar en caso que lo necesite. básicamente me siento seguro de mí y mis capacidades porque soy solo yo quien evalúa antes de enviarlo.

Being able to have full control of the test task and how to develop seems to make these participants feel more confident as the anxiety has decreased. As well as with P19, this participant also perceives a relation between improved performance and being less anxious due to the lack of audience watching, the possibility to re-record the video while performing the OP. These factors are perceived here as allowing test-takers to show abilities in a better way.

P12 also pointed out that not having an interview after the presentation was good, as the uncertainty of not knowing what was going to be asked triggered anxiety for her/him: “I think it decreased as it used to be a factor that triggered anxiety in the face-to-face classroom, and in the online format it was just a recording. This is just my particular case, though”. In this response, the online format is not only perceived as less anxiety-inducing but also the participant said that it is ‘just’ a recording. This is consistent with the idea that anxiety decreased because the new online format is easier.

Also, P12 stated in another response in relation to anxiety where he/she says: “It's positive for my anxiety for probably not for my development of English skills lol”. This response is revealing of the participant’s perception of the construct validity of the online test, as it makes it clear that, even if for the participant feeling less anxious was positive, she/he doubted if the test is really tapping on what it is supposed to be measuring.

Finally, P17 states that “El hecho de que no hayan preguntas al final de la presentación me hace sentir menos nerviosa, por ende, siento que me desempeño mejor”. Here the participant identified the lack of questions in the new format as a reason why he/she is less nervous while performing the task and therefore, has less anxiety while doing it and his/her performance is improved

Overall, for all the participants that provided answers about having less anxiety while performing for the Oral Project, the new Online format represented a positive change that allowed them to have a better performance.

There were also 7 of the 14 participants who reported that the new online version of the test was less stressful. Although they did not intendedly say it, after analyzing their answer we identified this effect as a consequence of changes such as the possibility of recording the video as many times as necessary to have a good performance. P23 stated that: “presenta más opciones para hacerlo, puede grabar cuantas veces quiera”.

P29 whose perception is identified under the theme of less anxiety and stress answered that: “I think they are being way more generous than before”. This perception is rather related to the way the evaluators grade your assignment and how this participant perceived a less strict evaluation in the online format in comparison to the face-to-face format, which allows the participant to be less stressed while performing the task. In sum, we believe that there are significant indications that the test is less strict than its previous version and, therefore, participants find it easier and, therefore, report less stress and anxiety.

More confidence

The final effect related to an attitudinal perception across participants is consistent with the picture of an easier test that is less stressful. 7 participants considered that the online format of the Oral Project made them feel more confident of their skills. Two of them relate this effect

somewhere between having more knowledge now and also being able to record the Oral Project by themselves, without the pressure of an audience.

For instance, P01 said: “siento que al momento de responder me muestro más segura y eso se refleja en mi postura, la forma en que hablo y también en la pronunciación”. Also, P20 stated that “I have more skills to develop the task now”. Both participants mentioned that they think they have more linguistic tools to develop the task appropriately. P01 further identified specific abilities that improved, such as speaking performance, pronunciation, and her/his posture in the new version and how that made him/her more confident.

Furthermore, P20 and P30 answered that they felt more confident in the new online format particularly because it lacks a questions section. P20 explained that “At that time I did not have the confidence to hold a fluent conversation in English. So, the questions made me feel really insecure”. This is also related to what the previous participants had stated about having developed more skills due to the new format such as fluency. Moreover, not having the round of questions was something good for the participant’s confidence as there was no need to improvise an answer. P20 indicated that “Since they don't ask, I don't feel nervous and I feel confident”.

Higher grades

Eight participants pointed out that the current version of the task allowed them to obtain higher grades. Participants explain that this is a consequence of having more opportunities to perform better and the decrease of test anxiety, as compared to the face-to-face version of the test. P22, for example, indicated that: “as it is a recording, it was easier for me to develop my ideas without the anxiety of facing a teacher or your classmates. Therefore, the grades were affected positively”.

Negative attitudes across participants

In this section, some negative perception across participants is described, in contrast to the first attitudinal effects across participants reported in the previous section.

More pressure

Six participants pointed out that the new online version of the test made them feel more pressured as the one-shot recording format. Here P07 explained that: “Since it has to be a one shot video, if you miss something you must start over, even if you are finishing the presentation”. Additionally, they report that the pressure is also caused by the perception of the risk of making a mistake. P25 indicates that it gives more pressure to this evaluation and makes the other ones seem like they are not worthy, or that they have no importance”. They refer to the fact of the current higher ponderation of the task and how that affects their tranquility as it has more weight in their final mark.

On the other hand, they also refer to the frustration produced when they do not get a perfect shot: “I think it has a positive side to be face-to-face and deliver your speech one time and you're done with it. Instead, you have to be obsessed with the way you record it in the online format, and finally, due to frustration you end up uploading something you know it is not the best representation of your skills”. The frustration here seems to stem from the expectation of having many opportunities that will lead to a perfect performance. In sum, the causes for feeling more pressure in the current version of the task are the higher ponderation and the opportunity of recording it several times, which may be perceived as a covert expectation to have a flawless performance and thus producing a sense of pressure to perform flawlessly and a sense of frustration when that does not happen.

Less frequent negative attitudes

The negative attitudinal effects described below were perceived by fewer participants than in the two previous ones.

More stress

This effect is mentioned only by P15 who mainly refers to the stress generated by the format change. He states that: “You have to find an opportune moment to record it and often this is made difficult by external noise; often making it necessary to record more than once”. So here, the factor of noise is identified as one that complicates the development of the task. Besides, in relation to the video recording, he adds: “As it's a recording, the student's performance is expected to be closer to "very good", because it has the "advantage" of being done several times. But often the video to be uploaded is the one you are most tired and exhausted”. This perception seems to be related to the pressure effect described above, because the higher level of stress is a result of feeling more pressure, which in turn is caused by the expectation that in the online format performance must be better than in the face-to-face version.

