Comment vérifier la dimensionnalité des outils d'évaluation ?

Conference Paper · March 2019

CITATION
1

READS
260

3 authors:

Dominique Casanova
Chambre de commerce et d'industrie de Paris Île-de-France
39 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Alhassane Aw
Chambre de commerce et d'industrie de Paris Île de France
9 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Marc Demeuse
Université de Mons
417 PUBLICATIONS **1,800** CITATIONS
SEE PROFILE

Comment vérifier la dimensionnalité des outils d'évaluation ?

Casanova Dominique, dcasanova@cci-paris-idf.fr, CCI Paris Ile-de-France, France
Kaddachi Lina[1], lina.kaddachi@eleve.ensai.fr, Ecole nationale de la statistique et de l'analyse d'information, France
Aw Alhassane, aaw@cci-paris-idf.fr, CCI Paris Ile-de-France, France
Demeuse Marc, marc.demeuse@umons.ac.be, Université de Mons, Belgique

---

Dans le domaine de l'éducation, les tests sont créés afin d'évaluer le degré de maîtrise d'une compétence donnée (ou habileté). Cette compétence n'est pas directement observable (on parle de trait latent – Béland, 2016) mais se manifeste dans la réalisation d'activités, soit, dans le cas d'un test, via les réponses données par le candidat à des stimuli variés (items) au sein de tâches. Les réponses aux items et les scores associés sont des indicateurs de l'habileté du candidat.
Pour des raisons de couverture de la compétence évaluée et de précision dans l'estimation de l'habileté des candidats, un test comporte en général plusieurs tâches Laveault et Grégoire, 2014 ; Lipscomb, 1992), chaque tâche pouvant être constituée de plusieurs items (réponses à des stimuli différents) et/ou analysée au moyen d'une grille comportant différents critères d'observation. Ces tâches et les items qui les constituent contribuent tous à la mesure d'une même compétence, mais sous des angles et dans des situations différentes. Dès lors ils ne donnent pas tous exactement le même résultat. Lorsque les résultats entre tâches ou entre items sont trop différents, la restitution d'un simple score total risque de manquer de pertinence et il pourra être utile de regrouper les items en sous-tests et de restituer des scores par sous-test.

Certains modèles de mesure couramment mobilisés requièrent, dans leurs conditions d'application, une unidimensionnalité des données collectées (Bertrand et Blais 2004). Cela n'est possible que si le test ou le sous-test est homogène, que tous les items/tâches évaluent la même compétence sous-jacente et qu'ils ne présentent pas de biais. Lorsque de tels modèles, comme le modèle de Rasch, sont mis en œuvre pour prendre des décisions, il est nécessaire de s'assurer que les conditions d'application sont suffisamment respectées, notamment la condition d'unidimensionnalité (Boadé, 2013). Dans la pratique cependant, l'unidimensionnalité est un idéal difficilement atteignable, et la question est plutôt de savoir si un instrument de mesure est suffisamment unidimensionnel pour l'usage envisagé des scores (Pini, 2012 ; Wu & Adams, 2007).

La dimensionnalité des instruments de mesure est souvent appréciée au moyen d'analyses factorielles des scores aux items (Bertrand et Blais, 2004) ou d'une analyse en composantes principales des résidus après application du modèle de mesure (Meyer, 2014). Mais l'interprétation des résultats obtenus est très subjective et il n'y a pas de règle établie pour décider du degré de satisfaction de l'hypothèse d'unidimensionnalité.

[1] Cette communication fait suite à un stage de deuxième année réalisé par Lina Kaddachi au sein de la Chambre de commerce et d'industrie de région Paris Ile-de-France.

Dans cette communication, nous proposons une méthodologie de mise en œuvre de modèles multidimensionnels de réponse aux items (Reckase, 2009) pour confirmer a présence d'une seconde dimension présumée et évaluer son impact sur les classements des performances au test.

