répondre aux défis de la révolution civilisationnelle provoquée par la présence des technologies numériques et leur expansion irrémédiable dans nos pratiques culturelles, sociales, économiques, politiques et pédagogiques quotidiennes. Le but de cette communication sera donc de stimuler le débat sur la question des relations mutuelles entre révolution numérique et évaluation en contexte scolaire, en termes d'innovations évolutives ou disruptives (Christensen, Johnson & Horn, 2008).

Références

Christensen, C. M., Johnson, C. W., & Horn, M. B. (2008). *Disrupting class*. New York : McGraw-Hill Professional Publishing.

Cuban, L. (2003). *Oversold and Underused : Computers in the Classroom*. Cambridge, Mass.: Harvard University Press.

Dräger, J., & Müller-Eiselt, R. (2015). *Die digitale Bildungsrevolution : Der radikale Wandel des Lernens und wie wir ihn gestalten können* (3 edition). München : Deutsche Verlags-Anstalt.

Puentedura, R. R. (2006). *Transformation, Technology, and Education*. Retrieved October 28, 2017, from http://hippasus.com/resources/tte/

Resnick, M. (2008). Sowing the Seeds for a More Creative Society. *Learning & Leading with Technology*, *35*(4), 18–22.

Serres, M. (2015). *Petite poucette*. Paris : Le pommier.

Trilling, B., & Fadel, C. (2009). *21st century skills : Learning for life in our times*. New Jersey : John Wiley & Sons.

Quand le numérique défie la mesure. Comment veiller à la qualité de certifications en langue professionnelle au format numérique ? (7602)

Dominique Casanova, Alhassane Aw* & Marc Demeuse***
**Chambre de commerce et d'industrie de Paris Île-de-France, France*
***Université de Mons, Belgique*

Mots-clés : tests informatisés, dimensionnalité, modélisation psychométrique

Introduction

Dans des sociétés de plus en plus connectées, où de très nombreuses personnes sont en contact avec des outils numériques pour l'échange d'informations, la tentation est grande de s'appuyer sur les possibilités offertes par le numérique dans le domaine de l'évaluation. Toutefois, l'introduction des outils numériques dans l'évaluation soulève inévitablement des questions concernant le construit du test, la standardisation de sa diffusion et la mesure de la compétence testée.

Pour les tests informatisés à grande échelle dans le domaine des langues, qui souvent existaient également au format papier, un des enjeux était de garantir l'équivalence entre la version électronique et la version papier-crayon (Mead et Drasgow, 1993 ; Houssemand et al., 2009 ; Casanova et al., 2011). Cela a souvent conduit à une utilisation *a minima* des possibilités numériques, pour que les deux versions restent les plus proches possibles et que les candidats puissent parcourir aisément le test au moyen d'une souris. L'objectif était de limiter l'altération du construit du test par les compétences numériques et l'apparition d'une variance non souhaitée dans les scores.

Plusieurs concepteurs de tests de langue ont fait le choix de maintenir l'épreuve d'expression écrite au format papier-crayon, de crainte que les différences de familiarité avec l'utilisation d'un clavier (et, dans le cas du français, les différences d'accessibilité des caractères accentués selon les claviers) affectent les résultats des candidats. Bennett (2003) fait état d'études qui montrent que des variations peuvent être constatées dans le cas d'épreuves de production écrite, dont une des sources supposées est la familiarité avec l'utilisation d'un clavier.

A l'inverse, Laurier et Diarra (2009) relatent des expérimentations qui montrent que les élèves habitués à utiliser l'ordinateur pour rédiger leurs écrits ont de meilleures notes lorsqu'ils sont soumis à des épreuves utilisant le traitement de texte que dans une évaluation en mode papier-crayon (notamment Russell et Haney, 2000). En effet, la manière d'écrire diffère à partir du moment où l'ordinateur permet une restructuration en continu du texte avec des fonctionnalités comme le copier-coller (Diarra, 20012). L'usage de l'ordinateur étant de plus en plus répandu, les candidats et la société civile s'étonnent de la persistance d'évaluations papier/crayon.

Dans la cadre de la refonte de ses certifications de français professionnel, la CCI Paris Ile-de-France a décidé de proposer les épreuves de la compétence « Comprendre et traiter de l'information » exclusivement par voie numérique et d'exploiter une variété de formats items (choix dans liste, glisser-déposer, boîtes de saisie…). Ce recours à l'outil numérique lui permet de proposer une évaluation par tâches plus réaliste, reflétant davantage les processus cognitifs mis en œuvre dans la réalisation d'activités langagières au sein de l'entreprise et qui sont souvent difficilement modélisables en format papier/crayon sans avoir recours à une évaluation humaine.

Ces certifications s'adressent à une population d'étudiants et de salariés, dont la familiarité avec l'outil informatique, aujourd'hui difficilement contournable dans le monde de l'entreprise, est un prérequis. Les candidats passent en général ces certifications à l'issue d'une formation durant laquelle ils ont la possibilité de se familiariser avec les tâches proposées. Un tutoriel interactif leur est également proposé en libre accès (https://www.lefrancaisdesaffaires.fr/ressources/les-tutoriels-d-entrainement/tutoriels-dfp/). Les productions écrites doivent être réalisées sur ordinateur, en correspondance avec la majorité des situations dans lesquelles les professionnels sont amenés à rédiger un écrit.