Less confidence

This effect was mentioned by P09 who responded, when he/she was asked if the change of format affected her evaluation that: “online I retake the video a lot of times, doubting myself and what am I am saying, face-to-face is more okay, let's bear this for a couple of minutes and we're done”. Therefore, the main causes of less confidence that we could identify were the round of

questions in the previous version of the task and the number of times the new version permits to record oneself. This can be perceived as positive for the majority of students, but for some, it may be negative, as they may feel more insecure.

4.3 Comparison of the researchers' and the participants' perspectives on face validity

As stated in section 4.1 the comparison matrix applied in this study included the elaboration of expectations of the researchers' on how the face validity could be affected by the significant changes in construct validity observed when comparing the two versions of the test (see section 4.1).

Overall, in almost every dimension, except in purpose, participants were able (to different extents) to show awareness of the changes observed by the researchers. However, the opinions regarding the benefits of the changes (positive, negative, neutral) were not as expected. For example, in the dimension of Delivery Format, the researchers found that these changes affected face validity negatively, as they decreased the capacity of the test to measure oral proficiency in English L2. However, participants considered this positively because it allowed them to record several times to achieve the desired performance, thus reducing their stress and improving performance and grades. In the dimensions of Directionality, Response Type, Stakes, Media Density, Target Skills, Task Type and Scoring Mechanism, there was no agreement among the participants about whether the changes were positive or negative. In contrast, researchers expected a negative perception on the part of students, based on the fact that the test is measuring a different construct of oral proficiency and other constructs unrelated to language use.

4.4 Conclusion to Results

The results of the study indicate that when asking students about the construct measured by the test, their responses express an awareness that the test had changed in several aspects of its construct validity. However, this awareness does not seem to affect the face validity of the test. In other words, participants perceive the changes, but still perceive that the test evaluates what it has to evaluate.

Alongside these overall results, in some particular aspects of change, face validity was affected. Responses across participants indicated that observed changes were mostly perceived positively rather than negatively. Regarding the positive perception of the changes, they generally pointed out that having a less demanding test has as a consequence a better test performance and higher grades. However, a feeling of not having learned enough is also present in some responses. The implications of these results will be discussed in the following chapter.

5. Discussion of Results

The results regarding the construct validity of the Oral Project test show that this is different in the online version. In particular, the new format of the test includes important format changes, such as the possibility to record performance several times, due to the new video format, the loss of a round of questions activity, and the addition of a handout as part of the test. The overall effect of these changes is that the test requires from students the performance of less language-related tasks. The online version of the test evaluates different aspects of the oral proficiency construct as some abilities were left out (for example, spontaneous conversation) and other abilities were incorporated (for example, memorization, technological skills). Evidence also indicates that, as expected, the new online test requires that test-takers recruit skills related to the use of computer devices and multimedia software.

Results regarding face validity indicate that students who have taken both versions of the Oral Project are aware of the changes that the online evaluation has brought in. These results are consistent with the changes that we observed during the analysis of the dimensions in our matrix. Somewhat surprisingly, despite recognising these changes in the construct of the test, students recognise at the same time that the test still evaluates the same construct. In particular, students perceive the online test is still measuring the same skill (i.e., oral proficiency) and has the same purpose as the original face-to-face test.

In this Chapter, we present the discussion of the results obtained concerning construct and face validity. To this purpose, the first discussion will be in relation to the construct validity of the test. Then, we will continue with the discussion on face validity, giving possible reasons for the apparent discrepancy between both. Moreover, in this section of the Chapter we will discuss the

relationship between the positive attitudinal effect with the better performance/grades of the participants as well as a discussion of the theme of less anxiety associated with the online version of the test. After that, we discuss the perceptions of participants who have considered the changes as something bad. Some final words on the relationship between construct and face validity are offered to conclude the chapter.

5.1 Construct Validity

5.1.1 The test changed

Evidence in this study clearly indicates that the versions of the Oral Project assessment are significantly different from each other. In turn, this suggests that the assessment construct of the test has changed in the transition of format versions. The change is characterized by the subtraction of skills that were present in the face-to-face version of the test and the addition of some new skills as part of what is being evaluated, including skills unrelated to the construct of L2 oral proficiency in English.

These changes must be interpreted in the first place as a logical consequence of the necessity of evaluating the English proficiency of students during the COVID-19 global pandemic. This situation forced teachers to change how the evaluation took place physically while measuring the same skills as the previous version of the test.

In the new version of the test any synchronicity and spontaneity was eliminated, resulting in a less complex test that is positively perceived by test-takers. This change also affected the way test-takers prepare the assessment, since now they are more aware of the result of having a deficient mark in their final grade.

Furthermore, the fact that the test changed to an online format, and that now it requires a video, could lead to extra work for test-takers and teachers. This has been pointed out as one of the disadvantages of CBA (see section 2.5.2 above). Evidence here suggests that even though a video task may still be evaluating oral proficiency, it lacks the interactional nature of L2 natural performance, which negatively affects the construct validity of the new version of the test. However, and as it is going to be furtherly discussed in the next section, the change of format and the new task of recording a video did not seem to negatively affect the face validity for the surveyed students.

5.2 Face Validity

In many of the dimensions that the researchers perceived a change and expected test-takers to perceive those changes as negative, findings of this study indicated the opposite. This is very interesting, as we expected that the changes observed in the test construct should conceivably have led to a lower sense that the test is measuring its construct adequately on the part of test-takers. This discrepancy may be explained in several ways.

Firstly, researchers' expectations may have been better informed by a specific notion of the changes made to the OP and the implications of such changes from a technical perspective of assessment. If so, it makes sense that the researchers were able to find more changes and face validity issues during our analysis, as we were more conscious than participants about how these changes affected the evaluation and its associated tasks.

On the contrary, as participants may not have yet the skills or knowledge to perceive that the changes on the construct validity of the task are significant, even though the general skill of oral proficiency is still being evaluated.

