ANALYSE DE VALIDITE DES TESTS DE CLASSEMENT EN LIGNE POUR LES LANGUES A L'AIDE DU MODELE DE RASCH

Cornelia Gick*, Hervé Platteaux **, Sergio Hoein*, Catherine Blons-Pierre****, Patricia Kohler*******

** Université de Fribourg (Suisse), cornelia.gick@unifr.ch*

***Université de Fribourg (Suisse), herve.platteaux@unifr.ch*

**** Université de Fribourg (Suisse), sergio.hoein@unifr.ch*

***** Université de Fribourg (Suisse), catherine.blons-pierre@unifr.ch*

****** Université de Fribourg (Suisse), patricia.kohler@unifr.ch*

Mots-clés : *test de classement en ligne, IPT/Analyse d'item, Modèle/Analyse de Rasch, Apprentissage des langues, Plateforme Moodle*

Résumé*. L'Université bilingue de Fribourg (Suisse) a réalisé des tests de classement en ligne adaptatifs pour quatre langues. Avec la plateforme Moodle, les étudiants passent un test, qui évalue leurs compétences langagières, afin de pouvoir s'inscrire à un cours de langue étrangère adapté à leur niveau. Depuis quatre semestres, les tests de classement en ligne ont été utilisés avec succès (très peu d'étudiants ont dû changer de niveau de cours). Pour augmenter encore la qualité de cette procédure, nous analysons la validité et de la fidélité des items des différents sous-tests conçus pour l'allemand (DaF) et le français (FLE) à l'aide du modèle de Rasch (Winsteps). Cela pour deux actions : la constitution des tests et de leurs sous-tests (mode adaptatif) et l'amélioration des items (qui participe à la formation continue des compétences des équipes de développement des items).*

---

1. Introduction

Le Centre de Langues et le Centre Nouvelles Technologies et Enseignement, de l'Université bilingue de Fribourg (Suisse), ont réalisé des tests de classement (placement tests) en ligne adaptatifs pour l'allemand, l'anglais, le français et l'italien langues étrangères. Avec la plateforme d'apprentissage en ligne Moodle, les étudiants passent un test, qui évalue leurs compétences langagières, avant de pouvoir s'inscrire, via un système d'inscription en ligne, à un cours de langue étrangère adapté à leur niveau. Mis ensemble, notre test de classement et le système d'inscription permettent d'automatiser le processus d'inscription et d'admission aux cours de langues.

Notre test de classement en ligne est un dispositif d'évaluation formative puisque son objectif premier est de déterminer le niveau de compétence de chaque apprenant afin de lui proposer des cours correspondant à son niveau. Pour le Centre de Langues qui gère le test, le but est aussi de créer des classes de niveaux relativement homogènes.

Le choix d'une plateforme d'apprentissage en ligne pour mettre en œuvre nos tests de classement adaptatifs (Linacre, 2000) provient d'une volonté de mettre à profit plusieurs forces de ces technologies. En premier lieu, elles permettent de réaliser une économie organisationnelle dans le cadre d'un enseignement/apprentissage en grands groupes. Avec une plateforme en ligne, chaque

semestre, 1'500 étudiants passent le test où et quand ils veulent sans qu'ils aient besoin de faire le pied de grue dans les couloirs du Centre de Langues. Le stress des enseignants, et du secrétariat, en est diminué d'autant durant la même période de la rentrée. L'évaluation de compétences réceptives permet un calcul automatique du niveau langagier de l'étudiant à partir de ses réponses au test en ligne.

Quelques principes résument le modèle de ces tests :

- Ils sont composés de trois catégories de sous-tests évaluant les compétences réceptives la compréhension orale et la compréhension écrite ainsi que les structures de la langue.
- Il intègre l'autoévaluation selon les descripteurs du Cadre Européen Commun de Référence pour les langues du Conseil de l'Europe (CECR).
- Pour chaque catégorie, plusieurs sous-tests sont proposés aux étudiants selon une modalité adaptative.
- Les deux tests utilisent un instrument qui permet à l'étudiant d'être dirigé de manière efficace dans un sous-test de son niveau. Pour le test FLE (Français Langue Etrangère), c'est l'autoévaluation sur la base du CECR ; pour le test DaF (Deutsch als Fremdsprache), ce sont 8-10 items de grammaire dont le résultat dirige vers un autre test de grammaire de 18-20 items.
- La création des items (questions et propositions de réponse) des sous-tests est basée sur le CECR (Blons-Pierre, 2010).

