Une pratique psychométrique pour la validité du test de positionnement SELF en espagnol

Mayra-Alejandra Siordia-García

Une pratique psychométrique pour la validité du test de positionnement SELF en espagnol

SIORDIA-GARCÍA
Mayra-Alejandra

Sous la direction de GÓMEZ Lucía

UFR LLASIC
Département Sciences du langage et Français Langue Étrangère
Section Sciences du Langage

Mémoire de Master 2 Professionnel 24 crédits

Parcours : Didactique des langues et Ingénierie Pédagogique Numérique (DILIPEM)

Année universitaire 2016-2017

Une pratique psychométrique pour la validité du test de positionnement SELF en espagnol

SIORDIA-GARCÍA
Mayra-Alejandra

Sous la direction de GÓMEZ Lucía

UFR LLASIC
Département Sciences du langage et Français Langue Étrangère
Section Sciences du Langage

Mémoire de Master 2 Professionnel 24 crédits

Parcours : Didactique des langues et Ingénierie Pédagogique Numérique (DILIPEM)

Année universitaire 2016-2017

Remerciements

Je tiens en premier lieu à remercier toute l'équipe d'*Innovalangues* de m'avoir donné l'opportunité de faire un deuxième stage au sein de l'équipe SELF espagnol, et pour les repas de midi que nous avons pu partager ensemble. Cela a été l'occasion d'échanger sur nos différentes tâches et nos cultures.

Toute ma reconnaissance à Madame Cristiana CERVINI, coordinatrice du lot SELF, pour sa disponibilité à m'éclairer sur toutes mes questions.

Je tiens aussi à exprimer toute ma gratitude à Madame Lucía GÓMEZ, responsable scientifique de l'équipe SELF espagnol et directrice du mémoire, qui m'a conseillé et encouragé tout le long de mon stage, et a lu attentivement mon mémoire.

Je remercie également l'équipe SELF espagnol de m'avoir gentiment accueillie pendant ce stage, principalement Monsieur Carlos Chávez, mon responsable de stage.

Toute ma gratitude à Triscia et à Sylvain qui ont aussi démontré une énorme disponibilité pour répondre à mes questions, telle qu'ils l'ont fait l'année passée.

Merci à Monsieur Gomez Montesinos, enseignant chercheur à l'Université de Sonora, au Mexique, spécialiste en psychométrie, qui a gentiment pris le temps de m'éclaircir sur des questions très pointues dans le domaine de la psychométrie.

Merci à mes parents et à mon frère de m'avoir encouragée inconditionnellement, et pour leur indéfectible présence malgré la distance géographique.

Un énorme merci à Gilles Verilhac qui m'a inconditionnellement soutenue tout au long de cette année scolaire.

Merci aux amitiés qui sont nées au cours de cette année scolaire, principalement à Isabelle Tournay, Régis Maldamé et Laura Mazzarella, qui m'ont beaucoup apporté sur les plans personnel et professionnel.

UNIVERSITÉ Grenoble Alpes

DÉCLARATION

Ce travail est le fruit d'un travail personnel et constitue un document original.

1. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
2. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
3. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
4. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : SIORDIA-GARCÍA

PRENOM :Mayra-Alejandra

DATE : 08/09/2017

SIGNATURE : 

Sommaire

Introduction

En janvier 2012, le projet SELF (Système d'Evaluation en Langues à visée Formative) a été officiellement lancé. Il s'agit du test de positionnement proposé par Innovalangues du Service des Langues de l'Université Grenoble Alpes.

Ce projet de type communic'actionnel se donne pour but, d'une part, de positionner les candidats dans un des niveaux[1] du CECRL et, d'autre part, de fournir un diagnostic formatif des apprenants pour améliorer leurs compétences/connaissances linguistiques en langue étrangère.

En mars 2016, j'ai eu l'occasion de rejoindre les concepteurs pédagogiques de l'équipe SELF espagnole en tant que stagiaire. À l'époque, le projet était en phase de conception des tâches d'évaluation (TE). Dans le cadre de ce projet, la banque des TE vise à tester « la compétence communicative, c'est-à-dire une capacité à utiliser la langue en contexte, pour agir avec la langue » (Cervini & Jouannaud, 2015).

Le travail de conception de SELF a été axé sur les trois piliers de l'évaluation : « la validité, la fiabilité et la faisabilité » (CECRL, 2001, p. 135). A présent, le projet a pour objectif la validation objective du test. Dans ce contexte, il m'appartient d'approfondir les concepts de validation et de fiabilité afin de mener à bien ce processus de validation : il s'agira d'un processus rigoureux mené à partir d'analyses psychométriques.

Ainsi, ce mémoire envisage de répondre aux questions suivantes : *Qu'est-ce qu'une tâche d'évaluation valide ? Qu'est-ce que les analyses psychométriques nous permettent d'observer sur la typologie des tâches proposées ? Quelles sont les contributions des logiciels TiaPlus et Winsteps à la validation du test ?*

Pour expliquer clairement ce processus de validation et répondre aux questions ci-dessus, ce mémoire se compose de trois parties. La définition du contexte du stage constituera la première partie de ce travail où la présentation de l'organisme

[1] A1, A2, B1, B2 et C1

d’accueil ainsi que des missions confiées dans le cadre du stage seront exposées. La deuxième partie portera sur le cadre théorique ; elle se donne pour objectif de faire une introduction à la psychométrie. Dans la troisième partie, le processus des analyses psychométriques sera approfondi afin d’évoquer les réponses à la problématique. La deuxième et troisième parties sont fortement liées l’une de l’autre. Nous attendons que certains passages puissent être éclairés lorsque sera expliquée la façon dont nous avons mené les concepts théoriques à la pratique. Enfin, une dernière partie sera consacrée aux conclusions tirées de ce travail.

Partie 1

-

Contexte Institutionnel

Chapitre 1. Présentation de l'organisme d'accueil

1. Projet Innovalangues

Figure 1 : logo du projet Innovalangues

Innovalangues est un projet qui a été lancé officiellement le 14 juin 2012. Il s'inscrit dans le cadre du programme IDEFI, Initiatives D'Excellences en Formations Innovantes, et financé par l'ANR, qui assure un suivi scientifique permettant la conception de dispositifs de formation performants en langues. La direction scientifique du projet est assurée par Madame Monica Masperi.

1.1. Objectifs

L'objectif principal d'*Innovalangues* est de contribuer, de manière significative, à améliorer les pratiques d'enseignement et de formation dans le domaine des langues. Il s'agit d'un projet au service de l'ensemble de la communauté éducative nationale, et au-delà, grâce à des actions diffusantes.

Selon le site du projet, les objectifs opérationnels d'Innovalangues sont :

I. Capitaliser les résultats de la recherche en didactique des langues,
II. Se doter d'un environnement numérique consacré à la formation des langues,
III. Former le public cible à travers cet environnement en s'adaptant à ses besoins afin de lui permettre d'acquérir le niveau B2 dans la langue cible,
IV. Étendre la recherche-action en langues réalisée sur le site grenoblois, et au-delà, grâce aux actions menées par le conseil de formation des formateurs en didactique des langues.

Cet écosystème numérique plurilingue se présente sous la forme de deux environnements *open sources*, distincts et complémentaires :

- **SELF : Système d'Évaluation en Langues à Visée Formative**. Il héberge un ensemble de modules d'évaluation diagnostique et formative, destinés à faire un bilan de compétences linguistiques d'après l'échelle du Cadre

Européen Commun de Référence en Langues (dorénavant CECRL). En effet, ce test proposé en six langues cibles : italien, anglais, mandarin, espagnol, français et japonais, vise à faire une synergie avec les parcours d'apprentissage hébergés sur la plateforme : Environnement Numérique Personnalisé d'Apprentissage en Langues. Ainsi, l'utilisateur pourra suivre une formation qui lui permettra de faire évoluer ses compétences en langue étrangère.

- **Environnement Numérique Personnalisé d'Apprentissage en Langues « ENPA-Langues ».** Hébergeur des parcours d'apprentissage (hybrides ou complètement à distance) calibrés grâce à son suivi scientifique, qui permet de capitaliser les avancées en didactique des langues.

Des équipes langues contribuent à la conception et au développement des ressources, des parcours et des dispositifs de formation de ces plateformes. Ces équipes langues travaillent à la fois sur le suivi d'un responsable scientifique de la langue ciblée, et sur le suivi d'un coordinateur du lot de travail, c'est-à-dire un référent scientifique du lot.

1.2. Les lots

D'après le site du projet Innovalangues, les lots de travail se donnent pour objectif de favoriser la créativité susceptible d'alimenter la plus-value innovatrice du projet. Les ressources et les outils conçus par eux convergeront progressivement dans l'ENPA Langues.

THEMPPO*, THEMatique Prosodie Production Orale :*

La mission de THEMPPO vise l'amélioration de la production orale en langue ciblée, et pour ce faire, met en avant la prosodie. Dans cette optique, l'équipe propose le développement d'outils, d'activités et de parcours d'apprentissage.

COCA*, Compétence Orale : Conception et Assistance :*

Le lot COCA se donne pour objectif la conception de solutions pour la compréhension de l'oral, tant pour l'étudiant en situation d'apprentissage, que pour l'enseignant en situation de concepteur des activités visant cette habileté.

GAMER*, Gaming Applications for Multilingual Educational Resources :*

Ce lot conçoit des jeux d'apprentissage pour les langues, et correspond donc à l'approche ludique de la pratique didactique et de l'acquisition de compétences langagières.

SELF*, Système d'Évaluation des Langues à visée Formative :*

Il s'agit d'un test de positionnement qui participe à la construction de l'écosystème Innovalangues, en synergie avec l'ENPA Langues. (Cf. Chapitre 2. Cadre du stage)

Chapitre 2. Cadre du stage

2. Projet Système d'Évaluation en Langues à visée Formative (SELF)

Le 'Système d'Évaluation en Langues à visée Formative' (SELF), est une plateforme qui héberge un test de positionnement plurilingue, numérique, autocorrectif et en ligne. Il est coordonné par Madame Cristiana Cervini, référente scientifique du lot SELF. Il est issu d'un vaste état des lieux qui ne se base pas seulement sur la recherche de l'existant des examens de positionnement, mais aussi des examens de certification.

À ce stade, il est déjà disponible en trois langues : italien (langue pilote), anglais et mandarin. Les équipes en espagnol, japonais et FLE, se préparent pour finir les dernières étapes de calibrage du test.

L'un des apports techniques du SELF à la communauté éducative est que son système d'édition et d'administration est très intuitif. Il est capable d'absorber un nombre élevé de connexions en simultané. À présent, des pics d'utilisation à hauteur de 160 connexions simultanées ont été constatés. De même, les résultats de positionnement sont fournis automatiquement à la fin de la passation.

Dans une perspective actionnelle, il envisage d'évaluer « la compétence communicative, c'est-à-dire la capacité à utiliser la langue en contexte, pour agir avec la langue » (Cervini & Jouannaud, 2015). De ce fait, afin de pouvoir répondre au principe d'authenticité, les tâches d'évaluation du SELF sont issues de ressources authentiques.

Dans le but de favoriser les pratiques de la didactique des langues, ainsi que d'orienter les élèves vers le niveau cible qui leur permettra de développer leur compétences linguistiques, SELF se traduit par « la mise en œuvre d'une démarche intégrée d'orientation d'utilité nationale[2], qui permettra à la fois d'évaluer de manière fiable les forces et les faiblesses de l'apprenant, et de faciliter la constitution des groupes de niveau » (Cahier des Charges[3], 2015, p. 2). (Dorénavant CC).

Ainsi, l'ensemble des tâches d'évaluation qui ont été validées lors du processus de validation (dorénavant : la banque des tâches d'évaluation) visent à positionner les candidats du niveau A1 au niveau C1 du CECRL, dans un temps de passation réduit (moins d'une heure), à travers l'évaluation de trois habiletés : la Compréhension de l'Oral (dorénavant CO), la Compréhension de l'Écrit (dorénavant CE) et l'Expression Écrite Courte (dorénavant EEC).

Enfin, comme produit final, SELF proposera un test de positionnement plurilingue issu d'une validation rigoureuse faite à partir d'analyses psychométriques.

2.1. La nomenclature employée par SELF

Toutes les tâches d'évaluation conçues dans le cadre du projet SELF suivent une nomenclature commune. Ainsi, une tâche d'évaluation peut être conformée par un, deux, trois, ou même quatre questions, que nous allons appeler **items**. Nous proposons une image illustrative des composantes d'une tâche d'évaluation. Chacun de ces éléments est décrit par la suite.

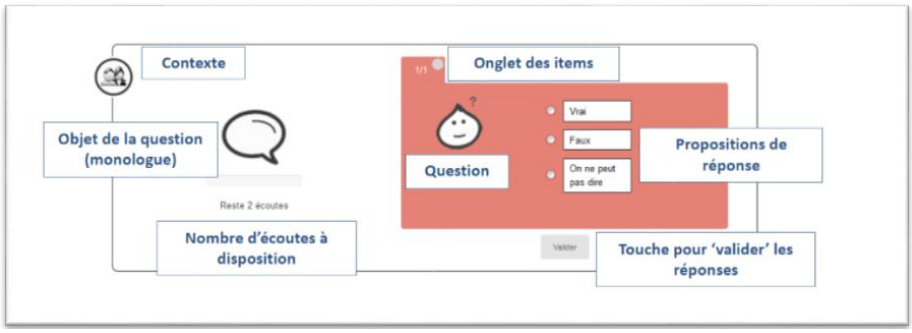


Figure 2 : tâche-type de la compréhension de l'oral.

[2] Souple, ouvert et libre de droits.
[3] L'information suivante est extraite du Cahier des Charges, (CC), Action 4 : Lot « SELF », V.8. Il s'agit encore d'un document scientifique, interne, et confidentiel. Cette partie n'est pas disponible publiquement dans la mesure où les recherches sur le projet Self d'INNOVALANGUES sont encore en cours de développement.

1. Contexte : une information permettant au candidat de se mettre dans le contexte de la situation d'écoute, de lecture, ou de rédaction proposée ;
2. Objet de la question : un texte oral, ou écrit, sur lequel porte l'effort de compréhension et la / les question(s) proposée(s) ;
3. Nombre d'écoutes à disposition. Seulement pour la CO ;
4. Onglet des items : dans le cas où plusieurs questions sont proposées ;
5. Question(s) / Item(s) ;
6. Proposition(s) de réponse(s) ;
7. Une touche pour valider : elle ne fonctionne qu'après avoir répondu à toutes les questions composant la tâche d'évaluation.

Dans cette nomenclature, plusieurs types de tâches d'évaluation sont proposés. Nous allons par la suite présenter celles qui ont été employées dans le cadre du stage.

1.- Vrai/Faux (V/F) : pour répondre à l'objet de la question, le candidat doit choisir entre vrai ou faux. Un seul des deux choix est possible.

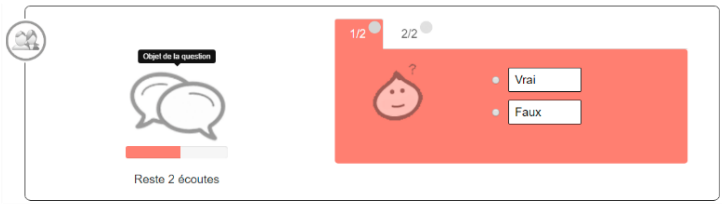


Figure 3 : tâche du type V/F

2.- Vrai/Faux/on ne peut pas dire -*non mentionné* - (V/F/NM) : cette tâche d'évaluation ressemble à la précédente. Elle propose un troisième choix de réponse : *on ne peut pas dire*. Au total, elle propose trois propositions de réponse, dont uniquement une est correcte.

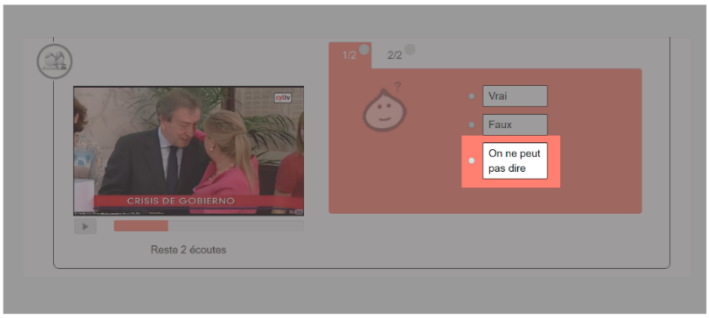


Figure 4 : tâche du type Vrai/Faux/NM

3.- Questionnaire à Choix Unique (QRU) : pour répondre à l'objet de la question, cette typologie de tâche contient plusieurs affirmations, dont une seule est correcte.

Figure 5 : tâche du type QRU

4.- Question à Réponses Multiples (QRM) : cette typologie de tâche contient plusieurs affirmations dont 2, ou plus, sont correctes.

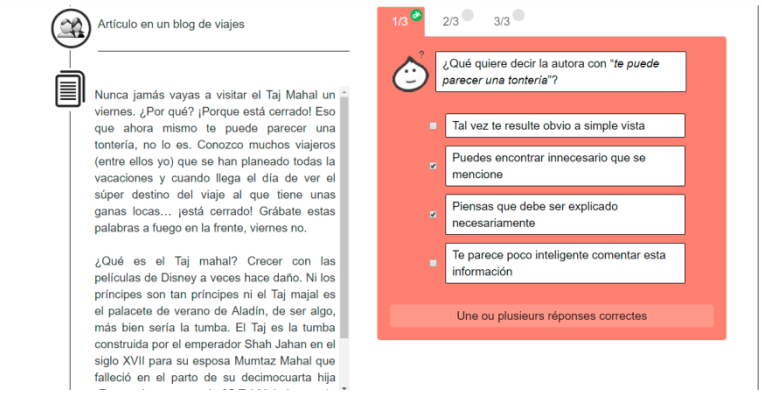


Figure 6 : tâche du type QRM

5. Question à Réponse Ouverte Courte (CROC) : aussi connue sous le nom de texte à trous. Le candidat doit compléter le texte avec les mots corrects.

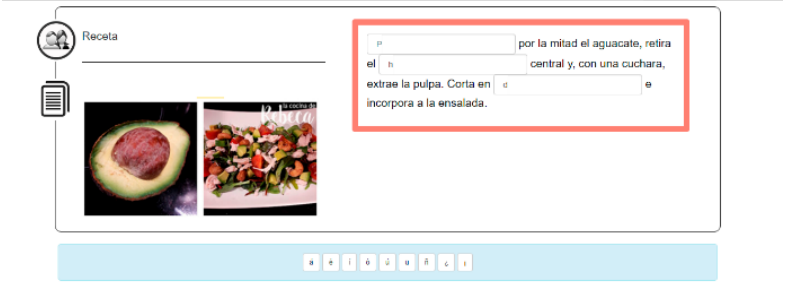


Figure 7 : tâche du type QROC

2.2. Contexte du projet SELF en espagnol

Le projet SELF en espagnol a été lancé en janvier 2015. L'équipe est composée par Mesdames Cristiana CERVINI, experte en *testing* et coordinatrice du lot SELF, Lucía GOMEZ, responsable scientifique du projet en espagnol, Patricia FRANCO, co-coordinatrice de l'équipe Espagnol d'Innovalangues, et Monsieur Carlos CHÁVEZ, co-coordinateur de l'équipe Espagnol d'Innovalangues.

L'équipe prépare actuellement les passations du pré-test. Cette étape constitue l'une des phases de la validation du test. Elle est faite à partir des analyses psychométriques. Nous proposons ci-dessous une image du cycle du *testing* de l'équipe SELF espagnol. Elle permet d'illustrer le stade actuel du projet.

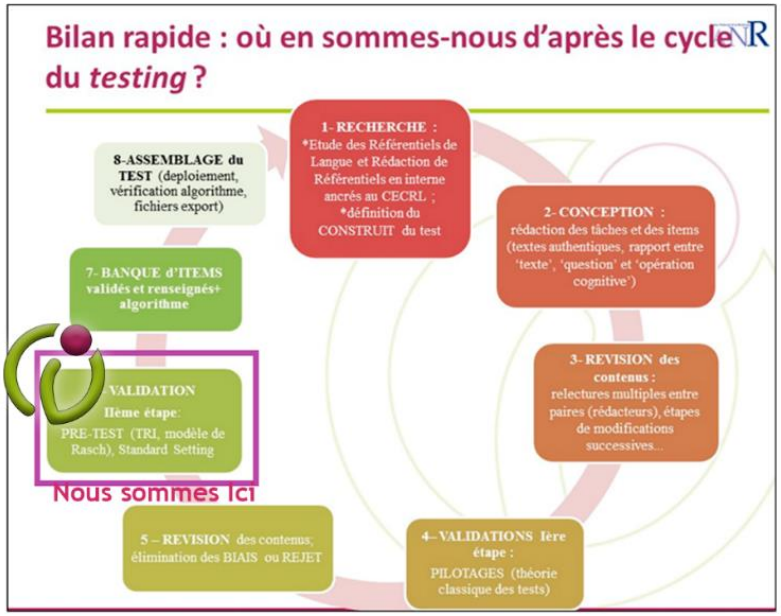


Figure 8 : cycle du *Testing*

Afin de mettre le lecteur dans le contexte, nous allons décrire les étapes précédentes qui ont mené SELF espagnol à son stade actuel :

1.- RECHERCHE : il s'agit de la recherche des référentiels de la langue cible, ainsi que de la création du syllabus synthétique. Ils doivent être ancrés au CECRL. Ils permettent de définir les points à évaluer par le test, ou, autrement dit, ils permettent de définir le construit à évaluer. Dans le cas de SELF en espagnol, le référentiel est le *Plan Curricular del Instituto Cervantes* (dorénavant PCIC). De son côté, le syllabus synthétique est une compilation issue de l'étude des méthodes de l'apprentissage des langues, ainsi que de l'expertise des concepteurs, qui en même temps sont des

enseignants de langue. Ceci qui permet de cibler le construit à évaluer d'après les besoins d'évaluation du public cible. À savoir, SELF en espagnol est un test qui vise à positionner un public francophone natif.

2.- CONCEPTION : il s'agit de la recherche des ressources authentiques à partir desquelles les tâches d'évaluation seront développées. Cette recherche est accompagnée de l'obtention des droits d'auteurs. Les ressources peuvent être modifiées si le concepteur le considère nécessaire. Cependant, toute modification faite à la ressource est accompagnée d'une réflexion qui permet au concepteur de se mettre dans la peau du candidat. Ce dernier est considéré comme acteur social qui devra accomplir une tâche dans une situation donnée. Ceci est le principe de conception du projet.

3.- REVISION DES CONTENUS : c'est l'une des étapes de validation faite à partir de la collaboration de la référente scientifique de la langue cible, Lucía Gómez, de la référente scientifique de SELF, Cristiana Cervini, et des concepteurs pédagogiques Carlos Chávez et Patricia Franco.

4.- VALIDATION 1ère ÉTAPE : elle constitue la première étape des analyses psychométriques. Elle est faite à partir de la méthode connue comme La Théorie Classique du Test[4] (TCT), et dans le cadre des tests que nous appelons PILOTAGES. Les étudiants d'Espagnol Langue Étrangère, sont convoqués à passer le test par niveau. Par exemple, le pilotage du niveau A1, est fait avec un échantillon d'étudiants qui font des études en espagnol A1.

5.- REVISION DES CONTENUS : c'est l'analyse des ressources. Elle est faite d'après les résultats tirés de la première analyse psychométrique. Cette étape permet soit de modifier les items, soit de les éliminer. Elle peut se traduire par la recherche des items qui ont été validés d'après la TCT, ainsi que par l' « affinage » des ressources d'après les indicateurs fournis par la TCT (Cf. le logiciel TiaPlus).

6.- VALIDATION 2ème ÉTAPE : elle permet de faire une deuxième analyse psychométrique à partir d'une autre méthode : la Théorie de la Réponse à l'Item (TRI)[5]. Elle est faite dans le cadre des tests que nous appelons : PRÉ-

[4] Analyses faites avec le logiciel TiaPlus
[5] Analyses faites avec le logiciel Winsteps.

TESTS. En ce moment, les étudiants sont convoqués à passer un test qui contient des tâches d'évaluation du niveau A1 au niveau C1. Par conséquent, l'échantillon est composé par des étudiants qui appartiennent à ces niveaux (Cf. le logiciel Winsteps). Cette étape est suivie et complétée par un processus, où les points de césure des niveaux du test seront déterminés à travers un processus qualitatif et quantitatif. Dans ce processus, connu comme *standard setting*, l'équipe fera appel à un panel de spécialistes en Espagnol Langue Étrangère.

3. Missions confiées

Les missions qui m'ont été confiées en entretien le 8 février 2017, sont les suivantes :

- Gestion des tests avant pilotage (éditeur)
- Modification des tâches (après pilotage)
- Appui au processus de pilotage
- Édition des ressources sur l'éditeur
- Appui au processus d'assemblage du test
- Soutien logistique et pédagogique aux différentes phases du projet

Ceci dans le but de mettre en place et de développer le pré-test. En effet, l'équipe avait comme objectif de déployer SELF en espagnol pour la rentrée 2017. Cependant, il s'avère important de présenter les contraintes auxquelles nous avons été confrontées, et qui nous ont conduites à modifier les tâches initialement prévues.

- En mars 2017, lors du début de ma collaboration comme stagiaire au sein du projet, les pilotages étaient pratiquement terminés. De ce fait, la première mission qui m'a été confiée a constitué à un bilan des résultats du pilotage (voir annexe 1). Ce qui a été l'occasion de faire une mise au point sur le nouveau stade du projet.
- Les dates de convocation aux pré-tests : les étudiants ont été convoqués pour les semaines 12 et 13 du calendrier universitaire, période des examens de fin d'année scolaire, ce qui n'a pas favorisé la participation des candidats.