Cas d'étude

L'épreuve de compréhension orale du Test d'évaluation de français (TEF)

Le premier cas d'étude concerne le Test d'évaluation de français (TEF), test conçu par la Chambre de commerce et d'industrie de région Paris Ile-de-France, qui est notamment utilisée dans le cadre des procédures d'immigration au Canada et au Québec. Une part importante des candidats passant le test au Québec, les concepteurs du test sont fortement encouragés à introduire davantage de référents québécois. Cela se traduit, pour l'épreuve de compréhension orale, par davantage de textes prononcés avec un accent québécois. Le TEF ayant été élaboré sur l'hypothèse d'un accent « neutre », l'introduction de tels items et de nature à altérer la dimensionnalité du test, en introduisant dans le construit une capacité à traiter des textes à l'accent varié. Il importe donc d'analyser l'impact que l'introduction de tels items peut avoir sur les résultats des candidats. Dans une première étude nous avions cherché à mettre en évidence un éventuel fonctionnement différentiel d'items à l'accent québécois modéré selon les pays de passation du test (Casanova et al. 2018a). Ici nous cherchons à mettre en évidence la présence d'une seconde dimension dans les données qui pourrait être interprétée comme la capacité spécifique des candidats à traiter des messages à l'accent québécois modéré.

Pour ce cas d'étude, les données sont constituées des réponses de 3093 individus à un même test de compréhension orale qui comportait 46 items à l'accent « standard » et 14 items à l'accent « québécois » (modéré). Tous les items sont dichotomiques et les candidats ayant réussi ou échoué à l'ensemble des questions se rapportant à un accent donné ont été préalablement retiré de l'échantillon. La fidélité du modèle unidimensionnel à 2 paramètres, estimée au moyen de l'indice de fidélité de séparation des individus est de 0,891.

Les items à correction automatisée du diplôme de français professionnel Affaires B1

Le second cas d'étude porte sur le diplôme de français professionnel Affaires B1, qui comporte 6 activités sur ordinateur à correction automatisée (listes déroulantes et glisser-déposer). Une étude préalable, au moyen d'analyses en composantes principales, avait mis en évidence une potentielle seconde dimension liée à la nature écrite ou orale du document support principal de ce sous-test (Casanova et al., 2018b), ce qui pose la question de la pertinence de restituer un score unique à ce sous-test. Cette éventuelle seconde dimension ne ressortant que faiblement des analyses, nous avons souhaité confirmer son existence et en apprécier l'importance au moyen d'un modèle à deux dimensions.

Pour ce cas d'étude, les données sont constituées des réponses de 193 individus à un même sous-test comportant 2 activités (13 items) dont le support principal est oral et 4 activités (37 items) dont le support principal est écrit. Les items sont dichotomiques pour la plupart. La fidélité du modèle unidimensionnel à 2 paramètres, estimée au moyen de l'indice de fidélité de séparation des individus est de 0,824.

Démarche adoptée

Pour chacun des deux cas d'étude, nous proposons de recourir à deux types de modélisation complémentaires, qui supposent que les items susceptibles de faire émerger une seconde dimension soient connus. Dans la première modélisation, ces items sont associés (via une matrice Q) à cette dimension est les autres items à une autre dimension. Cela permet d'obtenir les estimations les pour deux traits latents corrélés (au-delà de l'accent pour le TEF ou de la nature orale ou écrite du support principal pour le diplôme, chacun des deux traits évalue la compétence à traiter de l'information) et d'estimer la force de cette corrélation en la corrigeant pour atténuation au moyen des estimations de la fidélité sur chacune des deux dimensions.

Un second type de modélisation peut être mobilisé pour apprécier l'impact de la seconde dimension sur les données collectées (taille de l'effet). Cette modélisation fait porter l'ensemble des items sur une première dimension et également sur une seconde dimension les items susceptibles de présenter la spécificité présumée. On s'attend alors à ce que la première dimension capture ce qu'il y a de commun aux items du test et que la seconde dimension ne prenne en compte que la spécificité présentée par certains items (information contenue dans les résidus une fois éliminée la partie commune pour ces items). L'importance de la seconde dimension pourra être appréciée au moyen de son indice de fidélité. Si la fidélité s'avère faible et que l'erreur de mesure est comparable à l'écart-type des estimations pour cette seconde dimension, alors l'instrument ne permettra pas de distinguer efficacement les candidats selon cette dimension, qui peut dès-lors être négligée. Par ailleurs, l'estimation de la discrimination des items sur la seconde dimension (indice *a* du modèle*)* permet d'identifier ceux qui sont le plus porteur de la spécificité présumée. On peut ainsi voir si cette spécificité concerne une majorité des items (confirmation de la présence d'une seconde dimension commune) ou si elle est l'effet de quelques items isolés.

