Un test de classement en ligne pour évaluer les niveaux de compétence et constituer des groupes classes

Patricia Kohler[1], Hervé Platteaux[2], Catherine Blons-Pierre[1]

[1] Centre de langues, Université de Fribourg

[2] Centre NTE, Université de Fribourg

Résumé

L'Université de Fribourg développe un test de classement en ligne, avec la plateforme Moodle, pour évaluer les compétences des étudiants en langues étrangères (compréhension orale, compréhension écrite et production écrite) selon le Cadre Européen Commun de Référence du Conseil de l'Europe. Une démarche qualité nous aide à réfléchir sur les modèles de tests de classement en ligne pour que nos résultats soient transférables à d'autres contextes que l'enseignement des langues étrangères. Nous évaluons l'acceptabilité de notre test avec, d'une part, les données quantitatives produites par Moodle lorsque les étudiants passent le test. D'autre part, nous analysons les entretiens d'étudiants décrivant leur expérience avec le test. Certains résultats montrent la pertinence du modèle de test : la durée des sous-tests, l'utilité d'une auto-évaluation et la nécessité de sous-tests sur chaque compétence. D'autres résultats se dégagent sur la représentation qu'a l'usager d'un test en ligne qui remet en question des rapports fondamentaux de l'individu au temps, à l'espace, à l'identité et à ses sensibilités d'ordre culturel.

Mots-clé : Test de classement en ligne, référentiel de compétences, principes de fonctionnement, acceptabilité des utilisateurs

Abstract

The University of Fribourg develops an online placement test, with the Moodle learning platform, to evaluate competencies of students in foreign languages (competencies for listening, reading and writing) according to the European language referential of the Europe

Council. A quality process helps us thinking over the models of an online placement test in order that our results can be transferred to other contexts that foreign language teaching. We evaluate the acceptance of our test with, on one hand, quantitative data that are generated by Moodle when students use the test. On the other hand, we analyse the interviews of students who describe their experience with the test. Results show the validity of the test model: the duration of sub-tests, the utility of an auto-evaluation and the necessity of sub-tests for each competence. Other results appear about the mental representation a user has from an online placement test which change fundamental links of individuals with time, space, identity and cultural sensibilities.

Keywords: Online placement test, Competence referential, working principles, users' acceptance

1. Introduction

Le modèle de test de classement en ligne, pour l'enseignement/apprentissage des langues au Centre de langues de l'Université de Fribourg est axé sur l'enseignement/apprentissage des langues étrangères dans un contexte universitaire. Dans cette contribution, nous considérons les langues étrangères comme une étude de cas et notre expérience des tests de classement en ligne comme un modèle transférable à d'autres disciplines et d'autres contextes.

En amont de la décision de mettre en place un test de classement en ligne, dans n'importe quelle discipline et dans n'importe quel contexte universitaire, se pose la question « pourquoi ? » avant la question « comment ? ».

Dans notre cas, la situation initiale était la suivante :

- il existait un centre d'enseignement et de recherche pour les langues avec quatre unités d'enseignement pour quatre langues : allemand, anglais, français et italien ;
- des tests de classement sur support papier dont l'objectif était de placer les étudiants dans la classe la plus appropriée étaient organisés dans la première semaine du semestre universitaire selon des modalités différentes dans chaque unité ;
- le nombre total d'étudiants intéressés par les cours de langues dans notre université dépasse le millier, ce qui impliquait, avec un support papier, un temps de correction relativement long et grandement consommateur de ressources humaines.

L'un des objectifs premiers de la mise en place de ces tests de classement en ligne est d'unifier les modèles de tests. Les langues conservent leur spécificité dans l'enseignement/apprentissage mais les compétences évaluées sont globalement identiques et elles sont testées à travers des tâches communicatives permettant d'évaluer les compétences des apprenants en compréhension orale, compréhension écrite, production écrite et structures de la langue, selon les descripteurs du Cadre Européen Commun de Référence (ci-après, CECR) pour les langues du Conseil de l'Europe (2001).

Un deuxième objectif est de réduire le temps d'évaluation du niveau de l'apprenant. En effet, avec le test de classement en ligne, le feedback peut être quasi immédiat si l'on se contente d'une évaluation automatique, qui peut être mise en œuvre pour la compréhension orale, la compréhension écrite et les structures de la langue. Une évaluation de la production écrite suppose, au contraire, une évaluation en différé, qui ne doit pas excéder 48 heures, et débouche sur un feedback du niveau global.

Un troisième objectif est de parvenir à une automatisation de l'inscription des étudiants et à une gestion individuelle des inscriptions en reliant la plateforme servant à la passation du test (Moodle) à celle permettant la gestion électronique des inscriptions.

Une première mouture de test de classement en ligne a été proposée à certains étudiants en septembre 2008 et une seconde a été ouverte à tous les étudiants en février 2009 pour le français langue étrangère. C'est cette phase expérimentale qui nous permet de présenter, ici, le prototype de nos tests de classement en ligne, les analyses que nous avons pu en dégager et les évolutions que nous nous proposons de développer.

Partant des objectifs globaux définis ci-dessus, le présent article décrit et analyse les résultats relatifs aux conditions de test dans lesquelles ont été placés les étudiants. On y privilégie le point de vue des étudiants, principaux utilisateurs du test de classement en ligne. Le but est d'évaluer l'acceptabilité des étudiants vis-à-vis de ce dispositif automatique. Nous utilisons le terme acceptabilité au sens où l'entendent Tricot et ses collègues (2003), c'est-à-dire l'utilisation effective faite du dispositif mais aussi la façon dont est ressentie cette automatisation à distance. Nous évaluons cette acceptabilité du point de vue quantitatif, à l'aide des résultats et durées des sous-tests, dont est constitué l'ensemble du test de classement, et du cheminement global et du point de vue qualitatif à l'aide des feedbacks des étudiants après le test.

2. Le test de classement en ligne

Avant de décrire le fonctionnement de ce type de test, il convient de définir clairement les notions de test de classement en ligne, les types d'évaluation auxquels il se rapporte et les publics et institutions concernés.

Beaucoup de termes relatifs à l'évaluation proviennent de la littérature anglo-saxonne. Une fois traduits en français, ils subissent fréquemment des évolutions, notamment des spécialisations, qui nécessitent des mises à jour régulières de la terminologie utilisée. Dans la suite du texte, nous utilisons le terme *test de classement* comme synonyme de *placement test* en anglais et *Einstufungstest* en allemand.

2.1. Le test de classement, une proposition de définition

Le mot *test* apparaît dans la langue française à la fin du XIXème siècle et est défini par Pichot comme une « situation standardisée servant de stimulus à un comportement qui est évalué par la comparaison avec celui d'individus placés dans la même situation, afin de classer le sujet, soit quantitativement, soit typologiquement » (Pichot (1954) cité dans De Landsheere, 1979, p. 295). De Landsheere ajoute : « pour mériter le nom de test, un examen doit être standardisé, fidèle, valide et étalonné » (1979, p. 295). Le test sert à évaluer le niveau de connaissance ou de compétence acquise par un individu dans un domaine donné.

Un parcours d'apprentissage et son évaluation peuvent se découper en trois grands moments (Tourneur & Vasamillet, 1982), numérotés ci-dessous selon leur chronologie habituelle :

- M1 : accueil et placement des apprenants en début d'apprentissage ;
- M2 : déroulement de la séquence d'apprentissage orientée par des objectifs et un projet avec des évaluations intermédiaires ;
- M3 : évaluation sommative finale.

La fonction la plus connue et la plus répandue de l'évaluation est, sans conteste, la fonction sommative, qui intervient à la fin d'une séquence d'apprentissage plus ou moins longue et qui a pour objectif de vérifier une compétence pour sélectionner les apprenants, voire les classer. L'évaluation sommative peut déboucher sur la délivrance de diplômes et de certifications. On peut considérer que le Test de Connaissance du Français (TCF) et le Test d'Evaluation de Français (TEF) relèvent de ce type d'évaluation dans la mesure où ils ont pour objectif la

délivrance d'une certification de niveau international comportant une mesure quantitative individuelle des compétences testées. Des versions électroniques de ces deux tests, attestant d'un niveau de français reconnu valable sur une durée de deux ans (TCF) et d'un an (TEF), existent ; elles s'adressent avant tout au public des apprenants chinois.

La fonction formative de l'évaluation se différencie de la fonction sommative en ce qu'elle relie le processus de l'évaluation avec celui de l'enseignement/apprentissage pour permettre une régulation de celui-ci. Elle ne se limite donc pas à reconnaître ou sanctionner les acquis ou les manques à la fin d'une séquence d'enseignement/apprentissage.