- La formation en psychométrie qui devait être menée par le Centre International d'Etudes Pédagogiques (dorénavant CIEP) en mai 2017, n'a pas eu lieu. Une autre formation aura lieu le 12 septembre 2017.

Dans ce contexte, nous avons repris la conception des tâches d'évaluation du niveau C1. En effet, le premier déploiement du SELF espagnol allait être composé par des tâches d'évaluation du niveau A1 au niveau B2.

Nous sommes donc passés par une étape où il fallait faire des choix. En attendant la prise de décision, j'ai proposé de travailler sur les tutoriels du SELF, et les adapter aux besoins de la langue espagnole constatés pendant les passations des pilotages[6]. Ceci a débouché sur une vidéo tutoriel SELF espagnol (storyboard en annexe 2), et un guide d'utilisation SELF en espagnol (voir annexe 3). Ensuite, un guide d'utilisation SELF interlangue a été créé (voir annexe 4), avec la collaboration de Sylvain Coulange, de l'équipe SELF Japonais, et de Cristiana Cervini, avec le suivi de Mónica Masperi.

J'ai également décidé de proposer un mémoire sur les analyses psychométriques, car ce sujet représente la suite de ce que j'ai appris durant mon stage de M1. La maquette du Master DILIPEM est riche en conception, ceci me permet donc de compléter ma formation et de découvrir une partie très importante de la conception pédagogique : les pilotages. Il s'agit de la découverte des pilotages à travers des modèles statistiques.

Avec une optique très personnelle et sans avoir aucune base scientifique, j'ai constaté qu'il est très compliqué pour des membres qui travaillent sur la conception d'un dispositif numérique, de se mettre tous d'accord. En effet, chacune des conceptions exige un énorme investissement. C'est pourquoi il est très difficile de s'autoévaluer de façon objective. Ainsi, le début d'une évaluation objective de ce que nous concevons, surtout dans le cadre de l'évaluation, est possible à travers la psychométrie.

---

[6]Par exemple, dans le cas des Questionnaires à Réponse Ouverte Courte, il s'agissait de mettre l'accent sur écrire une majuscule après un point (faire remarquer qu'il y a un menu qui propose les caractères spéciaux), proposer un mot pour chacun des espaces proposés. Ici, nous avons envisagé de réduire le temps de la durée de la vidéo de 4 minutes 34 secondes à 3 minutes 20 secondes. En effet, la vidéo proposée sur la plateforme SELF est une vidéo qui a été faite au début du projet. Ainsi, les pilotages ont permis de repérer les points à affiner par rapport aux besoins de la langue.

Ainsi, nous allons, dans la deuxième partie, présenter les principes théoriques des modèles psychométriques adoptés comme méthodologie par SELF. Ensuite, Dans la troisième partie, le processus de l'analyse psychométrique sera illustré à travers des exemples.

Partie 2

-

Cadre Théorique

Chapitre 3. L'évaluation et la psychométrie

1. Les trois piliers de l'évaluation

Cette section se donne pour objectif de reprendre des notions mentionnées dans mon mémoire de M1, mais surtout d'approfondir ces concepts et de les comprendre dans une optique psychométrique. Il s'agit des trois piliers autour desquels la conception d'un dispositif évaluatif doit être axée : « la validité, la fiabilité et la faisabilité » (CECR, 2001, p. 135).

Le CECR 2001 nous dit qu'un test est valable lorsqu'il donne le reflet exact de la compétence évaluée parmi tous les candidats. Autrement dit, un test est valable si ses items, ou tâches d'évaluation, évaluent, dans notre cas, une habileté langagière, et pas une connaissance antérieure, ou une connaissance du monde, par exemple.

La fiabilité, « c'est la mesure selon laquelle on retrouvera le même classement des candidats dans deux passations (réelles ou simulées) des mêmes épreuves. » (CERL, 2001, p. 135).

Enfin, la faisabilité fait référence au temps de passation d'un test ainsi qu'aux ressources et moyens dont un établissement dispose pour la mise en œuvre du dispositif d'évaluation.

Cependant, dans un contexte psychométrique, la validité et la fiabilité ne sont pas seulement des concepts, mais aussi des processus. D'après ALTE 2011, la validité d'un test se concentre sur l'interprétation des résultats du test, il s'agit du : « degré de preuves et de théorie sous-tendant l'interprétation des scores entraînée par les utilisations données des tests » (AERA, APA, NCME 1999). (Cité par ALTE, 2011, p. 16). De son côté, Guilford (1937), définit la validité comme « la corrélation entre les scores d'un test et la mesure objective de ce que le test est en train de mesurer » [Traduction libre] (cité par Martinez, Hernández & Hernández, 2014, p. 221)]. Ainsi, nous pouvons dire que le processus de validité d'un test fait l'objet d'une analyse, ou interprétation, objective des scores d'un test.

D'après Morales Vallejo 2007, la **fiabilité** d'un test se traduit par des méthodes de vérification qui, en même temps, se traduisent par des coefficients de fiabilité. Ce principe de fiabilité exprime le degré de précision de la mesure. Si le degré de fiabilité descend, l'erreur augmente. Selon l'auteur, la marge d'erreur nous permet de calculer le coefficient de fiabilité. Pour ce faire, il faut aussi étudier l'erreur typique de la moyenne.

Ainsi, un test peut être valide, mais pas fiable. Il s'agit donc de processus différents. Nous pouvons ainsi trouver d'autres référence sur ce sujet, comme celle-ci : « un instrument peut être valide parce qu'il mesure ce que l'on dit qu'il mesure, néanmoins, il peut le mesurer avec une grande marge d'erreur ». [Traduction libre] (Morales Vallejo, 2007, p. 4)

De son côté, Bachman (2004) nous dit que lorsque l'on mesure la compétence communicative de la langue, il faut penser à retrouver les similarités entre la validité et la fiabilité :

> *Mais au lieu de les considérer tous deux comme des concepts différents, je pense qu'ils peuvent être reconnus comme des aspects complémentaires d'un même intérêt pour mesurer-identifier, estimer et contrôler les effets ou facteurs qui peuvent affecter le score du test.* [Traduction libre] (Bachman 2004 p. 160).

2. La Psychométrie

La psychométrie est une science de la psychologie. Elle est aussi une étape-clé et essentielle permettant la validation du construit qui a été à posteriori créé. Dans le cadre de ce projet, SELF en espagnol, nous allons préciser que nous parlons du construit de la compétence communicative, c'est-à-dire :

> « Le trait, ou les traits, qu'un test est destiné à mesurer. (…) La capacité, ou un ensemble de capacités, qui seront reflétées dans la performance du test, et autour desquelles des déductions peuvent être faites par rapport à la base du score de test. » [Traduction libre] (Multilingual glossary of language testing terms, 1999) (Cité par INNOVALANGUES 2014).

D'après Martinez, et al., (2014), la mesure de ces traits doit se faire de façon objective et fiable, afin d'aboutir à la validation du test. Pour ce faire, il faudra donner une valeur à chacun des éléments du construit à évaluer, c'est-à-dire les items qui ont été créés à cet effet, afin de les quantifier.

Nous pouvons donc dire que ce qui est envisagé dans la psychométrie, c'est l'objectivité, car d'après les auteurs, cette branche de la psychologie se donne pour objectif de justifier, en suivant une méthodologie, le degré de précision avec lequel nous pouvons évaluer un candidat lors de sa passation du test en question. De ce fait,

tout jugement provenant des attentes du concepteur, lors de la création du test, sera justifié, ou bien révoqué.

Dans le cadre de SELF, il s'agit de mettre en lumière que l'outil d'évaluation, dit instrument de mesure dans la psychométrie, mesure la performance communicative à plusieurs niveaux (A1, A2, B1, B2, C1) avec un degré d'erreur minimum. Ainsi, ce processus de validation, dans le cadre du projet SELF d'Innovalangues, est fait à travers deux modèles : la Théorie Classique du Test (dorénavant TCT) et la Théorie de Réponse à l'Item (dorénavant TRI).

2.1. La Théorie Classique du Test (TCT)

Le modèle de la Théorie Classique du Test se donne pour objectif d'« obtenir la notation correspondant à une personne dans une dimension, ou trait donné, tel que son intelligence, le niveau d'un trait de sa personnalité, sa maîtrise dans une matière, etc. » [Traduction libre] (Olea, Ponsoda, et Revuelta. 1998, p. 1).

Le principe de la TCT se représente à travers la formule suivante :

$$Xi = Vi + Ei$$

> « Ce modèle expose, tout simplement, que la notation observée d'un candidat *Xi* provient d'une notation vraie *Vi*, qui est la quantité que l'individu possède de l'attribut, plus une erreur de mesure *Ei* » (Martinez et al., 2014, p. 38).

D'après Crocker et Algina (1986), l'une des caractéristiques de ce modèle est que le score de **la notation vraie est obtenu à partir de la moyenne des scores observés**. Ce qui explique quelques-unes de ses limitations : « la TCT met l'accent sur la notation globale du test obtenue à partir de l'ensemble des items» (Martinez, Hernandez et Hernandez, 2014, p. 38).

Selon Olea et al., (1998), les caractéristiques du test et la notation d'une personne ne peuvent pas être séparées, car la notation d'une personne est définie par rapport au nombre des questions atteintes. D'un autre côté, le niveau de difficulté de l'item est défini par rapport au nombre de personnes qui lui répondent correctement. Par conséquent :

- « Les propriétés psychométriques importantes du test, telles que la difficulté des items, ou la fiabilité du test, sont fortement liées au type de personnes

employées pour les calculer » [Traduction libre] (Muñiz, 2010, p. 62). Autrement dit, les scores obtenus à travers le modèle de la TCT seront fortement affectés par l'échantillon dont on dispose. C'est-à-dire que si dans notre échantillon, nous avons majoritairement des candidats ayant un niveau faible de maîtrise du sujet à évaluer, les scores obtenus seront affectés par cet échantillon. Alors que si l'on dispose d'un échantillon ayant un niveau fort du sujet à évaluer, les scores seront aussi affectés par le degré de compétence de l'échantillon, c'est-à-dire des candidats.

- D'après Olea et al., (1998), la notation d'une personne dépend, en même temps, de l'ensemble des items administrés. Cela veut dire que la notation obtenue par une personne sera différente si l'on administre deux tests contenant des items qui mesurent le même construit. Dans ce contexte, les erreurs faites dans la première passation n'auront pas de corrélation avec les erreurs faites dans la deuxième passation. Ainsi, les notations pourront seulement être mises en relation avec les tests d'où elles proviennent.

Enfin, une autre limitation de la TCT dont nous parlons provient du traitement donné à l'erreur. Puisqu'elle considère que l'erreur de mesure est une propriété du test, l'erreur est, par conséquent, une variable attribuable à tous les candidats sans considérer leur notation.

Autrement dit, dans notre quotidien, il nous parait pertinent d'évaluer un étudiant à partir de sa moyenne. La moyenne représente la modalité d'évaluation la plus courante. Elle est calculée à partir d'une échelle qui, dans le système éducatif français, oscille entre 0 et 20. Une note attribuée à un élève sera, dans un premier temps, limitée par cette échelle, et ensuite, la note finale sera fortement affectée ou favorisée par une série de valeurs attribuées à chacun des objectifs qui conforment la maquette.

Si à la fin de l'année scolaire 2015-2016, un professeur trouve que les étudiants d'une promotion donnée ont des notes hétérogènes, mais en même temps qu'elles sont toutes au-dessus de la moyenne, l'enseignant pourrait croire que ceci est un phénomène tout à fait normal. Alors il décide d'animer le même cours avec les mêmes ressources pendant l'année scolaire 2016-2017. Cette fois-ci, il trouvera qu'également les notes obtenues par la promotion sont hétérogènes, mais qu'en

revanche, une grande proportion de la promotion a obtenu des notes en-dessous de la moyenne. La réflexion la plus logique serait que la promotion 2016-2017 est moins avantagée que celle de l'année précédente. Ou bien, le professeur en question peut aussi se dire qu'il a peut-être des améliorations possibles à faire aux examens avec lesquels il a essayé de mesurer les compétences acquises par les étudiants. Comme nous pouvons l'observer, l'analyse sur les acquis des étudiants est faite à partir d'un examen. Ce dernier est en même temps évalué à partir des résultats obtenus par les étudiants. Cependant, il est généralement compliqué d'identifier s'il y des améliorations à apporter à l'examen, ou s'il y a des étudiants qui effectivement n'ont pas acquis les compétences définies dans la maquette.

2.2. Le Coefficient Alpha de Cronbach et l'univocité du test

Nous avons précédemment dit que l'une des limitations de la TCT est le manque de corrélation entre les résultats obtenus de deux tests qui envisagent de mesurer le même construit. Nous pouvons donc nous demander en quoi ceci pourrait être un problème lorsqu'il s'agit effectivement de deux tests différents. En effet, nous parlons de deux tests mesurant le même construit, c'est-à-dire que le concepteur attend que les candidats aient des comportements similaires au moment de passer soit deux tests différents, soit le même test plusieurs fois. Ceci est seulement possible si l'on réussit effectivement à mesurer le même construit à travers les items que nous avons conçus, car un test est valide « s'il mesure ce qu'il a l'intention de mesurer » (ALTE, 2011, p. 16).

Ce processus précis de classement est possible dès lors que l'on propose des items homogènes ; nous comprenons ici par homogénéité le fait que chacun des items évalue le même construit. Nous pouvons donc dire que le test est unidimensionnel, c'est-à-dire qu'il possède une « dimension unique, qui est une supposition nécessaire pour construire une échelle permettant de mesurer des attributs (…) Propriété de l'instrument de mesure et non du processus psychologique sous-jacent » (Multilingual Glosary of Language Testing Terms, p. 244).

Ainsi, vu les limitations de la TCT, des chercheurs ont continué à développer cette théorie. C'est ainsi que le coefficient alpha de Cronbach a été développé : « α est une estimation de la corrélation entre deux échantillons aléatoires d'items comme

ceux qui conforment le test. L'index r_{ij}, dérivé de α, représente un index d'homogénéité interne » [Traduction libre] (Cronbach, 1951, p. 1).

Le coefficient Alpha de Cronbach nous permet donc d'estimer la fiabilité du test, car « la mesure de la fiabilité à travers l'alpha de Cronbach suppose que les items (…) mesurent un même construit, et qu'ils sont hautement corrélés » [Traduction libre] (Wlech et Comer, 1988) cité par Frías, s.d. p. 1.

Revenons, par exemple, au professeur qui souhaite connaître la qualité des examens avec lesquels il évalue ses étudiants. Cette fois-ci, nous allons dire que nous parlons d'un professeur d'Espagnol Langue Etrangère (ELE). Supposons que l'une des questions comprises dans son examen (instrument de mesure), propose une Question à Réponse Unique (QRU) qui a pour construit à mesurer : la fonction langagière « donner et demander information », niveau A1.

Quel est le prénom le plus courant chez les hommes au Mexique ?

a) Pepito

b) Gómez

c) Jimenez

La première question à se poser ici est : *est-ce qu'à travers cet item il est possible d'évaluer la fonction langagière « donner et demander information » ?* En effet, cette question est plutôt orientée vers l'évaluation d'une connaissance du monde, ainsi que de la connaissance de l'espagnol parlé au Mexique. Nous pouvons conclure que cette question n'évalue pas une habileté langagière, et pourtant, elle est loin d'évaluer le construit qui évalue (mesure) une habileté langagière. Ceci empêche d'aboutir au principe du coefficient du Cronbach : proposer un test homogène, c'est-à-dire un test qui ne mesure qu'une seule habileté langagière.

2.3. La Théorie de Réponse à l'Item (TRI)

D'après Olea et al., (1998), la Théorie de la réponse à l'Item constitue une nouvelle approche psychométrique permettant de combler les limitations présentées dans la TCT. Elle se concentre sur les propriétés des items de façon individuelle, tandis que la TCT se concentre sur les propriétés globales du test.

La caractéristique principale de la TRI est son invariance, c'est-à-dire l'invariance des items vis-à-vis de plusieurs distributions de l'habilité à mesurer. Ainsi, quel que soit le niveau d'habilité de l'ensemble des candidats, le calcul du degré de difficulté de l'item, ou de l'ensemble d'items, ne sera pas dépendant de l'échantillon. De la même façon, le niveau de l'habilité de la personne sera défini à partir des items qui correspondent à son seuil de compétences.

Il y a plusieurs modèles de la TRI. Cependant, nous allons nous focaliser sur le modèle de Rasch, car celui-ci fait l'objet des analyses psychométriques dans le cadre du projet SELF.

2.4. Le modèle de Rasch. Modèle à un paramètre

Le principe de ce modèle indique que « la probabilité de répondre correctement à un item dépend complètement du niveau de difficulté de l'item même, et du niveau de l'individu dans la variable de mesure (niveau d'habilité). » [Traduction libre] (Olea et al., 1998, p. 5).

D'après Prieto et Delgado (2003), son expression mathématique se représente comme suit :

$$\ln (P_{is}/1-P_{is})=(\emptyset_s-\beta_i)$$

In = fonction logarithme naturel

P = probabilité

Θs = Habilité de la personne

β_i = Difficulté de l'item

Les auteurs expliquent que cette équation indique que le rapport entre la probabilité d'une réponse correcte et la probabilité d'une réponse incorrecte à un item (Pis/1-Pis), est lié à la différence entre le niveau de la personne (θs) et la difficulté de l'item (βi).

Ainsi, quand une personne répond à un item correspondant à son niveau de compétence, elle aura la même probabilité de produire une réponse correcte et une réponse incorrecte (Pis/1-Pis = 0,50/0,50).

Dans ce cas, le logarithme naturel de Pis/1-Pis est alors égal à 0[7], ce qui reflète que la difficulté de l'item équivaut au niveau de compétence de la personne (θs-βi=0).

Si la compétence de l'individu est supérieure à celle demandée par l'item (θs-βi > 0), la probabilité d'une réponse correcte sera supérieure à celle d'une réponse incorrecte.

En revanche, si la compétence de l'individu est inférieure à celle qui est demandée par l'item (θs-βi < 0), la probabilité d'une réponse correcte est inférieure par rapport à la probabilité d'une réponse incorrecte.

Nous pouvons donc conclure que le modèle postule qu'un candidat devra être capable de répondre aux items qui appartiennent à son seuil de compétence.

2.4.1. L'unité de mesure du degré de difficulté de l'item et du niveau du candidat.

L'une des raisons pour laquelle comprendre les principes de la TRI demande des efforts supplémentaires pour une personne qui n'a pas de connaissance en statistique, est que ses analyses sont faites à partir d'unités de mesure que ne sont pas des nombres naturels. La plupart des tables de résultats fournies par Winsteps (logiciel employé dans le cadre du projet SELF) nous donnent des chiffres qui proviennent d'un calcul mathématico-statistique. Nous n'allons pas expliquer chacun des calculs fait par le logiciel. Nous partons du principe que si nous expliquons la raison et les finalités des unités de mesure employées par le logiciel, l'interprétation des tables de résultats pourra être faite de façon plus claire.

Nous allons ainsi expliquer que, pour pouvoir aboutir à un processus d'analyse plus détaillé et moins global, le modèle de Rasch ne propose pas seulement une nouvelle formule. Il calcule le degré de difficulté de l'item, ainsi que le niveau de compétence du candidat, avec une unité de mesure qui considère le succès ainsi que l'échec de réponse à un item donné. Son échelle couvre tous les nombres réels, depuis l'infini négatif jusqu'à l'infini positif. Dans la même voie, une autre particularité de cette mesure est qu'elle est linéaire, ce qui veut dire que l'estimation

[7] Le logarithme naturel est une fonction mathématique qui a pour propriété remarquable d'être égale à 0 pour une valeur d'entrée égale à 1. Ainsi, si Pis/1-Pis = 0,50 /0,50 = 1, alors ln (1- Pis/1-Pis) = 0

de l'habilité du candidat, ainsi que l'estimation de la difficulté d'un item sont proportionnelles (Tristan López, 2002). Cette unité de mesure est désignée sous le nom de 'logit'.

Voyons un exemple nous permettant de comprendre ce concept. Si nous administrons un examen d'histoire, avec un total de 126 questions, nous pouvons dire que notre échelle est de 0 à 126. Imaginons que nous avons un étudiant qui a répondu à 80 questions correctement, et un autre qui a 40 réponses correctes. La question à se poser est celle-ci : est-ce que 80 est le double de connaissance que 40 ? La réponse est non, car cette échelle ne peut pas garantir que 4 questions peuvent représenter le double de traits que 2 unités. C'est pour cela qu'il n'est pas possible d'affirmer que 80 réponses justes représentent le double de connaissance que 40. Ainsi, nous pouvons dire que l'échelle n'est pas linéaire (Tristant López. 2002). De ce fait, le « logit » est une unité de mesure qui permet de combler l'une des limitations de la TCT, dont l'estimation du degré de difficulté d'un item est égale à la moyenne des scores observés.

Ainsi, les propriétés du « logit » sont qu'il permet de :

- couvrir la totalité du trait à évaluer
- considérer le succès ainsi que l'échec au travers d'une seule quantité
- évaluer de façon réelle le niveau de compétence du candidat

2.4.2. L'ajustement au modèle d'évaluation

Si nous reprenons l'exemple du professeur d'ELE qui veut savoir d'où provient l'hétérogénéité des notes de ses élèves : le niveau des élèves, ou la conception de ses examens. Il peut le savoir à travers un processus de qualité. Ce processus de qualité, d'après Tristan López 2002, peut être désigné comme le processus de calcul qui permet d'estimer la qualité des résultats. Il est fait à partir de l'analyse de l'erreur d'ajustement, ou FIT. Ce processus consiste à vérifier qu'il s'adapte au modèle d'évaluation. Autrement dit, il s'agit de vérifier qu'il s'adapte au modèle de Rasch. Cela est fait à travers l'observation des résultats obtenus, et les estimations faites par le modèle de Rasch.

Le FIT se présente sous la forme de deux mesures. « La première mesure d'ajustement est désignée comme l'OUTFIT (de l'anglais *outlier-sensitive fit*

statistics), et qui peut être traduit par : ajustement externe. C'est une valeur sensible aux comportements inattendus qui affectent les items dont la difficulté est éloignée du niveau d'une personne ». (Tristán López, 2002, p. 111).

La deuxième façon de calculer l'ajustement est possible avec l'INFIT (de l'anglais *information-weighted fit statistics*) qui peut être librement traduit par : ajustement interne. Il s'agit d'une valeur sensible aux comportements inattendus qui affectent les items dont la difficulté est proche du niveau de capacité d'une personne. » [Traduction libre] (Tristán López, 2002, p. 111).

2.4.3. La courbe caractérisant l'item

D'après Olea et al., (1998), c'est un concept clé de la TRI. Il s'agit d'une courbe qui prétend estimer la probabilité de répondre correctement à un item en tant que trait latent. Normalement, elle est présentée sous forme de 'S', et elle vise à illustrer que si le score du trait latent augmente, les probabilités de répondre correctement augmenteront aussi. Elle veille aussi à démontrer que si l'on répond correctement, cela dépendra uniquement du trait latent, du construit à évaluer.

Ci-dessous, l'image d'une courbe empirique à l'item.

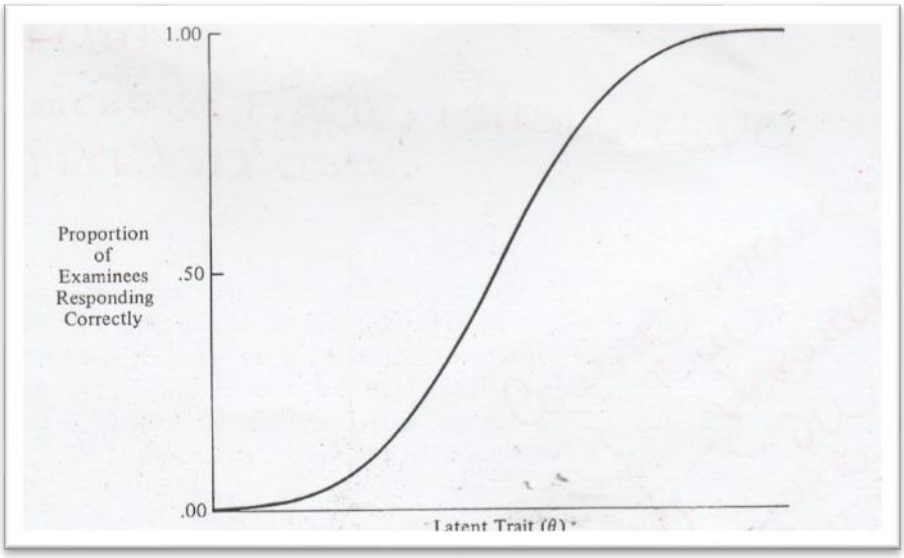


Figure 9 : courbe caractérisant l'item (CCI)

L'axe vertical du graphique représente la probabilité de réponse, et l'axe horizontal le niveau de difficulté d'un item, ou bien le niveau d'habilité d'un candidat.