Résultats

Nous avons appliqué nos modèles au moyen de la librairie mirt de R (Chalmers, 2012). Nous présentons ici les résultats obtenus lorsque l'estimation de la compétence des individus est réalisée au moyen d'une estimation du maximum de vraisemblance[2]. L'ensemble des indices de fidélité évoqués ci-après correspondent à la fidélité de séparation des individus.

1re modélisation

Dans le cas du TEF, la corrélation brute entre les deux traits est de 0,690 et la valeur corrigée pour atténuation s'élève à 0,881. Le modèle à deux dimensions s'ajuste mieux aux données que le modèle unidimensionnel.

Dans le cas du diplôme, la corrélation brute entre les deux traits est de 0,515 et la valeur corrigée pour évaluation s'élève à 0,736. Il semble donc que les sous-tests évaluent des choses partiellement différentes. Là encore, le modèle à deux dimensions s'ajuste mieux aux données que le modèle unidimensionnel.

[2] Lorsque la méthode utilisée est l'espérance a posteriori, on ne constate pas d'effet significatif.

2de modélisation

Dans le cas du TEF, la fidélité du second trait latent, porteur de la spécificité, est nettement plus faible (0,552) que celle du trait principal (0,875). Comme le montre la figure 1, les cas où l'estimation de la compétence sur cette dimension est significativement différente de zéro sont rares. Par ailleurs, un seul item présente un indice de discrimination sur la seconde dimension supérieur à 0,3. Il ne semble donc pas que l'introduction d'items à l'accent québécois modéré n''entraîne l'apparition d'une seconde dimension notable.

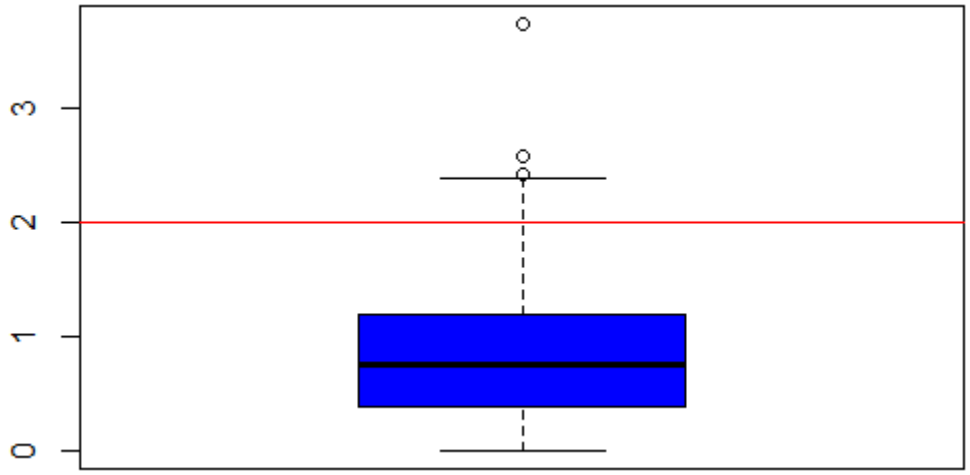

Figure 1 : distribution pour le TEF du ratio de l'estimation de la compétence sur l'erreur-type (2de dimension). Les valeurs supérieures à 2 sont significativement différentes de 0.