Ce type d'évaluation a principalement pour objectifs :

- d'identifier les points forts et les difficultés de l'apprenant ;
- d'optimiser les stratégies d'apprentissage ;
- de permettre au formateur d'analyser les procédures d'enseignement d'après un retour sur les objectifs atteints et non atteints.

Cette fonction formative, qui peut être proactive, interactive et rétroactive (Allal, 1988) débouche sur l'autoévaluation et sur une troisième fonction de l'évaluation : la fonction diagnostique.

Ce dernier type d'évaluation a pour objectif, dans un premier temps, d'établir un constat des connaissances de l'apprenant pour aller ensuite vers la construction autonome ou guidée d'un parcours d'apprentissage. L'évaluation diagnostique remplit plusieurs fonctions :

- elle permet de faire le point sur les connaissances acquises de l'apprenant ;
- elle fournit des indications pronostiques sur les progrès possibles et les orientations probables de l'apprenant (Veltcheff & Hilton, 2003) ;
- elle permet d'opérer des classements à visée formative, par exemple, pour répartir les apprenants en groupes opérationnels.

Notre test de classement en ligne participe de ce type d'évaluation puisque son objectif premier est de déterminer le niveau de compétence de chaque apprenant afin de l'orienter en lui proposant des cours correspondant à un niveau N déterminé à un moment M de son parcours d'apprentissage. Le test a également pour objectif, au regard de l'institution qui le gère, de créer des classes parmi les apprenants, constituant des groupes de niveaux

relativement homogènes. Le test de classement en ligne présenté dans cet article est donc conçu initialement pour le moment M1 du parcours d'enseignement/apprentissage.

2.2. Un test en ligne, modalité de notre test de classement

Il existe plusieurs modalités de passation de tests : tests écrits avec support papier, entretiens oraux, tests en ligne, etc. Le test en ligne est une modalité qui utilise le support informatique en dématérialisant les supports papier et audio du test. Il peut s'agir d'une simple digitalisation d'un test papier adapté aux exigences de l'outil informatique. Mais, le plus souvent, il s'agit d'items (éléments unitaires dont la réunion forme l'ensemble du test) spécialement conçus pour ce type d'outil et d'environnement.

Le choix de cette modalité de passation électronique n'est pas étranger au nombre de candidats. Se profile ainsi une autre caractéristique du test en ligne : son économie dans le cadre d'un enseignement/apprentissage en grands groupes, ce qui est le cas le plus fréquent lorsque l'on aborde l'enseignement dans un cadre universitaire. Nous choisissons les outils de la plateforme d'apprentissage Moodle pour le réaliser.

La plupart des tests de classement en ligne participe de la fonction formative de l'évaluation.

2.3. Différents modèles pour un test de classement

Un premier résultat transférable à d'autres contextes est la structure du test de classement. Les modèles que les enseignants du Centre de langues ont construits peuvent certainement inspirer d'autres praticiens et nous explicitons ici leurs logiques.

Dans notre optique, l'objectif du test de classement est d'évaluer le niveau des compétences d'un étudiant afin de placer celui-ci dans un cours adapté au niveau évalué. La structure globale d'un tel test de classement se construit donc autour des compétences à évaluer. Dans notre cas, il y a en a quatre : 1) la compréhension orale (CO), 2) la compréhension écrite CE), 3) la production écrite (PE) et 4) le lexique et les structures de la langue (LS). Plusieurs niveaux, définis dans chaque compétence, articulent une sous-structure. Dans notre cas, il y en a six : A1, A2, B1, B2, C1 et C2. Ces niveaux correspondent aux niveaux du CECR et vont du niveau A1 débutant au niveau très avancé C2 (Conseil de l'Europe, 2001).

Sur cette base commune, le test de classement proposé aux étudiants peut utiliser différents modèles. Notons qu'un test de classement, servant à évaluer un niveau de langue est constitué de plusieurs sous-tests, dédiés à une évaluation particulière. Dans ce qui suit, pour plus de

clarté, le terme *test* désigne le test dans son ensemble, parfois dénommé aussi le dispositif. Le terme *sous-test* désigne les tests dont est constitué le test.

Un premier modèle fait débuter le test par l'évaluation des compétences en compréhension orale, ce qui permet d'évaluer les compétences langagières de l'étudiant à un premier ordre de grandeur : A, B ou C. Ensuite, l'évaluation de la compréhension écrite donne accès à un second ordre de grandeur : A1 ou A2, par exemple. La même opération a lieu pour les niveaux B et C.

Ainsi, lorsqu'un étudiant commence l'évaluation, il entre dans un premier sous-test de compréhension orale. S'il ne dépasse pas le niveau de seuil fixé, il est d'un niveau global A et est dirigé vers un sous-test de lecture qui détermine s'il est de niveau A1 ou A2. Mais, s'il dépasse le niveau de seuil du premier sous-test oral, il est d'un niveau global B et passe un second sous-test oral. Ne pas dépasser le seuil de celui-ci, confirme le niveau B de l'étudiant et un sous-test de lecture le place au niveau B1 ou B2. Si, au contraire, il dépasse le niveau de seuil du second sous-test oral, un dernier sous-test de lecture confirme un niveau global C à l'étudiant qui obtient une note supérieure ou égale à 75% et un retour à B2 dans le cas contraire. Comme nous le voyons, la conception d'un test de placement repose en particulier sur des niveaux de seuil dans les sous-tests afin de permettre un cheminement similaire à celui que nous venons de décrire. Des résultats récents du Conseil de l'Europe (North & Jones, 2009) indiquent 66% comme une bonne base de départ pour des scores de césure. Une mise à l'essai des sous-tests doit cependant les calibrer plus précisément.

Un second modèle analysé correspond à la digitalisation d'un ancien test progressif sur papier. Il fonctionne comme un test de puissance. Un sous-test est toujours associé à une compétence mais les questions qui sont posées correspondent à différents niveaux de cette compétence. En répondant à un tel sous-test, l'étudiant glane les points des questions auxquelles il sait répondre mais est empêché d'augmenter son total de points face aux questions d'un niveau supérieur au sien. Le nombre de points obtenus ainsi reflète le niveau de ses compétences. Si on considère un ensemble d'étudiants, ils se répartiront de cette façon en groupes de niveaux selon leurs notes. Au cours de l'évaluation, deux compétences sont testées successivement : la compréhension écrite et la compréhension orale. L'évaluation automatisée avec les outils de la plateforme Moodle de ces deux compétences est complétée par une évaluation de la production écrite ainsi que par un entretien (évaluation de la production orale). Ces deux dernières évaluations sont traitées de façon non automatique et donnent lieu à une saisie manuelle des résultats.

Enfin, une autre équipe enseignante a construit un troisième modèle pour son test de classement sous la forme d'un test adaptatif. Les différentes tâches sont centrées sur l'évaluation d'une compétence mais elles correspondent toutes à un seul niveau de cette compétence. Dans ce modèle, pour notre dispositif, il y a cinq tâches communicatives pour chaque compétence. L'idée est alors d'adapter les tâches proposées, pour une certaine compétence, à un étudiant au fur et à mesure que se précise son niveau au cours du test. L'étudiant commence par auto-évaluer sa compréhension orale, en utilisant des descripteurs des différents niveaux de compétence fournis par le CECR du Conseil de l'Europe (2001).

L'étudiant choisit le niveau qui semble lui correspondre le mieux et il active un lien vers les tâches de ce niveau rassemblées dans un sous-test. Si ce niveau est confirmé, par une note supérieure au niveau de seuil, on cherche à savoir si l'étudiant ne s'est pas sous-estimé et les tâches du niveau supérieur lui sont proposées. Au contraire, si sa note n'atteint pas le niveau de seuil, on lui propose les tâches de niveau inférieur. Et ainsi de suite. Le niveau final de l'étudiant est alors déterminé d'après le niveau le plus haut où il a atteint le niveau de seuil. L'étudiant peut alors s'auto-évaluer pour la compréhension écrite et entre dans la suite des tâches adaptées au niveau déterminé par l'auto-évaluation. Il fait ensuite de même avec la compétence suivante du test.

Des enseignants d'autres disciplines peuvent construire leur propre test de classement sur des structures analogues dont l'élément fondamental est d'avoir ou de bâtir un référentiel des compétences à tester à l'entrée d'une formation. Les autres principes essentiels peuvent être résumés comme suit:

1) distinction de deux compétences, l'une permettant une première évaluation que vient affiner l'évaluation de l'autre compétence ;
2) tests de puissance automatiques complétés par un entretien oral ;
3) auto-évaluation et conception de tâches centrées sur une compétence et un niveau.