2.5. La validation d'après Innovalangues

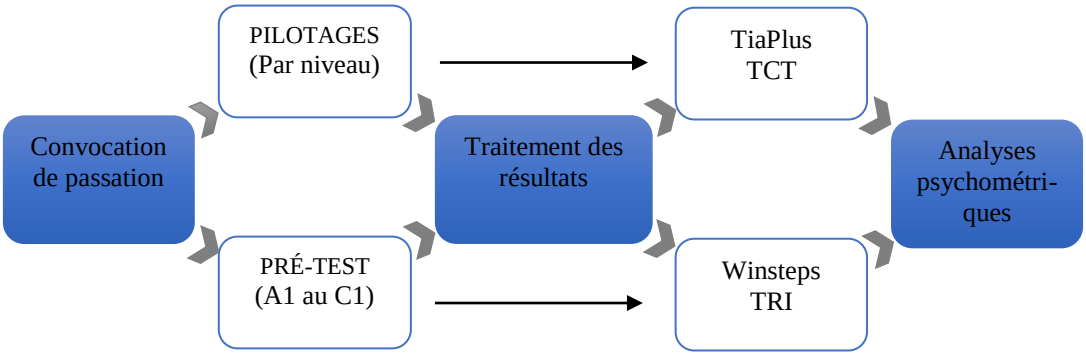
Cette partie a pour objectif de parler de la méthode adoptée pour valider le projet SELF d'Innovalangues. Nous entendons ici par validation la voie amenant le test vers la validité, ainsi que vers la fiabilité.

Pour ce faire, la constitution de la banque d'items du test SELF est choisie une fois effectuée la première étape et la deuxième étape de validation décrites dans le Cycle du *testing.* Comme nous l'avons dit, des étudiants d'un même niveau (A1, par exemple) sont convoqués à passer le test pilote niveau A1 (un seuil minimum de 50 candidats par pilotage est nécessaire). Ceci, dans le cadre du projet, est désigné sous le nom de « PilotageA1 ». Les convocations aux pilotages sont donc faites par niveau. Ensuite, les résultats des tests sont traités avec le logiciel TiaPlus. Il nous permettra de faire les analyses psychométriques d'après la TCT (Cf. le logiciel TiaPlus).

Cette première analyse nous permet d'identifier les items considérés comme les meilleurs, c'est-à-dire les items du niveau A1 au C1 qui ont été validés d'après la TCT, et qui ont mis en évidence un niveau de difficulté et un pouvoir discriminatif acceptables (Cf. partie 3 : Interprétation des tables de résultats du logiciel TiaPlus).

Une fois identifiés ces items, et une fois après avoir modifié les items qui ont eu besoin d'être modifiés, nous procédons à la convocation des étudiants du niveau A1 au C1 pour qu'ils passent le PRÉ-TEST. Il s'agit d'un test qui contient tous les niveaux à évaluer. Les résultats obtenus lors de la passation du pré-test sont traités avec le logiciel Winsteps (Cf. le logiciel Winsteps). Il nous permet de faire les analyses psychométriques d'après la TRI : « étant donné que les résultats des analyses basées sur la TCT sont fortement dépendants de l'échantillon, lorsqu'on souhaite effectuer une généralisation des résultats, il faut effectuer des analyses psychométriques qui se basent sur la Théorie de Réponse à l'Item (TRI). » (Biagiotti et Cervini, 2015, p. 3).

Ce processus peut être illustré comme suit :



Les résultats observables à travers la TCT

D'après les pilotages (c'est-à-dire les passations de test évaluant un seul niveau) préalablement effectués par le projet SELF, nous avons pu constater que ce processus nous permet de vérifier et d'observer :

- La fidélité du test
- L'homogénéité[8] des items, et par conséquent l'univocité du test
- L'indice de facilité, ou de difficulté, des items
- Le score des candidats
- Le comportement des distracteurs, autrement dit les propositions de réponse incorrectes.

Les résultats observables à travers la TRI

Nous avons précédemment dit que la TRI se base sur plusieurs modèles. La différence principale entre ces modèles est la quantité de paramètres à considérer lors des calculs mathématiques. Néanmoins, cette variété de paramètres ne fait pas l'objet de cette partie.

Cependant, il est important de signaler que tous les modèles de la TRI « se basent sur l'existence d'un trait latent, c'est-à-dire une caractéristique non observable qu'on va essayer d'estimer en utilisant des variables sur lesquelles nous avons un moyen de contrôle. Ces variables sont : les items et tout le contenu du test » (Biagiotti et Cervini, 2015, p. 4-5).

[8] Caractéristique d'un test qui évalue un même construit, dans notre cas, la compétence langagière à plusieurs échelles. Nous comprenons par *plusieurs échelles* les niveaux du CECRL testés.

Il s’agit de mesurer la performance des candidats à travers les items, dont la caractéristique est celle d’évaluer une compétence à travers des items ayant un degré de difficulté différent. Ainsi, nous allons « à la fois estimer la compétence du candidat, mais aussi la difficulté des items » (Biagiotti et Cervini, 2015, p. 5).

De ce fait, les résultats observables par le modèle de Rasch sont :

- La mesure de la difficulté des items (estimation de la difficulté des items en logits) ;
- La mesure de l’habileté des candidats (estimation en logits) ;
- L’indice de discrimination (plus ou moins équivalent au RIR en TCT) ;
- Les indicateurs sur la qualité de l’ajustement des données modèles (INFIT et OUTFIT).

Partie 3

-

Les analyses psychométriques de SELF en Espagnol

Chapitre 4. Le déroulement de la validation quantitative du SELF en espagnol

1. Un bref bilan du stade actuel du SELF en espagnol

À la fin du stage, l'équipe SELF espagnol avait convoqué les étudiants aux pilotages, et réalisé les analyses au travers de la TCT. À partir de cette étape, la sélection des items pour le pré-test a eu lieu.

Comme nous l'avons précédemment dit, le pré-test évalue tous les niveaux, de A1 au C1. Aussi, le pré-test est fait en deux versions : PRÉ-TEST A et PRÉ-TEST B. De ce fait, la banque d'items du pré-test est constituée de 102 items pour le pré-test A, et de 99 items pour le pré-test B. Ces items sont issus du pilotage. Il s'agit d'items qui ont été validés d'après la première analyse psychométrique.

Les deux pré-tests ont 30% d'items en commun, ce que nous appelons dans le cadre du projet : des items ancres. Ils ont pour caractéristique d'avoir démontré les meilleurs comportements lors des pilotages, c'est-à-dire un degré de difficulté acceptable d'après la TCT, ainsi qu'un pouvoir discriminatif (Cf. Le logiciel TiaPlus).

Entre autres, l'échantillon qui a participé aux pilotages provient des centres suivants :

- Centre de langues Vivantes de Grenoble
- Le Service des Langues de l'UGA
- L'Institut Universitaire de Technologie de l'UGA de Valence
- L'Université Picardie Jules Verne d'Amiens

La finalité est de travailler avec un échantillon de candidats représentatifs du public cible. C'est pourquoi l'échantillon est composé par des francophones natifs qui font des études supérieures.

Afin de garder un équilibre entre les niveaux à évaluer, à la date de la fin du stage, la composition du pré-test était la suivante : 20% d'items du niveau A1, 30% d'items du niveau A2, 30% d'items du niveau B1, et 20% d'items du niveau B2. Enfin, le niveau C1 sera rajouté.

Dans le cadre du projet, le seuil nécessaire pour procéder aux analyses psychométriques sur Winsteps est de 200 candidats pour chacune des versions de pré-test. Ils doivent être répartis de façon équilibrée dans le but de représenter tous

les niveaux à évaluer. Ainsi, 20% des candidats doivent représenter le niveau A1, 20% le niveau A2, et ainsi de suite.

2. Système d'édition SELF

Cette partie a pour objectif de présenter l'éditeur SELF dans ses grandes lignes. Nous allons également approfondir les fonctionnalités nous permettant d'aborder le fil conducteur du présent travail : l'analyse psychométrique. Ainsi, la fonction *export de résultats*, constituera l'objet principal de cette section.

L'éditeur SELF est un outil qui présente cinq fonctionnalités principales :

1. Création des tâches et des items
2. Création des tests par l'assemblage des tâches et des items
3. Ouverture et gestion des sessions[9] de test
4. Administration des tests
5. Export des résultats

2.1. Export des résultats, SELF :

Lorsque nous allons mener des analyses psychométriques, il est fondamental de stocker certaines données :

1. Le contenu des items : type d'item, questions et propositions de réponse (question correcte ou clé, ainsi que les propositions de réponse incorrectes : distracteurs).
2. Les données de l'échantillon sont aussi importantes : parcours académique actuel, adresse électronique, entre autres.

En effet, la fonctionnalité export de résultats du logiciel permet justement de récupérer toutes les données nécessaires pour pouvoir procéder aux analyses. Ces données peuvent être téléchargées, soit sous fichier .csv, soit sous fichier PDF.

[9] Une session est la création de codes différents pour passer le test. Cette fonctionnalité permet, par exemple, d'avoir un registre d'un test passé en ligne, ou en présentiel, ou d'un test passé par des candidats provenant de différentes universités. Ainsi, un seul test peut être passé par un échantillon varié.

Figure 10 : SELF export de résultats

L'image nous permet de voir que le système SELF permet de télécharger trois types de fichiers : export Stats (tia+, etc.), export liste de tâches, export livret PDF.

Figure 11 : SELF types de fichiers exports

2.2. Fichier Export Stats (tia+)

Comme son nom l'indique, il s'agit du fichier .csv, un script à joindre sur TiaPlus. Nous pouvons classer les données contenues dans ce fichier en deux parties : celles qui fournissent des informations sur la globalité du test, et celles qui fournissent des informations précises.

Dans le détail, les informations fournies sont les suivantes :

DONNÉES GLOBALES :

- Le pseudo du candidat
- Prénom
- Date de passation
- Temps total de passation dont chacun des candidats a eu besoin

DONNÉES PRÉCISES:

- Le nom attribué à chacune des tâches (Cf. la nomenclature employée par SELF, dans la première partie).
- La typologie de tâche (QRU, QRM, VF, etc.)
- Le niveau de difficulté estimé pour chacun des items
- Le temps que le candidat a mis pour répondre à l'item

- La proposition de réponse choisie par le candidat
- La note, ou les points attribués à la réponse choisie par le candidat

2.3. Fichier export liste de tâches

C'est un fichier .csv qui contient les informations suivantes :

- Numéro de la tâche
- N° d'item
- Les réponses correctes
- Le type d'item
- Le nom de l'item

2.4. Fichier export livret PDF

Il s'agit d'un fichier qui contient toutes les tâches d'évaluation.

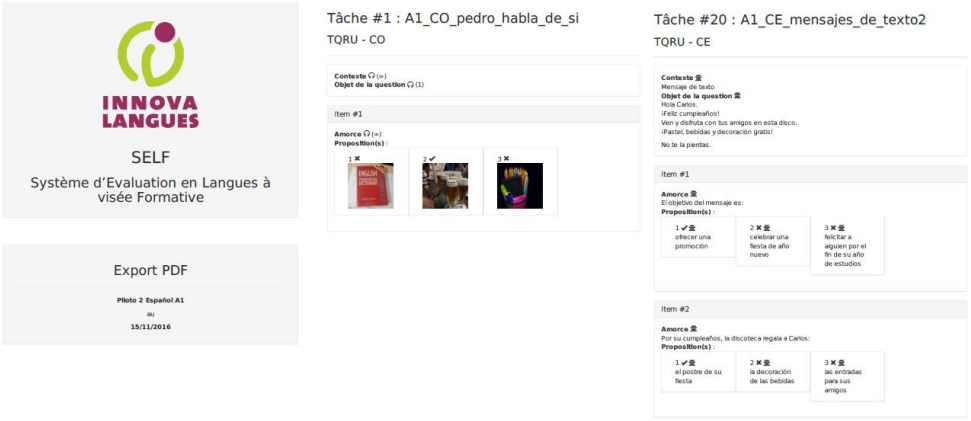


Figure 12 : livret PDF

SELF stocke les données suivantes des candidats :

- Nom d'utilisateur
- Nom et Prénom
- Email
- Etablissement (p.ex. CLV, UGA, etc.)
- Filière (p.ex. LEA, MEEF etc.)

- Date de passation
- Heure initiale de la passation
- Durée approximative de la passation
- Score agrégé (en route, ou vers tel ou tel niveau)
- Score CO
- Score CE
- Score EEC
- Si la personne a ou n'a pas fini le test
- Langue maternelle
- Autre langue de référence (une langue étrangère apprise)

Toutes ces données sont précieuses à garder. Elles permettent tout d'abord de faire les analyses psychométriques, et elles peuvent également faire l'objet du corpus d'une recherche. Cependant, nous allons nous focaliser sur les données dont nous avons besoin pour aboutir au calibrage du test. Dans le cadre de ce projet, les fichiers : ***Export Stats (tia+), export liste de tâches*** et ***export livret PDF*** retiennent notre intérêt.

Les fichiers .csv ***Export Stats (tia+)*** et ***export liste de tâches*** doivent passer par un processus de « nettoyage » que nous n'allons pas approfondir ici. Dans ses grandes lignes, ce processus est essentiel pour que les logiciels TiaPlus et Winsteps puissent interpréter les données (Cf. les particularités de TiaPlus dans la partie suivante).

3. Le logiciel TIAPLUS

Le logiciel TiaPlus, *Test and Item Analys*, est un logiciel dont la démarche statistique suivie se fonde sur la Théorie Classique du Test :

> *The basic assumption underlying TiaPlus is that the observed score (X) of one person on one item is actually the sum of two components: the so-called true score (T), and an additional error component (E): X = T + E.* (Cito, 2013, p.3)

D'après le manuel du logiciel, il a une double perspective : il peut fournir des informations concernant les items, informations qui seront fortement liées à l'échantillon. Il fournit également des informations sur l'échantillon même.

Cependant, l'un des inconvénients du logiciel est que « dans la réalité, ces informations ne sont pas clairement séparées » [Traduction libre. Cito, 2013, p. 3)]

En suivant le manuel du logiciel, nous avons pu repérer ces particularités du logiciel TiaPlus :

1. Il faut avoir un minimum de 5 candidats pour chacun des items proposés.

   *A rule of thumb for a minimum of observations is the **1 : 5 ratio**. That is: Try to have at least 5 times the number of persons, compared to the number of items in a test.* (Cito, 2013, p. 5)

   De ce fait, si nous avons un test composé de 50 questions, il nous faudra un minimum de 250 candidats pour aboutir à un traitement de données fiable (Cf. l'analyse factoriel). Le total de candidats demandé peut être obtenu par plusieurs passations du même test, c'est-à-dire que nous pouvons faire passer le même test (la même session, ou plusieurs) sur plusieurs créneaux.
2. Le logiciel peut gérer le texte ASCII/ANSI sous deux types de fichiers :
   a. Le premier concerne les fichiers d'extension .**csv** (*coma separated value)* et se présente de la façon suivante :

```
Johnsson M. Gloucester College MA 1252413,E,D,A,B,A,D,A,E,D,(...)D,A,E,D,A,B,A,D
Arnesen B. Gloucester College MA 1555417,A,D,A,B,A,D,A,D,D,(...)D,A,A,D,A,A,A,A
Dison G. Lipton College NY 4442413,A,A,A,B,D,D,A,D,D,(...)D,A,A,D,B,B,B,B
.
.
```

Figure 13 : exemple de fichier .csv

Ce type de fichier peut être généré soit avec un éditeur de texte, soit avec Excel. La figure 13 illustre le fait que les réponses des items commencent à partir de la première virgule.

b. Le deuxième type de fichier concerne les fichiers .txt, où la valeur des réponses aux items sera attribuée à partir du code ASCII ou ANSI.

c. TiaPlus peut aussi interpréter des fichiers .xls et .xlsx. Il faut toujours considérer que pour aboutir à l'interprétation du fichier en tant que donnée d'entrée, il ne doit pas contenir d'espace vide, ni d'image. La figure 14 permet de comprendre à quoi doit ressembler un fichier .xls ou .xlsx figurant comme fichier à interpréter par TiaPlus :

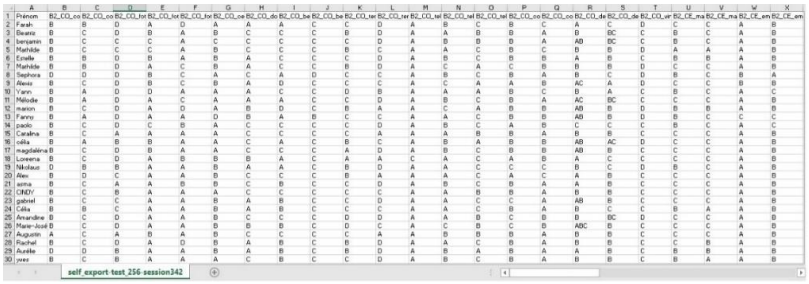


Figure 14 : exemple de fichier .xls ou .xlsx

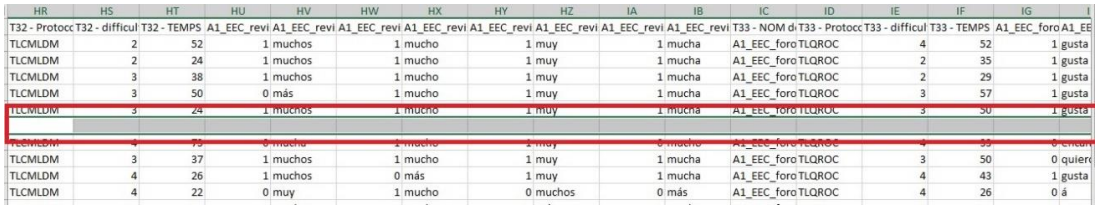
Il en résulte, dans le cas de ce projet, que les réponses aux items seront représentées par des lettres : ABCDEF.

3. Pour avoir un script comme le précédent, les réponses doivent être présentées sous forme de lettres. X est égal à un item non-répondu. A, dans le cas des Questions à Réponse Ouverte Courte (QROC) est égal à la réponse correcte, ou clé. B représente les distracteurs, ou la(les) réponse(s) incorrecte(s). Dans le cas des questions de type Vraie/Faux, Questions à Réponse Unique (QRU) ou Questions à Réponses multiples (QRM), la clé restera la même. Elles n'ont pas besoin d'être traitées, parce que le fichier export fourni par l'éditeur SELF renferme ces items sous la nomenclature : ABCD (Cf. annexe 5).
4. TiaPlus n'interprète que des nombres entiers. Ceci est un point à considérer au moment d'attribuer la valeur des items dans l'étape « paramétrage global des items ». Il ne faut pas attribuer de note décimale aux items.
5. TiaPlus peut traiter plusieurs types d'items : questions ouvertes, questions à échelles, questions à réponse unique, et questions à réponses multiples.

Avant de conclure cette partie, et à propos de la préparation du script que nous avons mentionnée dans ses grandes lignes dans le point numéro 3, nous considérons

pertinent de dire que l'un des facteurs qui peut entrainer un impact négatif sur la fiabilité du test, provient des candidats qui ont participé aux pilotages, mais qui n'ont pas fini le test, ou qui ont commencé à répondre au hasard à cause de plusieurs facteurs : ennui, fatigue, ils étaient arrivés à un niveau de difficulté qui leur demandait un effort plus considérable, ou à cause d'autres facteurs. Ainsi, ces participations doivent être éliminées avant de joindre le fichier sur TiaPlus.

Ci-dessous une image nous permettant de visualiser quand un candidat n'a pas fini le test :

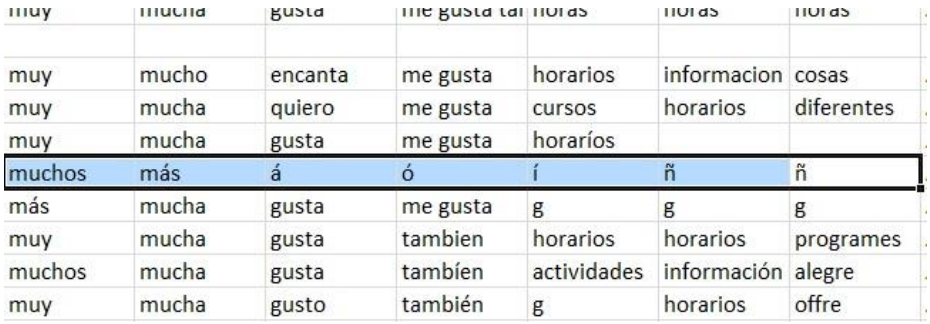


HR	HS	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID	IE	IF	IG	I
T32 - Protocc	T32 - difficul	T32 - TEMPS	A1_EEC_revi	A1_EEC_revi	A1_EEC_revi	A1_EEC_revi	A1_EEC_revi	A1_EEC_revi	A1_EEC_revi	A1_EEC_revi	T33 - NOM di	T33 - Protocc	T33 - difficul	T33 - TEMPS	A1_EEC_foro	A1_EE
TLCMLDM	2	52	1	muchos	1	mucho	1	muy	1	mucha	A1_EEC_foro	TLQROC	4	52	1	gusta
TLCMLDM	2	24	1	muchos	1	mucho	1	muy	1	mucha	A1_EEC_foro	TLQROC	2	35	1	gusta
TLCMLDM	3	38	1	muchos	1	mucho	1	muy	1	mucha	A1_EEC_foro	TLQROC	2	29	1	gusta
TLCMLDM	3	50	0	más	1	mucho	1	muy	1	mucha	A1_EEC_foro	TLQROC	3	57	1	gusta
TLCMLDM	3	24	1	muchos	1	mucho	1	muy	1	mucha	A1_EEC_foro	TLQROC	3	50	1	gusta
TLCMLDM	3	37	1	muchos	1	mucho	1	muy	1	mucha	A1_EEC_foro	TLQROC	3	50	0	quiere
TLCMLDM	4	26	1	muchos	0	más	1	muy	1	mucha	A1_EEC_foro	TLQROC	4	43	1	gusta
TLCMLDM	4	22	0	muy	1	mucho	0	muchos	0	más	A1_EEC_foro	TLQROC	4	26	0	á

Figure 15 : exemple d'un candidat qui n'a pas fini le test

Ce type de comportement est majoritairement observable à partir de l'évaluation de l'EEC, car sur ce niveau, les efforts à faire par les candidats augmentent de façon considérable. En effet, les activités de production, soit orale, soit écrite, demandent aux locuteurs non-natifs des efforts supplémentaires vis-à-vis des activités de compréhension.

Ci-dessous une image nous permettant de visualiser quand un candidat répond au hasard :



muy	mucha	gusta	me gusta tal	horas	horas	horas
muy	mucho	encanta	me gusta	horarios	informacion	cosas
muy	mucha	quiero	me gusta	cursos	horarios	diferentes
muy	mucha	gusta	me gusta	horarios		
muchos	más	á	ó	í	ñ	ñ
más	mucha	gusta	me gusta	g	g	g
muy	mucha	gusta	tambien	horarios	horarios	programes
muchos	mucha	gusta	también	actividades	información	alegre
muy	mucha	gusto	también	g	horarios	offre

Figure 16 : exemple d'un candidat ayant répondu au hasard

3.1. Interprétation des tables de résultats

Cette partie constitue un vrai défi. D'un côté, l'interprétation des tables de résultats représente la clé de la validité et de la fiabilité du test. D'un autre côté, plus nous approfondissons l'analyse des fichiers extraits du TiaPlus, plus la transposition des

notions théoriques à la pratique est complexe. C'est ainsi que j'ai pu me formuler les premières questions de ce travail : *Quels sont les critères à considérer lors de l'interprétation des données afin d'éviter une validité apparente*[10] *?* En effet, le logiciel nous donne des résultats bruts, mais *comment faut-il les interpréter ?* Et une fois interprétés, *que doit-on faire avec les items qui n'ont pas provoqué le comportement attendu sur les candidats lors de la conception* ?

Pour pouvoir répondre à ces questions, nous allons, dans un premier temps, présenter de façon générale les éléments fournis par TiaPlus. Dans un deuxième temps, nous allons adapter ces explications au projet à partir de la présentation de quelques exemples de la méthode de travail adoptée tout au long du stage. Cela permettra enfin de décrire les modifications faites, et les critères que nous avons considérés au moment d'effectuer ces modifications.

Pour pouvoir aboutir à cette explication, la consultation du manuel de TiaPlus a été nécessaire. Mais ce manuel ne donne pas de précision sur la façon d'interpréter les tableaux de résultats qu'il fournit. Ainsi, l'ensemble des explications exposées ici proviennent de documents faits dans le cadre du projet SELF, des documents fournis par le CIEP (en 2014 , celui-ci a mené à bien des formations permettant d'aboutir à la maîtrise des logiciels, ainsi qu'à l'interprétation des résultats). Enfin, la consultation d'autres bibliographies a été également nécessaire.

Ci-dessous un exemple du tableau contenant l'analyse principale : *TiaPlus main analysis (ANA)* :

Item Label	Item nr.	Weight	Key	P- and A- values A	B	C	D	E	F	O/D	Mis-sing	Weighted Max	Mean	P	Sd	RSK	Rit	Rir	AR
nevera_ro	1	1	B	27	71*	0	0			2	1	1	0,71	71	0,45	0,45	45	40	85
nevera_ro	2	1	B	58	40*	0	0			2	1	1	0,40	40	0,49	0,49	27	20	86
nevera_ro	3	1	A	89*	9	0	0			2	1	1	0,89	89	0,31	0,31	48	44	85
segunda_m	4	1	C	25	16	56*	0			2	1	1	0,56	56	0,50	0,50	39	33	85
segunda_m	5	1	A	71*	18	9	0			2	1	1	0,71	71	0,45	0,45	61	56	85
danza_cad	6	1	A	91*	2	5	0			2	1	1	0,91	91	0,29	0,29	46	42	85

Tableau 1 : TiaPlus main analysis (ANA)

Le tableau ANA (tableau 1) comprend toutes les données, ou les chiffres, que le logiciel, après avoir fait une démarche mathématique-statistique, a recueillis. C'est pourquoi le nom attribué par défaut est : *main analysis*. D'après le manuel

[10] *Qualité d'un test, ou toute autre mesure, qui semble correcte et adéquate à l'objet mesuré. Il s'agit d'un jugement subjectif plus que d'un jugement basé sur une analyse objective du test. La validité apparente est considérée comme une fausse forme de validité.* (ALTE Members, 1999, p. 244)

d'utilisation de TiaPlus, ce premier extrait du tableau contient les valeurs ***'P'*** et ***'A'*** ; elles désignent l'indice de difficulté des items. *'P'* est égal au pourcentage des personnes qui ont choisi la réponse correcte. '*A'* représente la proportion des candidats qui ont choisi la réponse incorrecte (distracteur). La lecture de la sélection des réponses peut se faire de la façon suivante : les chiffres montrés en dessous d'ABCD représentent le pourcentage des personnes qui ont choisi les réponses correspondantes, parmi les alternatives proposées (ABCD). Pour pouvoir identifier l'alternative correcte, c'est-à-dire la réponse correcte, la clé ou *key*¸ il suffit de repérer la lettre qui a une étoile sur la droite, par exemple A*.