Il convient néanmoins de s'assurer que les différentes modalités de réponse proposées (formats d'items) n'introduisent pas une variance non souhaitée dans les résultats, certains candidats pouvant être moins à l'aise avec certaines modalités. Ce peut être le cas notamment pour les items au format glisser-déposer, pour lesquels la modalité de réponse peut apparaître moins intuitive que le choix d'une option dans une liste, et qui est susceptible d'introduire une seconde dimension dans le test.

L'évaluation par tâches soulève une autre question, qui est l'identification d'un modèle de mesure approprié pour rendre compte des propriétés psychométriques du test et constituer une banque calibrée d'activités réutilisables. Les activités à correction automatique du Diplôme de français professionnel Affaires B1, qui fait l'objet de cette étude, s'appuient en effet sur un ou plusieurs documents supports (graphiques et/ou écrits et/ou oraux) à partir desquels les candidats doivent compléter en plusieurs endroits un document de réponse (formulaire, tableau, commentaire, courriel…). Il y a donc plusieurs « items » se rapportant à un même document, ce qui est susceptible d'introduire une dépendance entre les réponses à ces items. Or, un des postulats de la théorie classique des tests est que la corrélation entre les erreurs aux différents items vaut zéro (Demeuse et Henry, 2004) et l'indépendance locale est une des conditions d'application des modèles de réponses à l'item (Grondin et al., 2017).

Dans cette étude, menée sur les résultats à la première version du Diplôme de français professionnel Affaires B1, nous nous sommes attachés à vérifier la présence éventuelle d'une seconde dimension

induite par les activités au format glisser-déposer et à identifier un modèle de mesure pouvant s'appliquer aux données recueillies.

Le diplôme de français professionnel Affaires B1

Le diplôme de français professionnel Affaires vise à certifier le niveau de compétence en français des personnes qui souhaitent exercer des tâches de communication professionnelles. Il s'agit d'un examen ancré dans des pratiques professionnelles et qui s'adresse aux étudiants ou professionnels qui travaillent ou seront appelés à communiquer en français dans un contexte professionnel (francophone ou non) et qui souhaitent valider leurs acquis par un diplôme en référence à un niveau donné du Cadre Européen Commun de Référence pour les langues – CECR (Conseil de l'Europe, 2001).

L'adéquation du diplôme aux réalités professionnelles transparait non seulement dans le choix des documents supports, l'authenticité de leur forme et de leur contenu, mais également dans le caractère réaliste des mises en situation et des tâches de communication à réaliser indexées sur le CECR. En conséquence, les activités proposées placent toujours les candidats dans la situation d'acteurs du monde des affaires en relation avec les différents interlocuteurs de l'entreprise (collègues de travail, responsables hiérarchiques, services internes et fournisseurs, clients et prospects, etc.).

Les activités du diplôme renvoient à des situations de communication transversales, communes aux domaines d'activité les plus courants du monde de l'entreprise et des affaires : ressources humaines, management, marketing, finances, logistique, etc.

L'évaluation porte sur des compétences intégrées. La tâche de communication réalisée par le candidat prend la forme d'une production (écrite ou orale) conditionnée par la compréhension de documents professionnels (écrits ou oraux) et par la sélection des informations nécessaires à la réalisation de la tâche. L'évaluation s'inscrit également dans une démarche actionnelle (Richer, 2014), c'est-à-dire que le candidat doit réaliser, dans chaque activité du diplôme, une tâche de communication professionnelle définie par un contexte (situation professionnelle) et un ou plusieurs objectifs de communication. Pour réaliser cette tâche, il doit mobiliser, de manière stratégique, ses compétences de réception, de médiation, de production et/ou d'interaction. Il ne s'agit plus de lire/écouter pour comprendre, mais de comprendre (ce qu'on lit/écoute) pour agir. Ainsi, l'évaluation porte non seulement sur la maîtrise des moyens langagiers mais aussi sur le degré de réalisation de la tâche professionnelle. Quel que soit le niveau du diplôme choisi, les deux mêmes compétences sont évaluées :

- Comprendre et traiter de l'information
- Interagir à l'oral

La première compétence correspond à des tâches où l'interaction est en temps différé et dont le contenu des échanges est davantage contrôlé. Le candidat construit seul son discours, sur la base des messages à traiter et des consignes de réalisation et en ne perdant pas de vue son interlocuteur qui n'est cependant pas incarné et qui n'intervient pas directement dans l'échange. C'est sa capacité à traiter une variété et/ou une masse d'information, à la mettre en relation et à produire un discours en respectant des contraintes qui est évaluée.

La seconde compétence se distingue par sa dimension interpersonnelle et interactive « en temps réel », avec nécessité d'adapter son discours et son attitude aux réactions de l'interlocuteur. Le candidat a par ailleurs une plus grande latitude dans sa production langagière : il y a une plus grande variété de productions correctes possibles, plus de créativité, plus de stratégies mobilisables, comme le recours au non verbal et/ou à l'accentuation du discours.

Le tableau 1 présente les activités de la compétence « Comprendre et traiter de l'information » proposées pour le diplôme de niveau B1.