3. Questions analysées

En mettant en œuvre un test de classement en ligne, nos objectifs sont de parvenir à une automatisation de l'inscription des étudiants et de réduire le temps d'évaluation nécessaire pour déterminer le niveau de l'apprenant. Ils sont aussi de baser notre test sur un modèle de test reconnu, par exemple un modèle adaptatif reposant sur le CECR.

Selon notre définition, un test doit être standardisé, fidèle, valide et étalonné (De Landsheere,

1979). Ce sont ces caractéristiques qu'il nous faut vérifier pour notre dispositif. Tagliante (2005) en donne le sens précis en termes simples. L'étalonnage permet qu'un test situe un individu par rapport aux autres avec plus ou moins de sensibilité, c'est-à-dire de discrimination entre des petites différences. La fidélité est liée à la stabilité des différences vues par un test. La validité d'un test assure qu'il mesure bien ce qu'il est censé mesurer. La standardisation est relative aux conditions dans lesquelles se passe le test et qui doivent être semblables pour tous les usagers. Pour celle-ci, il s'agit d'analyser si les conditions d'utilisation du test conviennent ou doivent être améliorées.

C'est la question centrale que nous nous posons dans le présent article, celle de l'acceptabilité de notre test de classement avec ses caractéristiques : être automatisé, en ligne, basé sur un modèle adaptatif et le plus court possible. Et nous voulons donc déterminer l'utilisation effective que les étudiants font du dispositif mais aussi la façon dont ils perçoivent cette automatisation à distance. Se centrer essentiellement sur cette question nous semble important pour notre dispositif à distance lorsqu'on constate que les étudiants d'aujourd'hui sont de grands utilisateurs des technologies dans leur vie quotidienne mais peuvent avoir des difficultés avec de tels outils dans des situations d'apprentissage (Barbot & Pugibet, 2002).

Pour l'utilisation effective, différents paramètres du cheminement de l'étudiant dans le test sont analysés. Quel est le temps nécessaire pour faire un test composé de 11 sous-tests (5 pour CO, 5 pour CE et 1 pour PE) ? Quels sont les niveaux de seuil à fixer dans les sous-tests qui permettent un cheminement cohérent, pour chaque compétence, à partir de l'auto-évaluation ? Combien de sous-tests faut-il à un étudiant pour que son niveau de langue soit estimé ? Quelles sont les pistes permettant de réduire la durée totale du test ?

Les perceptions du test en ligne par son usager (satisfaction, difficulté ressentie, etc.) sont analysées à travers les principaux aspects liés aux représentations que l'étudiant se fait de l'utilisation de ce nouveau type de dispositif. La modalité du test remet-elle en question l'individu dans ses habitudes fondamentales d'apprentissage, son identité et ses sensibilités d'ordre culturel ? L'analyse de ces éléments permettra de mettre en relief les modalités et conditions d'acceptabilité d'un tel dispositif.

4. Aspects méthodologiques

Etant donné le calendrier interne de développement des tests dans les différentes langues, c'est le dispositif pour le français qui est le plus avancé et nous avons choisi de limiter l'analyse présente aux données récoltées pour cette langue. Une analyse similaire pour les

dispositifs mis en œuvre dans les trois autres langues sera faite ultérieurement. Elle permettra d'analyser les mêmes questions avec des dispositifs de test de classement basés sur d'autres modèles.

En cherchant à estimer l'acceptabilité des étudiants vis-à-vis du test, nous nous inscrivons dans un processus d'amélioration continue du dispositif qui utilise une « mise à l'essai » pour comprendre le vécu des utilisateurs du test (Platteaux, 2004). Celle-ci est une situation d'apprentissage que l'on veut la plus proche possible des conditions réelles d'usage. L'acceptabilité dépend des préconceptions de l'utilisateur, de ses expériences antérieures, de ses habitudes et de ses compétences. Nous l'évaluons, dès lors, du point de vue quantitatif, à l'aide des résultats et durées des sous-tests et du point de vue qualitatif, à l'aide des feedbacks des étudiants à un questionnaire et/ou un entretien passé après le test.

4.1. Les publics ayant testé notre dispositif

Le prototype de notre dispositif pour le français a été testé avec trois types de public à la rentrée d'automne 2008 et en février 2009.

- En 2008, un premier groupe, constitué d'une vingtaine d'étudiants en séjour Erasmus en cours intensif préparatoire de français pendant les trois semaines précédant le début du semestre, a effectué le test sous contrôle, à une date et dans un lieu donnés avec réinvestissement du résultat obtenu au test pour leur inscription au cours de langue pendant l'année du séjour même.
- En 2008, un second groupe formé d'étudiants germanophones, futurs juristes, du cours de terminologie juridique française avait accès librement au test en ligne, via la plate-forme éducative Moodle du cours sur laquelle il leur était proposé d'effectuer le test à titre d'auto-évaluation.
- En février 2009, un troisième groupe formé de 121 étudiants arrivant à Fribourg a utilisé le test de classement en ligne dans des conditions presque réelles : la semaine avant la rentrée avec le choix de faire le test sur la plateforme Moodle, en classe ou à distance.

Dans ce dernier groupe, 104 étudiants ont suivi le cheminement proposé en faisant les sous-tests pour chacune des trois compétences CO, CE et PE. Quelques informations ont été récoltées sur leur profil (cf. Tableau 1).

Genre	Faculté	Langue maternelle	Types de public
Féminin : 78 Masculin : 26	Droit : 13 Lettres : 31 Sciences : 4 SES : 16 Théologie : 1 Pas de réponse : 35	Allemand : 44 Anglais : 3 Italien : 3 Espagnol : 4 Autres : 15 Pas de réponses : 35	BA : 19 MA : 13 Doctorat : 5 Autres : 28 Pas de réponses : 39

Tableau 1 : éléments du profil des étudiants passant le test (Université Fribourg – FLE 02.2009).

Nous voyons que le public auquel s'adresse un test de langue comme le nôtre est extrêmement varié. Dans la colonne « Types de public », la catégorie « Autres » regroupe notamment les étudiants inscrits temporairement à Fribourg car au bénéfice d'un programme « Mobilité ». Ces profils nous montrent que l'analyse menée, en particulier avec la population de février 2009, prend en compte une pluralité d'avis, avec de nombreuses personnes de langue maternelle allemande.

4.2. Traitement des données quantitatives

L'ensemble des données quantitatives est recueilli automatiquement par la plateforme Moodle durant l'utilisation du test par les étudiants. Pour estimer les conditions du test, nous traitons essentiellement des données simples - comme la durée des sous-tests, les notes obtenues, etc. – et procédons à des calculs de moyennes et d'écarts-types. Dans cette analyse, nous appliquons des méthodes de statistique descriptive (Albarello, Bourgeois, & Guyot, 2007). En ce qui concerne les recherches d'éventuelles relations entre les résultats obtenus pour différentes compétences, nous faisons un test du Chi^2, en suivant la méthode décrite par Guéguen (1998).

Précisons que pour analyser l'acceptabilité par les utilisateurs des conditions du test de classement, nous ne traitons que les données des étudiants ayant fait le test au complet. En effet, il s'agit de ne pas estimer une durée moyenne du test, par exemple, en incluant les étudiants qui ont écourté celui-ci, volontairement ou non, ou ceux pour lesquels le test n'a pas fonctionné en calculant un niveau de langue erroné. En conséquence, parmi les 121 étudiants ayant fait le test, nous avons retiré successivement :

- ceux n'ayant pas atteint le seuil pour au moins un sous-test par compétence = 17 ;
- ceux ayant été changés de classe par rapport à l'inscription automatique = 6 ;

- ceux ayant fait plus d'un sous-test de production écrite = 4[1].

Tous les calculs sont donc faits avec les données des 94 étudiants restants. Pour analyser certains sous-tests, on élimine la tentative d'un étudiant si sa durée dépasse le temps maximum imparti. Mais on garde les tentatives des étudiants ayant obtenu 0 point avec un temps inférieur à la durée maximale.

4.3. Traitement des données qualitatives

Les données qualitatives proviennent des réponses que les étudiants ont données à un questionnaire et/ou un entretien mis en place afin de recueillir leur expérience avec le test.

Certains des membres des deux groupes d'étudiants de septembre 2008, décrits plus haut, ont participé à des entretiens non systématiques pendant cette phase des réajustements didactiques et technologiques, qui s'est déroulée tout au long de l'année académique 2008-2009. Ces entretiens, de type semi-directif, ont duré environ 45 minutes et les étudiants y ont participé sur base volontaire.