Les tests ont été utilisés avec succès : très peu d'étudiants ont dû changer de niveau de cours. Depuis le lancement des tests en ligne, chaque semestre dès février 2009, une récolte systématique de données a permis de mener diverses analyses que nous résumons dans une première partie de cet article. Sur cette base, pour augmenter encore la qualité des banques d'items, nous nous attachons à analyser plus en détails la validité et la fidélité (Tagliante, 2005) des items des différents sous-tests conçus pour les tests DaF et FLE.

Pour ce faire, nous utilisons la théorie de réponse aux items (Lord, 1980 ; Wright & Stone, 2004), en particulier le modèle de Rasch (1980), et le logiciel Winsteps, qui nous permettent de déterminer le pouvoir de discrimination d'un item et son degré de difficulté par rapport aux autres, ainsi que de formuler des indications concernant leur qualité. Nous disposons maintenant d'une population suffisante pour effectuer un tel traitement. Pour le test FLE, 400 étudiants environ ont passé le test entre les sessions de septembre 2010 et de février 2011. Pour le test DaF, 1000 étudiants environ, entre février 2009 et septembre 2011, dont 650 environ durant les sessions septembre 2010 et 2011 et février 2011 avec un test inchangé.

Ces résultats sont utilisés avec deux finalités qui forment le cœur du présent article :

- En premier lieu, l'analyse de Rasch intervient dans la conception d'une structure adaptative des tests. Elle permet de positionner les items dans les sous-tests, en fonction de leurs degrés de difficulté, de vérifier si les degrés de difficulté des items (et des sous-tests) correspondent à la population testée (Gick, 2011) et de vérifier la qualité des items utilisés.
- En second lieu, l'analyse de Rasch sert au développement continu de la production des items. Elle permet d'identifier des items problématiques dans les différents tests ciblés sur un niveau. Ces items sont alors soumis à l'équipe de conception qui propose des améliorations. Les analyses selon le modèle de Rasch servent ensuite à vérifier si ces changements portent leurs fruits lors de la passation suivante.

2. Résultats des analyses précédentes

La base des deux tests en ligne DaF et FLE sont les tests papier préexistants et validés. Depuis le lancement des tests en février 2009, diverses analyses ont été menées, à partir d'une récolte systématique de données, dans le but de vérifier la validité de notre modèle de test adaptatif et son opérationnalisation dans Moodle.

On a regardé le cheminement de l'étudiant dans le test (nombre sous-tests faits, durée totale, etc.) et la perception qu'il en construisait (Kohler, Platteaux, & Blons-Pierre, 2012). Ces analyses ont permis de régler les seuils de réussite des sous-tests, leurs limites en temps et le nombre d'items qui les composent. En parallèle une analyse quantitative, faite à partir des résultats obtenus aux tests en ligne par les premiers étudiants selon le modèle de Rasch, a permis de faire un premier tri sur la qualité des items, de supprimer des items estimés problématiques, ainsi que d'ajouter et de valider de nouveaux items (Gick, 2012).

Les feedbacks des étudiants sur les tests en ligne ont aussi permis de connaître et d'améliorer l'acceptance des tests en ligne (Kohler, Platteaux, & Blons-Pierre, 2010). Cette analyse se poursuit tant sur l'acceptance que sur le degré d'accord entre l'autoévaluation effectuée par les testés et le résultat de l'évaluation proposé par le test de classement (Kohler, 2012 ; Blons-Pierre, 2011 ; Blons-Pierre, Kohler, Gick, Hoein et Platteaux, 2012).

On a également vérifié, qu'à partir des réponses au test en ligne de l'étudiant, un calcul automatique pouvait être mis en place dans Moodle en utilisant les fonctions de tests de cette plateforme et les éléments d'évaluation associés à son tableau de notes pour évaluer des compétences réceptives langagières (Platteaux & Hoein, 2010).