Ci-dessous l'explication de ce que chacune des entrées veut dire :

Item label, est égal au nom attribué à l'item.
Item nr, est le numéro de l'item
Weight, est le score attribué à l'item. Par principe, tous les items du projet SELF ont la même valeur : 1,0. De ce fait, la note maximale du test, correspondra au nombre total d'items.
Key, ou clé, représente la réponse correcte.
ABCDEF, toutes les possibilités de réponse.
O/D omitted or double: est le pourcentage de personnes qui n'ont pas répondu à l'item.
#Missing: le nombre de personnes qui n'ont pas répondu à l'item**.**
Max: la note maximale attribuée à l'item. Comme nous l'avons dit précédemment, la note maximale est égale à 1.
Mean : la moyenne pondérée de l'item. Elle est obtenue à partir de l'addition des items traités, divisée par le nombre total de candidats formant l'échantillon.
P : est l'indice de difficulté de l'item. Sa valeur maximale est de 100. Dans le cadre de ce projet, '***P'*** ou **l'indice de difficulté** ne doit pas dépasser 97.5 %, ou être inférieur à 2.5 %. L'indice de difficulté se traduit comme suit :

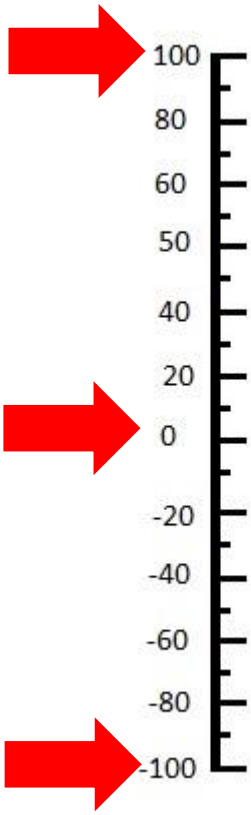
Item difficile	***Item acceptable***	***Item facile***
P < 30	*30 < P < 90*	*P > 90*
Moins de 30% de l'échantillon a choisi la réponse correcte.	*Entre 30% et 90 % de l'échantillon a choisi la réponse correcte*	*Au moins 90% de l'échantillon a choisi la réponse correcte*

Tableau 2 : indice de difficulté

SD : est la déviation standard ou l'écart-type de l'item. Il s'agit d'une donnée indiquant dans quelle mesure les scores individuels des items sont répartis autour du score moyen de tous les items. Plus l'écart-type est important, plus l'écart est large, et plus l'écart-type est large, moins fiable est notre instrument de mesure. Si toutes les personnes ont la même réponse à un item, alors la déviation standard sera égale à zéro. Plus elle est proche de 1, plus grande est la déviation de l'item de sa moyenne. En d'autres termes, si elle se rapproche de 1, il faudrait analyser les autres alternatives de réponses (distracteurs), car cela veut dire que les distracteurs semblent être des réponses correctes, et qu'ils ont été prioritairement choisis.

RSK: est une mesure relative à la déviation standard qui permet de comparer avec d'autres items ayant un score différent. Comme dans ce cas nous avons attribué les mêmes notes à tous les items, la mesure *RSK* aura la même valeur que l'écart type.

RIT : D'après le manuel du logiciel, il se focalise sur la fiabilité du test. Il part du principe qu'un RIT élevé veut dire que l'item est cohérent avec le test. Son rang (sa valeur) provient du coefficient de corrélation du Pearson. Il oscille entre -1,0 et +1,0. (Crocker et Algina, 1986, p, 32). Il est important de considérer que TiaPlus affiche son RIT et RIR (Cf. RIR) multipliés par 100. L'échelle oscille donc entre -100 et +100 (Cito, 2013, p, 36). Il fournit le degré de corrélation existant entre l'item et la totalité du test, y compris l'item qu'on est en train d'analyser. Il vise aussi à prédire la probabilité de répondre correctement à un item à partir du classement de deux groupes de candidats : ceux qui ont tendance à répondre correctement, que nous allons appeler le groupe « fort », et ceux qui ont tendance à répondre de façon incorrecte, que nous allons appeler le groupe « faible ». Nous allons illustrer cette catégorisation par le graphique suivant :



RIT ELÉVÉ : Un candidat donné avec un score élevé pour la totalité du test a tendance à répondre correctement à l'item. En revanche, une personne avec un score bas pour la totalité du test, a tendance à répondre à l'item de façon incorrecte. Si le phénomène précédemment décrit se produit, nous pouvons dire que l'**item est cohérent avec le test.** [Traduction libre] (Cito, 2013, p. 36)

RIT ÉGAL À ZÉRO : Il n'existe pas de corrélation entre la totalité du test et le comportement de l'item. [Traduction libre] (Cito, 2013, p. 36)

RIT NÉGATIF : Un candidat donné avec un score élevé à la totalité du test, a tendance à répondre à l'item de façon incorrecte, et une personne avec un score bas à la totalité du test, a tendance à répondre à l'item de façon correcte. (Cito, 2013, p. 36)

Graphique 1 : RIT

Ce classement permet de vérifier le comportement (le fonctionnement) de l'item par rapport au niveau du candidat. En d'autres termes, **discriminer** peut être interprété comme « différencier entre ceux qui connaissent plus et ceux qui connaissent moins » [Traduction Libre] (Morales Vallejo 2009, p, 10). C'est à partir de la recherche de la discrimination du niveau d'habileté du candidat que nous envisageons d'analyser le fonctionnement de l'item, pour analyser ensuite le niveau de difficulté de l'item. **La recherche de ce rapport entre habilité chez le candidat et degré de difficulté de l'item nous permet finalement de déterminer si l'item est cohérent avec le test.** En effet, celui-ci est l'un des principes de la TRI, ce qui nous permet de conclure que le logiciel TiaPlus n'envisage pas seulement d'analyser les items à partir de la TCT, mais aussi de la TRI. Cependant, il faut aussi considérer que le manuel du logiciel recommande d'employer d'autres logiciels si l'on envisage de faire des analyses à partir de la TRI.

RIR : il montre la corrélation entre l'item et le test sans considérer l'item que l'on est en train d'analyser. Il s'agit d'une alternative pour mesurer l'indice de discrimination de l'item. Sa valeur est comprise entre -1 et 1. Par recommandation du CIEP, les

items avec une valeur supérieure à 0,15 sont acceptables, ce qui peut se présenter de la façon suivante :

Item pas discriminant	***Item acceptable***	***Item discriminant***
RIR < 0.15	*0.15 < RIR < 0.30*	*RIR > 0.30*

Tableau 3 : RIR

Nous pouvons nous demander maintenant quel est l'intérêt d'analyser le même item depuis l'ensemble des items, et de l'analyser en dehors de l'ensemble des items ? Nous pouvons dire que le RIT fournit une mesure qui est influencée par le comportement de l'item, tandis que le RIR nous permet de voir le fonctionnement de la globalité du test sans considérer l'item qui fait l'objet de l'analyse. Autrement dit, le RIR « mesure la corrélation entre la réussite à l'item, et la réussite à l'ensemble des items du test sans prendre en compte l'item en question dans le score total » (Innovalangues, 2016, p. 21)

Ar : c'est *l'Alpha Rest Value* d'un item. Il s'agit de la fiabilité calculée de l'item. « Il faut vérifier que l'AR est inférieur à l'alpha du test (coefficient Alpha) » (Innovalangues 2016).

Le coefficient Alpha de Cronbach : comme nous l'avons présenté dans le cadre théorique, il représente un critère de la recherche de la fiabilité du test ; il s'agit d'une mesure qui veille à ce que tous les items conformant le test évaluent un même construit. Comme critère général, George et Mallery (2003, p. 231) cité par Fíras D. s.d.p 1, suggèrent les recommandations suivantes pour évaluer le coefficient Alpha de Cronbach :

Coefficient Alpha de Cronbach > 0,9	**Excellent**
Coefficient Alpha de Cronbach > 0,8	**Bon**
Coefficient Alpha de Cronbach > 0,7	***Acceptable***
Coefficient Alpha de Cronbach > 0,6	***Contestable***
Coefficient Alpha de Cronbach > 0,5	**Pauvre**
Coefficient Alpha de Cronbach < 0,5	**Inacceptable**

Tableau 4 : coefficient Alpha de Cronbach

C’est pourquoi, dans le cadre de ce projet, le Coefficient Alpha de Cronbach (CAC) est acceptable à partir de 0,70. Le CAC se présente de la façon suivante à la fin du tableau ANA.

```
-------------------------------------------------------------------------------
90% Confidence limits for Coefficient Alpha: (0,81 =< 0,86 =< 0,90)
-------------------------------------------------------------------------------
Estimated Coefficient Alpha if this test had a standard
norm length of 40 items: 0,83 (Spearman-Brown)

Cito, Measurement and Research Department. Arnhem, the Netherlands. © 2013.
```

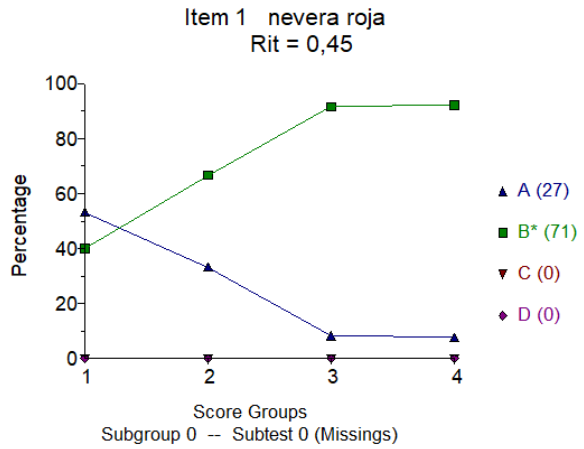
Tableau 5 : coefficient Alpha de Cronbach, TiaPlus

3.2. Les graphiques fournis par Tiaplus :

TiaPlus propose deux graphiques permettant de visualiser les données fournies par le tableau ANA. Le premier est la Courbe Caractéristique de l’Item (Cito, 2013, p. 20) que nous avons mentionnée dans le cadre théorique. Le deuxième est l’analyse factorielle. Nous allons par la suite présenter ces graphiques.

3.2.1. La courbe caractérisant l’item :

Elle permet de visualiser si la probabilité de répondre correctement à l’item est ascendante ou descendante. Elle montre aussi le comportement des distracteurs (alternatives de réponses incorrectes) ; cela est fait par rapport au niveau d’habilité des candidats. Ainsi, le graphique de la CCI est en relation avec le RIT que nous avons expliqué précédemment. Il est présenté de la façon suivante :



Où

B* (71) représente la clé (réponse correcte) pour laquelle la probabilité de répondre correctement dans le groupe fort est ascendante.

A (27) représente le distracteur qui a été majoritairement choisi par le groupe faible.

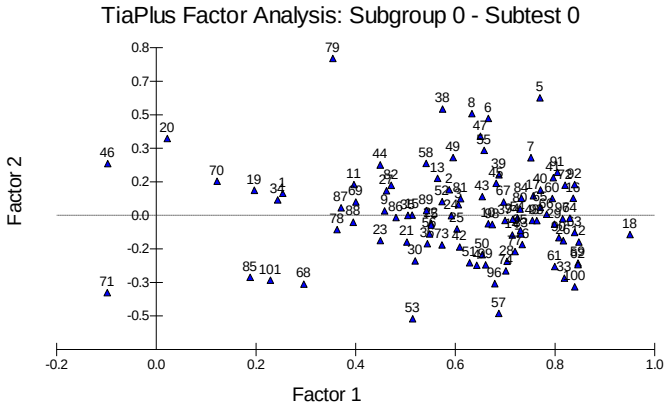
Graphique 2 : la courbe caractérisant l’item (CCI)

Enfin, nous avons ▼ C (0) et ◆ D (0) qui dans ce cas ne présentent pas de propositions de réponse, car l'item en question est du type VF. Ainsi, les données C et D seront seulement considérées lorsque nous analysons un item contenant trois ou quatre propositions de réponse.

Par contre, si la probabilité de répondre correctement à un item par les candidats « forts » n'est pas atteinte, le graphique sera présenté de façon inverse. C'est-à-dire que la probabilité de répondre de façon incorrecte à l'item, sera croissante. Nous proposons un exemple de cette situation dans le graphique 4.

3.2.2. L'analyse factorielle

Comme nous l'avons mentionné dans le cadre théorique, l'une des caractéristiques qui rend notre instrument de mesure valide est la fiabilité, c'est-à-dire la mesure grâce à laquelle, lors d'une ou plusieurs passations, les candidats seront positionnés de la même façon avec un degré d'erreur minimum. Le graphique de l'analyse factorielle vise justement à illustrer un classement univoque. En d'autres termes, il vise à illustrer si notre test mesure un seul construit : *Factor analysis can be used to check the dimensionality of your data: to determine whether all items are measuring the same trait (one factor) or not* (Cito, 2013, p .20). Si l'on réussit à mesurer un seul construit, les candidats seront donc classés ou positionnés de la même façon, après plusieurs passations. Ainsi, le graphique de l'analyse factorielle doit ressembler à ceci, si notre test est effectivement univoque :



Graphique 3 : l'analyse factorielle

Sur le graphique 3, nous avons un exemple de graphique de l'analyse factorielle qui montre que « les items 71 et 46 invalident l'hypothèse de l'unidimensionnalité, car ils se situent de l'autre côté de l'axe (Factor 2) par rapport à la majorité des items (Factor 1). L'item 20 (Copie de B1_EEC_casaniers (item 2)) est assez proche de l'axe vertical, et normalement si le test est unidimensionnel, les items doivent se situer du même côté du graphique, et le plus loin possible de l'axe vertical. » Biagiotti et Cervini 2015 P 16.

Cependant, pour aboutir à un graphique univoque, il faut aussi considérer les points suivants :

1. Le nombre de candidats aura une influence sur la fiabilité que nous pouvons attribuer à ce graphique. D'un côté, il est habituellement recommandable d'

   employer un échantillon 10 fois plus grand que le nombre de variables ou items (N =10k où k est le nombre d'items ou variables (Nunnally, 1978; Thorndike, 1982). D'autres auteurs (Guilford, 1954; Kline, 1986, 1994) estiment qu'un échantillon mineur peut suffire, deux ou trois fois le nombre de variables (N = 2k à 3k), à condition que le nombre des candidats ne soit pas inférieur à 200. Des échantillons plus petits peuvent être acceptables si nous refaisons l'analyse plusieurs fois » [Traduction libre] (Morales Vallejo, 2012, p. 14).

2. « Il y a d'autres cas où il est difficile de maintenir ce principe, par exemple quand il s'agit de tests très longs (à cause de la présence de la fatigue, ou de l'ennui) » [Traduction libre] (Martinez et al., 2014, P 169).
3. Enfin, d'après les travaux fait précédemment par le projet SELF en Italien, nous pouvons déduire que si le fichier à interpréter par TiaPlus contient des items avec des indices de difficulté qui ne sont pas acceptables, et qui ne sont pas discriminants, ils doivent être inactivés afin d'aboutir à ce graphique : « Après la 1ère analyse, nous ne pouvions pas avoir accès au graphique de l'analyse factorielle ; le logiciel ne pouvait pas effectuer les calculs à cause des items 31 (Copie de B2_EEC_post_forum (item 1)) et 32 (Copie de B2_EEC_post_forum (item 2)). Nous avons donc désactivé ces items et relancé l'analyse. » (Biagiotti et Cervini 2015. P 15).

3.3. Exemple d'analyse des items

Au moment d'effectuer les analyses psychométriques à Innovalangues, nous nous focalisons sur les valeurs suivantes :

- Le degré de difficulté/facilité de l'item
- Le RIR : indice de discrimination de l'item.
- Le RIT / *Factor analysis*
- Le Coefficient Alpha de Cronbach.

Nous allons par la suite présenter deux exemples d'analyse nous permettant de comprendre ce qu'est un item valide, et ce qu'est un item non-valide. Nous allons respectivement faire appel à ces exemples : exemple d'un item valide et exemple d'un item non valide.

3.3.1. Exemple d'un item valide

Voici un item qui démontre un comportement acceptable. Il s'agit de l'item numéro 34 du test pilote A2, il évalue la CO sous la forme d'un QRU.

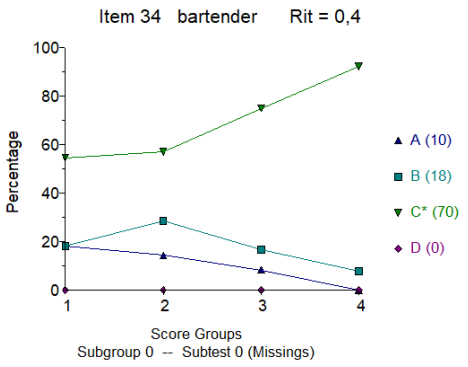
Item Label	Item nr.	Weight	Key	P- and A- values A	B	C	D	E	F	O/D	Mis-sing	Weighted Max	Mean	P	Sd	RSK	Rit	Rir	AR
bartender	34	1	C	10	18	70*	0			2	1	1	0,70	70	0,46	0,46	40	35	86

Tableau 6 : tableau ANA, exemple d'un item valide

La valeur de P, niveau de difficulté est de 71, c'est-à-dire que 71% de l'échantillon a choisi la réponse correcte.

Le distracteur A a été choisi par 10% de l'échantillon, tandis que le distracteur B a été choisi par 18% de l'échantillon. Ce qui nous permet de constater que les trois propositions de réponse ont atteint leur objectif.

Ensuite, le **RIT** qui vaut 40, pourrait, dans un premier temps, nous dire qu'il est un peu bas si l'on considère que sa valeur oscille entre 0 et 100. Cependant, le graphique du RIT nous permet d'observer qu'entre 50% et 60% de l'échantillon (le group fort) a choisi la clé, ou la réponse



Graphique 4 : RIT exemple d'un item valide

correcte, et que la probabilité de répondre correctement à cet item par une population conséquente de notre échantillon est ascendante, tandis que les distracteurs sont minoritairement choisis par le groupe faible.

Ceci est aussi un exemple graphique d'un item discriminatif, car il discrimine entre les candidats forts et les candidats faibles.

En outre, le **RIR,** ou indice de discrimination, dont la valeur est de 35 indique qu'il s'agit d'un item discriminant. C'est-à-dire qu'il y a une corrélation entre la réussite à cet item et la réussite au reste d'items conformant le test.

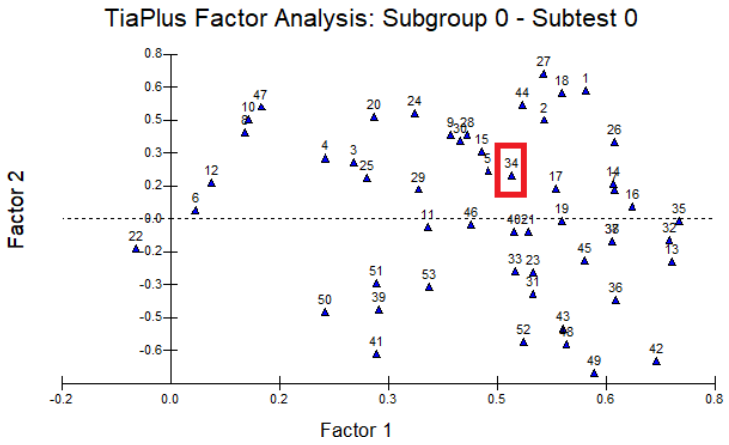
90% Confidence limits for Coefficient Alpha: (0,81 =< 0,86 =< 0,90)

Tableau 7 : coefficient Alpha de Cronbach. Exemple d'un item valide

Enfin, la totalité du test a obtenu un **Coefficient Alpha de Cronbach** de 0,86.

Si nous souhaitons connaître la fiabilité de l'item, il suffit de repérer la valeur de l'**AR : *Alpha Rest***, qui nous permet de vérifier la cohérence de cet item avec le test. Comme nous l'avons dit précédemment, cette donnée doit être inférieure au Coefficient Alpha de Cronbach. L'AR de cet item est de 0,86 et le AC est de 0,87. Il reste donc dans les valeurs permissibles.

A part le coefficient Alpha de Cronbach, nous pouvons vérifier si le comportement de cet item a abouti à la notion d'univocité grâce au graphique du *factor analysis.*



Graphique 5 : analyse factorielle. Exemple d'un item valide

Comme nous pouvons l'observer, l'item 34 est bien positionné du côté droit de l'axe (factor 1).

3.3.2. Exemple d'un item non valide :

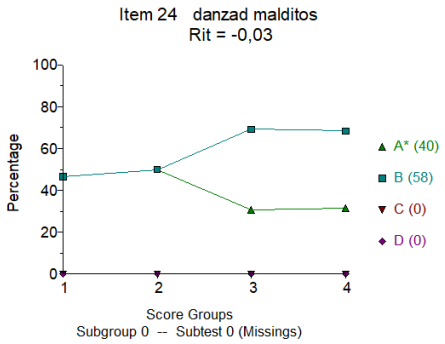
Voyons maintenant l'analyse d'un item qui n'a pas été validé. Il s'agit d'un item de type Vrai/Faux qui évalue la CE.

Item Label	Item nr.	Weight	Key	P- and A- values A	B	C	D	E	F	O/D	Mis-sing	Weighted Max	Mean	P	Sd	RSK	Rit	Rir	AR
danzad_ma	24	1	A	40*	58	0	0			2	1	1	0,40	40	0,49	0,49	-4	-11	86

Tableau 8 : coefficient Alpha de Cronbach. Exemple d'un item non valide

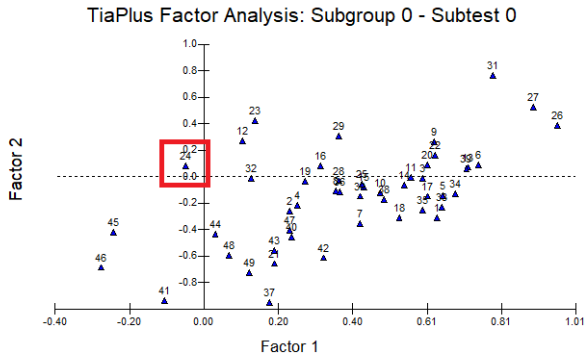
Son degré de difficulté est de 40. Par rapport à l'échelle, il a une valeur P acceptable. Cependant, nous pouvons aussi observer que 58 % de l'échantillon a choisi le distracteur B.

Ensuite, nous observons un RIT négatif de -4. Ce qui veut dire qu'entre 40 et 45% de l'échantillon peut répondre de façon correcte ou incorrecte, mais aussi que la probabilité que les candidats classés dans le groupe des candidats forts répondent incorrectement est croissante, tandis que la probabilité de répondre correctement est descendante.



Graphique 6 : CCI Exemple d'un item non valide.

Le **RIR** est aussi négatif, ce qui veut dire que cet item n'est pas discriminant. Par conséquent, il n'y a pas de corrélation entre la réussite à cet item, et la réussite au reste d'items conformant le test.



Graphique 7 : analyse factorielle. Exemple d'un item non valide

De même, l'analyse factorielle nous permet d'observer que cet item n'évalue pas le même construit que le reste des items.

Son ***Alpha Rest*** est de 0,86 de la même façon que l'**AC** estimé pour la totalité du test est de 0,86.

Quand nous repérons un item comme celui-ci, il est important de vérifier l'item et d'essayer de trouver la raison pour laquelle il n'est pas valide. L'item que nous sommes en train d'analyser est un item qui évalue la CE du niveau A2. Il s'agit d'un item du type VF. Il propose deux questions qui se ressemblent, et les deux réponses sont vraies. Elles se focalisent sur la même information que le candidat doit repérer pour qu'il puisse répondre correctement, ce qui peut le faire hésiter. Analysons les items en question :

Obra de teatro

Una pieza de danza, teatro y competición
En la época de la Gran Depresión, en Estados Unidos se organizaban concursos dónde parejas bailaban de manera continuada, día y noche, hasta acabar extenuados. Ganaban los que más resistían y recibían dinero en metálico. Basado en la célebre película de Sidney Pollak esta propuesta reproduce aquellos maratones de resistencia.

Ítem: VERDADERO/FALSO
Los bailarines que ganaban tenían un premio.

Ítem: VERDADERO/FALSO
Los bailarines que más bailaban ganaban dinero.

Début de la traduction :

Pièce de théâtre :

Une pièce de dance, théâtre et compétition
Pendant l'époque de la Grande Dépression, on organisait aux États-Unis des concours où les couples dansaient en continu, du matin au soir, jusqu'à finir épuisés. Les vainqueurs étaient ceux qui restaient, et ils gagnaient de l'argent liquide. Inspiré du célèbre film de Sidney Pollak, cette mise en scène remonte à ces marathons de résistance.