Tableau 1 : activités de la compétence « Comprendre et traiter de l'information » du Diplôme de français professionnel Affaires B1

Habiletés	Activités	Modalités de réponse	NB réponses attendues
Traiter l'information écrite	1 : Commenter un graphique	Choix dans listes	5
	2 : Apporter une réponse adaptée dans une situation problématique	Glisser-déposer	10
	3 : Réserver un espace d'exposition sur un salon, en tenant compte des instructions données	Choix dans listes	10
	4 : Compléter une fiche récapitulative de projet, établir des conclusions opérationnelles à partir des informations données	Glisser-déposer	12
Traiter l'information orale	5 : Organiser ses notes	Glisser-déposer	5
	6 : Transmettre la teneur du message d'un client et des instructions à un collègue	Choix dans listes	8
	7 : Rédiger un courriel de réponse à la demande, en tenant compte d'informations complémentaires	Rédaction libre	1
Interagir à l'écrit	8 : Rédiger une lettre de candidature	Rédaction libre	1

Les six premières activités sont à correction automatique et une pondération est utilisée de sorte que le score maximum de chaque activité soit identique. Les modalités de réponse des activités 1, 3 et 6 consistent en des choix d'option dans des listes (les listes sont différentes pour chaque item) alors que la réponse aux activités 2, 4 et 5 s'effectue sous forme de glisser-déposer. Pour l'activité 2, il s'agit de glisser-déposer chaque option pertinente (il y a aussi des distracteurs) dans une des trois rubriques d'un tableau, alors que pour l'activité 4, il y a une zone de destination spécifique à chacune des options pertinentes. Pour l'activité 5, les candidats utilisent les glisser-déposer pour ordonner après sélection les options pertinentes.

Le tableau 2 présente les activités de la compétence « Interagir à l'oral » pour le diplôme de niveau B1. Ces activités prennent la forme de jeux de rôles en présentiel où l'examinateur est l'interlocuteur du candidat. Les échanges sont enregistrés au moyen d'une application mobile, sur laquelle l'examinateur reporte le résultat de son évaluation à la fin de la passation. L'ensemble des informations est ensuite transmis automatiquement au système d'information du Français des affaires de la CCI Paris Ile-de-France.

Tableau 2 : activités de la compétence « Interagir à l'oral » du Diplôme de français professionnel Affaires B1

Activités	Modalités de réponse
Présenter le parcours de candidats à un poste et argumenter son choix auprès de la direction	Interaction orale avec un examinateur (jeu de rôles)
Argumenter auprès d'un décideur / d'un client lors d'un entretien/d'une vente	Interaction orale avec un examinateur (jeu de rôles)

Dans le cadre de cette étude, nous avons formulé deux hypothèses que nous avons cherché à vérifier à partir des données d'évaluation de la première version du diplôme :

- Hypothèse 1 : compte-tenu de la nature du public visé (étudiants, professionnels en activité, passant l'examen à l'issue d'une formation), en dépit de la différence des formats d'items utilisés, le sous-test constitué des items à correction automatique peut être considéré comme unidimensionnel.
- Hypothèse 2 : il est possible d'identifier un modèle de réponse à l'item approprié pour le traitement du sous-test constitué des items à correction automatique en vue de la constitution d'une banque d'activités et de versions équivalentes du diplôme.

Description des données

L'échantillon à notre disposition comportait les réponses de 192 individus, 64 % de femmes et 36% d'hommes. 57% d'entre eux avaient entre 19 et 26 ans et 94% entre 16 et 50 ans. 46% d'entre eux ont déclaré une motivation académique à leur inscription, 41% une motivation professionnelle et 13% une motivation individuelle. Les épreuves ont été organisées dans 18 pays, principalement aux États-Unis (28% des candidats), en France (14%), en Suisse (14%) et en Italie (12%). Les principales langues maternelles des candidats étaient l'anglais (30%), l'allemand (15%), l'italien (13%) et l'arabe (10%), parmi un total de 27 langues maternelles représentées.

Un quart des candidats a répondu en début de test à un questionnaire concernant leur profil. 51% se sont déclarés élèves ou étudiants, 43% professionnels en activité. 73% ont déclaré un niveau d'étude « enseignement supérieur », 11% « enseignement secondaire 2nd niveau » (soit l'équivalent du lycée en France), 9% « enseignement secondaire 1er niveau » (l'équivalent du collège en France) et 7% « enseignement primaire ». Les candidats se répartissent relativement uniformément (entre 6 et 12%) selon une liste de 8 domaines d'activité pré-identifiés, 26% d'entre eux ayant choisi la catégorie « Autre ». 29% des candidats ont déclaré n'avoir aucune expérience professionnelle, 25% une expérience de moins d'une année, 27% entre 1 et 4 années d'expérience et 27% plus de 4 années d'expérience. Dans 40% des cas, leur objectif principal était de « valider un parcours de formation en français professionnel », dans 36% des cas « faire reconnaître leurs compétences en français professionnel », dans 17% des cas « favoriser leur insertion dans le monde du travail » et dans 7% des cas « préparer une mobilité professionnelle ». Le prescripteur de l'examen était dans 45% des cas une école ou une université, dans 45% des cas une administration et pour les 10% restants, il s'agissait d'une démarche personnelle.