3. Analyse de Rasch dans le contexte de nos tests

Dans la théorie de la réponse aux items (IRT), il y a une estimation pour les personnes, à partir des items répondus vers les items sans réponse, en partant d'une probabilité de résolution de 50 % (Rost 2004). Ceci permet de formuler des observations sur les items et la totalité du test même si, pour une partie du test, nous avons un nombre limité de participants. Cette analyse nous donne des informations sur le degré de difficulté, l'adaptation au modèle et la qualité des items. Le LMS Moodle permet, avec un module supplémentaire, d'obtenir les réponses détaillées de chaque étudiant à chaque item. Avec le logiciel Winsteps, nous obtenons relativement rapidement les résultats que nous cherchons. Si ce logiciel rend les analyses très rapides et permet également d'obtenir un grand nombre de formats de sortie et de représentation des résultats, il est nécessaire de préparer les données en les codant en chiffres.

En raison de la structure adaptative de nos tests, il nous faut également réunir toutes les réponses de tous les étudiants aux différents sous-tests dans un seul tableau. Nous traitons uniquement les questions de type QCM, qui constituent la grande majorité des items du test FLE et sont le seul type dans le test DaF. Pour que l'analyse soit assez fiable, nous ne prenons en compte que les tests pour lesquels nous avons un nombre suffisant de réponses. A cause du modèle adaptatif, il arrive que nous n'obtenions pas assez de réponses en une seule session. Si cela se produit, nous décidons de n'analyser que les sous-tests avec un nombre de réponses suffisant, ou d'attendre un semestre ultérieur avant de procéder à des changements. A partir de 70 réponses les résultats semblent être significatifs. Pendant le développement du test DaF, on a utilisé sous réserve les résultats à partir de 30 réponses.

Même si les deux tests FLE et DaF sont semblables, il existe des différences qui se manifestent également lors de l'utilisation de l'analyse de Rasch : les niveaux du CECR (Cadre européen commun de référence pour les langues) sont au centre du test FLE ; chaque sous-test se positionnant par rapport à un des niveaux. Cela nécessite une discussion approfondie sur ces niveaux dans l'équipe de développement, ainsi qu'une phase d'expérimentation (Kohler, 2012). Ces deux actions ont pu être mises en œuvre et il a été possible de construire sur des travaux préalables. L'auto-positionnement, qui sert d'entrée dans chaque compétence, a un rôle important. Pour les étudiants, le niveau testé dans un sous-test est toujours clair. L'étudiant est dirigé vers un sous-test d'un niveau inférieur, d'un niveau supérieur ou vers une autre compétence. Le chemin à travers le test passe d'abord par la compréhension orale (CO) et écrite (CE), pour arriver ensuite à la composante « lexique et structure de la langue ». Dans les tests de CO et CE, plusieurs questions sont posées sur la base d'un texte (écrit ou oral). Pour l'utilisation de l'analyse de Rasch, cela

signifie que si un item est d'un niveau inférieur ou supérieur, il ne pourra pas simplement être déplacé dans un autre sous-test, mais devra être adapté. Cela justifie donc l'utilisation de l'analyse de Rasch dans une étape ultérieure, pour vérifier les résultats de cette adaptation.

Le test DaF est basé sur une progression et un continuum. Il a ainsi été possible de commencer avec un nombre réduit d'items (18 courtes tâches de compréhension écrite) et peu de sous-tests (trois tests de compréhension écrite). L'analyse de Rasch a permis le placement des items sur une échelle de difficulté et l'identification des items problématiques. Le niveau de cours est déterminé par un calcul, invisible aux étudiants, fait à partir de leurs résultats aux sous-tests. La difficulté majeure était de trouver les seuils de césure entre les niveaux. Le test ne pouvant être mis à l'essai avant son utilisation en situation réelle, le développement d'un tel test en si peu de temps impliquait un certain risque. Par conséquent l'attribution des niveaux de cours a été faite, pendant la première semaine de la première passation du test, par le biais d'une confirmation manuelle. Pour cela, les résultats calculés ont été comparés avec l'auto-évaluation des étudiants, en tenant compte des renseignements personnels, et le calcul a été en partie ajusté. C'est seulement après cette première semaine que tout a été automatisé. L'auto-évaluation du niveau de langue basée sur le CECR se fait à la fin du test, et continue à servir uniquement comme une aide à la décision en cas d'hésitation entre deux niveaux.