Dans le cas du diplôme, la fidélité du second trait latent (0,691) est proche de celle du premier trait (0,709). Il semble donc bien que les items dont le document principal est oral soient porteurs d'une spécificité. Cette fois-ci, la part des candidats dont d'estimation de la compétence sur la seconde dimension est significativement différente de 0 est de plus d'un quart (Cf. figure 2). Au vu de leur indice de discrimination sur cette dimension, la plupart des items sont porteurs de la spécificité de cette dimension.

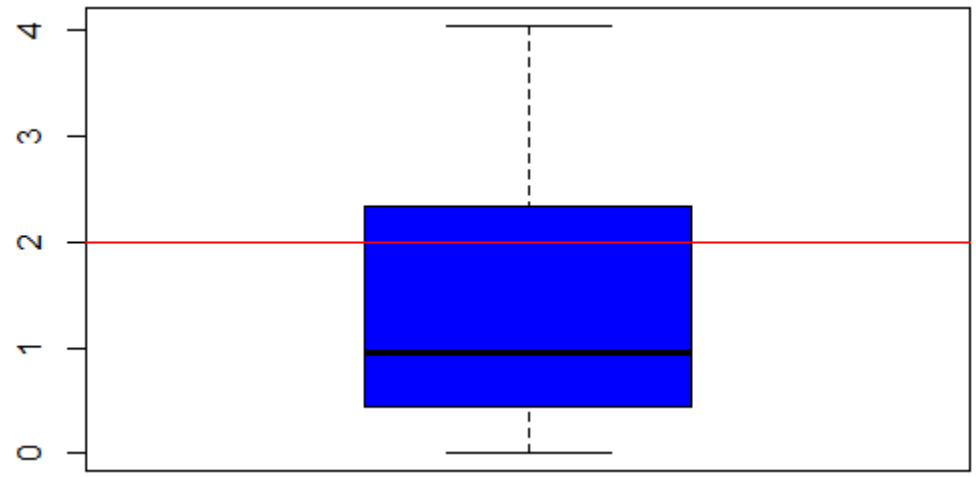

Figure 2 : distribution pour le diplôme du ratio de l'estimation de la compétence sur l'erreur-type (2de dimension). Les valeurs supérieures à 2 sont significativement différentes de 0.

On peut toutefois s'interroger sur le sens à donner à cette seconde dimension étant donné que les activités portant sur un document principal se trouvent en fin de questionnaire. Les candidats étant libres de répartir le temps de passation du sous-test à leur guise selon les activités, il est tout aussi plausible que cette dimension soit également liée à une capacité à gérer le temps d'examen. Lorsqu'on regarde où se situent, sur le continuum du second trait latent, les candidats dont l'estimation est significativement différente de zéro, on constate qu'il s'agit essentiellement des candidats pour lesquels l'estimation est la plus faible (Cf.figure 3). Or quand on considère les candidats dont l'estimation sur le second trait latent est la plus faible (inférieure à -1,5), la corrélation entre cette estimation et le temps consacré aux items dont le support principal est oral est relativement élevée (0,716).