Les données ont été analysées au moyen du logiciel jMetrik (Meyer, 2014), qui fournit un ensemble étendu de fonctionnalités prêtes à l'emploi et accessibles par interface graphique pour l'analyse d'items, tant dans le cadre de la théorie classique des tests que pour une analyse avec des modèles de réponse à l'item. La figure 1 représente la répartition du score brut des candidats pour les activités à réponse automatique (50 items).

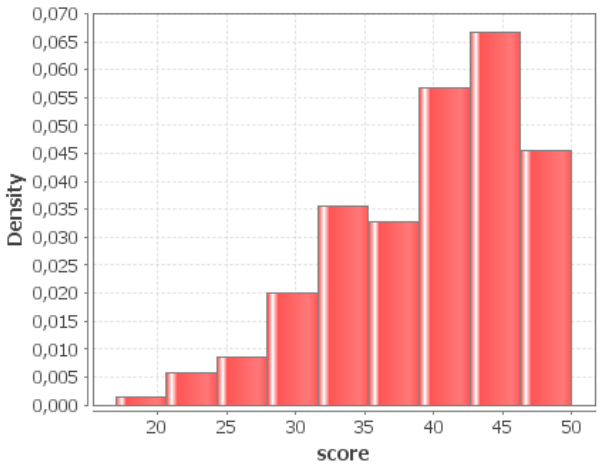

Figure 1 : distribution du score brut des candidats pour les activités à correction automatique

La moyenne des candidats est de 39,6 points (écart-type de 6,8 points), soit 79,2%, sachant que le seuil de réussite pour cette version du questionnaire était fixé à 35 points. La facilité des items (valeur p) de ce questionnaire (pour cet échantillon) varie entre 0,52 et 0,98, avec une moyenne de 0,78 et un écart-type de 0,13. La consistance interne du questionnaire (sous l'hypothèse d'une absence de corrélation entre les erreurs des différents items) peut être estimée au moyen d'un alpha de Cronbach à 0,85 et l'erreur-type correspondante, liée à l'échantillonnage des items, est de 2,68 points.

Aspects de dimensionnalité

Pour analyser la dimensionnalité de l'examen, nous avons considéré le score obtenu par les candidats à chacune des activités et mené une analyse factorielle multiple à partir de ces sous-scores, au moyen de la librairie *FactoMineR*. La figure 2 présente le diagramme en éboulis de l'inertie de chacune des dimensions de l'analyse.

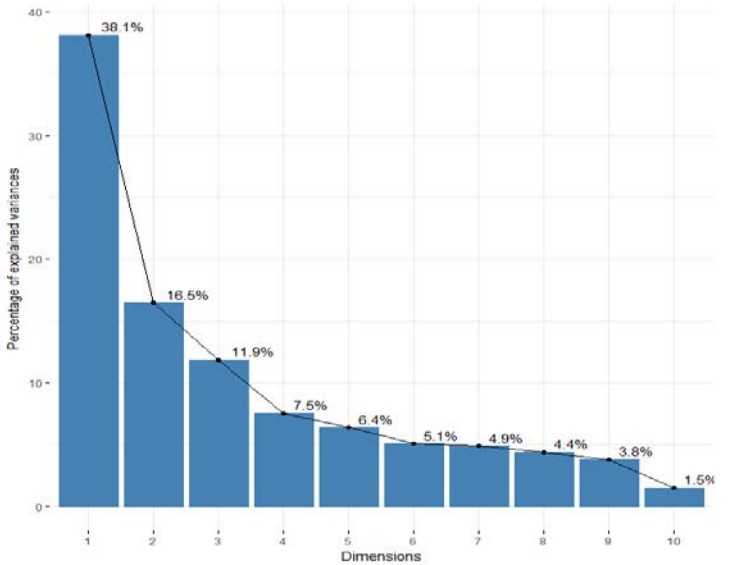

Figure 2 : pourcentage d'inertie expliquée par chacune des dimensions de l'analyse factorielle

On constate trois dimensions dominantes, expliquant à elles seules 66,5% de la variance des scores (38,1% pour la première d'entre elles). Le cercle de corrélation de la figure 3 montre que la première dimension est une dimension commune à laquelle contribuent chacune des activités. Cette dimension peut être interprétée comme la compétence à communiquer en français en contexte professionnel. La seconde dimension oppose clairement les activités de la compétence *Interagir à l'oral* (IO_1 et IO_2) de celles de la compétence *Comprendre et traiter de l'information*. Cela est conforme à la structure de l'examen et à la particularité mentionnée pour la compétence Interagir à l'oral, qui se distingue par sa dimension interpersonnelle et interactive « en temps réel », avec nécessité d'adapter son discours et son attitude aux réactions de l'interlocuteur.