Item : VRAI /FAUX
Les danseurs vainqueurs gagnaient un prix

Item : VRAI /FAUX
Les danseurs qui dansaient le plus gagnaient de l'argent

Fin de la traduction

En effet, les deux propositions de réponse sont vraies. Cependant, le fait que la deuxième soit plus précise par rapport au type de prix que les participants pouvaient remporter, peut faire hésiter les candidats.

Cet item a donc été éliminé. Il faisait partie d'une tâche d'évaluation conformée par 4 questions. Cependant, cette façon de procéder pour repérer un item avec un comportement non attendu, n'a pas été une concession. Ceci peut répondre à l'une des questions présentée comme problématique : *que doit-on faire avec les items qui n'ont pas provoqué le comportement attendu sur les candidats lors de la conception* ? En effet, les items sont créés à partir d'une ressource authentique. Quand le contenu de la ressource permet d'aboutir, soit à la création d'un autre item, soit à la modification d'un item, l'équipe de concepteurs fait les modifications nécessaires. Bien évidemment, la modification, ou la création, d'un item demande toujours au concepteur de se mettre en situation, c'est-à-dire qu'il est important de réfléchir sur les compétences existantes chez un candidat dans un niveau donné.

3.4. L'apport des analyses psychométriques par TiaPlus au SELF espagnol

Lors de l'analyse des 4 niveaux : A1, A2, B1, et B2, j'ai pu repérer les données suivantes :

Une autre mesure intéressante à considérer lors des analyses est l'écart type, ou *SD*. J'ai pu observer qu'il n'a jamais été supérieur à 0,50, et qu'au moment de trouver un écart type de 0,50, les distracteurs étaient choisis entre 43% et 55% de l'échantillon. Cela veut dire que seulement 57% et 45% des candidats de l'échantillon ont choisi la réponse correcte.

J'ai pu aussi constater que plus le niveau à évaluer est élevé, plus bas sera l'indice de difficulté des tests. De la même façon que plus le niveau est élevé, plus difficile sera de trouver des candidats. En effet, si nous reprenons le principe de la TRI, où l'indice de difficulté est estimé à partir des personnes qui ont répondu correctement à un item quelconque, et que la population des candidats à partir du niveau B est moins nombreuse, la probabilité de réponse correcte sera plus faible aussi.

INDICE DE DIFFICULTÉ (P)				
COMPÉTENCE	A1	A2	B1	B2
CO	64,64	61,85	59,65	52,63
CE	61,2	77,78	55,94	51,61
EEE	39	29,78	42,11	26,36
MOYENNE P	54,94	56,47	52,56	43,53

J’ai pu aussi repérer les temps minimum et maximum pour chacune des passations :

TEMPS DES PASSATIONS			
A1		A2	
TEMPS MIN	TEMPS MAX	TEMPS MIN	TEMPS MAX
26 min	1h	31 min	1h

TEMPS DES PASSATIONS			
B1		B2	
TEMPS MIN	TEMPS MAX	TEMPS MIN	TEMPS MAX
35 min	6h	40 min	6 jours

Concernant le comportement stimulé chez les candidats par certains types de tâches, et pour répondre à la problématique : *Qu’est-ce que les analyses psychométriques nous permettent d’observer sur la typologie des tâches proposées ?* nous avons aussi pu remarquer que :

- Les tableaux *Vrai/Faux/on ne peut pas dire* peuvent induire à l’erreur, car l’option « on ne peut pas dire » est fortement choisie, même si elle ne correspond pas à la réponse correcte.

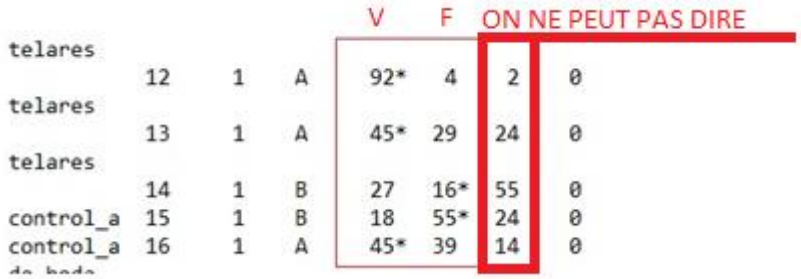


Tableau 9 : analyse de tâches VFNM

Comme nous pouvons observer, sur les items contenus sur le tableau 9, l’option « on ne peut pas dire », qui est la troisième des options comprises dans le rectangle rouge, est tout le temps choisie. Pour l’item 12, elle a été choisie par 2% de l’échantillon, pour l’item 13 par 24%, pour l’item 14 par 55%, pour l’item 15 par 24%, et enfin pour l’item 16 par 14% (La tâche d’évaluation « telares » est proposée en annexe 6).

- Dans la même voie, les Questions à Réponses Multiples (QRM) ont aussi été l'objet de confusion. Voyons un exemple de ce que l'analyse faite sur TiaPlus nous permet d'observer sur les QRM :

			CLÉ	A	B	C	D
carta_dir	24	1	A	47*	0	27	4
carta_dir	25	1	D	6	10	4	76*
carta_dir	26	1	BD	22	37*	12	55*

Tableau 10 : analyse de QRM

Le tableau 10 nous permet d'observer, tout d'abord, que les items 24 et le 25 sont des QRU. La seule question à choix multiples est l'item numéro 26. Cependant, nous pouvons observer que tous les distracteurs ont été choisis, à l'exception du distracteur B (en jaune). L'item 24 a eu 47% de réponses correctes, et 31% de réponses incorrectes. L'item 25 a eu 76% de réponses correctes, et 20% de réponses incorrectes. Enfin, l'item 26 a eu respectivement 55% et 37% de réponses correctes. Cependant, il a eu aussi 34% de réponses incorrectes. Ceci nous permet d'observer que plus d'un candidat a choisi les deux réponses correctes, et aussi au moins un distracteur. Cette tâche d'évaluation est présentée en annexe 7 .

De ce fait, ils ont été majoritairement éliminés dans la version de Pré-test. Jusqu'à présent, il y a uniquement 2 items du type QRU dans la version de PRE-TEST. C'est pourquoi l'équipe SELF en espagnol est en train de décider s'il est pertinent d'éliminer ces typologies de tâches du pré-test. En effet, cette décision sera prise sous le contrôle de Madame Cristiana Cervini, référente scientifique du projet SELF, et Madame Monica Masperi, responsable scientifique de l'IDEFI Innovalangues.

4. Le logiciel WINSTEPS

Winsteps est un logiciel avec une licence payante[11] qui fonctionne sous Windows. Il a été conçu sous la direction de John Michael Linacre. Le site web www.winsteps.com a mis à disposition une version d'essai : MINISTEP.

[11]Le cout de la licence est de $149 dollars américains.

La démarche statistique du logiciel se base sur le modèle de Rash, ou modèle à un paramètre. À la différence du logiciel TiaPlus, il « peut opérer avec un minimum de deux personnes par item ». [Traduction libre] (Linacre, 2017, p. 596). De même, les données fournies par Winsteps permettent de faire des analyses sur les items, et sur les candidats conformant l'échantillon. Ceci est possible, car le modèle de Rasch vise justement à identifier à travers une série de « pré-requis », quels sont les items et quels sont les candidats qui s'adaptent au modèle, c'est-à-dire au modèle de Rasch.

En effet, nous avons dit dans la partie précédente que le logiciel TiaPlus, à travers son graphique de la courbe caractéristique de l'item, nous permet de discriminer les candidats forts des candidats faibles. Cependant, il nous reste encore à déterminer, à travers une méthode de vérification en profondeur, si les résultats des comportements non attendus par les items proviennent soit de l'objet de mesure, soit de l'échantillon.

Cela est possible, comme nous l'avons dit dans le cadre théorique, au travers du processus de qualité qui cherche à connaître le degré d'erreur, ou degré d'ajustement, au travers du FIT.

Cette partie présentera d'abord quelques-unes des données d'entrée conformant le tableau d'analyse de départ du projet SELF. Ensuite, nous allons présenter un item qui s'adapte au modèle de Rasch à deux paramètres. Finalement, nous allons présenter un item qui ne s'adapte pas au modèle. Ce dernier nous permettra d'illustrer la façon dont nous pouvons identifier d'où peut provenir le désajustement : de l'item, ou du candidat.

Pour aboutir à cette partie, il a été nécessaire de travailler sur Ministep, la version d'essai de Winsteps. En effet, les réflexions sur cette partie du mémoire ont été faites à partir des deux dernières semaines du stage, dû aux contraintes présentées dans la première partie de ce travail. Cependant, cela fut l'occasion de découvrir pendant un mois de travail sur Ministep, que ce dernier met à disposition de l'usager les mêmes fonctionnalités que celles proposées par Winsteps. La seule différence repérée sur Ministep est qu'il permet d'analyser un total de 25 items et 75 candidats. En revanche, Winstep permet d'analyser un total de 60,000 items et 10,000,000 de candidats. Ainsi, pour aboutir à cette partie du mémoire, j'ai choisi d'analyser

seulement 24 des items figurant comme les items ancres[12], et nous avons choisi au hasard 75 candidats qui ont participé au Pré-test B.

4.1. Interprétation des tables de résultats

Ci-dessous un exemple de la table[13] qui fait l'objet des analyses psychométriques du projet : *Item Statistics Entry Order* :

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MODEL MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
1	63	75	-.97	.35	1.29	1.3	1.56	1.2	.20	.39	81.1	86.0	A1_CO_carmen (item 1) PROP
2	65	75	-1.24	.38	.96	-.1	1.01	.2	.39	.37	87.8	88.2	A1_CO_amigos_cafe_cerveza
3	73	75	-3.28	.75	.83	-.1	.18	-.8	.34	.22	97.3	97.3	A2_CO_nevera_roja (item 1)
4	50	75	.26	.28	1.21	1.6	2.34	4.1	.27	.44	66.2	73.8	A2_CO_nevera_roja (item 2)
5	71	75	-2.46	.56	1.07	.3	.57	-.3	.30	.29	94.6	94.6	A2_CO_nevera_roja (item 3)
6	51	75	.18	.28	.94	-.4	.87	-.4	.48	.44	78.4	74.6	B1_CO_jabon_rey (item 1) P
7	43	75	.77	.27	.95	-.5	.88	-.5	.49	.45	70.3	70.2	B1_CO_jabon_rey (item 2) P

Tableau 11 : tableau 14, Winsteps

Le tableau 11 présente tous les items analysés dans l'ordre, du premier au dernier. Si nous suivons le guide du logiciel, nous trouvons que :

Entry number, est le numéro de l'item.

Total score, est le total des personnes qui ont répondu correctement à l'item.

Total count, est le total de réponses données à l'item.

Measure, est le degré de difficulté estimé de l'item. Cette donnée est présentée en logits. Comme elle est estimée pour chacun des items tout en considérant leur degré réel de difficulté, il n'existe pas une échelle limite (les logits sont compris entre plus ou moins l'infini). Cependant, « il est commun qu'elle oscille entre -3 et +3. Pour des raisons pratiques, elle ne varie pas au-delà de l'intervalle -5 et +5 logits. Dans un cas extrême, elle peut avoir une variation de -8 au +8 logits. » [Traduction libre] (Tristán López, 2002, p. 16).

Model S.E. est l'écart type de chacune des mesures (de degré de difficulté de l'item). Cette mesure est donnée en logits.

INFIT MNSQ[14], « est la mesure qui nous permet d'effectuer le contrôle de qualité de l'ajustement interne du test. Il s'agit d'une valeur sensible au comportement

[12] Les items ancres sont les items qui ont en commun les deux versions du pré-test (pré-test A et pré-test B). Ils ont été choisis comme items ancres, car ils ont démontré les meilleurs indices de discrimination une fois effectuées les analyses des pilotages.

[13] À savoir, le logiciel Winsteps fournit un total de 120 tables. Cependant, nous allons seulement présenter deux tables. Elles font partie de la méthode d'analyse psychométrique adoptée par le projet. Nous tenons à préciser que cette méthode a été adopté sous le conseil du CIEP.

inattendu qui affecte les items dont la difficulté est PROCHE du niveau d'habilité d'une personne. » [Traduction libre] (Tristán López. 2002. P 111). Par exemple, il «permet de détecter les incohérences d'un candidat qui échoue à un item de son niveau » (Biaggioti & Cervini, 2015, p. 4.) De son côté, Linacre (2017) dit que les infit mean-squares élevés indiquent que les items ont eu une mauvaise performance, quand ils ont été répondus par les candidats cibles.

OUTFIT MNSQ, est la mesure qui nous permet d'effectuer le contrôle de qualité d'ajustement externe du test. Il s'agit d'une « valeur sensible aux comportements inattendus qui affectent les items dont la difficulté est éloignée du niveau d'une personne ». [Traduction libre] (Tristán López. 2002. P 111). « Par exemple, un candidat fort qui a échoué à un item très facile, ou l'inverse. » (Biaggioti & Cervini, 2015, p, 4.)

Dans le cadre du projet, les indicateurs de ces deux données sont les suivants :

MeanSquare	
> 2.0	Dégrade l'instrument de mesure
1.5 – 2.0	Non productif pour l'instrument de mesure, mais ne le dégrade pas
0.5 – 1.5	Utile et productif pour l'instrument de mesure
< 0.5	Moins productif pour la mesure, mais ne la dégrade pas. Peut conduire à de mauvaises interprétations

Tableau 12 : valeurs du FIT MeanSquare

Cependant, pour les items ancres, il est recommandé que ces valeurs soient comprises entre 0,75 et 1,35 (Innovalangues s.d. p.1).

INFIT/OUTFIT ZSTD, nous pouvons trouver que l'INFIT MNSQ, ainsi que le l'OUTFIT MNSQ débouchent sur des mesures productives pour l'instrument de mesure. C'est-à-dire, que d'après elles, le contrôle de qualité est fait correctement. Cela peut entrainer des doutes. En effet, la mesure standardisée a pour objectif de vérifier s'il y a un « ajustement raisonnable entre les candidats et les items » [Traduction libre] (Gonzalez Montesinos, 2008, p. 26)

---

[14] Mean Square, ou moyenne quadratique. Elle fait la somme des données élevées au carré « afin d'éliminer les nombres négatifs. » [Traduction libre] (Gonzalez Montesinos, 2008, p. 24)

Dans le cadre du projet, les indicateurs de ces deux données sont les suivants :

Standardized	
>= 3,0	Données très inattendues par le modèle
2,0 – 2,9	Données partiellement imprévisibles
-1,9 – 1,9	Le modèle ajuste bien les données
<= -20	Données trop imprédictibles. Eventuelle présence d'autres dimensions

Tableau 13 : valeurs du FIT ZSTD

Ptmeasure-Al Corr, Il s'agit du coefficient de corrélation. D'après le CIEP (s.d.a), un coefficient de corrélation acceptable doit osciller entre 0,15 et 0,30. D'après González Montesinos (2008), plus élevée est la valeur de la corrélation, plus élevé sera l'indicateur d'univocité. Voici ci-dessous une table nous permettant de voir les valeurs données au coefficient de corrélation.

PTMes	<0	<0.15	>0.15	>0.30

Par contre, Linacre (2017) nous dit que lorsqu'on trouve une corrélation négative, nous pouvons considérer qu'elle indique que la réponse à l'item ne favorise pas l'univocité du test. Par conséquent, l'item peut être omis.

Ptmeasure-Al Exp, « C'est la valeur attendue du point de corrélation par le modèle de Rash. » [Traduction libre] (citar sitio web, pie de página)

4.2. Exemple d'analyse des items

Lorsque nous analysons quels items et quels étudiants s'adaptent au modèle de Rash, nous nous focalisons sur les valeurs suivantes :

- La mesure de difficulté des items
- L'OUTFIT et l'INFIT, modèle Mean Square.
- L'OUTFIT et l'INFIT, modèle Standardisé.
- Le Z score[15], ou résidu, qui est la soustraction entre les valeurs attendues par le modèle, et les résultats observés. (Gonzalez Montesinos, 2008)

[15] Cette donnée peut être interprétée à partir de la table 11 : *table of poorly fitting item*

Un exemple d'analyse est présenté par la suite afin de comprendre ce qu'est un item qui s'adapte au modèle de Rash, et par conséquent valide.

4.2.1. Exemple d'item qui s'adapte au modèle de Rasch.

ITEM STATISTICS: ENTRY ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
2	65	75	-1.24	.38	.96	-.1	1.01	.2	.39	.37	87.8	88.2	A1_CO_amigos_cafe_cerveza (item 1) PROPOSITION

Tableau 14 : exemple d'item qui s'adapte au modèle de Rasch.

Il s'agit de l'item numéro 2, qui évalue la CO. Il a été répondu par 65 personnes sur 75. Nous pouvons observer que le degré de difficulté estimé, *measure*, est un nombre négatif. Cependant, ce n'est pas un facteur négatif. Tout d'abord, il s'agit d'une mesure estimée en *logits*, dont sa valeur peut osciller entre l'infini négatif et l'infini positif.

Repèrerons la mesure la plus basse et celle la plus élevée :

L'échelle de mesure calculée pour cet item oscille de -3,28 à +5,43 logits. La valeur -3,28 représente le degré de difficulté le plus bas, tandis que +5.43 représente le degré de difficulté le plus élevé. En effet, « un item difficile obtiendra une calibration supérieure à 0, tandis qu'un item facile obtiendra une calibration inférieure à 0 » [Traduction libre] (Tristán López, 2002, p. 17). D'après Tristan Lopez (2002), nous pouvons conclure que cet item a un degré de difficulté facile. Néanmoins, il est aussi important de garder en tête, que nous sommes en train de mesurer un construit à plusieurs niveaux de difficulté : A, B et C. De ce fait, lors de la validation de l'item, il est important de vérifier que les candidats qui sont capables de répondre à cet item y répondent correctement. Nous allons donc vérifier ce processus d'ajustement en partant des mesures d'ajustement : FIT.

INFIT		OUTFIT	
MNSQ	ZSTD	MNSQ	ZSTD
.96	-.1	1.01	.2

Nous allons d'abord interpréter les valeurs MNSQ, puis les valeurs ZSTD. D'après les bornes du projet SELF Innovalangues, avec les OUTFITS et INFITS classés entre 0,5 et 1,5, un item est « utile et productif pour l'instrument de mesure » (Innovalangues, sd, p. 1)

Ainsi, cet item est utile et productif pour l'instrument de mesure, c'est-à-dire le test SELF en espagnol. Mais, qu'est-ce que ces données veulent dire ? D'abord, l'OUTFIT MNSQ nous dit que lors de la passation, cet item n'a pas eu de comportement inattendu. Et son INFIT MNSQ nous indique que cet item a été répondu par les candidats qui ont le niveau pour y répondre.

Ceci est aussi justifié par la mesure standardisée, qui doit être comprise entre -1,9 et +1,9 pour un item qui s'ajuste bien au modèle.

De plus, si nous vérifions sur la table des résidus : *Table 11, table of poorly fitting item*, nous pouvons constater que cet item ne figure pas parmi les items qui ont démontré un « ajustement pauvre ». (Cf. tableau 15)

```
    4  A2_CO_nevera_roja (item     .26     1.2   A    2.3
  OBSERVED:  1:  1  0  1  0  0   1  1  0  0  0   1  1  1  0  1   1  0  1  1  1
Z-RESIDUAL:

  OBSERVED: 21:  1  1  1  1  0   1  0  0  1  0   1  1  1  1  0   0  1  1  1  1
Z-RESIDUAL:                     X

  OBSERVED: 41:  1  0  1  1  1   1  0  1  1  0   1  1  1  1  0   0  1  0  1  1
Z-RESIDUAL:              3        -9                                    -2

  OBSERVED: 61:  1  1  1  0  1   1  1  1  0  1   0  0  1  0  1
Z-RESIDUAL:

    1  A1_CO_carmen (item 1) P    -.97     1.3   B    1.6
  OBSERVED:  1:  1  1  1  1  1   1  1  0  1  1   1  1  1  0  1   1  1  1  1  1
Z-RESIDUAL:                              -2                 -2

  OBSERVED: 21:  1  1  1  1  0   1  1  1  1  1   1  1  1  1  1   0  1  0  1  1
Z-RESIDUAL:                   -2   X

  OBSERVED: 41:  1  1  1  1  0   1  1  1  1  1   1  1  1  0  1   1  1  1  1  0
Z-RESIDUAL:              2 -4                                 -2              -5

  OBSERVED: 61:  1  1  1  1  1   0  1  1  0  1   0  1  0  1  1
Z-RESIDUAL:                                       -3    -4
    9  B2_CO_control_alcoholem     .77     1.1   C    1.2
  OBSERVED:  1:  0  0  0  1  1   1  0  1  1  0   0  1  1  1  1   1  0  1  1  0
Z-RESIDUAL:

  OBSERVED: 21:  0  0  0  0  1   1  0  0  1  0   0  1  0  1  0   0  1  1  1  1
Z-RESIDUAL:                     X

  OBSERVED: 41:  1  0  1  0  1   0  1  1  1  1   0  1  1  1  0   1  1  1  0  1
Z-RESIDUAL:

  OBSERVED: 61:  1  0  1  0  1   0  1  0  1  1   1  0  1  0  0
Z-RESIDUAL:                           5
```

Tableau 15 : table of poorly fitting item

Enfin, nous trouvons que le coefficient de corrélation estimé pour cet item est supérieur à 0,30. Il a une valeur de 0,39 ce qui veut dire qu'il contribue à l'univocité du test.

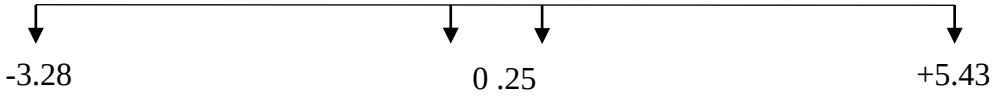
4.2.2. Exemple d'un item qui ne s'adapte pas au modèle de Rasch

Il s'agit de l'item 4, qui évalue la CO.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
4	50	75	.26	.28	1.21	1.6	2.34	4.1	.27	.44	66.2	73.8	A2_CO_nevera_roja (item 2) PROPOSITION

Tableau 16 : exemple d'item qui ne s'adapte pas au modèle de Rasch.

Il a été répondu par 50 personnes sur 75. Nous voyons que Winsteps a calculé un niveau de difficulté de 0,26 pour cet item.



INFIT		OUTFIT	
MNSQ	ZSTD	MNSQ	ZSTD
1.21	1.6	2.34	4.1

Nous pouvons observer que les indicateurs d'ajustement interne sont acceptables, tandis que les indicateurs d'ajustement externe ne rentrent pas dans le rang d'un item productif pour le modèle.

Maintenant, nous allons analyser la table 11, *le table de poorly fitting item* :

TABLE OF POORLY FITTING ITEM (PERSON IN ENTRY ORDER)
NUMBER - NAME -- --------------- MEASURE - INFIT (MNSQ) OUTFIT

4 A2_CO_nevera_roja (item .26 1.2 A 2.3

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
OBSERVED: 1:	1	0	1	0	0	1	1	0	0	0	1	1	1	0	1	1	0	1	1	1
Z-RESIDUAL:																				
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
OBSERVED: 21:	1	1	1	1	0	1	0	0	1	0	1	1	1	1	0	0	1	1	1	1
Z-RESIDUAL:						X														
	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
OBSERVED: 41:	1	0	1	1	1	1	0	1	1	0	1	1	1	1	0	0	1	0	1	1
Z-RESIDUAL:				3			-9											-2		
	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75					
OBSERVED: 61:	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1					
Z-RESIDUAL:																				

Tableau 17 : table of poorly fitting item

Nous observons que cet item est le premier de la liste des items que comprend cette table : *Items are listed in descending order of misfit* (Linacre, 2017, p. 356), ce

qui veut dire que cet item présente un désajustement important. Cependant, nous voyons que 4 candidats font l'objet de notre attention dans cette table[16] :

- Le candidat 26[17] est un candidat extrême: « 'X' indicates that the person obtained an extreme score » (Linacre, 2017. P 356). D'après Tristant 2008, un candidat extrême[18] est soit un candidat qui a répondu à toutes les questions de façon correcte, soit un candidat qui a répondu à toutes les questions de façon incorrecte. Dans ce cas, l'auteur nous recommande de désactiver les candidats extrêmes. D'après Linacre (2017), les ponctuations extrêmes représentent les niveaux le plus bas et le plus élevé d'un candidat par rapport à un item quelconque.
- Nous avons ensuite les candidats 44, 47 et 58. Nous pouvons observer que le résidu du candidat 47 est de -9. D'après le CIEP (s.d.b), « dans les cas d'outfits trop importants, il peut arriver que seuls quelques candidats présentent des résidus trop grands (>6). Il peut alors être utile de les retirer de l'analyse ». En effet, les outfit mnsq et zstd sont trop importants. Enfin, Linacre (2017) nous dit qu'un outfit mean-square élevé peut être le résultat de plusieurs réponses données au hasard.

Nous pouvons donc désactiver les candidats 26 et 47. Et une fois faites les analyses de chacun des items, nous pouvons relancer une deuxième analyse, ensuite une troisième analyse. Ceci se refait autant de fois que le concepteur/évaluateur le considère nécessaire, afin d'aboutir à un instrument de mesure avec un degré minimum d'erreur. Néanmoins, il est important de signaler que dans le cas particulier de cet item, les données inattendues proviennent plutôt de l'ajustement externe, c'est-à-dire des candidats. Il est possible que lors de la deuxième analyse, si nous ne considérons pas les candidats 26 et 47, nous aurons un OUTFIT qui s'adapte au modèle.