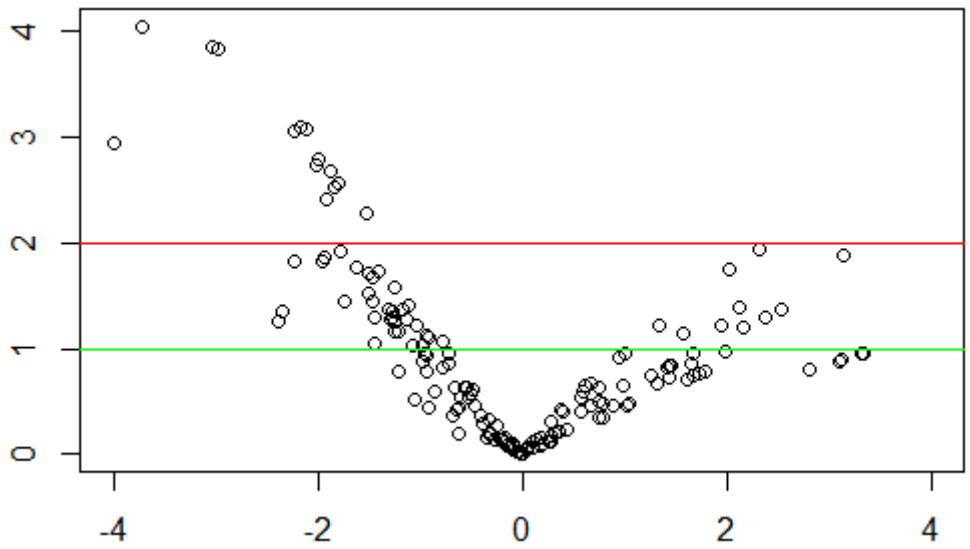

Figure 3 : valeur du ratio estimation / erreur-type (en ordonnée) en fonction de l'estimation pour le second trait latent (abscisse)

Conclusion

L'application de modèles de réponses aux items bidimensionnels dans une perspective confirmatoire, quand on soupçonne un sous-ensemble d'items d'être porteur d'une même spécificité, permet d'éclairer les résultats obtenus plus classiquement au moyen d'analyses en composantes principales.

Dans le cas du test d'évaluation de français, les résultats obtenus confirment que l'introduction d'items à l'accent québécois modéré n'a pas eu pour conséquence l'apparition d'une seconde dimension marquée qui serait liée à la capacité à traiter des textes ne présentant pas un accent « standard ». Le Français des affaires peut donc en confiance poursuivre sa stratégie d'utilisation d'accents (modérés) variés de la francophonie dans son test. A l'opposé, les résultats obtenus pour le diplôme (mais sur un échantillon limité et en privilégiant le maximum de vraisemblance pour les estimations) semblent confirmer la présence d'une seconde dimension non négligeable. Cela contredit les conclusions d'une analyse en composante principale des résidus menée sur les mêmes données (Casanova & al., 2018b). Cependant, étant donnée l'emplacement des items dans le questionnaire, on ne

peut pas déterminer à l'heure actuelle si cette seconde dimension est liée à la nature écrite ou oral du support principal ou à la gestion du temps par les candidats.

Mots-clés

Théorie de réponse aux items, modèles multidimensionnels, français langue étrangère

Références bibliographiques

Béland, S. (2016). Réflexion : est-il possible de mesurer une compétence en contexte d'évaluation à grande échelle? *Revue canadienne des jeunes chercheures et chercheurs en éducation, Vol. 7(1)*

Bertrand, R. & Blais, J.-G. (2004). *Modèle de mesure : l'apport de la théorie de la réponse aux items*. Sainte-Foy, Québec : Presses de l'Université du Québec.

Boadé, G. (2013). Robustesse du modèle de Rasch unidimensionnel à la violation de l'hypothèse d'unidimensionnalité. https://papyrus.bib.umontreal.ca/xmlui/handle/1866/9942 (consulté le 25/03/2019)

Casanova, D., Aw, A. & Demeuse, M. (2018a). Nos items à l'accent québécois difR-ils ? *Actes du 30e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe)*, Esch-sur-Alsette, Luxembourg.

Casanova, D., Aw, A. & Demeuse, M. (2018b). Quand le numérique défie la mesure. Comment veiller à la qualité de certifications en langue professionnelle au format numérique ? *Actes du 30e colloque international de l'Association pour le Développement des Méthodologies d'Évaluation en Éducation (ADMEE-Europe)*, Esch-sur-Alsette, Luxembourg.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of statistical Software, vol. 48(6)*

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

Lipscomb, M. S. (1992). *A comparison of domain sampling procedures for test construction*. Brooks Air Force Base, Tex : Armstrong Laboratory, Air Force Materiel Command

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

Pini, G. (2006). *A propos de la théorie des réponses aux items (TRI – IRT). Le cas d'items dichotomiques.* Groupe Edumétrie : Qualité de la mesure en éducation

Reckase, M. D. (2009). *Multidimensional Item Response Theory (Statistics for Social and Be-havioral Sciences).* New York: Springer..

Wu, M. & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Educational Measurement Solutions, Melbourne.