Under this assumption, it seems important to highlight that some of the participants did show a more critical perspective on this issue and more awareness of assessment issues. Even though there were students who did not perceive that the construct of the test changed, there were some others that realized this change. They distinguished, for example, that for measuring language proficiency more comprehensively, it is necessary to evaluate the capacity of L2 learners of speaking fluently and spontaneously in English..

Hence, students who are more aware of the change, recognize that the loss of improvisation and spontaneity that was present in the final round of questions of the face-to-face Oral Project (which was explicitly related to oral proficiency) leads to the test being perceived as measuring the ability to memorize a script. For this reason, participants who demonstrated more awareness of assessment, were more able to notice that the changes are relevant to say that actually the test is no longer measuring the same construct of language proficiency.

5.2.1 Changes were not that serious

Also, participants may still consider that the test evaluates the same construct because some of the new evaluated skills and tasks did not cause an important problem to the task-takers, nor required more effort from them. For example, the skill of technology management and the tasks that require it were probably part of the test-takers' set of skills before performing the evaluation. We assumed that this happened because test-takers belong to a generation that is already very well-acquainted with the use of technological tools and devices and multimedia production. In this case, instead of presenting a problem, the possibility of using technological devices appears to have been beneficial or at least unproblematic for the students.

Furthermore, it is also true that the dimensions that changed from one version to the other version of the test did not include its main focus on the oral proficiency. It could thus be argued that participants did notice certain changes in the construct of the test but did not consider them as affecting the overall target skill or goal of the test. In other words, face validity was not affected because participants believed the test is still evaluating English L2 oral proficiency for the English Language course, but only using different tasks

5.2.2. Attitudinal effects and performance

Evidence also indicates a common perception of participants that the online Oral Project is easier than in its face-to-face version. This is related to participants' perception that teachers were more lenient (¨generous¨) when evaluating the performance in the new online format. Moreover, some changes were, in general, evaluated as positive by participants, but not in relation to their views of the assessment as a learning tool but rather to test-takers' perception of their own comfortability and convenience with the test.

These perceptions increase because of the change of the environment in which the participants had to engage with the presentation task. Evidence here indicates that the face-to-face oral project was perceived as stressful by participants. Moreover, this stress was frequently associated with potential negative effects in performance. However, it seems that this is not happening in the online Oral Project. Removing that condition in the online test, according to the students, has allowed them to deal with less pressure, decrease the levels of anxiety, and overall to engage in the task more confidently of their performance and the final results of the test.

This relief can be interpreted as the consequence of eliminating direct interactions from the test. It seems here that students felt unprepared to engage in spontaneous conversations -no matter

what the topic was- with their language instructors. From their perspective, answering open-ended questions to a person who is more knowledgeable and proficient in the language they were learning and practicing, was seen as a source of pressure that is not present in the online version of the test.

Some participants also stated that delivering a Handout was crucial for them in order to create a script that was easy to follow. We can assume here that using a script containing the adequate tenses, the necessary vocabulary for the specific topic and context, and reminders of particular linguistic expressions to use in their presentations, test-takers raised their level of security and confidence when it came to engaging with the tasks of the Oral Project.

Also, it seems reasonable to assume that the chance of recording themselves all the times they wanted to, was a major stress relief. This new online Oral Project appears to have helped significantly to the students regarding their production of English. It may thus be that the online version of the test provides a more relaxed and secure environment, where participants are able to repair their performance and learn from their own mistakes. Moreover, these may contribute actually to a better performance in several dimensions of the construct, whether it be content related, vocabulary, proficiency, pronunciation, grammar, etc.

In sum, the evidence seems to indicate that students seem to understand that the test changed, but they consider that the new materials, the new guidelines, and the new process of the Oral Project continues to focus on the proficiency in English and offers better conditions to perform in general.

5.2.3 Less Anxiety

An interesting finding of this study was the fact that participants reported that the new online test had the effect of producing less anxiety than the face-to-face version. This is interesting because this more subjective dimension was not addressed directly in the questions displayed in our survey. Among the reasons for this perception, students indicated the difference between having the audience in front of the test-taker versus being able to record oneself calmly at home. This condition results in being able to show what the student really knows and not leaving behind language performance problems derived from being anxious. We identify here an important educational problem, as presenting in front of an audience is indeed part of the normal activities of students and future academic and professional activities.

Part of the explanation for the stress factor seems indeed related to the idea that a face-to-face performance can be a factor that is detrimental to performance as it causes test- takers to feel "less secure". This feeling appears to be related to the ability to feel in control of the testing situation.

Some participants also seem to understand that the safety the recording format may produce is an element that may be good for their anxiety but not necessarily so for the development of their language skills. This suggests that although some students do not possess formal knowledge about assessment, they have the intuition that this format might not be providing the opportunity to develop their performance skills optimally as the previous format did.

5.2.4 The change was bad for the L2 learning process

As presented above, the most repeated response across participants indicated a generally positive perception of the change of format. However, some participants characterised the same changes as something that could be negative for them as L2 learners.

This insight is interesting, since they are consistent with the evidence of the study regarding the effects that the change of format had in the construct validity of the test. These results suggest that there is some level of comprehension from students regarding the way the new format might have affected the construct validity of the test. These participants noticed that the new test is measuring new abilities that are not necessarily related to language proficiency and consider that this affects them as L2 learners. In contrast with the majority of participants who perceived the change but do not consider it as something worrying, this other group can leave aside the fact that they are getting better grades in general and understand that the changes might be affecting their learning process. This perception, although not held by the majority of participants, is indeed supported by the evidence that indicates that the test evaluates new, non-linguistic abilities to complete it and that it does not any longer evaluate other skills that were important for their language development, such as spontaneous and interactive oral performance.

This should be considered by evaluators and test designers in further instances of online oral assessment, considering that there will probably not be a full comeback to the face-to-face format as soon as expected. Even if some participants distinguish what changed and do not think it is important, there are some others who can trust less in the evaluation. This effect is not only important in terms of face validity, but it is also an indicator of issues with construct validity and how the test is not measuring language proficiency the way it used to.