---

[16] Cette table se lit de gauche à droite. Chacun des 1 ou 0 correspond à la réponse donnée par chacun des candidats. 1 correspond à une réponse correcte, et 0, dans ce cas, correspond à une réponse incorrecte. Nous proposons en rouge le numéro qui correspond à chacun des candidats, d'après la succession qui commence par la gauche et continue par la droite.

[17] NB : Il y a deux aspects à ne pas oublier. Le premier est que cette analyse a été faite sur des items ancres. Ceci pourrait expliquer les différences entre les résultats obtenus par le candidat boulay-gweldann lors de la passation totale du test. Cf annexe n : 8.

[18] Le même terme EXTRÊME s'adapte à un item. Un item est extrême quand il n'a jamais été répondu correctement, ou quand il a été répondu correctement par tous les candidats.

L'analyse du logiciel Winsteps et de ses fonctions aurait pu être plus approfondie avec plus de temps. Ce que nous venons de présenter, représente les premières réflexions sur ce processus. En effet, une période de 4 mois ne suffit pas pour pouvoir maitriser un logiciel qui fournit des données sous 120 tables, et dont la maîtrise est liée à des connaissances dans un domaine de spécialisation tel que la psychométrie.

4.3. L'apport des analyses psychométriques par TiaPlus au SELF espagnol

Lors des analyses faites sur les items choisis, nous avons pu repérer que les mesures d'ajustement interne du test, c'est-à-dire l'INFIT, indiquent que les items sont utiles et productifs pour le test, car ils oscillent tous entre 0,5 et 1,5. De même, la mesure standardisée de l'INFIT nous permet de vérifier que ces items s'ajustent bien au modèle de Rasch, car cette mesure (INFIT ZSTD) a été estimée entre -1,9 et + 1,9.

L'item qui a démontré des comportements plus inattendus, est l'item que nous avons proposé ci-dessus, qui a pour titre : *un item qui ne s'adapte pas au modèle de Rasch.* Cependant, nous considérons que ceci peut provenir de la participation de deux candidats, et non pas de l'item lui-même. Pour pouvoir l'affirmer, il faudrait faire une deuxième analyse.

La présente analyse nous permet de conclure que les analyses faites sur TiaPlus, malgré les limitations de la TCT, ont permis de repérer les points d'amélioration de ces items qui conforment la banque des items ancres de SELF en espagnol, ce qui lors des analyses sur Winsteps se traduit par des résultats plutôt favorables. Bien entendu, cette partie se base principalement sur les conclusions que nous pouvons tirer à partir des mesures d'ajustement, FIT.

Conclusion

Une tâche d'évaluation valide est celle qui aura une probabilité de réponse correcte élevée par l'échantillon ciblé comme étant capable d'y répondre correctement.

Si chacune des tâches d'évaluation est validée à travers ce principe, la probabilité de positionner les candidats de la même façon après plusieurs passations sera aussi élevée. Cela peut se traduire par : un test qui mesure un seul construit, et par conséquent, un test fiable.

Comme le dit Bachman 2004, la fiabilité et la validité sont des aspects complémentaires. En effet, si la validation de chacune des tâches d'évaluation est réussie, une fois que chacune d'elles positionne de façon fiable, nous pourrions même dire qu'il s'agit d'un seul processus, où la validité et la fiabilité ont un lien indissociable.

Les apports psychométriques de SELF en espagnol à la didactique des langues, que j'ai pu repérer pendant quatre mois de stage, grâce au travail fait en équipe, est que la typologie des tâches a aussi un impact sur la validité du test. Nous avons pu constater que les tâches d'évaluation du type vrai/faux, ainsi que les questions à choix unique, font partie des tâches d'évaluation qui ont favorisé la validité du test.

J'ai aussi pu observer comment les analyses psychométriques peuvent devenir un juge ou médiateur du travail de conception, surtout dans le cadre de l'évaluation. Elles permettent au concepteur prendre de recul sur son travail, de voir au-delà, et de constater si ce qu'il avait initialement prévu a atteint les objectifs définis.

L'intérêt de ces pratiques psychométriques repose justement sur sa caractéristique médiatrice. Grâce au modèle de Rasch de la TRI, nous pouvons faire une analyse sur les comportements des candidats, mais aussi sur le comportement du test. Cela est bénéfique pour le candidat qui cherche à améliorer ses compétences en langue étrangère à travers un positionnement, mais aussi pour le concepteur qui s'intéresse à positionner de façon fiable.

Bibliographie

ALTE, (2011). *Manuel pour l'Élaboration et la Passation de Tests et d'Examens de Langue.* Division des Politiques linguistiques. Strasbourg : Conseil de l'Europe, DG II – Service de l'éducation.

Bachman, L. F. (2004). *Statistical Analysis for Language Assessment*. New York : Cambridge University Press.

Biagiotti, T. & Cervini, C. (2015) *SELF – Italien. Rapport analyses psychométriques Versions 1 et 2 du Pré-test*. Grenoble, France.

Cervini, C. & Jouannaud M.-P., (2015). Ouvertures et tensions liées à la conception d'un système d'évaluation en langues, numérique, multilingue et en ligne, dans une perspective communicative et actionnelle. *Revue Apprentissage des langues et systèmes d'information et de communication, Alsic, 18*. Repéré le 21 décembre 2015 à: https://alsic.revues.org/2821

CIEP (s.d. a) *La Théorie Classique des Tests & Tia Plus pour les nuls*

CIEP (s.d. b) *Mode d'emploi du logiciel Winsteps pour l'utilisation du modèle de Rasch Version 1.2* Paris

CITO. (2013), *TiaPlus Users manual,* Arnhem, NL. Repéré le 20 mars à http://tiaplus.cito.nl/

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334. Repéré le 20 mars à http://ieg.or.kr/include/file_down.php?save_path=/data1/ref&filename=11250016003000297.pdf&filename2=11250016003000297.pdf.

González-Montesinos, J.M. (2008) *Manual técnico A. Serie: Medición y Metodología.* México D.F. Repéré le 14 juillet à http://www.winsteps.com/a/recursos-offline.pdf

Innovalangues, (s.d.) *SELF – Italien Bornes pour les analyses psychométriques TCT & TRI*

Innovalangues. (2014). *Cahier des charges.* Action 4 : Lot « SELF », V.8.

Conseil de l'Europe. (2001). *Cadre Européen Commun de référence pour les langues : apprendre, enseigner, évaluer*. Strasbourg : Didier.

Frías D. (s.d.) *Alfa de Cronbach y consistencia interna de los ítems de un instrumento de medida* . Réperé le 25 mai à http://www.uv.es/~friasnav/AlfaCronbach.pdf

Linacre, J.M. (2017) *A user's guide to WINSTEPS. MINISTEP Rasch-Model Computer Programs*. Beaverton, Oregon. Repéré le 14 juillet 2017 à http://www.winsteps.com/manuals.htm

Martínez, Ma., Hernández, Ma., & Hernández Ma. (2014) *Psicometría*. Alianza Editorial. Madrid. Réperé le 20 mars 2017 à http://www.dandrosh.com.mx/books/Psicometria%20-20Alianza%20Editorial.pdf

Morales Vallejo P. (2007) *La fiabilidad de los tests y escalas.* Repéré le 27 avril 2017 à https://matcris5.files.wordpress.com/2014/04/fiabilidad-tests-y-escalas-morales-2007.pdf

Morales Vallejo P. (2009) *Análisis de ítems en las pruebas objetivas.* Réperé le 17 avril 2017 à https://educrea.cl/wp-content/uploads/2014/11/19-nov-analisis-de-items-en-las-pruebas-objetivas.pdf

Morales Vallejo P. (2012) *Tamaño necesario de la muestra: ¿Cuántos sujetos necesitamos?* Repéré 6le 14 juillet 2017 à http://www2.df.gob.mx/virtual/evaluadf/docs/gral/taller2015/S0202EAC.pdf

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo 31* 55-66. Réperé le 29 de mayo de 2017 à http://www.redalyc.org/articulo.oa?id=77812441006

Olea, J., Ponsoda, V., & Revuelta. J. (1998) *Psicometria I. Tema VI: Teoría de la respuesta al ítem* (matériel didactique). Facultad de Psicología. Universidad Autónoma de Madrid España. Madrid. Réperé le 17 avril 2017 à https://www.uam.es/personal_pdi/psicologia/cadalso/Docencia/PoliTRI/TRI4_v2.pdf

Projet Innovalangues. (s.d.). Repéré le 12 août 2017 à http://innovalangues.fr/

Prieto, G. & Delgado, A. (2003) Análisis de un test mediante el modelo de Rasch. *Psicothema*. *Vol. 15 n°1*, 94-100. Reperé le 25 mai à http://www.psicothema.com/pdf/1029.pdf

Tristán L.A. (2002). *Análisis de Rasch para Todos: una guía para evaluadores educativos* CENEVAL, México, D.F.

Sigles et abréviations utilisés

ANR : Agence Nationale de la Recherche
CA : Coefficient Alpha de Cronbach
CIEP : Centre International d'Etudes Pédagogiques
CE : Compréhension Écrite
CC : Cahier des Charges
CO : Compréhension Orale
COCA : Compréhension Orale, Conception Assistance
EEC : Expression Écrite Courte
ENPA : Environnement Numérique d'Apprentissage Personnalisé
GAMMER : *Gaming Applications for Multilingual Educational Resources*
IDEFI : Initiatives d'Excellence en Formations Innovantes
PCIC : *Plan Curricular del Instituto Cervantes*
QRM : Question à Réponses multiples
QROC : Question à Réponse Ouverte Courte
QRU : Question à Réponse Unique
SELF : Système d'Evaluation en Langues à visée Formative
THEMPPO : THEMatique Prosodie Production Orale
TCT : Théorie Classique des Tests
TRI : Théorie de Réponse à l'Item
VF : Vrai/Faux
VFNM : Vrai/Faux/Non Mentionné

Table des illustrations

Table des tableaux psychométriques

Table des graphiques

Table des annexes

Annexe 1
BILAN Pilotages de SELF Espagnol A1-B2

INTRODUCTION

Le présent bilan comporte les résultats des pilotages faits depuis décembre 2016 jusqu'au 13 avril 2017.

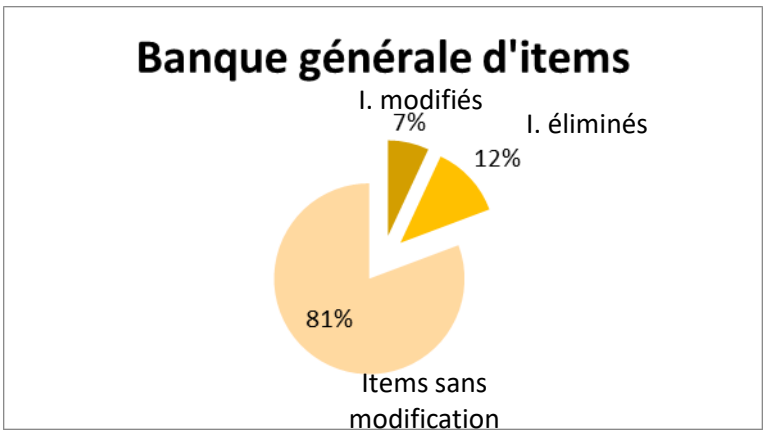
Afin de mener à bien cette étape du test, nous avons varié notre public. Les pilotages ont été donc passés à : Grenoble (CLV, MDL, Minatec, SDL), Valence (UGA), et Amiens.

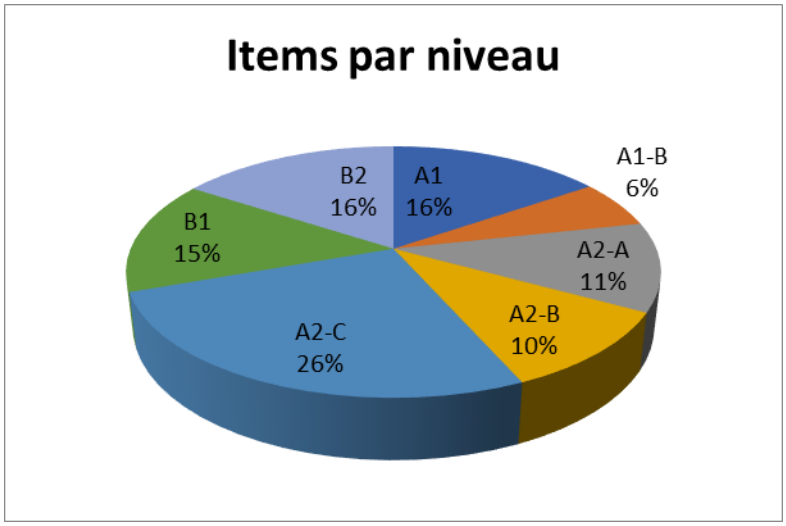
Nous pouvons classer le public envisagé pour répondre aux pilotages en ligne en trois grands groupes : ceux qui n'ont pas fini le test, ceux qui n'ont pas fait le test, et ceux qui l'ont passé en totalité. Les deux premiers groupes étant majoritaires, nous ne considérons pas fiables les données extraites de cette modalité du pilotage.

En ce qui concerne les passations présentielles, nous pouvons dire que nous avons des résultats assez satisfaisants. Nous allons donc présenter les bilans des passations en présentielle.

Bilan Général du pilotage A1-B2

Au total, l'équipe SELF-espagnol compte sur une banque de 372 items, dont nous n'en avons éliminé que 12% et modifié que 7%, ce qui est représenté par le graphique suivant :





Les pilotages ont été passés par 455 candidats, dont 403 figurent comme des candidats effectifs. Ceci représente 89% des candidats.

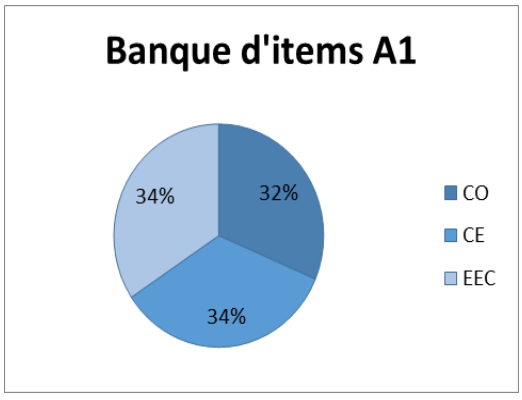
La moyenne du P repérée tout au long des pilotages est de 54,54.

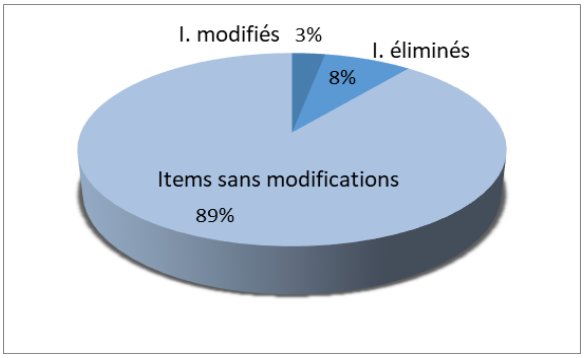
Bilan du niveau A1

Les pilotages du niveau A1 se subdivisent en A1 et A1-B. Le premier a été passé pendant le premier semestre du niveau, tandis que le deuxième vers la fin du deuxième semestre du niveau. Pour les deux groupes, on regroupe un total de 102 items différents, dont 3% ont été modifiés, et 8% éliminés. A savoir, la plupart des items éliminés (6 items) correspondent aux items de la CO.

Nombre de candidats effectifs = 106

Moyenne du P = 59,6



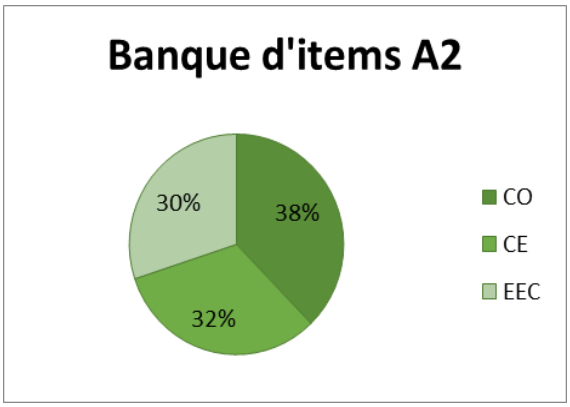


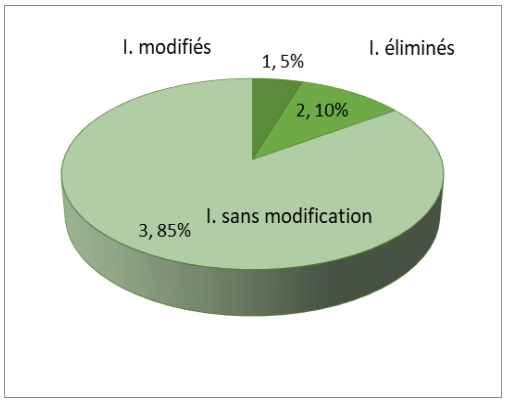
Bilan du niveau A2

Les pilotages du niveau A2 se subdivisent en A2-A, A2-B et A2-C. Pour les 3 groupes, on regroupe un total de 143 items différents, dont 5% ont été modifiés et 10% éliminés.

Nombre de candidats effectifs = 179

Moyenne du P = 55,19



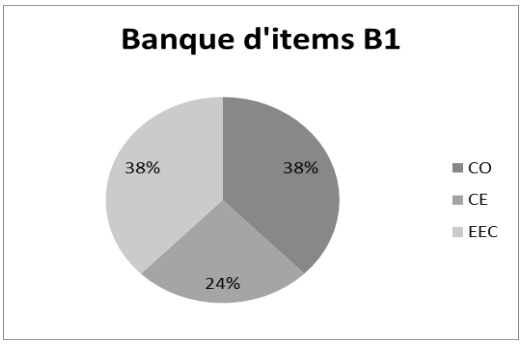


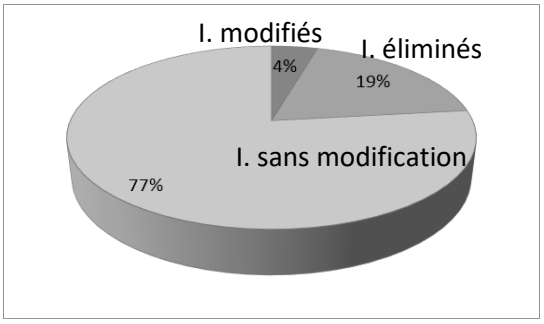
Bilan du niveau B1

Pour le niveau B1, on compte un total de 69 items différents, dont 4% ont été modifiés, et 19% éliminés.

Nombre de candidats effectifs = 69

Moyenne du P = 54,8



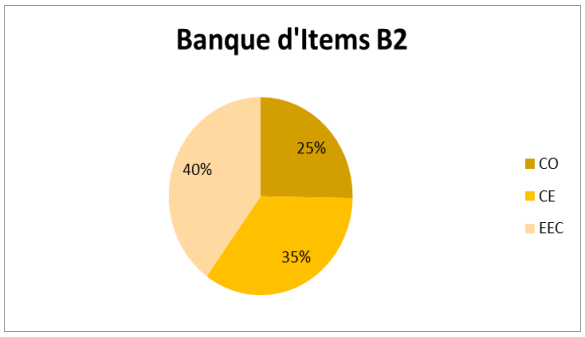


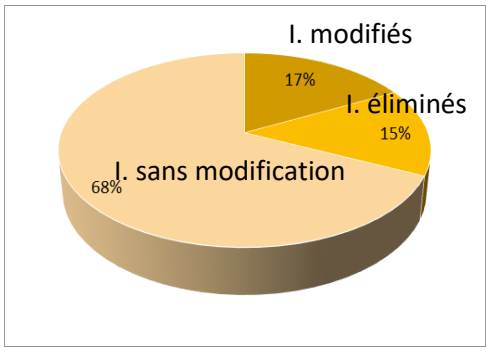
Bilan du niveau B2

Pour le niveau B2, on regroupe un total de 75 items différents, dont 17% ont été modifiés, et 15% éliminés.

Nombre de candidats effectifs = 49

Moyenne du P = 42,2





DÉTAILS DE CHAQUE NIVEAU

A1			
54 candidats effectifs			
54,41 moyenne du P			
	ITEMS	MODIFIES	ELIMINES
CO	17	0	5
CE	27	1	1
EEC	29	2	1
TOTAL	73	3	7
	%	4%	10%

A1-B			
52 candidats effectifs			
64,79 moyenne du P			
	ITEMS	MODIFIES	ELIMINES
CO	15	0	1
CE	8	0	0
EEC	6	0	0
TOTAL	29	0	1
	%	0	3%

A2-A			
50 candidats effectifs			
52,49 moyenne du P			
	ITEMS	MODIFIES	ELIMINES
CO	14	0	5
CE	20	1	1
EEC	20	1	1
TOTAL	**54**	**2**	**7**
	%	**4%**	**13%**

A2-B			
65 candidats effectifs			
60,75 moyenne du P			
	ITEMS	MODIFIES	ELIMINES
CO	17	0	1
CE	18	3	1
EEC	14	1	1
TOTAL	**49**	**4**	**3**
	%	8%	6%

A2-C			
64 candidats effectifs			
52,34 moyenne du P			
	ITEMS	MODIFIES	ELIMINES
CO	23	0	3
CE	8	0	1
EEC	9	1	1
TOTAL	**40**	**1**	**5**
%		**3%**	**13%**

B1			
69 candidats effectifs			
54,8 moyenne du P			
	ITEMS	MODIFIES	ELIMINES
CO	26	0	4
CE	17	0	5
EEC	26	3	4
TOTAL	**69**	**3**	**13**
%		**4%**	**19%**

B2			
49 candidats effectifs			
52,2 moyenne du P			
	ITEMS	MODIFIES	ELIMINES
CO	19	5	2
CE	26	5	5
EEC	30	3	4
TOTAL	**75**	**13**	**11**
%		**17%**	**15%**

Annexe 2
VIDÉO : SELF Espagnol

1. INTRODUCTION

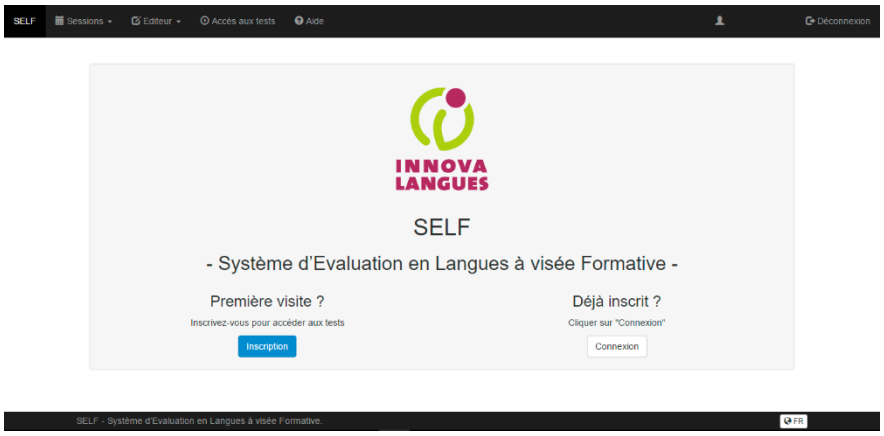
STORYBOARD 1

Bienvenue dans SELF, votre Système d'Evaluation en Langues à visée formative !

Vous allez être évalués en 3 habiletés :

- Compréhension de l'Oral
- Compréhension de l'Ecrit
- Expression Ecrite Courte

Chaque habileté est évaluée par une série de tâches auxquelles vous pouvez répondre sans contrainte de temps.
En moyenne, ce test a une durée de 50 minutes.



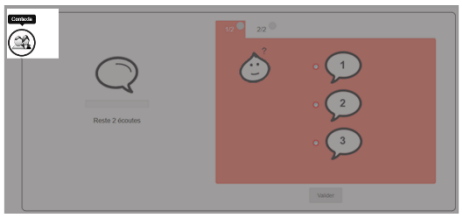
2. INTERFACE - tâche type

STORYBOARD 2

Voici un exemple d'une tâche d'évaluation.

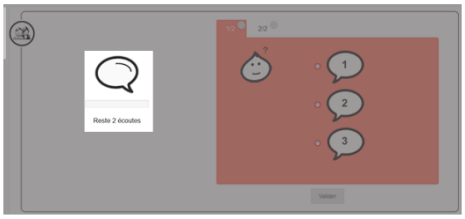
STORYBOARD 2.1

En haut à gauche, vous trouvez le contexte, c'est-à-dire une information qui vous permet de vous caler dans la situation d'écoute, de lecture, ou de rédaction proposée.



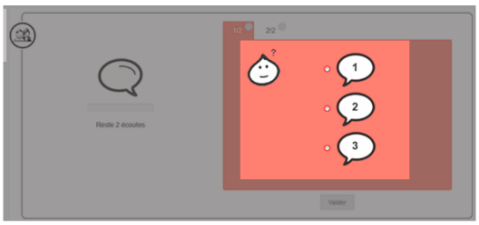
STORYBOARD 2.2

En dessous du contexte, vous trouvez l'objet de la question, c'est-à-dire un texte oral ou écrit sur lequel porte l'effort de compréhension, et la ou les questions proposées.



STORYBOARD 2.3

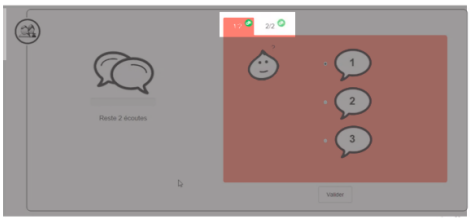
Et sur la droite, la ou les questions et les propositions de réponse.