Au cours des sessions suivantes, de nouvelles tâches ont pu être ajoutées. La répartition sur plusieurs sous-tests et l'intégration de nouveaux items ont montré des problèmes pendant la session de printemps 2010. Depuis le semestre d'automne 2010, nous pouvons utiliser un test fiable, qui peut être analysé plus en détail et affiné. Cela n'aurait pas été possible sans l'utilisation en parallèle de l'analyse de Rasch.

4. Rasch pour la conception des tests

Depuis le début, le développement du test DaF a été accompagné pas à pas par l'analyse de Rasch. Après chaque session, nous nous intéressons en premier lieu aux valeurs suivantes :

- Le degré de difficulté (Measure) des items par compétence. Les valeurs indiquées ici ont été utiles pour la distribution des items dans les sous-tests.
- La relation entre la difficulté des items et le niveau de la population testée, donc la question: est-ce que les items proposés couvrent suffisamment l'éventail des niveaux ? Cela est identifiable dans le « Item Map ».
- La relation des trois sous-tests entre eux.
- La qualité des items et des distracteurs par rapport à l'« Infit » et au « Point-Measure ».

Pour développer le test DaF on s'est basé sur les questions à choix multiples de grammaire du test de classement papier qui fournissaient des indicateurs utiles pour différencier les niveaux A2 - B2. Nous avons analysé les données des tâches de grammaire de la dernière passation du test papier, au semestre d'automne 2008. Cela a fourni des points de repère pour la première version électronique des items de grammaire. A partir du semestre de printemps 2009 les données obtenues dans Moodle ont été préparées pour être ensuite analysées avec le logiciel Winsteps. Ce qui nous a intéressés tout particulièrement, à part les valeurs de difficulté (Measure), a été le spectre de difficultés que ces tâches couvraient. Pour cela nous avons utilisé l'« Item map ». L'impression subjective qu'il manquait des items dans le segment des niveaux supérieurs a été confirmée. On a montré également ainsi que le test était globalement trop bas par rapport au niveau de notre population ; il fallait donc le compléter avec des items à des niveaux supérieurs. Les valeurs de la « Measure » et de l'« Infit » ont permis de sélectionner huit items de différents niveaux pour le sous-test de distribution qui représente la porte d'entrée du test. L'analyse a également permis d'implémenter une progression du degré de difficulté dans le test. Pour le semestre d'automne cela signifie qu'au moins un sous-test a pu être construit sur la base d'items validés, et qui a été utilisé comme point d'ancrage pour le développement futur.

Les items nouvellement développés pour les compétences réceptives ont été utilisés pour la première fois à la session d'automne 2009. Ils ont été regroupés, selon le degré de difficulté supposé, en trois sous-tests de compréhension écrite (CE) et quatre sous-tests de compréhension orale (CO). Pour obtenir un nombre suffisant de tentatives, les étudiants ont été envoyés, selon leur résultat à un sous-test, au sous-test de difficulté supérieur ou inférieur. Dans cette structure, il peut arriver qu'un étudiant obtienne un résultat bas dans le test de grammaire, et qu'il soit donc dirigé vers un test CE de bas niveau, qu'il réussisse avec un bon résultat et qu'il se trouve devoir passer tout les trois sous-tests CE. En réussissant le test CE de plus haut niveau, il est ensuite dirigé vers un test CO de haut niveau. S'il n'a pas alors un bon niveau dans cette compétence, et obtient un résultat bas, il est redirigé vers le test CO d'un niveau inférieur. Si son résultat est à nouveau bas, il est encore redirigé vers un niveau inférieur. Dans cette configuration, la durée du test est rallongée, mais cela est une contrainte à accepter si l'on veut obtenir des données fiables pour l'analyse de Rasch.

