Quelles méthodes pour l'évaluation standardisée des compétences des élèves ?

Thierry ROCHER
Bureau de l'évaluation des élèves, DEPP[1] , Ministère de l'Education nationale

Cet article présente les méthodes psychométriques qui sont généralement employées dans les programmes d'évaluations standardisées des compétences des élèves. Alors qu'elles sont largement partagées au niveau national et au niveau international, ces méthodes restent relativement méconnues en France que ce soit dans le monde académique, le monde éducatif ou encore celui de la statistique publique. Leurs fondements théoriques ainsi que leurs hypothèses sous-jacentes sont présentées. Nous montrons leur intérêt d'un point de vue pratique mais également leurs limites.

Introduction

Les programmes d'évaluations standardisées des compétences des élèves, tels que PISA (Programme International pour le Suivi des acquis des élèves) mené par l'OCDE ou CEDRE (Cycle des Évaluations Disciplinaires Réalisées sur Échantillons) mené par la DEPP, ont pour objectif de mesurer le niveau des acquis des élèves, à différents moments de la scolarité. Ces évaluations s'intéressent aux élèves comme éléments d'une population; elles n'ont pas vocation à rendre compte de leurs résultats au niveau individuel. Elles se situent donc à un niveau global et doivent permettre d'apprécier les résultats du système éducatif et leur évolution dans le temps.

Ces enquêtes se sont multipliées depuis le début des années 2000 (Trosseille et Rocher, 2015). Elles occupent aujourd'hui une place importante dans le domaine de l'éducation, notamment via la médiatisation de leurs résultats. La mise en œuvre de politiques éducatives se réfère ainsi souvent à ces évaluations, en particulier aux évaluations internationales qui, derrière la diffusion de palmarès globalisants, fournissent un éclairage très intéressant sur les forces et les faiblesses des systèmes éducatifs. A titre d'illustration, l'enquête PISA, à travers le prisme de la comparaison internationale, a permis de révéler l'ampleur des inégalités qui touchent le système éducatif français (Baudelot & Establet, 2009). Ce constat a récemment été confirmé par l'évaluation nationale CEDRE dont les résultats concernant les compétences des élèves de 3e en mathématiques montrent que le poids de l'origine sociale a augmenté depuis six ans (Arzoumanian et Dalibard, 2015). Ces évaluations fournissent ainsi des éléments qui alimentent concrètement les débats sur l'éducation (comme récemment ceux sur la réforme du collège ou sur la ségrégation sociale)

D'un point de vue méthodologique, ces évaluations reposent sur des échantillons représentatifs

1. Direction de l'évaluation, de la prospective et de la performance

et suivent des procédures standardisées afin de limiter l'erreur de mesure à tous les niveaux (passation, correction, etc.). Elles s'appuient sur un ensemble de méthodes relevant du domaine de la psychométrie, c'est-à-dire de la mesure de dimensions psychologiques, et qui a donné naissance au domaine de l'*édumétrie* dans le champ de l'éducation. Bien que le dispositif de test standardisé ait été inventé en France par Alfred Binet et ses collaborateurs au début du XX^{e} siècle, la psychométrie est un champ très méconnu en France, alors que ce domaine s'est considérablement développé dans d'autres pays, notamment aux Etats-Unis, à travers des thèmes comme la méritocratie scolaire (assurer un traitement équitable des élèves) ou bien comme l'intelligence (sujet ayant d'ailleurs conduit à certaines dérives idéologiques, cf. Gould, S. J., 1997).

C'est la nature de la variable mesurée qui distingue principalement les programmes d'évaluation d'autres enquêtes statistiques. En effet, il est convenu que les compétences des élèves ne s'observent pas directement. Seules les manifestations de ces compétences sont observables, par exemple à travers les résultats obtenus à un test standardisé. L'existence supposée de la compétence visée est alors matérialisée dans la réussite au test. D'une certaine manière, on pourrait avancer que c'est l'opération de mesure elle-même qui définit concrètement l'objet de la mesure, d'où le célèbre pied de nez d'Alfred Binet, en réponse à la question « qu'est-ce que l'intelligence? » : « c'est ce que mesure mon test ». Le terme de « construit » est alors souvent employé pour désigner l'objet de la mesure.