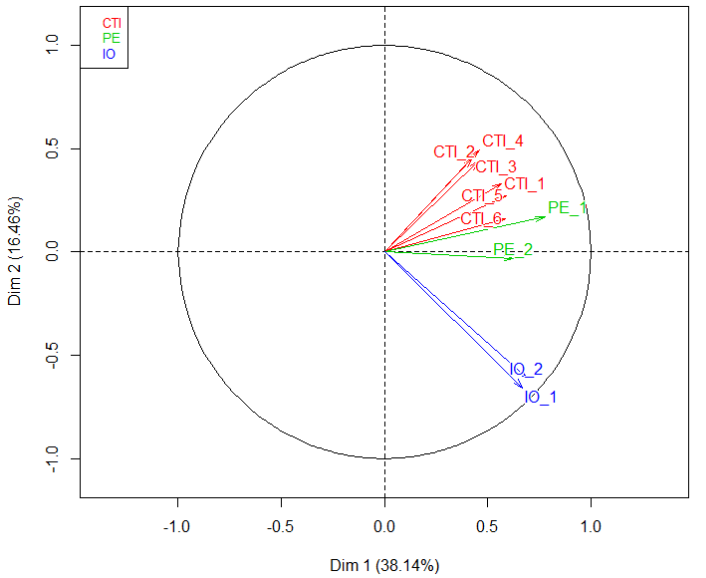

Figure 3 : Cercle des corrélations pour les deux premières dimensions de l'analyse factorielle

La troisième dimension met en opposition les activités à correction automatique (activités 1 à 6) et les activités conduisant à la rédaction d'une production écrite (activités 7 et 8) de la compétence « Comprendre et traiter l'information ». Pour mieux analyser cette opposition, nous avons procédé à une nouvelle analyse factorielle multiple, en nous limitant aux scores obtenus aux activités de la compétence « Comprendre et traiter l'information ». Deux dimensions principales se dégagent, qui expliquent 58% de la variance des scores. Le cercle de corrélation de la figure 4 montre que la seconde dimension oppose les activités de production écrite (PE_1 et PE_2) aux activités à correction automatique, qui sont par ailleurs bien regroupées entre elles.

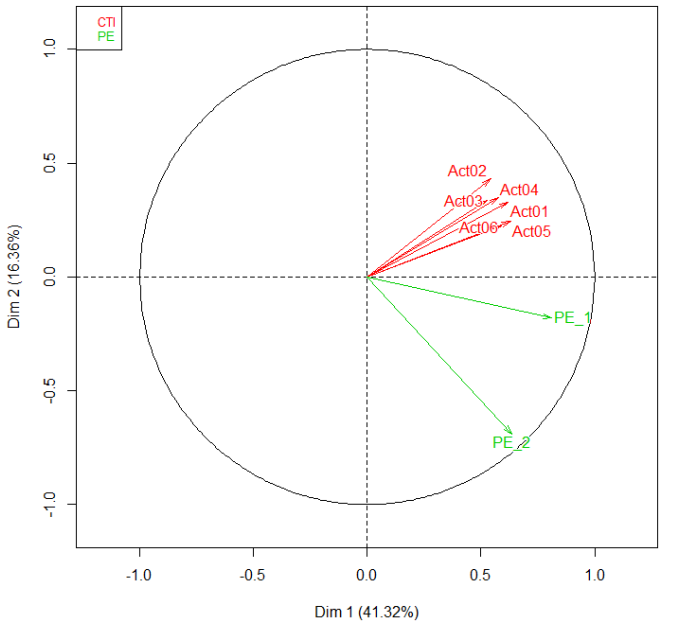

Figure 4 : Cercle des corrélations pour les activités de la compétence « Comprendre et traiter l'information »

Cette opposition concerne surtout la seconde production écrite et est peut-être due à un artefact. En effet, dans cette première version du test, le temps était partagé entre l'ensemble des activités et une partie des candidats est visiblement arrivée à cours de temps pour la seconde production écrite. Les résultats sont sensiblement plus faibles pour cette activité et les corrélations des scores de cette activité avec les scores de chacune des activités à correction automatique sont plus faibles que dans le cas de la première production écrite.

Une troisième analyse factorielle multiple, limitée aux scores obtenus aux activités à correction automatique permet d'apporter une première réponse à l'hypothèse 1. Cette analyse factorielle met en évidence un facteur dominant (44,1% de variance expliquée) et un second facteur explique 14,7% de la variance, les autres dimensions ayant une inertie comparable. Le cercle des corrélations montre que ce second facteur met en opposition les activités 2 et 3, d'une part, et 5 et 6, d'autre part. Or, ce qui oppose ces activités n'est pas la modalité de réponse aux items, mais la nature du document support principal, qui est écrit pour les activités 2 à 4 et oral pour les activités 5 et 6 (pour l'activité 1 il s'agit d'un graphique). Cela montre qu'il serait trop restrictif de vouloir évaluer la compétence à comprendre et à traiter de l'information en se satisfaisant d'un seul type de support. En comparaison, les différences de modalités de réponse ne semblent pas, pour le public concerné, avoir un impact notable sur les performances des candidats.