Outre l'intérêt manifesté pour l'amélioration du dispositif, ces témoignages ont permis de dégager un paramètre important pour l'étudiant dans l'utilisation ou la non-utilisation du test dans sa version en ligne, celui de la représentation qu'en a son usager. Il paraît, par ailleurs, incontournable de mettre en relation ce regard estudiantin porté sur le test, avec celui de l'équipe de ses concepteurs soit l'ensemble des enseignants du Centre de langues de l'Université de Fribourg.

L'analyse de ce double regard porté sur le nouveau dispositif d'évaluation met en relief l'impact de nouvelles valeurs de notre société en mutation dont les universités ne peuvent plus faire abstraction.

[1] Nous pensons ici à un problème de compréhension de la consigne. A l'entrée de l'activité de production écrite, il est précisé : « Faites UNE SEULE production écrite » et à la sortie : « Merci beaucoup pour votre texte. Vous avez maintenant terminé cette évaluation. » Mais ces 4 étudiants ont proposé une deuxième production écrite.

5. Résultats et discussion pour le test de français

5.1. Analyse des données quantitatives

5.1.1. Etalonnage et sensibilité du test

En ce qui concerne l'étalonnage de notre test, un premier résultat est que le niveau moyen d'un étudiant, calculé automatiquement par le système en fonction des réponses aux sous-tests passés, est validé. En effet, c'est seulement pour 6 étudiants, sur les 121 ayant fait des sous-tests, que les enseignants ont dû procéder à un changement de classe. Autrement dit, notre test parvient à situer les individus les uns par rapport aux autres puisqu'il les regroupe par niveaux de langues homogènes. Ainsi, par la suite, nous considérons comme adéquats ces niveaux moyens calculés automatiquement.

Etant donné notre but de mettre au point un test de classement, nous nous contentons de ce résultat actuel pour l'étalonnage. Par contre, nous procéderons ultérieurement à une analyse de sensibilité pour pouvoir mieux distinguer les étudiants entre eux et pouvoir faire un diagnostic individuel avec notre test. Pour ce faire, nous emploierons des techniques d'analyse de sous-tests spécialisées sur la détermination de la difficulté et la discrimination des items (Crocker & Algina, 1986).

5.1.2. Faire le test à distance

Le système en ligne permet à l'étudiant de faire le test de classement depuis l'endroit qu'il choisit. La population des étudiants de février 2009 était constituée d'étudiants arrivant à l'Université de Fribourg et nous ne les avons donc pas obligés à faire le test à distance, pour ne pas les stresser. Nous avons cependant constaté que deux tiers d'entre eux ont choisi cette option. Etant donné que la procédure du test proposé était nouvelle pour eux, tout comme la plateforme Moodle, la très faible quantité de messages faisant part d'incompréhensions quant à ce qu'il faut faire (objectif de l'activité) ou demandant une aide d'ordre manipulatoire (ergonomie de l'outil en ligne) nous montre la faisabilité d'un test totalement à distance. Nous poursuivons donc notre projet de mettre en place un test de classement entièrement fait en ligne quelques semaines avant la rentrée universitaire pour pouvoir organiser à l'avance les groupes selon leur répartition en niveaux.

5.1.3. Temps nécessaire au test

La plateforme moodle offre la possibilité de limiter un sous-test dans le temps, autrement dit, d'indiquer à l'étudiant qu'il doit le faire dans une durée prédéterminée. Si l'étudiant la dépasse, seules comptent les réponses données par l'étudiant avant.

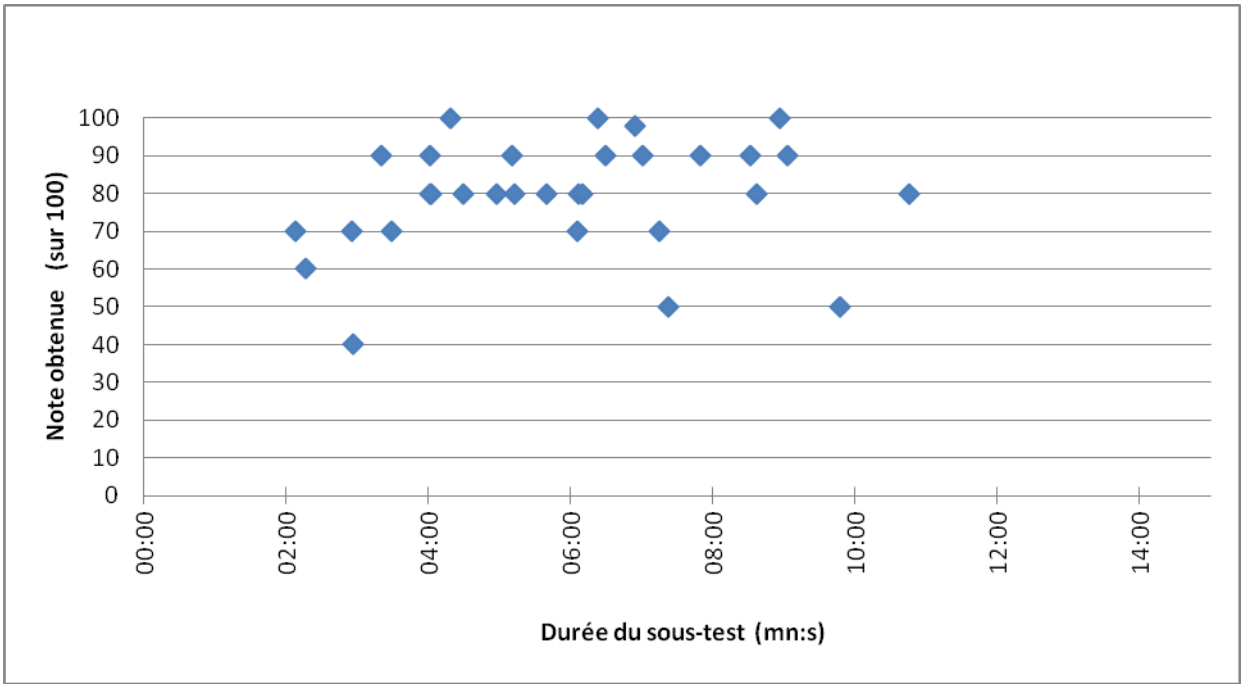

Figure 1 : temps et notes des 31 étudiants passant le sous-test COA1 (Université Fribourg – FLE 02.2009).

L'objectif de cette limite de temps dans les sous-tests est de réduire le temps global passé par l'étudiant pour effectuer le test de classement. On doit cependant vérifier que le temps laissé pour faire les sous-tests est assez long. Pour les sous-tests de niveaux A1, A2 et B1, les étudiants doivent avoir beaucoup de temps car la compétence testée à ce niveau est : « je comprends une annonce simple (orale ou écrite) et, en conséquence, je réponds juste ». Pour les sous-tests de niveaux B2 et C1, leur durée limitée doit permettre d'évaluer le temps de réponse des étudiants car la compétence testée est : « je comprends vite une annonce (orale ou écrite) ». Bien d'autres facteurs interviennent dans les conditions créées par un sous-test : la difficulté du texte ou de l'extrait sonore utilisé pour poser des questions qui peuvent, elles aussi, être plus ou moins complexes.

Les graphiques utilisant des nuages de points pour représenter les étudiants avec la durée de leur sous-test et le résultat obtenu permettent de voir si nous plaçons les étudiants dans les conditions adéquates (cf. Figures 1 et 2). La Figure 1 est exemplaire pour un sous-test sur une compétence « je comprends une conversation ». Les étudiants les plus lents ont mis 10:30 (mn:s) alors qu'ils disposaient de 15 minutes au total. Obtenir un nuage de points d'une forme similaire pour des sous-tests B2 et C1 indiqueraient au contraire un temps trop long. La Figure 2 montre une durée adéquate pour ces niveaux des sous-tests. Nous ferons évoluer notre dispositif selon ces résultats pour les durées prédéterminées des sous-tests.