5.3 Construct validity and face validity

The evidence of this study reflects a complex relationship between face and construct validity. While some test-takers do not perceive the effects of the change of format, others are aware of them, and they also notice that the new test might have an effect on the development of their L2 learning process. Despite that, they positively value the changes because they are getting better grades, feel less anxious and more confident and, overall, appreciate the fact that the test is now easier than before and the improvement that has brought to their performance and grades.

In contrast, only a few participants expressed negative perceptions about the changes and the value they gave to the new format. So, it is the case here that construct, and face validity were not equally affected by the change of format. For this case in particular, evidence indicates that participants recognised the changes in construct validity, but this did not change their perception of the tests capacity to measure oral proficiency in English L2.

The evidence in this study suggests that face validity, although is affected by changes in construct validity, is not a function of it. In the case reported in this study, face validity seemed to be constituted by certain attributes of construct and purpose that seemed to override the awareness of changes in other dimensions of the test. This suggests, in turn, that face validity is also a construct that requires more specification and attention from L2 assessment researchers.

6. Conclusion

6.1 Summary of the study

The study informed in this report aimed at establishing whether the construct and face validity of an oral proficiency test called Oral Project had been affected after the change of format from face-to-face to online format, forced by COVID-19 pandemic. On the one hand, the study attempted to establish whether construct validity was affected or not due to the format change by comparing the characteristics of the two versions of the test according to a set of predefined dimensions. On the other hand, we applied a survey with questions concerning the students' perceptions of the changes (see details in chapter 3) in order to examine possible changes in the face validity of the test.

Results indicated that construct validity was affected by the change to the online format in several dimensions, although not in all of them. Essentially, changes were observed in the skills that the test measured in the face-to-face format, such as improvisation skills. This was the direct result of omitting a final round of questions from the procedure in the online version of this evaluation. This also impacted the Media Density dimension, as the evaluation changed from face-to-face to online version. The Stakes of the test also changed, as the new version represented a higher percentage of the final score of the course. In relation to the test Target skills, the online format measures new, different abilities, such as memorization and technological abilities.

As to effects on the face validity of the test, participants were able to identify several changes in the construct validity of the test already identified by the researchers. Particular observations included changes in structure, the possibility of having more trials, the inclusion of less tasks, and the lack of opportunity for improvisation. The recognition of these changes did not seem to affect face validity, as participants still considered that the test is measuring the same

target skill (L2 oral proficiency) as before and had the same overall purpose in the course of English Language.

Participants also reported that with the online test they felt less anxiety, less pressure, less stress, and more confidence. All of this was associated with a general perception that changes from one test format to the other were positive as they allowed overall performance to improve which, in turn, allowed grades in the test to also improve. However, some participants did not perceive the changes as something necessarily positive, as they considered that the test may not be appropriately measuring oral proficiency. Some participants also missed features of the face-to-face format, such as the final round of questions, because that gave them more opportunities to show their proficiency.

Results seem to indicate that there is a biased perception in the participants regarding their positive view of the changes observed in the test- It seems, in this regard, that there is some awareness that getting better grades and having a better performance was convenient for them as students although not necessarily so for their development of L2 proficiency. Finally, and despite these particular points, face validity does not seem to have been affected by the changes in the test construct.

6.2 Limitations to the study

As with any case study, findings here cannot be generalized to other tests or test-takers. At the same time, the study reported here only collected and analyzed perceptions from test-takers and not from teachers or test designers. This is important for the findings on face validity reported here, as teachers may have different views and opinions about the validity issues related to the change of format of the OP test.

Moreover, the study raised an important limitation issue regarding the test-takers not having knowledge about L2 assessment issues. It would be interesting if they did, as this information would allow them to better distinguish between their convenience and their learning process in a more precise way. In this way, we consider that perhaps, if test-takers knew about evaluation, they could have identified better the changes in the new Oral Project assessment as something detrimental for their learning process, and the phenomenon of having participants who realize the changes but did not consider them as something substantial may have been different.

Lastly, we found that the test situation could not be fully described as it was not possible for us to confirm or deny if the handout and the round of questions were considered when grading the OP test. Some of the evidence, specifically the answers of some test-takers, seemed to confirm that these were included in the final grade for the test. However, test guidelines and rubrics did not establish that explicitly. This limitation is relevant because changes of construct and face validity depend on a complete description of the test situation.

6.3 Suggestions for further research

Further research on construct and face validity effects of the change to online format for tests should include data with the perception of teachers and test developers in order to have a complete view of how stakeholders perceive the validity of the test.

Moreover, it would be interesting to investigate face and construct validity potential effects in other L2 assessments that have changed their format. Having a study about the other assessments would allow us to obtain a clearer view of how construct and face validity interact with each other in the context of particular testing situations.

6.4 Final comments

The study presented in this report indicated several changes in construct and face validity as it was mentioned above. In relation to construct validity it was concluded there were abilities lost in the format change such as improvisation and spontaneity skills. Besides, changes in the media density of the assessment and also addition of new skills to measure were found. However, participants considered that the task was still achieving its main purpose. In relation to face validity several changes were found for the participants such as more structure and more trials, highlighting some of them such as less anxiety and more confidence in the new version. Nonetheless, just as in construct validity most participants think the OP is still assessing what it is supposed to measure.

We hope this study can contribute to clarify and comprehend in a better way an aspect of the many changes the outbreak of the pandemic had in the academic context of our programme and in general. We expect also to contribute to the necessary discussion regarding the impact these changes have in the student learning process. We also hope that the discussion in this report allows to inform the processes of changes in the evaluation format in order to create assessment tools which retain their testing qualities by explicitly ensuring the maintenance of validity. Finally, we expect that the study contributes to creating awareness amongst students and L2 learners of the importance of including knowledge about L2 assessment as an essential part of our expertise as language learners and researchers.