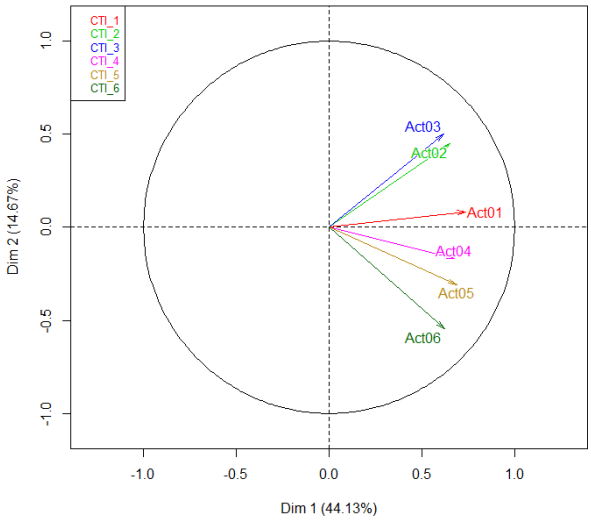

Figure 5 : Cercle des corrélations pour les activités à correction automatique

Une autre façon de questionner la dimensionnalité des activités à correction automatique est d'appliquer un modèle de Rasch aux données (en considérant cette fois-ci les scores par item) et de mener une analyse en composantes principales sur les résidus standardisés. Cela permet notamment de tester l'hypothèse d'unidimensionnalité qui est une des conditions d'application du modèle. L'analyse menée avec le logiciel jMetrik, en considérant les items comme dichotomiques, met en évidence 5 facteurs principaux qui expliquent la variance résiduelle, mais de manière comparable, ce que résume le tableau 3.

Tableau 3 : Inertie des facteurs principaux de l'analyse en composantes principales des résidus standardisés (jMetrik)

	F1	F2	F3	F4	F5
Eigen value	2,84	2,36	2,11	2,02	1,88
Proportion Var	0,06	0,05	0,04	0,04	0,04
Proportion Explained	0,25	0,21	0,19	0,18	0,17

Ce sont des items de l'activité 2 qui saturent le plus sur le premier facteur, en opposition avec des items de l'activité 5 et 6, dans une moindre mesure. Or ces deux activités sont toutes deux des activités au format glisser-déposer. L'activité 2 se rapporte cependant à un support principal écrit alors que l'activité 5 se rapporte à un support principal oral. Encore une fois, c'est davantage la différence de canal entre documents supports (écrit versus oral) qui est susceptible d'introduire une seconde dimension dans le test que les différences de modalité de réponse, du moins pour le public considéré. Les données d'ajustement des items au modèle de Rasch sont par ailleurs satisfaisantes, à de rares exceptions près.

Modèle de mesure

L'un des objectifs de la CCI Paris Ile-de-France est de constituer une banque d'activités calibrée pour favoriser la réutilisation des activités dans de nouveaux questionnaires de difficulté comparable et

faciliter la définition des points de césure en s'appuyant sur les caractéristiques empiriques des items. La théorie de réponses à l'item fournit un cadre approprié pour une telle entreprise, comme Hambleton et Swaminathan l'ont bien montré, dès 1985. Sa mise en œuvre nécessite toutefois de veiller au respect de conditions d'application que sont, en reprenant les définitions de Laveault et Grégoire (2014) l'unidimensionnalité (« tous les items doivent mesurer un seul et même trait ») et l'indépendance locale (« le trait qui fait l'objet de l'évaluation doit être le seul facteur qui détermine la variabilité des réponses aux items d'un test »). Compte-tenu de la taille de notre échantillon, nous privilégions l'utilisation d'un modèle de Rasch (Penta et al., 2005), qui ajoute une contrainte supplémentaire : la capacité discriminatoire des items[76] doit être comparable.

Les résultats de l'analyse en composantes principales des résidus standardisés, présentée en fin de section précédente, nous rassure sur l'unidimensionnalité du questionnaire. L'analyse ne met pas en évidence un facteur qui expliquerait de façon prépondérante la variance des résidus standardisés. Une vérification plus approfondie de cette hypothèse mériterait toutefois d'être entreprise en mobilisant des techniques plus spécifiques, comme l'analyse factorielle non-linéaire ou l'utilisation de la procédure DIMTEST (Laveault et Grégoire, 2014, p. 295), qui ne sont pas disponibles dans jMetrik.

Un moyen de tester l'hypothèse d'indépendance locale est d'observer les corrélations entre les résidus des candidats (Yen, 1984 ; Yen, 1993), ce que permet de faire aisément jMetrik. L'analyse des données du diplôme de français professionnel Affaires B1 montre (en considérant comme dépendants les items pour lesquels la corrélation entre résidus est supérieure ou égale à 0,25) une dépendance locale entre plusieurs items pour l'activité 2, l'activité 3 et une dépendance forte entre items pour l'activité 5. Cela confirme qu'en faisant porter différents items sur un même (ensemble de) document(s) support(s), il y a un risque élevé d'introduire une dépendance entre items. L'importance de cette dépendance pour les items de l'activité 5 est probablement due à la nécessité d'ordonner les options sélectionnées : si une option n'est pas à sa place, la suivante risque de ne pas l'être non plus. Or les dépendances entre items peuvent avoir des conséquences importantes sur la validité des estimations (Tuerlinckx et de Boeck, 2001) et conduisent à une surestimation de l'information apportée par les items, donc à une sous-estimation des erreurs de mesure. Cela peut également avoir un impact significatif sur les estimations des individus (Sideridis, 2011) et donc sur la définition de points de césures s'appuyant sur ces données empiriques.

Verhelst et Verstralen (2008) proposent comme solution à ce problème de mettre en œuvre le modèle à crédit partiels (généralisation du modèle de Rasch) proposé par Masters (1982), en regroupant les items d'une même activité en un item polytomique dont le score correspond au nombre de bonnes réponses données par le candidat aux différents items constituant l'activité. La mise en œuvre d'un tel modèle peut être réalisée au moyen de jMetrik. Le tableau 4 compare les indices statistiques des deux modèles : items dichotomiques (considérés, à tort, comme localement indépendants) d'une part, et mélange d'items dichotomiques et polytomiques (indépendants localement), d'autre part.