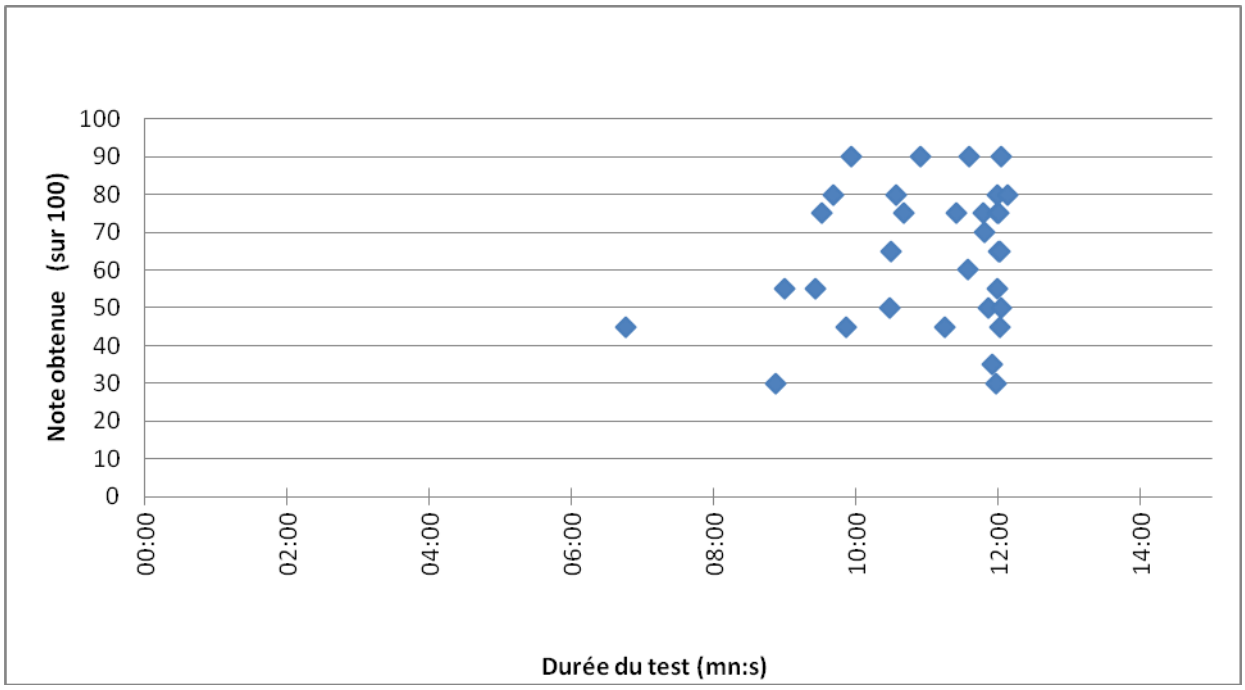

Figure 2 : temps et notes des 32 étudiants passant le sous-test COB2 (Université Fribourg – FLE 02.2009).

Voyons les durées moyennes du temps nécessaire aux étudiants pour faire les sous-tests proposés (cf. Tableau 2). Une très grande dispersion apparaît dans tous les sous-tests, en particulier pour ceux des bas niveaux, autour de la moyenne en temps nécessaire. Cette dispersion traduit sans doute une très grande variété des stratégies de résolution des sous-tests : tout faire très vite ou plus lentement parce que c'est difficile, revenir pour vérifier ses réponses avant de tout envoyer, s'arrêter longuement sur un item, etc. Par contre, les temps passés pour chaque compétence (cf. Tableau 3) sont très comparables.

Sous-tests	COA1	COA2	COB1	COB2	COC1
Nb étudiants	31	57	60	32	14
Temps moyen (mn:s)	05:53	08:12	12:31	10:59	10:35
Ecart-type σ (mn:s)	02:17	03:05	03:02	01:17	01:50
Sous-tests	**CEA1**	**CEA2**	**CEB1**	**CEB2**	**CEC1**
Nb étudiants	37	29	36	35	26
Temps moyen (mn:s)	05:09	09:44	09:41	10:28	10:09
Ecart-type σ (mn:s)	02:27	03:09	02:34	04:43	03:55
Sous-tests	**PEA1**	**PEA2**	**PEB1**	**PEB2**	**PEC1**
Nb étudiants	32	16	28	14	4
Temps moyen (mn:s)	08:25	17:16	21:14	26:02	28:43
Ecart-type σ (mn:s)	05:01	03:30	05:58	05:42	02:11

Tableau 2 : durée moyenne de chaque sous-test (Université de Fribourg – FLE 02.2009).

Sous-tests	CO	CE	PE	Temps Total
Temps moyen (mm:ss)	19:32	15:26	17:14	52:12
Ecart-type σ (mm:ss)	09:09	09:04	08:37	21:20

Tableau 3 : durée moyenne du sous-test, par compétence et au total (Université de Fribourg – FLE 02.2009).

5.1.4. Cheminement dans le test

Les étudiants entament l'évaluation d'une compétence au moyen d'une auto-évaluation qui leur permet de commencer par le sous-test qu'ils estiment le plus adapté à leur niveau de langue. Mais il nous semble que les étudiants ne le déterminent pas en toute connaissance de cause, malgré les descripteurs à leur disposition. Ils passent en moyenne une minute et demie pour auto-évaluer leur niveau de compréhension orale, première étape du test de classement. En revanche, 84 % des étudiants passent ensuite moins de 18 secondes en moyenne (σ = 13 s) pour auto-évaluer leur compétence de compréhension écrite et 71 % ont besoin seulement de 26 secondes en moyenne (σ = 38 s) pour leur compétence de production écrite. Le temps pour la compréhension orale peut être considéré comme adéquat car il leur permet de lire les

descripteurs des niveaux et de s'auto-évaluer. Au contraire, pour les deux autres compétences, ce n'est pas le cas.

Pourquoi cette différence ? Peut-être les étudiants pensent-ils que leur niveau est identique pour les trois compétences ? Mais cette hypothèse est rejetée par un calcul du $Chi^{2,}$ entre les résultats de CO et CE Auto-Eval et ceux de CE et PE Auto-Eval. Sans doute, ne perçoivent-ils alors pas l'impact d'une auto-évaluation hâtive sur la complexité accrue du cheminement dans le test. Nous décidons de communiquer plus sur ce point.

Considérons maintenant le nombre de sous-tests fait en moyenne par un étudiant, au total et par compétence (cf. Tableaux 4 et 5). Pour clarifier les résultats présentés, rappelons que, pour les étudiants considérés, la compétence PE ajoute toujours un sous-test au nombre total des sous-tests faits.

	Nombre sous-tests CO	**Nombre sous-tests CE**	**Nombre total de sous-tests**
Nombre moyen	2.12	1.76	4.87
Ecart-type	0.62	0.79	1.15

Tableau 4 : nombre de sous-tests faits, par compétence et au total (Université de Fribourg – FLE 02.2009).

Etudiants selon compétences	**Etudiants selon nombre total de sous-tests**	
1 sous-test CO & 1 sous-test CE : 12%	3 sous-tests : 12%	4 sous-tests : 26%
1 sous-test CO & > 2 sous-tests CE : 1%	5 sous-tests : 36%	6 sous-tests : 19%
> 2 sous-tests CO & 1 sous-test CE : 31%	7 sous-tests : 5%	8 sous-tests : 2%
> 2 sous-tests CO & > 2 sous-tests CE : 56%		

Tableau 5 : pourcentage des étudiants et nombre de sous-tests faits (Université de Fribourg – FLE 02.2009).

Selon notre principe de calcul automatique des niveaux, l'estimation d'un niveau nécessite que l'étudiant fasse au moins deux sous-tests dans une compétence. En effet, le niveau validé pour une compétence est le plus haut pour lequel le résultat de l'étudiant dépasse le niveau de seuil. En conséquence, si un étudiant ne fait qu'un sous-test et que son résultat est inférieur au seuil, le système ne peut faire de calcul. Dans le cas où un étudiant ne fait qu'un sous-test et que son résultat est supérieur au seuil, le système peut faire le calcul mais il se pourrait que le niveau de l'étudiant soit en fait supérieur.

La moyenne du nombre total des sous-tests faits semble indiquer que ce nombre minimal de 2 sous-tests n'est pas atteint, du moins pour la compétence CE. Le résultat le plus significatif est d'ailleurs la différence de comportement des étudiants entre les compétences CO et CE.

Notons que les 12% d'étudiants n'ayant fait qu'un sous-test CO et un sous-test CE et répertoriés dans le Tableau cinq sont dans un cas favorable : celui où ils ont atteint le niveau de seuil. Le système pouvait donc calculer automatiquement un niveau de langue. Au contraire, les 17 étudiants dont nous n'avons pas traité les données (cf. Section 4.2) étaient dans le cas défavorable où ils n'avaient pas atteint le niveau de seuil dans le seul sous-test fait par eux dans une compétence. En tous les cas, on peut faire l'hypothèse que ces deux groupes d'étudiants, soit 23% des personnes ayant fait des sous-tests, ont trouvé le test trop long ou se sont perdus dans le cheminement. Ces deux groupes sont constitués de personnes à tous les niveaux de langue.