References

Alghammas, A. (2020). Online Language Assessment during the COVID-19 Pandemic: University Faculty Members' Perceptions and Practices. The Asian EFL Journal, 27(4.4), 169–195. https://www.asian-efl-journal.com/wp-content/uploads/AEJ-Volume-27-Issue-4.4-October-2020.pdf

Alghammas, A. (2020). Web-Based Synchronous Speaking Platforms: Students' Attitudes and Practices. *International Journal of English Linguistics*, *10*(3), 21–31. https://doi.org/10.5539/ijel.v10n3p21

Akbari, R. (2012). Validity in Language Testing. En C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 30–36). Cambridge University Press. https://books.google.cl/books?hl=en&lr=&id=7STeAgAAQBAJ&oi=fnd&pg=PA30&dq=validity+in+language+testing&ots=_SfyV6F6ut&sig=dd5Sp8qT7WGoKaUaRg3LbpDDq68&redir_esc=y#v=onepage&q=validity%20in%20language%20testing&f=false

Bergdahl, N., & Nouri, J. (2020). COVID-19 and Crisis-Prompted Distance Education in Sweden. Technology, Knowledge and Learning, 26(3), 443–459. https://doi.org/10.1007/s10758-020-09470-6

Chapelle, C. A. (1999). VALIDITY IN LANGUAGE ASSESSMENT. *Annual Review of Applied Linguistics*, *19*, 254–272. https://doi.org/10.1017/s0267190599190135

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple. Language Testing, 29(1), 19–27. https://doi.org/10.1177/0265532211417211

Dooey, P. (2008). Language Testing and Technology: Problems of Transition to a New Era. *ReCALL*, *20*(1), 21–34. https://doi.org/10.1017/s0958344008000311

Gacs, A., Goertler, S., & Spasova, S. (2020). Planned online language education versus crisis-prompted online language teaching: Lessons for the future. *Foreign Language Annals*, *53*(2), 380–392. https://doi.org/10.1111/flan.12460

Holden, R. R. (2010). Face Validity. *The Corsini Encyclopedia of Psychology*. Published. https://doi.org/10.1002/9780470479216.corpsy0341

Long, M. H. (1997). Construct Validity in SLA Research: A Response to Firth and Wagner. *The Modern Language Journal*, *81*(3), 318–323. https://www.jstor.org/stable/329306

Meunier, L. (2013). Computer Adaptive Language Tests (CALT) Offer a Great Potential for Functional Testing. Yet, Why Don't They?. *CALICO Journal*, *11*(4). https://journals.equinoxpub.com/CALICO/article/view/23426/19431

MINEDUC. (2020, septiembre). *Abrir las escuelas, orientaciones para establecimientos educacionales en Paso 3 y 4*. https://sigamosaprendiendo.mineduc.cl/wp-content/uploads/2020/09/AbrirLasEscuelas-OrientacionesAnexos-09.09.pdf

Mosier, C. I. (1947). A Critical Examination of the Concepts of Face Validity. *Educational and Psychological Measurement*, *7*(2), 191–205. https://doi.org/10.1177/001316444700700201

Nevo, B. (1985). Face Validity Revisited. *Journal of Educational Measurement*, *22*(4), 287–293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x

Nottingham Trent University. (2007). About Computer-Based Assessment. Computer-Based Assessment. https://now.ntu.ac.uk/d2l/lor/viewer/viewFile.d2lfile/6605/62186/CBA%20Revised/page_01.htm

Palma, N. (2021, 11 noviembre). Ministerio de Educación anuncia retorno obligatorio de las clases presenciales a partir de marzo de 2022. *Diario UChile*.

https://radio.uchile.cl/2021/11/11/ministerio-de-educacion-anuncia-retorno-obligatorio-de-las-clases-presenciales-a-partir-de-marzo-de-2022/

Pathan,M. M. (2012). Computer Assisted Language Testing [CALT]: Advantages, Implications and Limitations. Sebha: The University of Sebha Press.

Priyo, A., & Nugroho, A. (2020). EFL Classes Must Go Online! Teaching Activities and Challenges during COVID-19 Pandemic in Indonesia. *Register Journal*, *13*(1), 49–76. https://doi.org/10.18326/rgt.v13i1.49-76

Roever, C. (2001). Web-Based Language Testing. *Language Learning & Technology*, *5*(2), 84–94. https://llt.msu.edu/vol5num2/roever/default.html

Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: a potential impact of face validity on student learning. *Language Testing in Asia*, *5*(1). https://doi.org/10.1186/s40468-015-0019-z

Sulaiman, Z.; & Khan, M. (2019). COMPUTER ASSISTED LANGUAGE TESTING (CALT): ISSUES AND CHALLENGES. 1-11.

Suvorov, R., & Hegelheimer, V. (2013). Computer-Assisted Language Testing. *The Companion to Language Assessment*, *II*(1), 594–613. https://doi.org/10.1002/9781118411360.wbcla083

Winke P.M., Isbell D.R. (2016) Computer-Assisted Language Assessment. In: Thorne S., May S. (eds) Language, Education and Technology. Encyclopedia of Language and Education (3rd ed.). Springer, Cham. https://doi.org/10.1007/978-3-319-02328-1_25-1

Zuskin, R. D. (1993). Assessing L2 Sociolinguistics Competence: In Search of Support from Pragmatic Theories. *Pragmatics and Language Learning*, *4*(1), 166–180.

Appendix

Appendix 1: Oral Project format change survey.

Prior Information

1. E-mail
2. Did you take the Oral Project in both face-to-face and online format?
   - ☐ Yes
   - ☐ No

Questions	Follow up 1	Follow up 2	Follow up 3
1. The Oral Project in the face-to-face format had a round of questions at the end of each presentation. In the online format, this round does not happen. Do you think that this has affected your evaluation? ☐ Yes ☐ No	If yes, how? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)
2. Do you believe the change of face-to-face format to an online format has affected the evaluation of the Oral Project? ☐ Yes ☐ No	If yes, how? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)
3. Comparing the face-to-face and online formats of the Oral Project, you think that the skills it evaluates... ☐ Have changed ☐ Have not changed	If it has changed, how? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)
4. Considering the characteristics of the new version of the Oral Project, do you think that the evaluators have changed the way they calculate your scores? ☐ Yes ☐ No	If yes, how? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)
5. Compared with the face-to-face format of the Oral Project, have your scores in the online format... ☐ Increased ☐ Decreased ☐ Remained the same	If it has increased/decreased, why? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative	Why? (No incorrect ideas!)