Bien entendu, toute statistique peut être considérée comme un construit, pas seulement celles ayant trait à l'évaluation. Cependant, des degrés sont sans doute à distinguer, en lien avec le caractère tangible de la variable visée. Par exemple, la réussite scolaire peut-être appréhendée par la variable « réussite au baccalauréat » qui est mesurable directement car elle est sanctionnée par un diplôme, donnant lieu à un acte administratif que l'on peut comptabiliser. Le « décrochage scolaire », quant à lui, est un concept qui doit reposer sur une définition précise, choisie parmi un ensemble de définitions possibles, ce choix faisant acte de construction. Une fois la définition établie, le calcul repose le plus souvent sur l'observation de variables administratives, telles que la non ré-inscription dans un établissement scolaire.
En comparaison, la mesure des compétences se présente comme une démarche de construction assez particulière. L'idée sous-jacente de la psychométrie consiste à postuler qu'un test mesure des performances qui sont la manifestation d'un niveau de compétence, non observable directement. Ainsi, l'objet de la mesure est une variable latente.

1. Quelques notions fondamentales

Nous présentons tout d'abord un exemple d'application qui a pour objectif d'illustrer de façon pédagogique les grandes notions de psychométrie. Il s'agit d'un questionnaire portant sur la taille des individus, passé par un échantillon d'adultes (cf. encadré 1).

Encadré 1 : Mesurer la taille des individus avec un questionnaire

La situation est la suivante : nous n'avons aucun moyen de mesurer directement la taille des individus d'un échantillon donné. Mais nous avons la possibilité de proposer un questionnaire, composé de questions appelant une réponse binaire (oui/non) et n'évoquant pas directement la taille. Nous nous plaçons ainsi artificiellement dans le cas de la mesure d'une variable latente que nous cherchons à approcher à l'aide d'un questionnaire, soit un dispositif de mesure apparemment comparable à celui d'une évaluation standardisée.

Cet exemple est beaucoup utilisé aux Pays-Bas dans les cours de psychométrie, car il permet d'introduire de façon pédagogique les concepts utilisés en psychométrie. Dans cet esprit, nous avons de notre côté élaboré un questionnaire de 24 items, nécessitant simplement d'indiquer l'accord ou le désaccord avec une série d'affirmations. Voici un extrait de ce questionnaire :

1 Je dois souvent faire attention à ne pas me cogner la tête
2 Pour les photos de groupe, on me demande souvent d'être au premier rang
3 On me demande souvent si je fais du basket-ball
4 Dans la plupart des voitures, je suis mal assis(e)
5 Je dois souvent faire faire les ourlets quand j'achète un pantalon
6 Je dois souvent me baisser pour faire la bise
7 Au supermarché, je dois souvent demander de l'aide pour attraper des produits en haut des gondoles
8 A deux sous un parapluie, c'est souvent moi qui le tiens ...

Ce questionnaire a été proposé via Internet à un échantillon composé de 276 adultes dans un réseau à la fois professionnel et personnel. L'échantillon est plutôt jeune (55 % sont âgés de moins de 30 ans) et féminin (65 % de femmes) mais la question de la représentativité n'est pas importante au regard de notre propos qui concerne les problématiques de mesure.

Une notion fondamentale en psychométrie est celle de la **validité** : le test mesure-t-il bien ce qu'il est censé mesurer?

Dans le cadre de notre exemple, nous pouvons approcher la validité assez directement puisque la dernière question demande aux enquêtés d'indiquer leur taille. Nous avons calculé un score de façon très simple à partir des 24 questions en attribuant 1 point pour chacune d'entre elle, en fonction de la modalité associée à une taille plus élevée : par exemple, les individus obtiennent 1 point s'ils répondent oui à la première question, 0 sinon; et inversement, pour la deuxième question. Il est alors possible d'analyser la relation entre ce score et la taille déclarée : le coefficient de corrélation linéaire de 0,85 indique un lien positif et fort entre le score construit et la taille. De ce point de vue, nous pouvons conclure à la validité de notre questionnaire, même si l'ampleur de la corrélation observée peut être largement discutée.