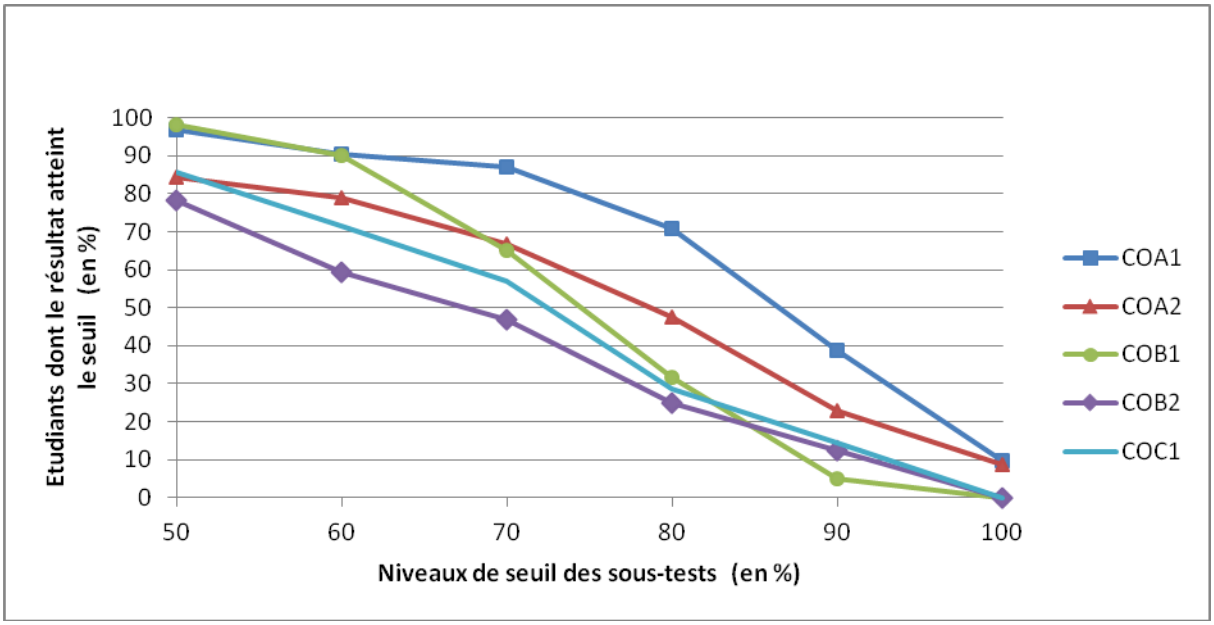

Figure 3 : atteinte du niveau de seuil selon sa valeur dans sous-tests CO (Université Fribourg – FLE 02.2009).

Le réglage des niveaux de seuil est déterminant dans le cheminement des étudiants après leur choix d'un point de départ avec le sous-test d'auto-évaluation. Avec les résultats des étudiants aux différents sous-tests, on peut représenter la proportion d'étudiants qui atteignent ou pas le niveau de seuil en fonction de la valeur fixée à celui-ci (cf. Figures 3 et 4).

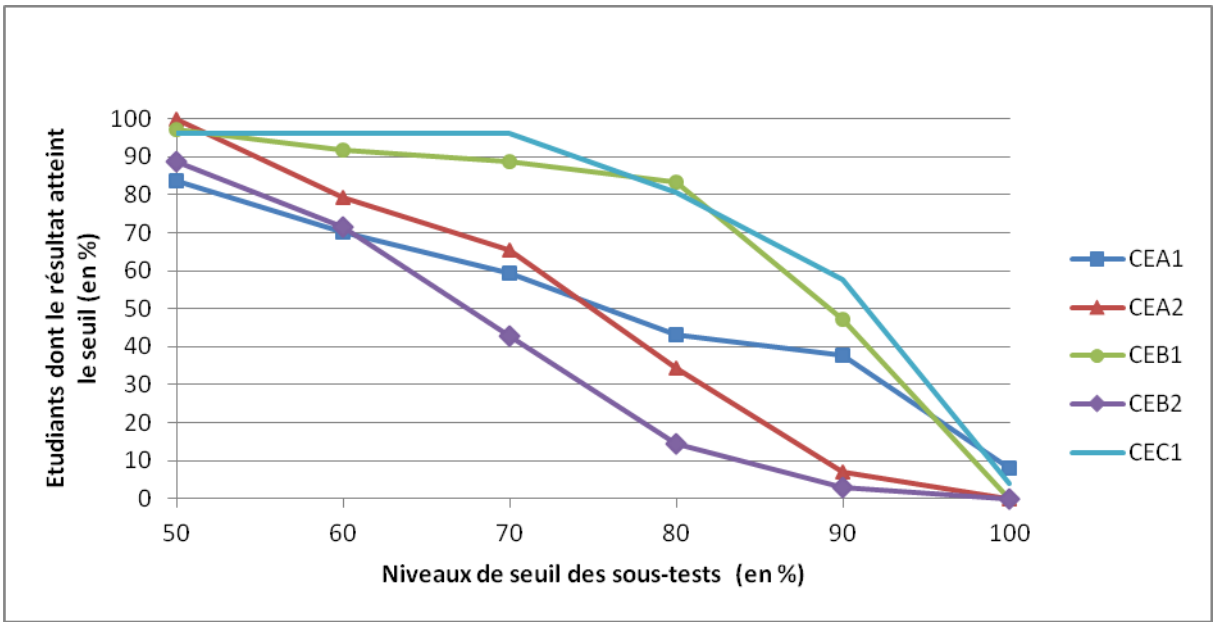

Figure 4 : atteinte du niveau de seuil selon sa valeur dans sous-tests CE (Université Fribourg – FLE 02.2009).

Pour les sous-tests sur la compétence CO, on voit d'abord que le sous-test COA1 est trop simple. En effet, il faut un niveau de seuil supérieur à 80% pour voir une nette diminution des étudiants dont les résultats atteignent ce seuil. Autrement dit, presque tous les étudiants entrant dans ce sous-test, avec un seuil fixé en dessous de cette valeur, atteignent au moins le seuil et le sous-test n'apporte que peu d'information supplémentaire sur le niveau de langue de l'étudiant. On fait le même constat pour les sous-tests CEC1 et CEB1 pour la compétence CE. Et, au contraire, il n'y a pas de sous-test trop difficile. Ainsi, le niveau de difficulté des autres sous-tests, pour CO et CE, est adéquat à un premier niveau d'analyse.

Comme précisé antérieurement, la détermination précise de la difficulté des sous-tests et de leur discrimination sera faite ultérieurement avec une méthode spécialisée sur ces questions, par exemple basée sur les modèles de l'*Item Response Theory* (Lord, 1980).

5.1.5. Pistes pour réduire la durée totale du test

Il faut considérer le test à la fois au niveau du temps nécessaire à l'étudiant pour le faire et au niveau du temps de correction nécessaire, avec le délai d'attente entraîné pour l'étudiant.

En particulier, les sous-tests basés sur une production écrite et/ou une production orale des étudiants nécessitent un temps de correction non automatique et différé entraînant une attente ou des incompréhensions de l'étudiant vis-à-vis des feedbacks. Ainsi, on a pu voir qu'une proportion non négligeable d'étudiants est retournée plusieurs fois sur la plateforme, après avoir terminé le test, afin de voir si leurs résultats étaient communiqués et s'ils pouvaient s'inscrire au cours adapté à leur niveau. Ils étaient informés du délai de 48 heures nécessaires à l'évaluation. Cependant, on peut imaginer la frustration de ces étudiants tant que ces résultats ne leur étaient pas accessibles.

De plus, le temps nécessaire à l'écriture des productions écrites est long (cf. Tableau 2). Pour ces raisons, il nous semble opportun de supprimer le sous-test de production écrite et d'introduire des sous-tests de lexique et structures de la langue corrigés automatiquement et moins longs à faire par l'étudiant.

Une autre idée pour réduire la durée totale est de mesurer si les résultats obtenus aux différentes compétences sont reliés et d'enlever la ou les compétences les plus longues à tester. Un test du Chi^2 est fait entre les résultats individuels obtenus pour les trois compétences testées et le niveau de compétence moyen calculé : CO et MOY, PE et MOY, CE et MOY. Celui-ci (MOY) est la simple moyenne arithmétique des niveaux des trois compétences évaluées. Un second test du Chi^2 est fait entre les résultats individuels obtenus pour les trois compétences testées : CO et CE, CO et PE, CE et PE. Mais ils nous forcent à abandonner l'hypothèse d'une relation entre les résultats obtenus. Seul le Chi^2 entre les compétences de production écrite et la compétence moyenne montre une possibilité de relation entre ces deux compétences. Nous n'approfondissons néanmoins pas cette piste puisque nous développons un test automatique. Tout semble donc indiquer que nous avons bien affaire à des compétences différentes. En conséquence, il ne peut être question de déterminer le niveau de langue d'un étudiant avec un test centré sur une seule compétence : cela réduirait la durée du test de classement mais ne permettrait pas d'évaluer son niveau moyen de manière satisfaisante.