TABLE 13.1 DaF-Einstufungstest-HS09 ZOU368WS.TXT Mar 11 17:54 2012
INPUT: 273 Student 114 Item REPORTED: 255 Student 18 Item 2 CATS WINSTEPS 3.71.0.1

Student: REAL SEP.: 1.04 REL.: .52 ... Item: REAL SEP.: 6.03 REL.: .97

Item STATISTICS: MEASURE ORDER

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
69	14	37	3.23	.41	.80	-1.1	.78	-.8	.68	.57	78.8	74.1	LV3_03
67	17	38	2.78	.39	.91	-.5	.84	-.7	.60	.54	73.5	70.4	LV3_01
72	16	33	2.64	.42	.81	-1.2	.74	-1.3	.66	.54	75.9	69.7	LV3_06
71	24	36	1.66	.40	1.09	.6	1.17	.8	.37	.44	62.5	69.2	LV3_05
66	36	157	1.58	.22	.93	-.6	.84	-.7	.49	.45	83.2	80.3	LV2_06
68	28	36	.98	.44	1.12	.7	1.09	.4	.29	.36	78.1	75.3	LV3_02
70	29	37	.88	.43	1.04	.3	.94	.0	.35	.36	72.7	76.0	LV3_04
62	62	145	.43	.19	1.13	1.6	1.20	1.4	.40	.49	66.2	71.7	LV2_02
63	65	148	.33	.19	1.09	1.1	1.12	.9	.42	.49	71.9	71.7	LV2_03
60	43	90	-.29	.25	.97	-.4	.90	-.7	.52	.49	65.0	67.3	LV1_06
65	89	135	-.72	.20	1.08	.9	1.06	.4	.39	.44	69.9	72.9	LV2_05
61	107	163	-.90	.19	.98	-.1	.88	-.7	.51	.49	72.8	74.1	LV2_01
58	71	117	-1.08	.22	.95	-.6	.91	-.8	.51	.47	69.5	67.7	LV1_04
59	83	121	-1.52	.22	1.01	.1	1.02	.2	.45	.46	70.4	72.2	LV1_05
64	107	140	-1.53	.22	1.12	1.1	1.12	.5	.36	.43	75.4	79.2	LV2_04
55	119	154	-2.36	.22	.84	-1.3	.73	-1.5	.58	.47	81.3	79.4	LV1_01
56	113	139	-2.57	.24	.87	-.9	.72	-1.2	.51	.42	83.3	81.5	LV1_02
57	124	136	-3.55	.33	1.06	.3	1.07	.3	.31	.35	91.1	91.2	LV1_03
MEAN	63.7	103.4	.00	.29	.99	.0	.95	-.2			74.5	74.7	
S.D.	38.1	50.1	1.89	.09	.11	.8	.15	.8			7.0	5.8	

Figure 1 : DaF HS09 – Item Statistics : Measure Order

L'analyse de Rasch pour la compréhension écrite (cf. Figure 1) donne l'image suivante :

- La relation entre la difficulté attendue (Entry Number: Item 55-72) et celle effectivement mesurée (Measure) était déjà bonne lors de la première passation du test.
- La distribution des items dans les trois sous-tests (LV1-LV3) montre également des bons résultats.
- Les items LV2-04, LV2-02 et LV3-02 ont attiré notre attention à cause de leur « Infit MNSQ », qui dans le cas idéal devrait se situer autour de 1.00. Mais nous pouvons continuer à utiliser ces items en constatant : 1) que la valeur z « Infit ZSTD », qui devrait se trouver entre -2.00 et +2.00, est bon pour les trois items ; 2) que les valeurs du « Point Measure » sont supérieures à .20 ; 3) et que les valeurs qui équilibrent les différences de niveau entre les sous-tests (PT-Measure Exp.) sont également bonnes.
- Pour l'item n. 63 (LV2-04), la valeur basse du « Measure » (-1.53) est surprenante. Dans ce cas précis, il a paru intéressant d'examiner de plus près.

Le graphique des « Expected Score ICC » pour la « Entry Number 63 » (LV2-04) montre que les étudiants avec un niveau bas ont réussi à résoudre la tâche. Cela est surprenant, car le texte de cet item de compréhension écrite contient beaucoup de mots qui ne sont probablement pas compris à ce niveau (cf. Figure 2). Nonobstant cela, l'item a pu être résolu correctement.

