Quels modèles de mesure pour l'évaluation par tâches ?

Cette note de vulgarisation fait suite à une communication présentée à l'occasion du 30e colloque de l'ADMEE-Europe, organisé par l'université du Luxembourg du 10 au 12 janvier 2018.

Casanova, D., Aw, A., Demeuse, M. (2018). Quand le numérique défie la mesure. Comment veiller à la qualité de certifications en langue professionnelle au format numérique ? In Actes du 30e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe). Belval, Luxembourg. https://admee2018.sciencesconf.org/

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. 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.

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. L'encart de la page suivante présente les activités de la compétence « Comprendre et traiter de l'information » proposées pour le diplôme de niveau B1. Les six premières activités sont à correction automatique.

L'évaluation par tâches au moyen d'activités à correction automatique soulève cependant une 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 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[1] est une des conditions d'application des modèles de réponses à l'item.

Faute d'une modélisation appropriée, 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.

Cela peut également avoir un impact significatif sur les estimations des individus (Sideridis, 2011) et donc sur la définition de seuils de réussite qui s'appuieraient sur ces données empiriques.

Un moyen de tester l'hypothèse d'indépendance locale est d'observer les corrélations entre les résidus des candidats, c'est-à-dire en retranchant au score obtenu à chaque item par chaque candidat la probabilité de réussite du candidat, déterminée par son score total au test (que l'on peut déterminer en calibrant les données au moyen d'un modèle de réponse probabiliste comme le modèle de Rasch[2]).

Encart : 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

[1] L'indépendance locale implique que « 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 » (Laveault et Grégoire, 2014).

[2] Le modèle de Rasch est une méthode d'analyse de données statistiques qui s'inscrit dans la théorie de réponse à l'item. Elle est particulièrement employée en psychométrie pour mesurer des éléments tels que les capacités, les attitudes ou des traits de personnalité de personnes répondant à des questionnaires.

L'analyse des données du diplôme de français professionnel Affaires B1 montre 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.

Verhelst et Verstralen (2008) proposent une solution à ce problème 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) et en mettant en œuvre le modèle à crédits partiels (généralisation du modèle de Rasch).

Cette modélisation permet notamment un meilleur ajustement des données au modèle. Sa mise en œuvre, sur les données du diplôme de français professionnel Affaires B1, conduit à des indices de fidélité légèrement plus faibles (mais plus fiables). En menant une analyse classique sur les items qui tient du regroupement polytomique, on obtient également une estimation de la fidélité par consistance interne plus faible, ce qui conduit à une erreur de mesure liée à l'échantillonnage plus élevée. Ces différences sont appréciables et montrent l'importance de contrôler la présence d'une dépendance locale entre items pour reporter des indices statistiques pertinents.

En conclusion, 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 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 veiller à 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. Pour le diplôme de français professionnel Affaires B1, la solution choisie consiste à regrouper les items interdépendants sous la forme de super-items et d'appliquer un modèle de Rasch à crédits partiels sur ces données.

Références

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].

Laveault, D. et Grégoire, J. (2014). *Introduction aux théories de tests en psychologie et en sciences de l'éducation* (3e édition). Bruxelles : De Boeck Université.

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