De même, nous cherchons des relations éventuelles entre les résultats de l'auto-évaluation et des sous-tests, pour les trois compétences. Un test du Chi^2 n'en montre aucune, sauf pour la compétence PE. Il n'est donc pas envisageable de remplacer les sous-tests automatiques par une simple auto-évaluation. Le calcul des différences pour chaque compétence, en termes de nombre de niveaux, entre l'autoévaluation et le test lui-même (cf. Tableau 6) montre un cheminement court pour la grande majorité des étudiants. Il n'est donc pas non plus

envisageable d'enlever l'auto-évaluation.

Différence de niveau entre l'auto-évaluation et les sous-tests	Compétence CO (% d'étudiants)	Compétence CE (% d'étudiants)	Compétence PE (% d'étudiants)
0 niveau	38	43	64
1 niveau	53	37	30
2 niveaux	9	19	4
3 niveaux	0	1	2

Tableau 6 : différence de niveaux entre auto-évaluation et sous-tests (Université de Fribourg – FLE 02.2009).

Enfin, l'idée est aussi de mettre en place un sous-test pour les débutants. Même si celui-ci ne réduira le temps que pour une fraction des étudiants, il semble en effet illogique de forcer les débutants à un cheminement dans trois compétences.

5.2. Aspects socio-culturels du test de classement en ligne

Outre ces données statistiques, les enquêtes qualitatives menées auprès d'un public volontaire ont permis d'analyser les deux grands axes de réflexion suivants : l'interaction entre apprentissage et innovation technologique dans l'apprentissage des langues et le rapport de l'usager au nouvel environnement spatio-temporel. L'équipe enseignante a aussi eu ses propres perceptions qu'il convient de présenter brièvement.

5.2.1. Mutation dans l'organisation du travail de l'équipe enseignante

L'enseignant-concepteur a généralement vécu le projet du test en ligne comme le lancement d'un domaine de recherche portant sur *la gestion de l'évaluation par les technologies de l'information et de la communication en enseignement (TICE)*. Tout le monde parle de l'évaluation en langues et tout le monde parle des TICE. Mais peu de « produits » se réalisent vraiment à la croisée des deux domaines (Mangenot & Louveau, 2006). Cette mutation vers le mode d'évaluation en ligne est donc passée par une phase de réflexion générale des objectifs mêmes de l'évaluation. La collaboration avec le Centre nouvelles technologies et enseignement (NTE) a permis l'approfondissement dans l'acquisition des savoir-faire techniques et des outils nécessaires à la mise en ligne du projet, comme par exemple les choix de type ergonomique. Que cherche-t-on finalement à évaluer, pourquoi et comment le réaliser

du point de vue technique? Comment interpréter les résultats ? En quoi les TICE permettent-elles d'améliorer le dispositif traditionnel ?

Tous ces questionnements - motivants par la remise en question permanente qu'ils engendrent - ont trouvé leurs solutions grâce à un travail traditionnellement collaboratif et les outils technologiques ont favorisé la visibilité des phases de travail. Cette mutation dans l'organisation du travail de l'équipe enseignante n'est pas nouvelle. Toutefois, elle s'est généralisée et banalisée en impliquant non seulement l'ensemble de ses membres mais aussi d'autres services de support technologique ou administratif. Le nouveau dispositif ne s'est, en effet, pas arrêté à l'évaluation : il a permis à l'équipe enseignante d'aller au-delà des finalités du test lui-même et de collaborer plus intensément avec les autres services impliqués, en termes d'administration et d'inscription au cours : l'inscription s'effectue en effet en ligne à la suite du test et de son résultat. Il y a donc eu mutation dans l'organisation du travail en termes de contenu mais aussi de moyens dans sa mise en œuvre.

5.2.2. L'innovation, une contrainte ou une évidence

L'étudiant parle-t-il lui aussi du test comme d'une innovation? En tant qu'usager, ce prétendu renouveau est toujours interprété par l'étudiant comme une tension entre contrainte légitime et évidence légitime. La légitimité, c'est le signal de la modernité, normal et positif pour une université qui vit avec son temps, voire le devance : « *c'est normal… et puis, dans une université bilingue, je viens pour ça* ». Pour certains étudiants, le savoir paraît donc indissociable d'un support technologique, nécessitant lui-même toute une série de savoir-faire facilement accessibles pour des générations qui possèdent toute leur vie dans leur Ipod. D'autres étudiants découvrent l'enchantement technologique, par exemple lors de leur séjour d'études dans notre université : « *Tous les profs ici, ils ont ça ! Chez nous, tu te rends compte…* » ; selon eux, c'est l'œuvre de la plateforme en ligne Moodle, clé magique pour ouvrir toutes les portes du savoir. Une troisième catégorie d'étudiants est également en pleine découverte de Moodle et d'autres outils logiciels en ligne qu'ils trouvent légitimes, s'ils en perçoivent l'utilité pédagogique, mais qu'ils perçoivent néanmoins comme une contrainte parce que, pour eux, le développement du savoir-faire nécessaire implique des efforts.

Ainsi trois jeunesses se côtoient : la première post-moderne, la plus importante, est soucieuse d'un parcours de formation, ancré dans un discours de solitude dans la performance individuelle qui se construit par l'acquisition de savoirs qui doivent être au dernier cri (Tardieu & Pugibet, 2006). Le savoir doit être à la mode pour être crédible, tel un produit de

consommation en continuelle mutation, à la disposition de l'individu qui vit d'abord avec son ordinateur, même si les modes d'apprentissage traditionnels restent encore très prisés : « *Le test ? Il est moderne, je peux faire le test quand je veux, avec mon ordinateur* ». Pour une autre communauté, le test en ligne sur une plate-forme d'apprentissage est vécu comme une aventure féérique, une chance d'ouverture à saisir vers l'accès à tous les savoirs. Le test en ligne constitue une vraie révolution de la pensée démocratique pour l'étudiant qui fait la découverte du « je » autonome en face d'une communauté qu'il choisit librement, passant parfois du « nous » collectiviste au « nous » fraternel (Sériot, 2000). Quand une erreur de manipulation se produisait, ces étudiants disaient à haute voix leur message d'espoir : « *ça ne marche pas… pas tout de suite… ça ne fait rien, ça va marcher, il faut réessayer…* ». Et enfin, pour une troisième catégorie, cette innovation bouleverse des habitudes d'apprentissage et elle est plutôt vue comme une contrainte dont l'utilité reste floue.

5.2.3. Un isolement motivant

Ce qui a changé, c'est aussi le rapport apprenant-enseignant, ce dernier devenant invisible ou tout au plus un facilitateur-passeur d'accès au savoir dont il fabrique une forme de banque de données en libre accès dans le supermarché du savoir pour un apprenant-consommateur qui doit pouvoir se débrouiller tout seul. L'étudiant construit son parcours, le choisit et tout commence par ce test. Comment ressent-il ce nouveau mode d'évaluation vécu en complète autonomie ?

Tous les usagers expriment toujours un très grand enthousiasme à réaliser le test en ligne car l'apprenant se sent valorisé à plusieurs titres : tout d'abord, ce test en ligne repose sur un choix personnel puisque l'étudiant, même s'il doit faire le test, le met en œuvre sur base volontaire. C'est un premier pas vers la construction du savoir linguistique. Seul en face de la machine, l'apprenant-usager est en face de lui-même, devant le miroir de sa personne dans l'évaluation instantanée de son savoir au moyen des nouvelles technologies. L'erreur n'est plus que virtuelle, sur la machine, donc invisible au professeur et donc plus confortablement surmontée. Ce confort génère une prise de confiance, doublée d'une stimulation de la concentration, de la volonté.

5.2.4. Une diversification des rapports au temps et à la personne

Ce qui change fondamentalement, c'est aussi le rapport de l'apprenant au temps : cette évaluation en ligne lui permet de construire son savoir à tout moment et depuis n'importe quel lieu. En effet, l'apprenant peut s'engager dans son apprentissage linguistique librement, à tout

moment de son cursus et en fonction de ses besoins langagiers. L'université lui donne ainsi accès à une forme de formation continue dont il se sent le propre acteur sur un mode d'auto-formation dont le test symbolise la première étape. Les étudiants ont en effet tous apprécié la souplesse du système : « *on peut le faire quand on veut, et puis, on peut aussi s'arrêter et continuer.* » ; « *on peut le faire pour toutes les langues* » ; « *on peut s'inscrire tout de suite quand on sent qu'on en a besoin* ».