Figure 2 : DaF Item LV

Avec les bonnes stratégies, cet item a pu être résolu par des étudiants avec un niveau bas, car il suffit de retrouver les mots phare « Umgebung » et « je … desto » dans le texte. De plus, les distracteurs 1 et 2 peuvent probablement être identifiés comme faux sur la base de connaissances générales. Cet item ne peut être considéré comme valide car il devrait tester la compréhension, et pas les stratégies évoquées. Ce qui est enseigné en tant que stratégie de lecture et compréhension pendant les cours de langue se montre problématique dans une situation de test de classement. Ici, des règles centrales du développement d'items n'ont pas été respectées. Une reformulation des distracteurs aurait été possible. Il a cependant été décidé de ne pas réutiliser cet item pour les tests à venir. L'analyse de Rasch en elle-même n'a pas reconnu cet item comme problématique. Par contre, uniquement l'écart entre le niveau évalué par une didacticienne de langue étrangère et le degré de difficulté effectivement calculé nous a rendus attentifs à cet item.

5. Modèle de Rasch et discussion pour le développement des tests

Les tests se développent en procédant à une analyse continue de la validité des items des sous-tests. Pour ce faire, nous voulons établir une analyse en boucle composée des étapes suivantes :

- Une analyse à l'aide du modèle de Rasch sert à identifier des items problématiques.
- L'équipe de conception des items discute ceux-ci et propose des améliorations.
- Les améliorations sont implémentées dans le test pour la session suivante.
- Le modèle de Rasch sert à vérifier si ces changements portent leurs fruits.
- L'équipe de conception discute des résultats obtenus.

D'autres analyses, menées en parallèle (satisfaction, changements de classe, …), servent à garantir que le test continue à fonctionner suite à ces modifications et à identifier les points avec une marge d'amélioration (Blons-Pierre et al., 2012 ; Gick, 2012).

Nous voulons mettre en place cette procédure en boucle pour améliorer les items des tests de façon continue. Actuellement nous sommes en pleine phase d'implémentation d'une première boucle. Nous avons fait une analyse des items, identifié les items problématiques, et analysons ces items dans l'équipe de conception, pour en faire ressortir les améliorations qui seront implémentées dans le test dès le prochain semestre.

5.1 Identifier les items potentiellement problématiques

Le but de notre première étape est d'identifier quelques items problématiques qui seront analysés et éventuellement modifiés. Comme nous l'avons vu plus haut, nos tests de classement remplissent déjà leur fonction, et nous voulons évidemment que cela reste ainsi. La procédure d'amélioration se faisant sur le test de classement en utilisation (pour générer des réponses en situation réelle), nous limitons donc la proportion d'items modifiés entre deux passations. Cela permet au test, qui a fait ses preuves sur le terrain, de continuer à atteindre son but (placer les apprenants dans des classes le plus homogènes possibles), tout en permettant d'améliorer les items.
L'analyse de Rasch nous aide à identifier les items potentiellement à modifier. Nous utilisons ici le modèle de Rasch de façon pragmatique : il nous sert à identifier les items problématiques mais ne nous sert pas pour l'identification de la source du problème. Pour identifier ces items problématiques, nous utilisons principalement deux représentations des résultats proposés par le logiciel Winsteps : l' « item : map» et le tableau de sortie selon Item. Nous regardons d'abord toutes les données ensemble, puis nous les séparons selon les compétences testées.

Les graphiques des « Item : Map » donne un premier aperçu de la position des items les uns par rapport aux autres selon leur degré de difficulté. Nous regardons si des items se placent en dehors de la « zone » de difficulté pour laquelle ils ont été créés. Idéalement, avec la structure adaptative de nos tests, nous aimerions trouver, du bas vers le haut, d'abord les items des tests des niveaux A, ensuite, au milieu, ceux des niveaux B et, tout en haut les items des niveaux C. Si la majeure partie des items se place bien dans cette structure, on en trouve aussi qui sont placés plus haut ou plus bas que leur niveau idéal. Ces items sont retenus pour la discussion dans l'équipe (étape 2).