		☐ Neither positive nor negative	
6. Comparing the face-to-face and online formats of the Oral Project, the importance of the oral project... ☐ Has increased ☐ Has decreased ☐ Has remained the same	If it has increased/decreased, how? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)
7. Comparing the face-to-face and online formats of the Oral Project, the purpose of measuring language proficiency in the Project... ☐ Has changed ☐ Has not changed	If it has changed, how? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)
8. Comparing the face-to-face and online formats of the Oral Project, the way you perform the task... ☐ Has changed ☐ Has not changed	If it has changed, how? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)
9. Comparing the face-to-face and online formats of the Oral Project. Do you believe that the inclusion of the handout in the online format of the Oral Project has affected the evaluation? ☐ Yes ☐ No	If yes, why?	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)
10. Do you believe that the change to video format has affected the evaluation of the Oral Project? ☐ Yes ☐ No	If yes, how? (No incorrect ideas!)	According to you, this is: ☐ Positive ☐ Negative ☐ Neither positive nor negative	Why? (No incorrect ideas!)

Appendix 2: Oral Project Guidelines and Rubrics used in 2019 and 2021.

Universidad de Chile

Facultad de Filosofía y Humanidades

Departamento de Lingüística

Advanced English I

2021

ORAL PROJECT GUIDELINES

This Oral Project has the purpose of linking the topics discussed in class (life lessons, human rights, body matters) to the context of modern Chilean society.

The modality of discourse should be **argumentative.** A description or narration will not be enough to obtain a passing grade. You should state the purpose of the presentation at the beginning of it. As any presentation, it should have an **introduction, a main claim**, **a development** where you support your main claim as well as warrants, and a **conclusion.**

Your claims should be supported by strong arguments and examples from books, news, films, articles, songs, class discussions, as well as information that is common knowledge. You should show a global and accurate understanding of the topic you are presenting. At least **2 different sources of information** should be used.

You will have to prepare an individual **5-to-6-minute video** in which you will develop the topic from both a general and a personal perspective. The presentation must evidence reflective thinking by using supporting ideas and examples.

Your presentation and the handout must be uploaded to U-tareas by Tuesday June 17th at 16:00. We strongly recommend that you do this much earlier so that you do not risk getting a failing grade. **Late submissions will not be accepted unless you have a valid justification, e.g., a note from your doctor or program coordination**.

The video must be **recorded in one take**, i.e., **you must not edit** the video after shooting it. You will be expected to know your topic and provide a clear development of ideas **without reading from a piece of paper or the screen.** Even though you can have some written material with you in case you forget something, you must address the camera when talking. You must also make sure that the **sound quality** of your video is acceptable.

If these instructions are not followed (that is, the video has been edited, you depend on your notes, and/or the sound quality is poor), your presentation will not be evaluated and you will be given **24**

hours to send a video that complies with the instructions and **1.0 point will be subtracted** from your project grade (that is, if you get "5.8" in the rubric, your grade will be "4.8").

The presentation **does not have to** include any **visual aids**. However, you must **prepare a handout** that includes:

- Relevant and sophisticated vocabulary (collocations, new words, proverbs, emotive language, proverbs, at least 50% of the suggested items from the handout template, i.e., 8 items);
- Sophisticated grammar structures with examples from your presentation (at least 50% of the suggested items from the handout template, i.e., 9 items);
- Your main arguments with references to identify the origin of your information in APA format (at least 2).

You may also include relevant pictures, tables and graphs, links, and other helpful information for your classmates to grasp your ideas.

The handout should be **uploaded by June 17th at 16:00 together with the video**. Failing to send the handout will result in **subtracting 0.5 points** from your project grade (that is, if you get "5.8" in the rubric, your grade will be "5.3".

The topics were allocated by means of *random.org*.

Topics

1. "Better a drop of wisdom than an ocean of gold." Do Chileans live up to this principle?	
2. "Learn from the mistakes of others. You can't live long enough to make them all yourself." Do Chileans live up to this principle?	
3. "Consideration for others is the basis of a good life, a good society." Does this apply to Chile?	
4. Should military service be compulsory in Chile?	--
5. Chileans have become too greedy and selfish. Do you agree?	
6. What factors are the most important in defining who you are? Is your nationality one of them?	--

7. “The grass is always greener on the other side (of the fence).” How does this apply to Chile?	
8. “A leopard cannot change its spots.” Do you think this proverb applies to Chileans when dealing with classism and/or sexism?	--
9. “The youth can walk faster but the elder knows the road.” How does this proverb apply to Chile?	
10. Is the phrase "Live and let live" a fair description of the concept of tolerance?	
11. What proverb(s) do you think describe Chileans’ outlook on life?	--
12. Is it true that people don’t learn from the past? What about Chileans?	--
13. Is there a possibility to get rid of the stereotypes that are associated with different countries and their inhabitants? Explain.	
14. When people move to another country, some decide to adopt the customs of the new country, while others decide to keep their own customs. Compare these two positions. Which one do you prefer?	
15. Do you think Chile should offer help to refugees?	--
16. Do you think that policies the government has adopted to tackle the immigration crisis in northern Chile have been effective?	--
17. Should Chileans be proud that so many immigrants would like to live in their country?	
18. How much racism is there in Chile? Is it disappearing or getting worse?	
19. Why might Chileans feel challenged by the arrival of immigrants in the country?	--
20. How can the traditions and customs of immigrants be better integrated into our society?	--