Cette dernière réflexion laisse entrevoir la naissance de la culture d'entreprise à l'université qui répond à la loi de l'offre et de la demande, au nom du savoir « utile » et de la formation tout au long de la vie. Dans le domaine des langues, le test de classement constitue le point de départ d'une nouvelle culture universitaire professionnalisante.

D'autre part, on constate que les étudiants interrogés sur leur expérience du test en ligne se présentent à travers un « on » impersonnel aussi souvent qu'avec un « je », ce qui peut être interprété de la façon suivante : l'interaction entre l'homme et la machine s'exprime à l'aide d'un singulier collectif, un repère sécurisant car universel, bien que délocalisé. On saisit alors tout l'impact de la globalisation sur les stratégies de travail de l'étudiant qui va faire un test en ligne dans le même état d'esprit qu'il se connecte à *Facebook*, membre d'une communauté virtuelle plus ou moins connue. Au-delà du temps individuel librement organisé auquel nous faisions allusion précédemment, surgit une deuxième strate temporelle particulière à l'interaction de l'étudiant avec la machine qui le met en relation avec une société virtuelle dans un temps global. Il n'est donc plus réduit à un simple individu coupé de la société mais bien reconnu comme une personne membre d'une communauté avec laquelle il partage une même tâche, le test.

La diversification des rapports au temps et à la personne est exprimée avec beaucoup de légèreté et de superficialité par l'étudiant qui ne semble pas être vraiment conscient de cette complexification. Pourrait-on même penser qu'un test en ligne constitue une activité presque ludique ? Pour tenter de répondre à cette question, il serait intéressant d'observer les réactions des étudiants à l'iconographie d'un test en ligne qui intègre une image drôle entre deux tâches. Les réactions dépendent fortement de données socio-culturelles liées au statut du rire dans un contexte universitaire. Par exemple, l'image d'un footballeur américain en pause, assis avec son ordinateur sur les genoux en train de faire le test, va détendre et amuser l'étudiant anglo-saxon. Par contre, il n'en va pas de même pour l'étudiant germanophone pour lequel le test n'est valide que s'il est sérieux. Comment le contenu d'un test peut-il provoquer le rire ? Cette réflexion émane surtout d'étudiants de culture du Nord pour lesquels il en serait

du test comme du paysage qui « *ne sera jamais risible.* » (Bergson, 1940, p. 2). Quant au francophone, il va longuement chercher la signification cartésienne de l'image par rapport au test. D'autres étudiants ne seront en rien dérangés par cette apparition ludique au milieu du test. Un autre exemple : l'ordre d'apparition des drapeaux nationaux pour signifier la langue des consignes peut vite créer l'incident diplomatique en blessant les identités nationales. Autant de réactions qui prouvent l'importance de données socioculturelles dans la construction visuelle et l'iconographie du test en ligne.

6. Conclusions

En conclusion, il convient de rappeler que le test pour lequel nous avons produit des analyses descriptives à partir de données quantitatives et qualitatives, est un test de classement (Placement Test/Einstufungstest/Test d'entrata). L'objectif premier de ce type de test est de constituer des groupes classes relativement homogènes mais non de statuer, comme le ferait une évaluation sommative, sur le niveau de langues effectif de chaque individu.

Pour parvenir à cet « enclassement » via le test en ligne, nous avons utilisé les niveaux et les descripteurs du Cadre Européen Commun de Référence (CECR) pour les langues, qui nous ont permis de déterminer une moyenne des compétences d'un individu pour la compréhension orale, la compréhension écrite et la production écrite. Conformément à l'esprit du CECR, on admet que les niveaux des apprenants appelés à constituer ces groupes classes sont inégaux dans les différentes compétences. Le but du cours de langue sera justement de chercher à équilibrer les compétences individuelles en s'appuyant sur la dynamique et les compétences diverses du groupe ainsi constitué.

De ce point de vue, le test de classement en ligne a pleinement rempli son office, dans la mesure où les groupes classes ont été constitués de manière efficace : nous avons eu à opérer des changements de classe (niveau inadapté de l'individu au groupe) inférieurs à 1% de la population totale des inscrits. Nos trois objectifs sont donc en voie de réalisation : validation d'un modèle de test unifié, utilisable dans un temps raisonnable et permettant une inscription automatique.

La conception du dispositif et l'analyse de son acceptabilité ont permis de dégager le rôle primordial joué, dans les modèles de tests élaborés, par certains paramètres, comme la durée des sous-tests, l'utilité que l'étudiant s'auto-évalue pour chaque compétence, ou encore la nécessité de sous-tests particuliers à chaque compétence différente.

De plus, les feedbacks des étudiants ont permis de voir à quel point l'expérience du test en ligne remet donc en question les rapports fondamentaux de l'individu au temps, à l'espace, à l'identité et à ses sensibilités d'ordre culturel. Le plus étonnant est le fait que l'usager parvienne à exprimer ces mutations de la façon la plus naturelle du monde mais ne semble ni s'en émouvoir, ni être très conscient des nouvelles possibilités qui lui sont offertes pour gérer ses apprentissages. Mais une chose est sûre, c'est que l'usager, comme le concepteur, sait qu'il ne peut plus faire l'économie des TICE dans l'acquisition et l'évaluation du savoir dont elles dépendent jusqu'à en modifier les fondements, si on en juge par les premières expérimentations de notre projet.

Les résultats que nous trouvons semblent ainsi largement transférables à d'autres domaines que celui des langues. La suite de notre travail s'inscrit dans la perspective que notre test puisse participer à diverses formes d'évaluation, selon les besoins de l'apprenant ou de l'institution en M2 pour mesurer la progression des apprenants et également en M3 pour vérifier les acquis des apprenants à l'issue d'une séquence d'apprentissage. Il s'agirait ainsi de permettre un diagnostique des connaissances des étudiants pour un pronostique de leur orientation dans l'apprentissage.

7. Bibliographie

Albarello, L., Bourgeois, E., & Guyot, J.-L. (2007). *Statistique descriptive : Un outil pour les praticiens-chercheurs*: De Boeck.

Allal, L. (1988). Vers un élargissement de la pédagogie de maîtrise : processus de régulation interactive, rétroactive et proactive In M. Huberman (Ed.), *Assurer la réussite des apprentissages scolaires ? Les propositions de la pédagogie de maîtrise* (pp. 86-126). Neuchâtel: Delachaux et Niestlé.

Barbot, M.-J., & Pugibet, V. (2002). Apprentissage des langues et technologies : usages en émergence. *Le Français dans le Monde, recherches et applications (Numéro spécial - janvier 2002)*. Paris: Edition Clé international.

Bergson, H. (1940). *Le rire, essais sur la signification du comique*. Paris: Quadrige/PUF.

Conseil de l'Europe. (2001). *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Paris: Didier.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont CA: Wadsworth.

De Landsheere, G. (1979). *Dictionnaire de l'évaluation et de la recherche en éducation*. Paris: PUF.

Guéguen, N. (1998). *Manuel de statistique pour psychologues*. Paris: Dunod.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Erlbaum.

Mangenot, F., & Louveau, E. (2006). *Internet et la classe de langue*. Paris: Edition Clé international.

North, B., & Jones, N. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe.

Platteaux, H. (2004). Regard sur l'accompagnement pédagogique de cours e-Learning à l'université. *J. Viens & A. Wyrsch (Eds.) eLearning: Concepts d'évaluation et applications. Revue Suisse des Sciences de l'Education*, *26*(2), 249-264.

Sériot, P. (2000). *Structure et totalité*. Paris: PUF.

Tagliante, C. (2005). *L'évaluation et le Cadre européen commun*. Paris: CLE International.

Tardieu, C., & Pugibet, V. (2006). Les TIC, enseignement et apprentissage. *Revue Langues et cultures de l'ALSIC, 9*, 237-247.

Tourneur, Y., & Vasamillet, C. (1982). *L'évaluation au service de la formation : situations, techniques, résultats*. Mons: Université de Mons.

Tricot, A., & al. (2003). *Utilité, utilisabilité, acceptabilité. Interpréter les relations entre trois dimensions de l'évaluation des EIAH.* Paper presented at the Actes de la Conférence Environnements Informatiques pour l'Apprentissage Humain, Strasbourg 15-17 avril.

Veltcheff, C., & Hilton, S. (2003). *L'évaluation en FLE*. Paris: Hachette.