Avant cela, ces items sont regardés dans les tableaux d'analyse par item pour savoir s'ils sont également identifiés comme problématiques selon d'autres facteurs (p.ex. Pt.Measure, Infit, Outfit) et la courbe de score attendu (Expected Score ICC). Cela est souvent le cas. Ces tables de résultat permettent également de ressortir des Items n'attirant pas l'attention dans l'item map, mais pour lesquels d'autres indications font penser qu'ils sont problématiques : p.ex. un Pt.Measure bas, un Infit ou Outfit élevé (ou bas). « Measure », « Point Measure » (Wright et Stone, 1979). Nous identifions ainsi un certain nombre d'autres items potentiellement problématiques, qui seront soumis à l'équipe de conception des items.
Par exemple, pour les données du test FLE des deux semestres 2011 (nous avons regroupé ces deux semestres pour obtenir un nombre suffisant de réponses), nous identifions une quinzaine d'items sur un total de 210.

5.2 Identification des problèmes, proposition et implémentation des modifications

Dans cette phase, l'équipe de conception des items va :

- Regarder les réponses des étudiants à chaque item potentiellement problématique (résultats des tentatives dans Moodle)
- Discuter et poser des hypothèses sur : Est-ce que il y a effectivement un problème avec cet item ? Si oui : quelles pourraient en être les causes ? Si non, en quoi l'analyse l'identifie comme problématique ?
- Proposer des possibles améliorations pouvant aller du changement d'une virgule à l'élimination/substitution complète de l'item.
- Décider des items à changer en cherchant le maintien du fonctionnement des tests.

Nous sommes actuellement dans la phase d'identification des problèmes et de proposition de modifications. Pour le test FLE nous avons par exemple commencé une discussion sur certains items. Tous les items problématiques ont été ou seront discutés de cette façon, et l'équipe décidera lesquels modifier. Voici quelques exemples de discussion :

Un exemple d'item de compréhension écrite :
- Question : Elle s'intéresse à la planche à voile depuis la pré-adolescence.
- Réponses : Vrai, Faux, La réponse n'est pas dans le texte.
- Analyse de Rasch : Cet item a un pt. measure très bas, et un infit haut (avec zsdt haut), la courbe ICC montre également un problème.

Hypothèse formulée :
- Problème avec formulation/interprétation: « LT monte première fois » dans le texte à lire n'est pas égale à « s'intéresse » dans la question.
- Possible solution: reformuler « intéresse ». Changer avec : « ****** à la planche à voile ».

Ce changement va être implémenté.

Un exemple d'item de lexique et structure :
- Question : Il s'agit d'une phrase dans laquelle on cherche un accord verbal.
- Réponses : a choisi, choisira, ait choisi, avait choisi.
- Analyse de Rasch : Cette question a un pt. measure bas, et sa courbe montre également un problème. La distribution des réponses sur les distracteurs donne un indice : la majorité des personnes choisissent une même mauvaise réponse (distracteur 3).

Hypothèse formulée :
- Item trop difficile, distracteur interprété comme juste.
- Solutions possibles : a) Il s'agit d'un test de niveau élevé, donc il doit être difficile. b) Il est trop difficile, on pourrait changer le distracteur en « choisissait ».

Ce changement demande qu'il soit encore discuté avant d'être implémenté.

5.3 Analyse avec les items modifiés

Lorsque les modifications seront implémentées dans les tests des prochaines sessions, nous pourrons collecter une nouvelle série de données avec lesquelles nous effectuerons une nouvelle analyse sur la base du modèle de Rasch. Nous pourrons ensuite regarder avec l'équipe si la modification a porté ses fruits. Si oui, on aura appris que l'hypothèse sur le problème de l'item était fondée. Autrement nous pourrons décider si une modification additionnelle s'impose ou si l'item doit être substitué. Dans les deux cas on aura un indice de l'efficacité des modifications proposées.

6. Conclusions et perspectives

L'analyse de Rasch nous apparaît utile pour la conception et le développement de tests adaptatifs. Elle contribue à la transformation rapide, pour les tests DaF et FLE, d'un test papier progressif en un test adaptatif en ligne valide en se servant de données récoltées en situation réelle. Dans une phase ultérieure, cette analyse soutient également l'optimisation des tests en visualisant les degrés de difficultés des items les uns par rapport aux autres. De plus elle donne des indices sur la validité de chaque item individuel.