[76] Dans les modèles de réponse aux items, le paramètre « a », qui correspond à la pente de la courbe caractéristique de l'item en son point d'inflexion est interprété comme la capacité discriminatoire de l'item. Le modèle de Rasch contraint cette valeur à 1.

Tableau 4 : statistiques relatives à la qualité des échelles des modèles

	Items dichotomiques considérés comme localement indépendants		Mélange d'items dichotomiques et polytomiques	
	Items	Individus	Items	Individus
Variance observée	1,0592	1,2093	0,8453	1,0088
Écart-type	1,0292	1,0997	0,9194	1,0044
Erreur quadratique moyenne	0,0541	0,2247	0,0490	0,2085
Racine carrée de l'erreur quadratique moyenne	0,2326	0,4740	0,2214	0,4566
Variance ajustée	1,0051	0,9846	0,7963	0,8003
Écart-type ajusté	1,0025	0,9923	0,8924	0,8946
Indice de séparation	4,3104	2,0934	4,0311	1,9591
Nombre de strates	6,0805	3,1246	5,7082	2,9455
Fidélité de la séparation	0,9489	0,8142	0,9420	0,7933

Le modèle tenant compte de la dépendance locale entre items conduit à une moindre variance dans les données et à des indices de fidélité légèrement plus faibles, notamment en ce qui concerne la séparation des individus. En menant une analyse classique sur les items qui tient du regroupement polytomique, une nouvelle estimation de la fidélité par consistance interne peut être obtenue (alpha de Cronbach de 0,76, contre 0,85 dans le modèle dichotomique), ainsi que de l'erreur de mesure liée à l'échantillonnage (3,31 points contre 2,68 points). Ces différences sont appréciables et montrent l'importance de contrôler la présence d'une dépendance locale entre items.

La dernière condition d'application du modèle de Rasch est l'hypothèse de capacité discriminatoire équivalente des items. Cette hypothèse peut être vérifiée en appliquant un modèle à 2 paramètres aux données (pour les items dichotomiques) et en analysant la dispersion des valeurs du paramètre de discrimination. La figure 6 représente la distribution des valeurs de ce paramètre pour notre cas de figure. La discrimination des items varie entre 0,47 et 1,62 et elle est comprise entre 0,79 et 1,24 pour la moitié des items.

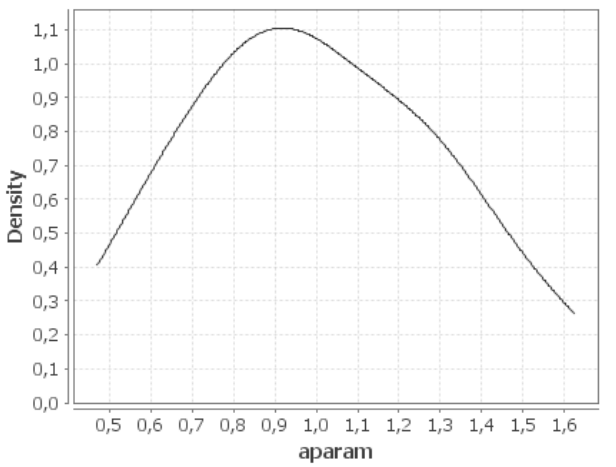

Figure 6 : Cercle des corrélations pour les activités à correction automatique

Si la condition d'égale discrimination des items ne semble qu'en partie satisfaite, il n'y a pas de présence de cas vraiment extrêmes, et compte-tenu de la taille de l'échantillon, les estimations du paramètre « a » comportent une erreur-type importante (entre 0,16 et 0,50). Cela ne nous semble pas justifier un rejet du modèle proposé.

Conclusion

La CCI Paris Ile-de-France a fait le choix de tirer parti des possibilités de l'outil informatique dans le cadre de son projet de refonte des diplômes de français professionnels, en exploitant différents formats d'items et en proposant une évaluation par des tâches plus complexes.

Compte-tenu de la nature du public ciblé par les diplômes de français professionnel, les différences éventuelles de familiarité des candidats avec l'outil informatique ne semblent pas, pour les activités proposées dans le Diplôme de français professionnel Affaires B1, introduire une variance non souhaitée dans les données. C'est davantage la nature écrite ou orale des documents supports, sur lesquels s'appuient les tâches, qui explique les différences de performance entre individus selon les tâches.

La vérification de cette hypothèse a été menée au moyen d'une analyse factorielle multiple sur les scores aux activités, d'une part, et d'une analyse en composantes principales des résidus standardisés après application d'un modèle de Rasch. Une façon supplémentaire de questionner la présence d'une dimension liée aux différences entre modalités de réponse serait d'appliquer un modèle multidimensionnel, en faisant pour cela l'hypothèse que les items des activités utilisant des glisser-déposer comme modalité de réponse contribuent à une seconde dimension et éventuellement que les activités utilisant des listes de choix contribuent à une dimension supplémentaire. Si un tel modèle s'avérait significativement mieux ajusté aux données, cela témoignerait de la présence de dimensions liées aux modalités de réponse.