En matière d'évaluation standardisée de compétences, nous ne disposons évidemment pas d'une variable de référence, telle que la taille réelle, puisque précisément les compétences sont inobservables directement. La question de la validité d'une évaluation devient alors une question complexe. La littérature abonde de références dans ce domaine. En résumé, différents types de validité sont généralement distingués : validité de contenu, de construit, *critériée*, etc. Dans le cas de CEDRE par exemple, la validité est principalement assurée à travers une validité dite de contenu : un groupe de concepteurs composé d'enseignants, d'inspecteurs, de formateurs est garant, sur la base de leur propre expertise, de l'adéquation du contenu de l'évaluation avec les programmes scolaires, les instructions officielles et les pratiques de classes. Ainsi, un niveau de performance observé à l'évaluation de mathématiques est censé traduire un niveau de compétence, au regard des attendus en mathématiques.

Une question centrale de psychométrie est celle de la **dimensionnalité** d'un ensemble d'items. Nous calculons un score, mais cela n'a de sens que sous l'hypothèse que les items mesurent la même dimension, que le test est unidimensionnel. Cependant, il est clair que les items présentés ici ne mesurent pas purement la dimension taille, mais interrogent chacun une multiplicité de dimensions. L'idée est qu'un facteur commun prépondérant relie ces items, facteur lié à la

taille. Ainsi, la majorité des évaluations rend compte des résultats à travers un score global, selon un cadre unidimensionnel.

L'exemple nous permet également d'illustrer la notion de **fonctionnements différentiels d'items** ou FDI, qui est liée à la question de la dimensionnalité. Un FDI apparaît entre des groupes d'individus dès lors qu'à niveau égal sur la variable latente mesurée, la probabilité de réussir un item donné n'est pas la même selon le groupe considéré. Cela signifie qu'une autre variable, liée au groupe, est intervenue, au-delà de la dimension visée. Un fonctionnement différentiel se traduit souvent par une différence de réponse entre les groupes plus importante à l'item considéré qu'en moyenne sur l'ensemble des items. Par exemple, à la question « A deux sous un parapluie, c'est souvent moi qui le tiens », 89 % des hommes répondent oui contre 52 % des femmes, soit un écart de 37 points, alors qu'en moyenne sur l'ensemble des items, la différence entre les hommes et les femmes est de 20 points. Cet écart de 20 points renvoie à ce qu'on appelle l'impact, c'est-à-dire la différence entre les deux groupes sur la variable latente, en l'occurrence la différence de taille entre hommes et femmes. Un écart additionnel renvoie à un fonctionnement différentiel. À taille égale, les hommes disent tenir le parapluie plus souvent que les femmes. Une autre dimension que la taille, liée au genre, a joué dans la réponse. La question est alors dite « biaisée » selon le genre.

De manière pratique, un concept important est celui de la **fidélité** du test. Le score calculé comporte une part d'erreur de mesure. En effet, on peut considérer que les items d'un test ont été échantillonnés dans l'« univers » possible des items censés mesurer la dimension visée par le test. Dès lors, un autre ensemble d'items n'aurait pas conduit exactement aux mêmes scores. Le test est dit fidèle lorsque l'erreur de mesure est réduite. Le coefficient α de Cronbach est un indicateur de fidélité du test[2]. En l'occurrence, pour le questionnaire sur la taille, il a pour valeur 0.80, ce qui est considéré comme satisfaisant.