Cette analyse comporte aussi des limites. Notre expérience montre notamment le besoin de continuer les analyses quantitatives tout en y associant d'autres méthodes plus qualitatives pour développer des tests de classement en langues étrangères adaptatifs. Ainsi, l'analyse de Rasch ne permet pas de détecter tous les items « problématiques ». Le regard des experts en didactique des langues étrangères reste incontournable. En effet, les réflexions de ces experts sur les items

problématiques permettent d'expliciter les problèmes liés à certains items. Pour tirer avantage des forces des deux méthodes, nous avons décidé de mettre en place une procédure en boucle, alliant analyse de Rasch et regard des experts.

Les premières expériences avec cette procédure portent leurs fruits. Nos perspectives de travail sont de la rendre plus systématique. Ensuite, il s'agira d'effectuer un bilan de la procédure, en termes des choses apprises sur le développement des items par les équipes des tests (formation continue au sein du Centre de Langues).

7. Bibliographie

Blons-Pierre, C. (2010). *Tensions between change and quality assurance : A dynamic necessary for innovation in university language centres/Tensions entre politique du changement et assurance qualité : une dynamique nécessaire pour l'évolution des centres de langues universitaires.* Communication présentée à la 11th International CercleS Conference, Helsinki, 4 September.

Blons-Pierre, C. (2011). *L'impact du CECR/GeR/CEFR sur le fonctionnement d'un centre de langues dans une université bilingue suisse.* Communication présentée à la 4th International ALTE Conference, 7-9 July.

Blons-Pierre, C., Kohler, P., Gick, C., Hoein, S. & Platteaux, H. (2012). Analyse qualitative des tests de classement en ligne pour les langues : acceptance et rôle de l'autoévaluation. Dans *Actes du XXIVème Colloque de l'ADMEE Europe. L'évaluation des compétences en milieu scolaire et en milieu professionnel*, Luxembourg 11-13 janvier.

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

Gick, C. (2011). *Online-Einstufung und Online-Einschreibung in studienbegleitende Deutschkurse mittels adaptivem Einstufungstest: Entwicklungsschritte und Stand der Dinge.* Communication présentée au Bremer Symposion 2011 (Fremdsprachenzentrum der Hochschulen im Land Bremen), Bremen 4-5 März.

Gick, C. (2012). Einstufungstests im Spannungsfeld von Bologna-Reform, Referenzrahmen für Sprachen, Fachwissenschaft und Nutzern, In C. Blons-Pierre (dir.) : *Apprendre, enseigner et évaluer les langues dans le contexte de Bologne et du CECR / Sprachen lernen, lehren und beurteilen im Kontext von Bologna und dem GER*. Bern : Peter Lang, 189-217.

Kohler, P. (2012). Gestion de l'auto-évaluation dans un dispositif de test en ligne: limites et perspectives. In C. Blons-Pierre (dir.), *Apprendre, enseigner et évaluer les langues dans le contexte de Bologne et du CECR / Sprachen lernen, lehren und beurteilen im Kontext von Bologna und dem GER* (pp. 219-240). Bern : Peter Lang.

Kohler, P., Platteaux, H., & Blons-Pierre, C. (2010). *Réactions des étudiants à l'usage des tests en ligne pour évaluer les compétences langagières.* Communication présentée au XXVIème Congrès AIPU. Réformes et changements pédagogiques dans l'enseignement supérieur, Rabat 17-21 mai.

Kohler, P., Platteaux, H., & Blons-Pierre, C. (2012). Un test de classement en ligne pour évaluer les niveaux de compétence et constituer des groupes classes. *Revue RIPES - Numéro Spécial « Innover dans l'évaluation des apprentissages : pourquoi et comment ? », 27*(2), Récupéré du site de la revue : http://ripes.revues.org/532.

Linacre, J. M. (2000). *Computer-Adaptive Testing: A methodology whose time has come*. Chicago: MESA Psychometric Laboratory - University of Chicago.

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

Platteaux, H., & Hoein, S. (2010). *Tests de positionnement en langues étrangères sur Moodle.* Communication présentée à la Conférence Moodle Moot 2010, Troyes 28-30 juin.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Danish Institute for Educational Research Copenhagen 1960, expanded edition with foreword and afterword by B.D. Wright. Chicago: The University of Chicago press.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Zweite, überarbeitete und erweiterte Auflage, Bern, Göttingen, Toronto, Seattle: Hans Huber.

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

Wright, B. D., & Stone, M. H. (2004). *Making measure*. Chicago: The Phaneron Press.