21. How would you describe your relationship with elders? Is it true that Chileans don't show respect to their elders anymore?	
22. How are traditions passed down between generations in different Chilean communities?	
23. Do you think every person should have a guiding principle or to live up to?	--
24. Refer to a Chilean freedom fighter and discuss the rules of thumb they followed or have followed.	--
25. Has the Chilean law of food labelling changed the country's eating habits?	
26. Do you agree with the elimination of mandatory Physical Education for 3rd and 4th graders in Chilean high schools?	
27. Choose a Chilean freedom fighter and argue for/against his/her objective. Who are/were they and what do/did they do? Refer to personality traits and personal stories that made them fight for a cause.	
28. Are there rights and responsibilities so central to our way of life they should be entrenched in Chile's new constitution?	--
29. Nelson Mandela said: "Education is the most powerful weapon which you can use to change the world." Do you agree with this and how does this apply to Chile?	--
30. To use or not to use violence when fighting for freedom or for a cause. Where do you stand? Refer to icons such as Mandela, Gandhi or others you know.	
31. How are the elderly viewed in Chile? Do they face any specific challenges in terms of protecting their human rights? If so, what are being taken to protect them? Are they effective?	--
32. In Chile, women are on average paid less than men for doing the same job. Why is this and what can be done to change the situation?	
33. Do you think prisoners in Chile should be allowed to vote?	--
34. Can people under age 21 make a real impact on Chilean society? What examples from the past or present can you think of to support your opinion?	--

35. "Whenever you take a step forward, you are bound to disturb something." Do Indira Gandhi's words describe what has been happening in Chile since Oct. 18, 2019?	
36. How many human rights do you think are fully observed in your country? In the world?	--
37. Mental health must be a basic human right in Chile. Do you agree?	
38. What has been the Chilean government's response to the COVID19 pandemic? Has it been effective?	--
39. The pandemic's impact on mental health throughout Chilean society will likely outlive COVID-19. Do you agree?	
40. Currently, many public health institutions in Chile have refused to inoculate undocumented immigrants against COVID19. Do you agree with their position?	--

Rubric

	1	2	3	4
Relevance of information & organization	Information is not related to the topic at all. Topic shows no clear organization. Less than 2 sources of information are used.	One or two ideas and examples are not relevant and do not help to develop the topic. Organization is neglected; there is constant backtracking or hesitation. Less than 2 sources of information are used.	One or two ideas and examples are not relevant and do not help to develop the topic. Organization can be improved. 2 or more sources of information are used, but they are not varied.	All ideas and examples are relevant and help develop the topic. Good organization. It clearly serves the purpose of argumentation. 2 or more varied sources of information are used.
Discourse	The issue is described, but not presented as topic for argumentation. Arguments, if present, are weak and unsupported.	There is some attempt to present the issue as a topic for argumentation, but most of the arguments are presented, but not developed.	Half of the ideas presented supported the argument. These ideas are coherent and there is a fair use of cohesive devices. Ideas are sometimes supported by warrants.	Most or all of the ideas support the discussion presented. The ideas are coherent, cohesive devices are used, and supported by appropriate warrants.
Vocabulary	Very basic use of vocabulary. Use of true and false cognates. There is obvious interference with the L1.	Some of the expressions used show some degree of complexity, but this is not a constant behaviour.	The student used several sophisticated and new expressions that provided richness to his/her discourse, but at times it appeared contrived.	The student uses many sophisticated and new expressions that provided richness to their discourse making it appear natural.
Grammar	The student mostly produces simple grammatical structures and struggles with them. Major mistakes are made and none or only a few are self- corrected	The student produces some complex grammatical structures and struggles with simple structures. Some major mistakes are produced, and some are self-corrected.	Some complex grammatical structures are used but the student shows some difficulty to produce them. Some mistakes are made, and the student only sometimes is able to self-correct.	Complex grammatical structures are used to contribute to the points presented. Only minor mistakes are made, and the student is able to self-correct most of them or all of them.
Pronunciation	Intelligibility is hindered by pronunciation and intonation highly interferes with the L1.	Intelligibility is partially achieved because there are several sounds from the L2 that have not been acquired. Intonation can be improved.	Intelligibility is achieved and only a few sounds are not produced following the patterns of the L2. Intonation is appropriate.	Intelligibility is fully achieved and only on a few occasions sounds are not produced following the patterns of the L2. Intonation is definitely closer to the L2.
Pace & delivery	Highly hesitant. Pauses and hesitations interfere with communication.	Natural hesitations and pauses filled with English sounds and expressions. No interference with communication.		
Presentation length	Presentation is carried out within the stated period of time (+/- 30 seconds).			

Universidad de Chile
Facultad de Filosofía y Humanidades
Departamento de Lingüística
Advanced English I
2021

ORAL PROJECT

Student	
Topic	
Video link (optional)	

USEFUL GRAMMAR

(AT LEAST 9 ITEMS WITH EXAMPLES FROM YOUR PRESENTATION IN FULL SENTENCES)

verb patterns	+ infinitive	
	+ gerund	

conditionals	zero	
	first	
	second	
	third	
	mixed	

modals	+ infinitive	
	+ past participle	
	+ be + ing	
	for future deduction	

perfect aspect	past (simple / continuous)	
	present (simple / continuous)	
	future (simple / continuous)	

passive voice	
causative	

used to / would	
wishes and regrets	
reported speech	

USEFUL VOCABULARY

(AT LEAST 8 ITEMS WITH EXAMPLES FROM YOUR PRESENTATION IN FULL SENTENCES)

idioms & figurative language	
emotive adjectives & adverbs	
compound words	
phrasal verbs	

ARGUMENTS & REFERENCES

(AT LEAST 2)

	Arguments	References in APA format
1		
2		

UNIVERSIDAD DE CHILE
FACULTAD DE FILOSOFÍA Y HUMANIDADES
DEPARTAMENTO DE LINGÜÍSTICA
LENGUA INGLESA NIVEL INTERMEDIO I- 2019

ORAL PROJECT

Name: ……..... Final Mark: ……

Part 1. Presentation (25% of the final mark) ...../16 points Mark:..... 25%:.....