L'évaluation par tâches complexes, où les candidats ont à compléter en plusieurs endroits un document de réponse (formulaire, tableau, commentaire, courriel…), sur la base d'un même (ensemble de) document(s) support(s), est cependant susceptible d'introduire des dépendances locales entre items, là où les modèles de mesure habituels font l'hypothèse de mesures indépendantes les unes des autres. Si aucune précaution n'est prise dans l'application de ces modèles, les qualités métriques rapportées risquent d'être surestimées et les informations empiriques, sur lesquelles s'appuie la prise de décision concernant l'établissement de points de césures, erronées.

Il convient donc de détecter les cas de dépendance locale entre items et, lorsque de telles dépendances existent, d'identifier un modèle de mesure plus approprié pour le traitement des données. Une solution envisageable est de regrouper les items dépendants en items polytomiques et, lorsqu'on souhaite s'appuyer sur la théorie de réponse à l'item, d'appliquer un modèle à crédits partiels. On obtient alors un meilleur ajustement des données au modèle et des estimations plus fiables des propriétés métriques de l'échelle et des paramètres des items et des individus.

Références

Bennett, R. E. (2003), Online Assessment and the Comparability of Score Meaning, Educational Testing Service Research Memorandum RM-03-05, [en ligne] [www.ets.org/Media/Research/pdf/RM-03-05-Bennett.pdf]

Casanova, D., Crendal, A., Holle, A., Demeuse, M. (2011). Élaboration d'une version électronique équivalente à la version papier-crayon d'un test de français langue étrangère à enjeux critiques. In J.G. Blais et J.L. Gilles (éds*). Évaluation des apprentissages et technologies de l'information et de la communication*. Québec : Les presses de l'Université Laval. (pp. 245-266).

Demeuse, M., et Henry, G. (2004). Théorie (classique) des scores de test (chap.5). In Demeuse (Dir.) *Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation*. Notes de cours, Version janvier 2004, mise à jour janvier 2008, format PDF [http://iredu.u-bourgogne.fr/images/stories/Documents/Cours_disponibles/Demeuse/Cours/racine.pdf].

Diarra, L. (2012). *Comparabilité entre modalités d'évaluation TIC et papier-crayon : cas de productions écrites en français en cinquième secondaire au Québec*. Thèse de doctorat, Université de Montréal.

Grondin, J., Dionne, E., Savard, J., et Casimiro, L. (2017). Démonstration d'une méthodologie mettant à profit les modèles de Rasch : l'exemple d'une échelle de mesure de l'offre active de services de santé en français (chap.1). In E. Dionne et G. Raîche (éds.). *Mesure et évaluation des compétences en éducation médicale : Regards actuels et prospectifs*. Presses de l'Université du Québec.

Hambleton R.K., Swaminathan H. (1985). Item Banking. In R.K. Hambleton & H. Swaminathan. *Item Response Theory*, p. 255-279. Dordrecht : Springer.

Houssemand, C., R. Meyers et R. Martin (2009). « L'évaluation du profil psychosocial de demandeurs d'emploi, une population peu familiarisée à la technologie informatique ». In J.-G. Blais (éd.). *Évaluation des apprentissages et technologies de l'information et de la communication. Enjeux, applications et modèles de mesure*, p.137-158. Québec : Les Presses de l'Université Laval.

Laurier, M. D. et Diarra, L. (2009). « L'apport des technologies dans l'évaluation de la compétence à écrire ». Dans J.-G. Blais (éd.). *Évaluation des apprentissages et technologies de l'information et de la communication. Enjeux, applications et modèles de mesure*, p. 77- 104. Québec : Les Presses de l'Université Laval.

Mead, A., D. et F. Drasgow (1993). « Equivalence of computerized and paper-and-pencil cognitive ability tests : a meta-analysis ». *Psychological Bulletin, 114*(3), p. 449-458.

Meyer, J.P. (2014). *Applied Measurement with jMetrik.* New-York : Routledge.

Penta, M., Arnould, C., Decruyanaere, C. (2005). *Développer et interpréter une échelle de mesure : applications du modèle de Rasch.* Sprimont : Mardaga.

Richer, J.J. (2014). Conditions d'une mise en œuvre de la perspective actionnelle en didactique des langues ». *Recherche et pratiques pédagogiques en langues de spécialité* [En ligne], Vol. XXXIII N° 1, mis en ligne le 03 mars 2014, consulté le 29 mars 2018. URL : http://journals.openedition.org/apliut/4162 ; DOI:10.4000/apliut.4162

Russell, M. et Haney, W. (2000). Bridging the Gap Between Testing and Technology in Schools. Education Policy Analysis Archives, 8(19).

Sideridis, G.D. (2011). The Effects of Local Item Dependence on Estimates of Ability in the Rasch Model. *Rasch Measurement Transactions, 2011, 25:3*, 1334-6

Tuerlinckx, F., et De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*, 181-195.

Verhelst, N.D. et Verstralen, H.H.F.M. (2008). Some Considerations on the Partial Credit Model. *Psicologica, 29,* 229-254.

Yen, W.M. (1984). Effects of local item dependance on the Fit and Equating Performance of the Three-Parameter Logistic Model. *Applied Psychological Measurement, 53*, 125-145.

Yen, W.M. (1993). Scaling performance assessments : Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.