3. MODALITES D'INTERACTION, réponse-valider

STORYBOARD 3

Vous devez répondre à plusieurs questions. Ces questions se trouvent sur différents onglets.



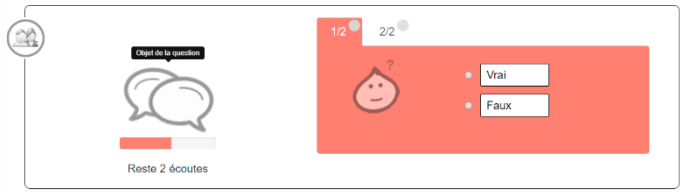
STORYBOARD 3.1

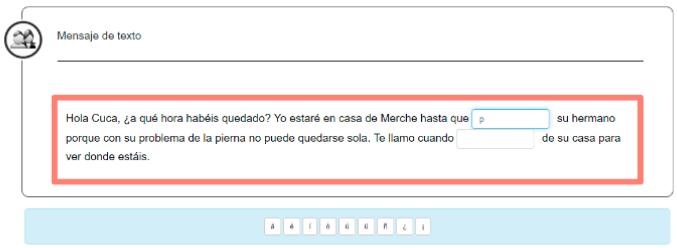
Vous ne pouvez valider votre, ou vos réponse(s), qu'après avoir répondu à toutes les questions.

4. LES TYPES DE TÂCHES

STORYBOARD 4

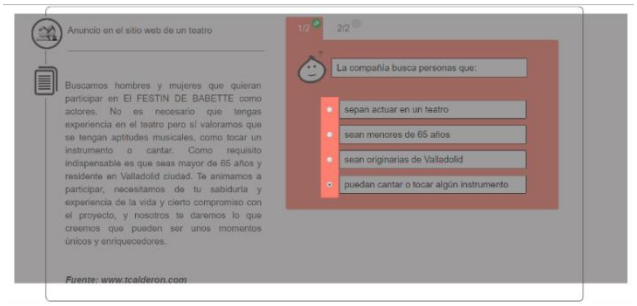
Plusieurs types d'exercices sont proposés : des vrai/faux, des textes lacunaires, des questionnaires à choix multiples.





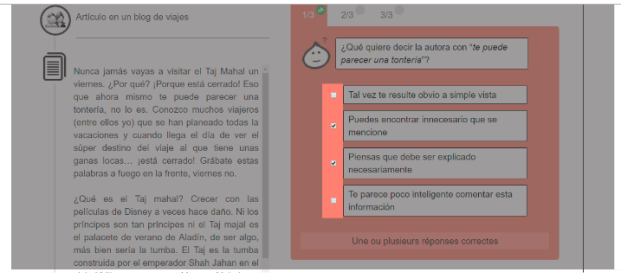
STORYBOARD 4.1

Les questions à choix multiples peuvent admettre soit une seule réponse correcte,



STORYBOARD 4.2

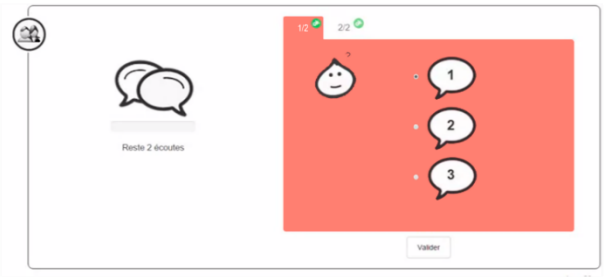
soit plusieurs réponses correctes.



5. Tâches CO

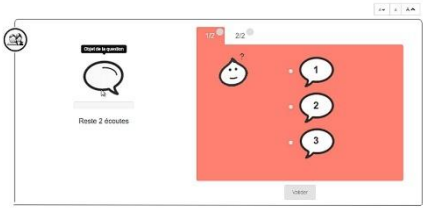
STORYBOARD 5

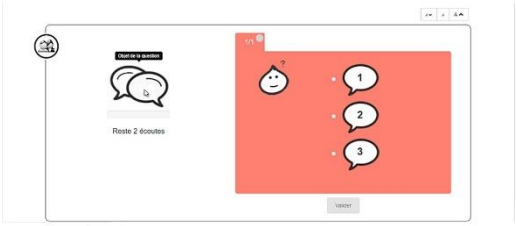
Dans les tâches de compréhension de l'oral, tous les éléments sont oralisés, donc à écouter.

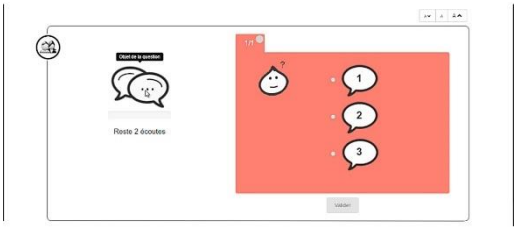


STORYBOARD 5.1

L'objet de la question peut se présenter sous la forme d'un monologue, d'un dialogue, ou d'un dialogue interrompu (audio ou vidéo).

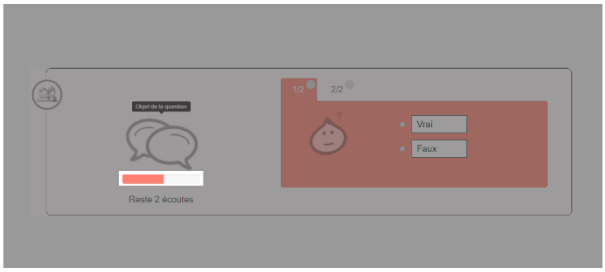






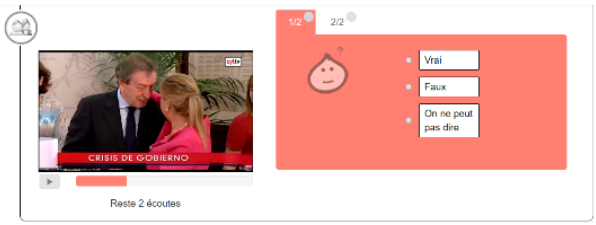
STORYBOARD 5.2

Attention : il ne peut être écouté qu'une ou deux fois. Le nombre d'écoutes restantes est indiqué sous l'objet de la question.



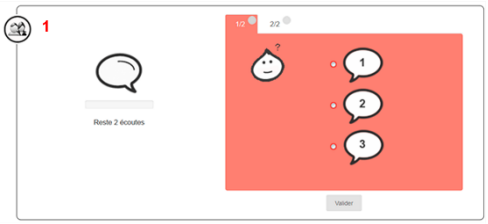
STORYBOARD 5.3

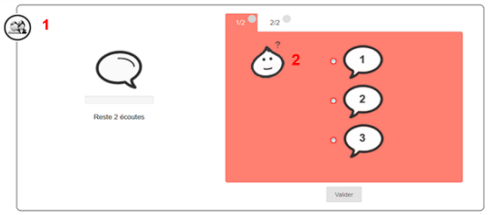
Une fois démarrée la lecture d'une vidéo, ou l'écoute d'un audio, vous ne pourrez pas mettre en pause, ni revenir en arrière. Soyez donc bien concentré.

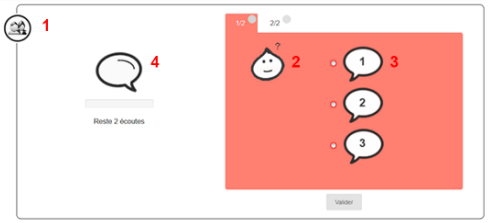


STORYBOARD 5.4

C'est pourquoi nous vous suggérons d'écouter d'abord le contexte, puis la question, les propositions de réponse, et enfin l'objet de la question.







6. Tâches CE

STORYBOARD 6.1

Les tâches proposées en compréhension de l'écrit présentent à l'écran les mêmes éléments que celles de compréhension de l'oral : un contexte, l'objet de la question, des questions, et des propositions de réponse.

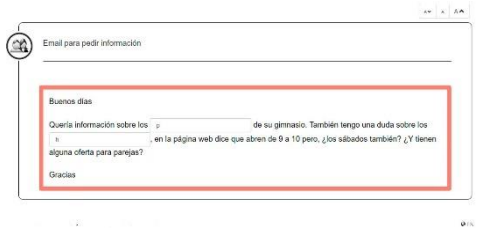
STORYBOARD 6.2

L'objet de la question peut aussi se présenter sous forme d'image. Vous pouvez cliquer dessus pour l'agrandir.

7. Tâches EEC

STORYBOARD 7.1

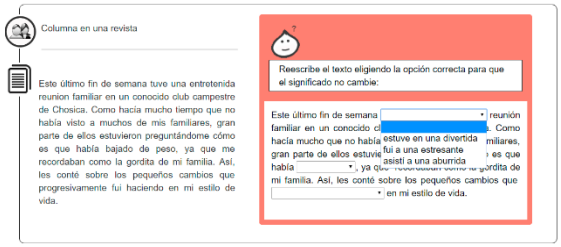
En expression écrite courte, des textes lacunaires vous sont proposés.

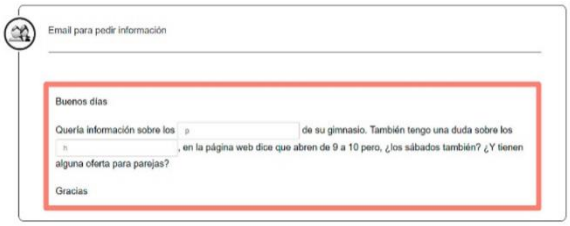


STORYBOARD 7.2 et 7.3

Vous saisissez vos réponses :

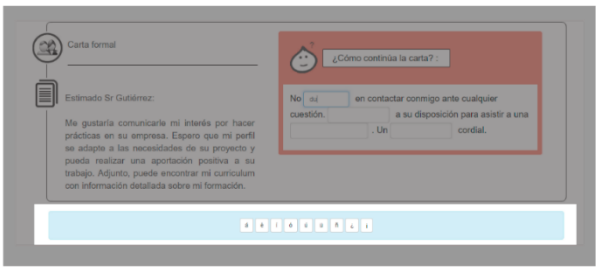
- soit en sélectionnant votre réponse à partir d'un menu déroulant.
- soit directement, en vous servant du clavier.





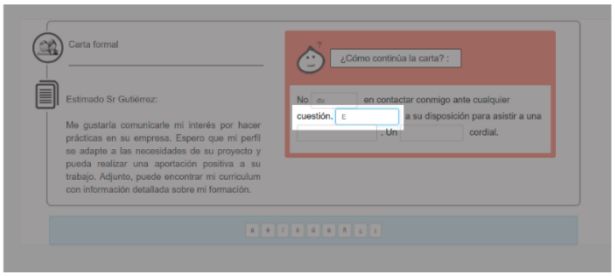
STORYBOARD 7.4

Les caractères spéciaux peuvent être saisis en cliquant directement sur les caractères dont le menu s'affiche en bas de l'écran.



STORYBOARD 7.5

Faites aussi attention aux signes de ponctuation. Ils peuvent vous donner des pistes. Par exemple, si vous trouvez un point, il faut commencer avec une majuscule.

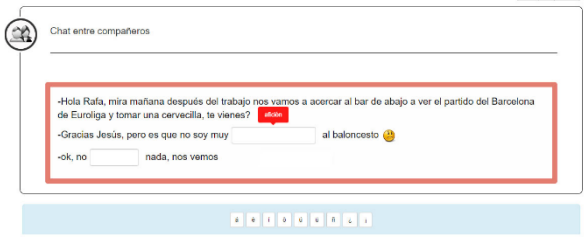


STORYBOARD 7.6

Dans les textes lacunaires, des indices graphiques sont parfois donnés, comme la première lettre, ou la première syllabe du mot.

STORYBOARD 7.7

Vous trouverez parfois des info-bulles qui apparaissent par un survol de la zone à compléter.
Soyez rassurés : si vous ne trouvez pas d'indice graphique, ni d'indication dans les info-bulles, il est prévu que vous puissiez vous en passer.

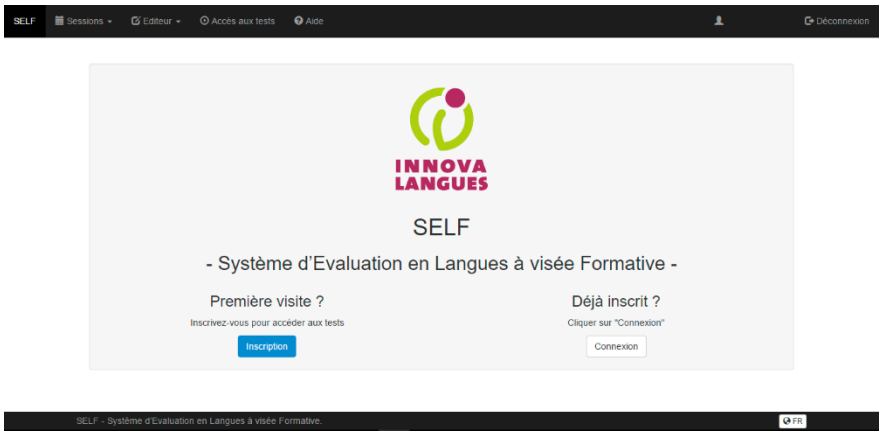


8. Fin

STORYBOARD 8

Et maintenant, concentrez-vous !

Bon courage et bonne chance ! ☺



Logiciels utilisés :

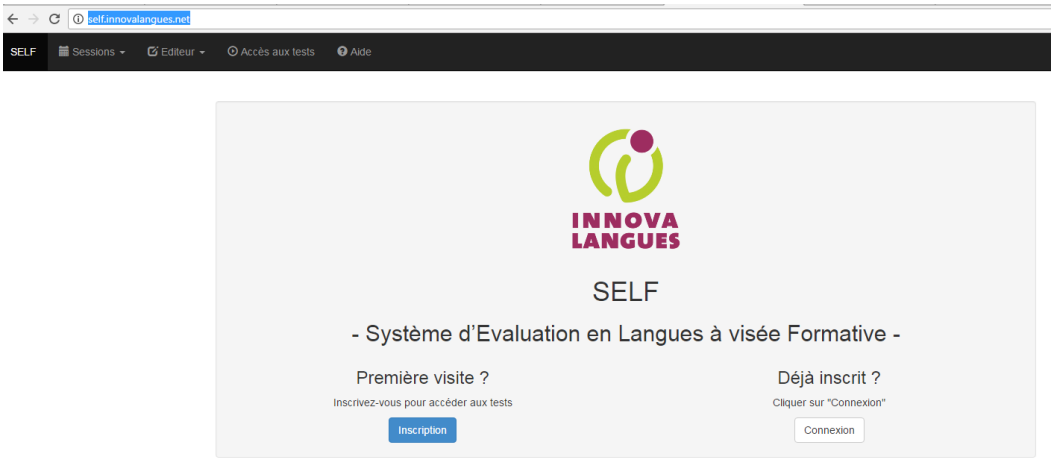
Camtasia Studio (Capture d'écran et montage)
Adobe Photoshop (Traitement d'images)
Sony sound forge (Traitement d'audio)

Voix :

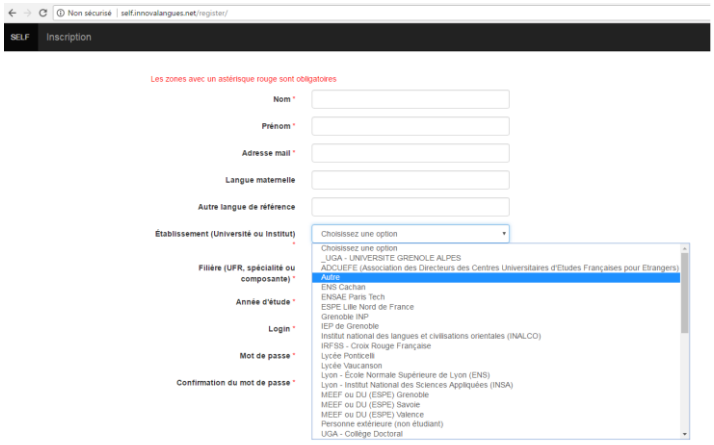
Laura MAZZARELLA

Annexe 3
Guide de passation du test SELF en Espagnol

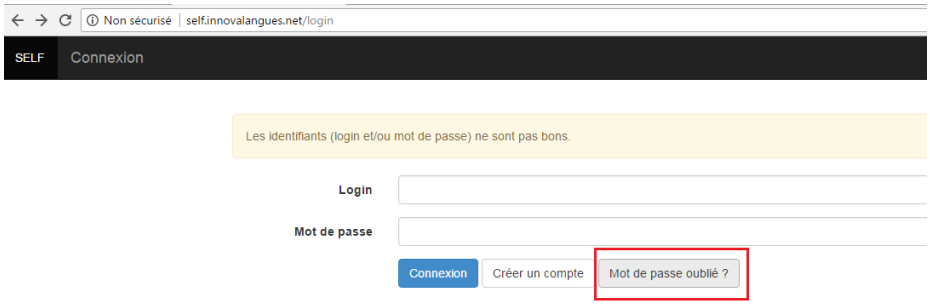
Le test que vous êtes en train de passer est un test de positionnement qui évalue la Compréhension de l'Oral, la Compréhension de l'Ecrit, et l'Expression Ecrite Courte.
Pour passer le test, vous devrez aller sur le lien suivant : http://self.innovalangues.net/ en utilisant les navigateurs **Mozilla ou Chrome**. Veillez à ne pas utiliser Internet Explorer, ou Zafari (Mac).
Le système affichera deux choix possibles :



1. Sélectionnez « première visite » si vous n'avez jamais passé ce test dans une autre langue (anglais, italien, japonais, mandarin, espagnol, ou français). Puis, remplissez le formulaire. (Si votre université ne figure pas dans la liste du choix « option établissement », choisissez « autre »).

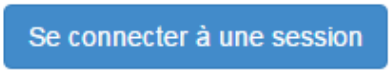


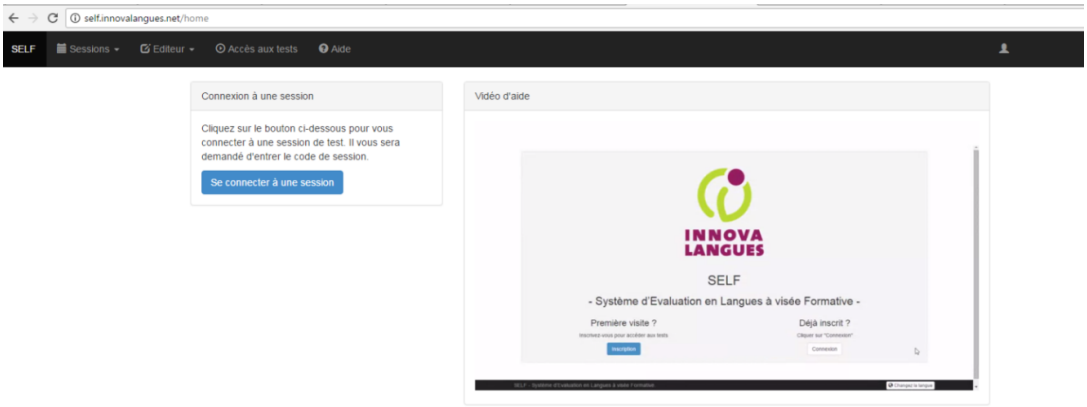
2. Sélectionnez « déjà inscrit » si vous avez déjà passé ce test dans une autre langue. Votre écran ressemblera à ceci. Si vous avez oublié votre identifiant ou mot de passe, cliquez sur « mot de passe oublié ». Vous recevrez un mail pour réinitialiser votre mot de passe. Attention, le mail peut être reçu sur votre boîte SPAM.



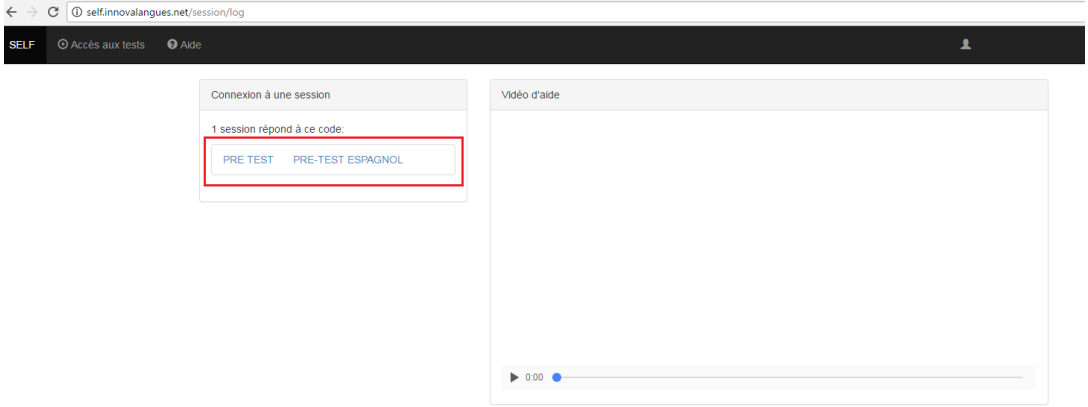
Il est très important de regarder la vidéo contenant les consignes.

Une fois que la vidéo est finie, cliquez sur le bouton « se connecter à une session » en haut à gauche. Puis, vous devez insérer le code fourni, soit par mail, soit par votre professeur(e).



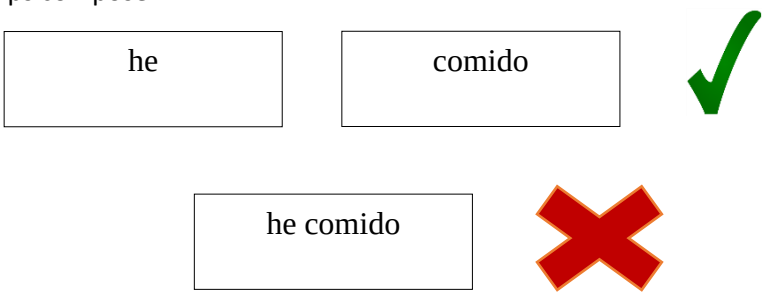


Puis, cliquez sur l'option en haut à gauche pour commencer.

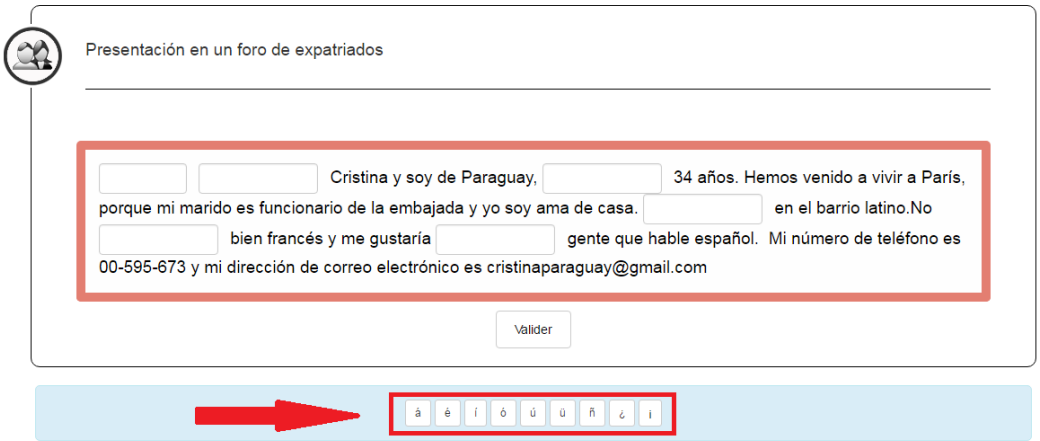


Attention, certaines particularités ne sont pas spécifiées dans cette vidéo :

- L'évaluation de l'expression écrite courte consiste à remplir des trous avec **un seul** mot. Si vous trouvez deux trous consécutifs, cela veut dire qu'il faut taper **un seul** mot par trou. Vous pouvez, par exemple, trouver cette situation quand il s'agit d'un temps composé.



- Toujours sur l'expression écrite courte, la longueur du trou ne correspond pas à la longueur du mot à saisir.
- Pour les caractères spéciaux comme les accents, vous pouvez les insérer soit avec les codes du clavier, soit avec une barre proposée en dessous de la tâche d'évaluation.



DES CONSEILS POUR UNE PASSATION OPTIMALE :

- Si le système affiche « erreur », il suffit de cliquer sur la flèche « revenir en arrière du navigateur », puis appuyez sur la touche F5 du clavier.

Annexe 4

Guide de passation du test SELF Interlangue

Test de Positionnement en langues

http://self.innovalangues.net/

Connexion à SELF

Quelques instructions importantes avant de commencer :

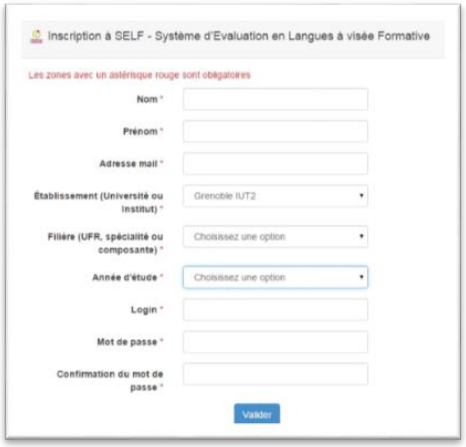
- Utilisez un ordinateur (Windows, Mac ou Linux), **n'utilisez pas** de tablette ou de téléphone.
- Utilisez une version récente de **Mozilla Firefox**, SELF n'est pas adapté à Chrome et IE.
- Contrôlez que les **paramètres audio-vidéo** sont bien réglés sur votre ordinateur.