Organization	Lacks organization. Difficult to follow. Continuous backtracking and/or hesitation that interrupt flow of speech. 0-1 p	Organization needs improvement. Continuous backtracking and/or hesitation that interrupt flow of speech sometimes. 1.5- 2 p	Organization is mostly good. Minor lapses. Backtracking and/or hesitation do not interrupt flow of speech most times. 2.5 – 3 p	Well-organized presentation. Very few or no lapses. Easy to follow. Backtracking and/or hesitation do not interrupt flow of speech. 3.5 -4 p
Topic Mastery	No evidence of practice. Student reads at all times or topic has been clearly memorized. 0-1 p	Some evidence of practice but frequent topic-related hesitation disrupts flow. Constant reading or topic seems to have been memorized. 1.5-2 p	Good control of topic despite minor lapses. Reading at times or topic seems to have been memorized to some extent. 2.5 – 3 p	Very good. Dominates topic at all times. Fully at ease with topic. No reading is needed. Topic does not seem to have been memorized. 3.5- 4 p
Perspective	Unable to look at the topic beyond personal experience. 0-1 p	Some attempts at a broader view of the topic but it remains self referent. 1.5 – 2 p	Good. A couple of personal examples are used but topic remains mainly impersonal. 2.5-3 p	Very good. Presents the topic from an objective point of view. Personal opinion is clearly distinguished from facts presented as part of the topic. 3.5- 4 p
Strategies (Self-monitoring, Approximation, Time-gaining strategies)	Completely unaware of his/her mistakes. No clear use of strategies at all. 0-1 p	Some attempts at using strategies. An important number of words/phrases are left uncorrected. 1.5 – 2 p	Frequent use of strategies when needed. 2.5-3 p	Strategies are present in all cases when needed. 3.5- 4 p

Part II. Language (75% of the final mark) ...../15 points Mark:..... 75%:.....

Grammar	**Feature is absent.** Completely inconsistent in the use of even the simplest verb tenses. Elementary mistakes are present. Completely inconsistent in the use of syntactic structures. Completely inconsistent use of determiners relative clauses, adjectives, adverbs, etc. 0-1 p	**Poor.** Some attempts at using verb tenses correctly and consistently but most cases are incorrect. Student uses a limited range of verb tense but these are mostly simple tenses. Elementary mistakes are present. Some attempts are made at using correct syntax and other grammatical features but a large number of mistakes remain. 2 p	**Satisfactory.** Frequently uses verb tenses correctly. Student uses a limited range of verb tenses in a correct way most times. Frequently uses correct sentence syntax and other grammatical features but several mistakes remain. Enough for the level but needs improvement. 3 p	**Good.** Most verb tenses are correctly used but one or two mistakes remain. Student uses different verb tenses in a correct way most times. A couple of mistakes still remain in syntax and other grammatical features, such as concord 4 p	**Very Good.** Consistent use of all verb tenses studied *in the program* so far. Sentences are syntactically well-structured. SV concord is observed throughout discourse. Consistent use of determiners, adjectives (comparative/superlative), adverbs, nouns, relative clauses and prepositional phrases to modify/ complement the head in the noun phrase. 5 p
Vocabulary	**Feature is absent.** Completely inadequate use of vocabulary in terms of topic, meaning and collocation. Mother tongue interference most times. Wrong register/ frequent shift. No variety of vocabulary or linking words at all. 0-1 p	**Poor.** A few attempts are made at using the appropriate words for the context. Meaning and collocation of words often incorrect. Frequent mother tongue interference. Frequent register shifts. Very few synonyms, no variety in the use of linking words. 2 p	**Satisfactory.** Some use of appropriate vocabulary in the context of the topic presented. Some mother tongue interference. Frequent use of correct words with the right collocation. Some register shifts. Some use of synonyms and different linking words. Enough for the level but needs improvement. 3 p	**Good.** Correct word meaning and collocation most times. Only one or two register shifts. Interference from mother tongue does not affect use of vocabulary. Uses different linking words and synonyms but some repetition is still noticeable. 4 p	**Very Good.** Appropriate use of vocabulary in the context of the topic presented. Appropriate use of vocabulary in terms of meaning and collocation. No interference from mother tongue. Adequate register. Variety in the use of synonyms and linking words. 5 p

Pronunciation	**Feature is absent.** Completely inconsistent production of sounds in the English phonological system. Strong mother tongue interference. Spanish intonation at all times. 0-1 p	**Poor.** Most sounds are interfered but some attempts are made at producing English sounds. Mother tongue interference is noticeable. Spanish intonation is often noticeable. 2 p	**Satisfactory.** Some sounds remain interfered or are produced incorrectly. Mother tongue interference is still noticeable, though the student seems to be aware of it. Spanish intonation is often noticeable, though the student seems to be aware of it. 3 p	**Good.** Most sounds are correctly pronounced. Specific sounds are interfered at times. Non-systematic errors. Various attempts at avoiding Spanish intonation. 4 p	**Very Good.** Consistent production of all sounds in the phonological system of English (vowels and consonants) in most cases. No remarkable mother tongue interference Clear and consistent attempt at avoiding Spanish intonation. 5 p

Comments:

UNIVERSIDAD DE CHILE
FACULTAD DE FILOSOFÍA Y HUMANIDADES
DEPARTAMENTO DE LINGÜÍSTICA
LENGUA INGLESA NIVEL INTERMEDIO I

ORAL PROJECT 2018
GUIDELINES

1. The students will randomly pick a topic from a list. Once picked, the topic can not be changed.

2. Students will have to prepare a 5-minute presentation in which they will develop the topic from both a general and a personal perspective. The use of supporting ideas and examples is highly recommended.

3. This presentation will need visual aids to help the students develop their talk. However, if the students do not want to use Power Point presentations, they can use the board or any other kind of material they want.

4. The visual aids are not supposed to contain the whole presentation so that the students can read from them. The number of words used in each slide should be limited (10-15 words).

5. The day of the presentation, the students will be expected to know their topic and provide a clear development of ideas without reading from a piece of paper. Even though they can have a written material with them in case they forget something, they must address the class when talking.

Appendix 3: Survey Answers.

The following link leads to a downloadable excel spreadsheet, in which the answers of the survey shown in Appendix 1 were organized:

Oral Project format change survey answers