- Si c'est la première fois que vous rentrez dans le système SELF, cliquez sur « **Inscription** », en dessous de « Première visite ».

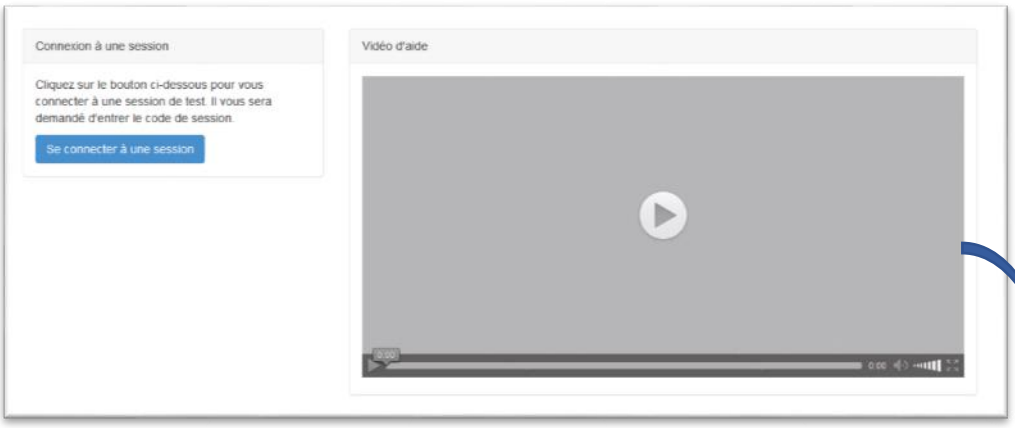
- Si besoin, cliquez sur « mot de passe oublié » dans « déjà inscrit » pour que SELF vous envoie un email à l'adresse que vous lui avez donnée lors de votre dernière inscription (vérifiez vos SPAM).

- Remplissez le formulaire d'inscription.

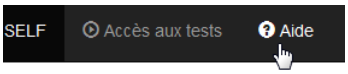
  Login par défaut : ***nom.prenom*** *(modifiable).*
  Indiquez bien votre établissement, filière et année d'études.
  Notez bien vos identifiant et mot de passe !
  (SELF ne vous envoie pas d'e-mail de confirmation).



Vidéo d'aide & connexion à une session



- Regardez attentivement la vidéo d'aide. Cette vidéo reste visionnable à tout moment, en cliquant sur le bouton **« aide »** du menu.



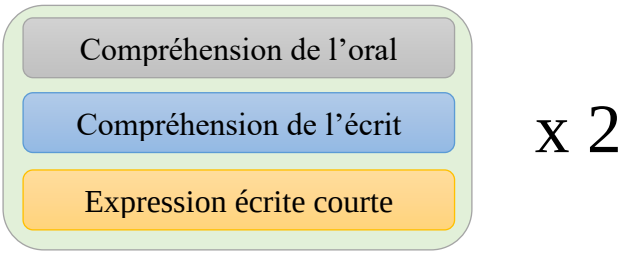
- Connectez-vous à une session de test avec **le mot de passe de session** qui vous a été fourni.

Le test

La durée moyenne est de **50 minutes.**

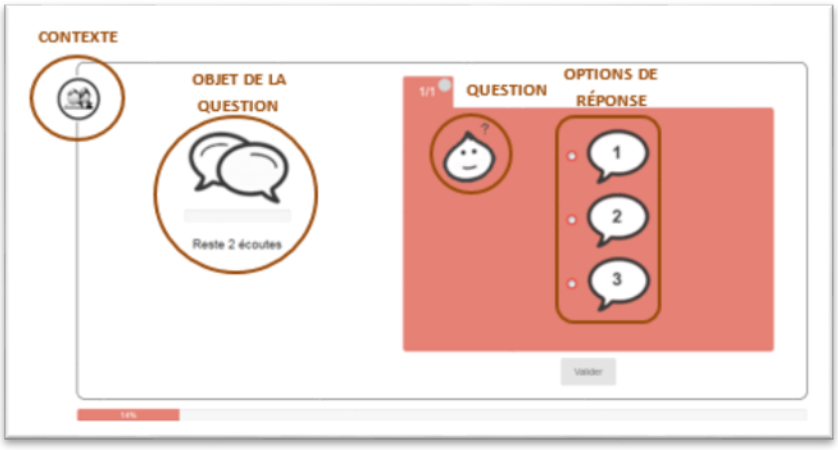
- Le test se compose de 2 parties, chacune composée de 3 compétences.
- La première partie est **générale**, la deuxième est **adaptée à votre niveau**.
- La complexité va croissante.
- Tout va s'enchaîner automatiquement.

Composition du test :



Rappels

- Cliquez sur les icônes pour écouter l'audio



- Il y a parfois plusieurs questions !

- Il faut répondre à toutes les questions, et vous ne pouvez pas revenir en arrière.

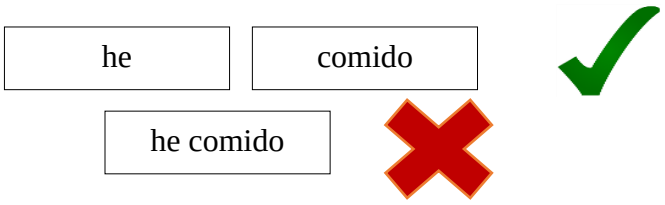
Résultats

- À la fin du test, vous pouvez télécharger vos résultats au format pdf.

- Pas de panique en cas de déconnexion, votre progression est sauvegardée, et sera restaurée quand vous vous reconnecterez avec les mêmes identifiants.

SELF Espagnol

- L'évaluation de l'expression écrite courte consiste à remplir des trous avec **un seul** mot. Si vous trouvez deux trous consécutifs, cela veut dire qu'il faut taper **un seul** mot par trou. Vous pouvez, par exemple, trouver cette situation quand il s'agit d'un temps composé.



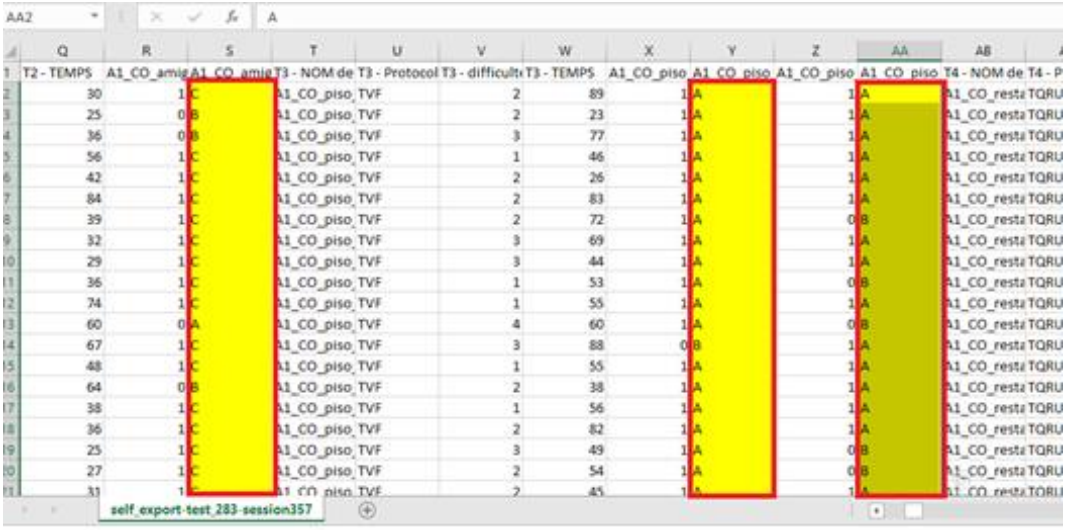
- La longueur du trou ne correspond pas à la longueur du mot à saisir.
- La réponse sera incorrecte si elle n'est pas **correctement accentuée** ! Vous trouverez une barre de caractères spéciaux quand cela est nécessaire.

SELF japonais

S'il y a des tâches où vous devez écrire en japonais, utilisez le clavier IME. Vous pouvez consulter une aide à tout moment du test, vous expliquant la démarche à suivre (Windows et Mac).

Annexe 5
Fichier export

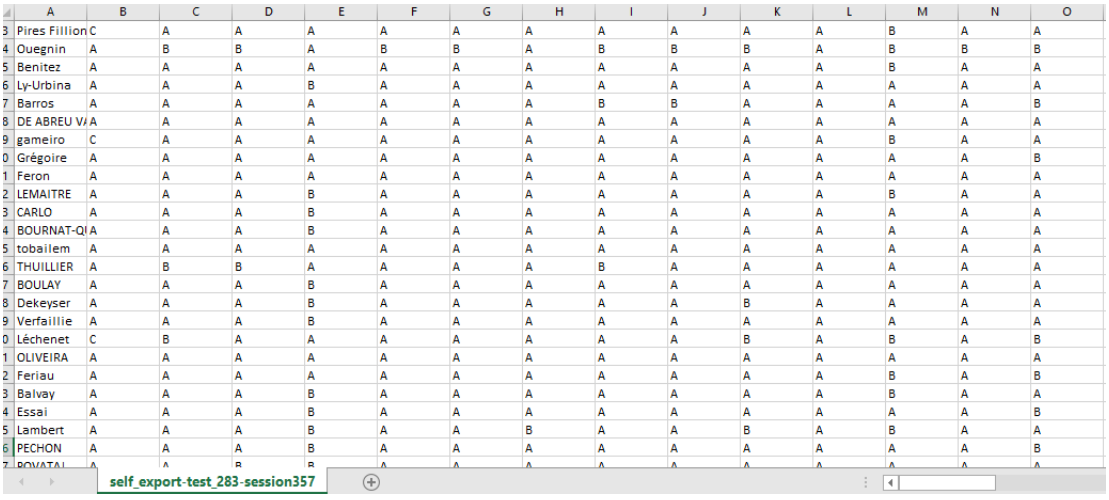
Fichier export sans traitement : CO et CE



	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	T2 - TEMPS	A1_CO_amig	A1_CO_amig	T3 - NOM de	T3 - Protocol	T3 - difficult	T3 - TEMPS	A1_CO_piso	A1_CO_piso	A1_CO_piso	A1_CO_piso	T4 - NOM de
2	30	1	C	A1_CO_piso	TVF	2	89	1	A	1	A	A1_CO_resta
3	25	0	B	A1_CO_piso	TVF	2	23	1	A	1	A	A1_CO_resta
4	36	0	B	A1_CO_piso	TVF	3	77	1	A	1	A	A1_CO_resta
5	56	1	C	A1_CO_piso	TVF	1	46	1	A	1	A	A1_CO_resta
6	42	1	C	A1_CO_piso	TVF	2	26	1	A	1	A	A1_CO_resta
7	84	1	C	A1_CO_piso	TVF	2	83	1	A	1	A	A1_CO_resta
8	39	1	C	A1_CO_piso	TVF	2	72	1	A	0	B	A1_CO_resta
9	32	1	C	A1_CO_piso	TVF	3	69	1	A	1	A	A1_CO_resta
10	29	1	C	A1_CO_piso	TVF	3	44	1	A	1	A	A1_CO_resta
11	36	1	C	A1_CO_piso	TVF	1	53	1	A	0	B	A1_CO_resta
12	74	1	C	A1_CO_piso	TVF	1	55	1	A	1	A	A1_CO_resta
13	60	0	A	A1_CO_piso	TVF	4	60	1	A	0	B	A1_CO_resta
14	67	1	C	A1_CO_piso	TVF	3	88	0	B	1	A	A1_CO_resta
15	48	1	C	A1_CO_piso	TVF	1	55	1	A	1	A	A1_CO_resta
16	64	0	B	A1_CO_piso	TVF	2	38	1	A	1	A	A1_CO_resta
17	38	1	C	A1_CO_piso	TVF	1	56	1	A	1	A	A1_CO_resta
18	36	1	C	A1_CO_piso	TVF	2	82	1	A	1	A	A1_CO_resta
19	25	1	C	A1_CO_piso	TVF	3	49	1	A	0	B	A1_CO_resta
20	27	1	C	A1_CO_piso	TVF	2	54	1	A	0	B	A1_CO_resta
21	31	1	C	A1_CO_piso	TVF	2	45	1	A	1	A	A1_CO_resta

self_export-test_283-session357

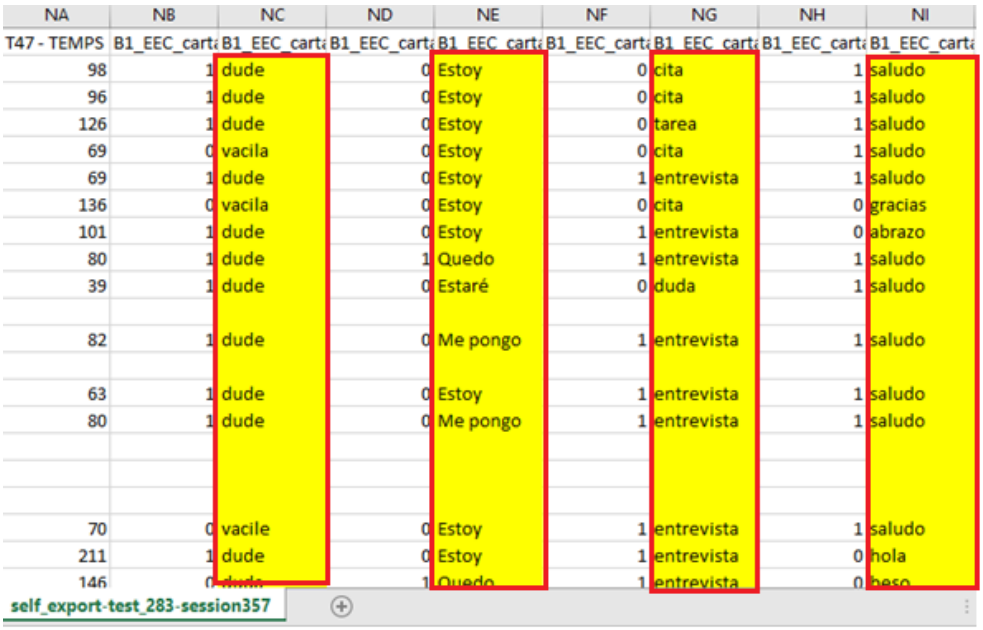
Fichier export sans traitement : EEC



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
3	Pires Fillion	C	A	A	A	A	A	A	A	A	A	A	B	A	A
4	Ouegnin	A	B	B	A	B	B	A	B	B	B	A	B	B	B
5	Benitez	A	A	A	A	A	A	A	A	A	A	A	B	A	A
6	Ly-Urbina	A	A	A	B	A	A	A	A	A	A	A	A	A	A
7	Barros	A	A	A	A	A	A	A	B	B	A	A	A	A	B
8	DE ABREU VA	A	A	A	A	A	A	A	A	A	A	A	A	A	A
9	gameiro	C	A	A	A	A	A	A	A	A	A	A	B	A	A
0	Grégoire	A	A	A	A	A	A	A	A	A	A	A	A	A	B
1	Feron	A	A	A	A	A	A	A	A	A	A	A	A	A	A
2	LEMAITRE	A	A	A	B	A	A	A	A	A	A	A	B	A	A
3	CARLO	A	A	A	B	A	A	A	A	A	A	A	A	A	A
4	BOURNAT-QI	A	A	A	B	A	A	A	A	A	A	A	A	A	A
5	tobailem	A	A	A	A	A	A	A	A	A	A	A	A	A	A
6	THUILLIER	A	B	B	A	A	A	A	B	A	A	A	A	A	A
7	BOULAY	A	A	A	B	A	A	A	A	A	A	A	A	A	A
8	Dekeyser	A	A	A	B	A	A	A	A	A	B	A	A	A	A
9	Verfaillie	A	A	A	B	A	A	A	A	A	A	A	A	A	A
0	Léchenet	C	B	A	A	A	A	A	A	A	B	A	B	A	B
1	OLIVEIRA	A	A	A	A	A	A	A	A	A	A	A	A	A	A
2	Feriau	A	A	A	A	A	A	A	A	A	A	A	B	A	B
3	Balvay	A	A	A	B	A	A	A	A	A	A	A	B	A	A
4	Essai	A	A	A	B	A	A	A	A	A	A	A	A	A	B
5	Lambert	A	A	A	B	A	A	B	A	A	B	A	B	A	A
6	PECHON	A	A	A	B	A	A	A	A	A	A	A	A	A	B
7	ROYATAL	A	A	B	B	A	A	A	A	A	A	A	A	A	A

self_export-test_283-session357

Fichier export traité



NA	NB	NC	ND	NE	NF	NG	NH	NI
T47 - TEMPS	B1_EEC_carta	B1_EEC_carta	B1_EEC_carta	B1_EEC_carta	B1_EEC_carta	B1_EEC_carta	B1_EEC_carta	B1_EEC_carta
98	1	dude	0	Estoy	0	cita	1	saludo
96	1	dude	0	Estoy	0	cita	1	saludo
126	1	dude	0	Estoy	0	tarea	1	saludo
69	0	vacila	0	Estoy	0	cita	1	saludo
69	1	dude	0	Estoy	1	entrevista	1	saludo
136	0	vacila	0	Estoy	0	cita	0	gracias
101	1	dude	0	Estoy	1	entrevista	0	abrazo
80	1	dude	1	Quedo	1	entrevista	1	saludo
39	1	dude	0	Estaré	0	duda	1	saludo
82	1	dude	0	Me pongo	1	entrevista	1	saludo
63	1	dude	0	Estoy	1	entrevista	1	saludo
80	1	dude	0	Me pongo	1	entrevista	1	saludo
70	0	vacile	0	Estoy	1	entrevista	1	saludo
211	1	dude	0	Estoy	1	entrevista	0	hola
146	0	dude	1	Quedo	1	entrevista	0	beso

self_export-test_283-session357

Annexe 6
Exemple de question à réponses multiples

Contexto
Carta al director
Objeto de la pregunta
Ayer me senté expectante ante la televisión a ver la ceremonia de los Premios Pizarro, con la esperanza de que se hiciera alguna alusión a las manifestaciones que han recorrido todo el país en contra de la reforma laboral del Gobierno y se estableciera una complicidad entre el mundo del cine y la sociedad.
Pero no hubo nada así, desafortunadamente. Quiero animarles a que sean conscientes de su responsabilidad en los momentos tan críticos que vivimos y decirles, con todo respeto, que siempre es compatible el agradecimiento a todos sus equipos, amigos y familiares, por el premio recibido, a la vez que denuncien las injusticias que toda la sociedad está padeciendo. Todos debemos alzar la voz contra la injusticia donde estemos…
Item 1
¿Por qué hubo manifestaciones?

Proposition(s) :
1.-Porque hay un cambio en la ley de trabajo* (Clave)
2.- para la convocatoria de nuevas elecciones
3.-Para denunciar las injusticias que sufre la sociedad
4.-Para la aprobación de una nueva ley relacionada con la cultura

Item 2 :
Después la gala, la mujer se sintió :

Proposition(s) :
1.- agobiada
2.-Ilusionada
3.-Avergonzada
4.- Decepcionada* (Clé)

Item 3 :
La persona que escribe :

Proposition(s) :
1.- le disgustó que los premiados no agradeciesen a la sociedad sus premios
2.- le fastidió que los artistas fuesen poco solidarios con los problemas de la gente *(Clé)
3.- le sorprendió que los participantes hablasen sobre injusticias en un acto cultural
4.- le dio rabia que los asistentes a la gala no mencionasen nada sobre las manifestaciones*(Clé)

DÉBUT DE LA TRADUCTION

Contexte
Lettre au responsable

Objet de la question
Hier, j'ai regardé la cérémonie des prix Pizarro à la TV. J'avais espoir que l'on fasse un commentaire sur les manifestations qui ont eu lieu dans le pays contre la réforme du travail du gouvernement, et ainsi créer une solidarité entre le monde du cinéma et la société.

Malheureusement, cela n'a pas été le cas. Je veux vous encourager à prendre conscience de la responsabilité que vous avez au moment où nous traversons des moments si critiques, et vous dire, respectueusement, qu'il y a toujours de l'estime envers vos équipes, amis et famille, grâce au prix que vous avez gagné. Cependant, vous devriez aussi dénoncer les injustices que toute la société traverse. Nous devons tous élever notre voix contre l'injustice partout où nous sommes.

Item 1 :
Pour quelle raison y a-t-il eu des manifestations ?

Proposition(s) :

1.-Parce qu'il y a eu un changement dans la loi travail* (Clé)
2.-Pour appeler à de nouvelles élections
3.-Pour dénoncer les injustices que traverse la société
4.-Pour l'approbation d'une nouvelle loi en lien avec la culture

Item 2 :
Après le gala, la femme était :

Proposition(s) :

1.- battue
2.- rêveuse
3.-embarrassée
4.- déçue* (clé)

Item 3 :
La personne qui écrit :

Proposition (s) :

1.- était mécontente que les gagnants ne remercient pas la société pour ses prix
2.- était agacée que les artistes aient peu de solidarité envers les problèmes du peuple* (clé)
3.- était surprise que les participants aient parlé des injustices dans un évènement culturel
4.- était en colère que les participants au gala n'aient pas mentionné les manifestations *(clé)

FIN DE LA TRADUCTION

Annexe 7
Exemple de tâche : vrai/faux/on ne peut pas dire

Contexto :

Documental

Objeto de la pregunta:

La industria de la seda llenaba de árboles de moreras los montes y hasta cuatro mil producían tejidos que generaban una gran riqueza a los alpujarreños. No quedaron casi moreras, ni tampoco telares, o casi ninguno, pues todavía funciona uno en Bubión al estilo tradicional. Encontró uno, una joya única, el telar, casi de casualidad, tras la jubilación de un artesano granadino. Es de madera de castaño y sus dos siglos de antigüedad lo convierten en el vestigio más remoto de lo que fue siglos atrás la industria de la seda.

Item 1:

> Antiguamente la industria de la seda fue muy abundante en la región de las Alpujarras (VRAI)

Item 2:

> En la actualidad se pueden encontrar escasas moreras en la región (VRAI)

Item 3:

> La mujer heredó el telar que había pertenecido a su familia (FAUX)

DÉBUT DE LA TRADUCTION

Contexte :

Documentaire

Objet de la question :

L'industrie de la soie remplissait les collines de mûriers, et jusqu'à quatre mille métiers à tisser produisaient des tissus, générant une grande richesse pour les *alpujarreños*. Il ne restait plus de mûrier, ni de métier à tisser, ou pratiquement plus, sauf encore un qui fonctionnait à Bubion, de type traditionnel. Elle a trouvé une perle rare, un métier à tisser. C'était presque par hasard, après la retraite d'un artisan de Grenade. Il est fait en bois de châtaignier, et ses deux siècles d'âge font de lui le vestige le plus lointain de ce qui fut autrefois l'industrie de la soie.

Item 1 :

Autrefois, l'industrie de la soie était très abondante dans la région des *Alpujarras* (VRAI)

Item 2 :

À l'heure actuelle, il existe peu de mûriers dans la région (VRAI)

Item 3 :

La femme a hérité du métier à tisser qui avait appartenu à sa famille (FAUX)

FIN DE LA TRADUCTION

Annexe 8
Résultats du candidat boulay-gweldann

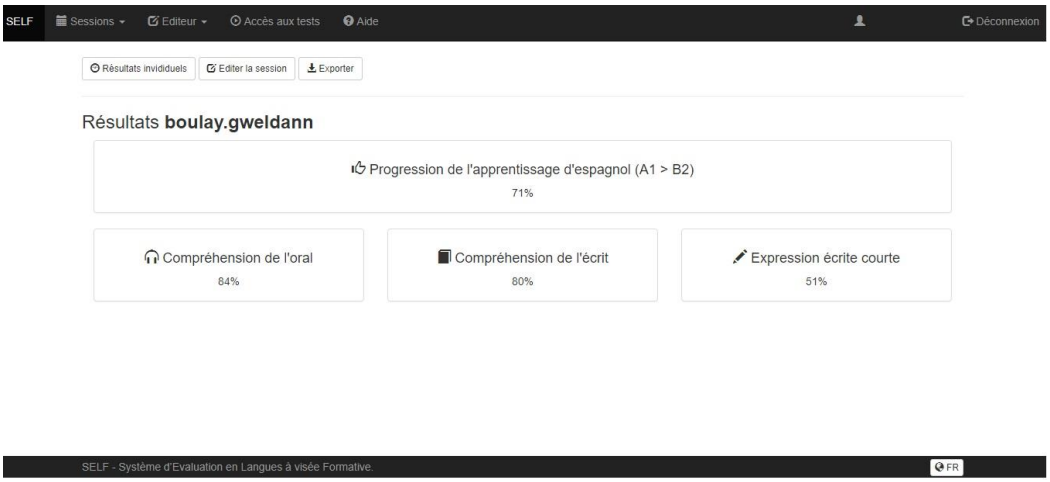


Table des matières

MOTS-CLÉS : SELF, testing, évaluation, validité, fiabilité, théorie classique du test, (TCT), théorie de la réponse à l'item, (TRI), calibrage, analyse psychométrique, psychométrie, construit, item, tâche d'évaluation, TiaPlus, Winsteps.

RÉSUMÉ

Le présent mémoire a été effectué dans le cadre d'un stage de 5 mois au sein de l'IDEFI Innovalangues, plus précisément au sein du processus de validation du test de positionnement SELF en espagnol. Il décrit la méthode de validation faite à partir des analyses psychométriques basées sur la Théorie Classique du Test (TCT), et la Théorie de la Réponse à l'Item (TRI), au moyen des logiciels TiaPlus et Winsteps.