Au-delà de cet indice global, il est intéressant d'étudier les items eux-mêmes. Les taux de réponse observés aux différentes modalités proposées – ici, oui ou non – sont bien entendu des indicateurs essentiels. Par exemple, dans le cas d'une évaluation, les items peuvent être comparés en termes de **difficulté**, qui est appréciée par le pourcentage de bonnes réponses. Une autre notion importante est celle de **pouvoir discriminant** de chaque item, qui renvoie au lien avec les résultats obtenus à l'ensemble du test. En effet, si l'item mesure bien la dimension qu'il est censé mesurer, alors il discriminera bien les personnes selon cette dimension. Une manière de vérifier qu'il mesure bien la dimension supposée est d'examiner les corrélations de l'item avec d'autres items censés mesurer la même dimension. Concernant le questionnaire sur la taille, les corrélations items-test, c'est-à-dire les corrélations entre la réussite à un item donné et le score aux autres items, sont assez élevées, à l'exception d'un item dont la corrélation item-test est nulle. Il s'agit d'un item repris du questionnaire néerlandais : « Dans un lit, j'ai souvent froid aux pieds. ». Utilisé aux Pays-Bas, cet item doit donc être discriminant selon la taille des Néerlandais, mais ce n'est pas le cas sur notre échantillon français. Nous supposons qu'il s'agit d'une différence culturelle liée aux habitudes de border les draps ou la couette, forte en France et absente aux Pays-Bas où le problème d'avoir froid au pied la nuit se pose sans doute pour les personnes de grande taille. Ainsi, cet item ne mesure pas la dimension taille en France mais plutôt une autre dimension décorrélée, telle que la frilosité…

Pour finir avec le cas d'école, nous abordons la notion d'**échelle**. Avant tout, notons que le questionnaire ne nous permet pas de connaître la taille des individus. Il nous permet simplement de classer avec plus ou moins de fiabilité les individus selon leur taille, et d'introduire une métrique. Ainsi, le score simple que nous avons calculé, compris entre 0 et 24, de moyenne 11,0 et d'écart-type 4,3, est une échelle de mesure, sur laquelle il est possible d'établir un classement

2. Ce coefficient généralement désigné par alpha se définit par la formule :
alpha = (k/(k-1))(1-((somme des variances internes aux items)/(variance totale))) où k représente le nombre d'items

des individus ainsi que des distances entre eux. Il s'agit d'une échelle dite d'intervalle, qui autorise la comparaison des intervalles de scores entre individus. Autrement dit, les rapports entre intervalles ne sont pas modifiés par transformation linéaire[3]. L'origine et l'unité peuvent donc être transformées, et ce de manière arbitraire. Dans notre exemple, nous pouvons rendre compte des résultats sur l'échelle des scores observés, de moyenne 11,0 et d'écart-type 4,3, mais également sur une échelle standardisée, de moyenne 0 et d'écart-type 1, ou de moyenne 250 et d'écart-type 50 comme dans CEDRE, ou encore de moyenne 500 et d'écart-type 100 comme dans PISA. Autrement dit, les valeurs elles-mêmes n'ont pas de significations, au-delà du classement et de la distance entre individus.

Le lecteur intéressé trouvera ces notions décrites plus en détails, avec une application concrète à l'évaluation de compétences, dans le rapport technique associé à chaque évaluation du cycle CEDRE (DEPP, 2015).

2 Les modèles de réponse à l'item (MRI)

En matière d'analyse des résultats, une première approche – dite classique – se concentre sur l'analyse du score observé, c'est-à-dire du nombre de bonnes réponses obtenues aux items d'un test donné. Dans la pratique, cette approche révèle vite des limites. En effet, les résultats observés dépendent de l'ensemble des items considérés. Il n'est donc pas possible de distinguer ce qui relève de la difficulté du test de ce qui relève du niveau de compétence des élèves. Le recours à une modélisation plus adaptée, qui se situe au niveau des items eux-mêmes et non au niveau du score agrégé, est apparu nécessaire. En particulier, les modèles de réponse à l'item (MRI), nés dans les années 1960, se sont imposés dans le champ des évaluations standardisées à grande échelle. Les MRI sont une classe de modèles probabilistes. Ils modélisent la probabilité qu'un élève donne une certaine réponse à un item, en fonction de paramètres concernant l'élève et l'item. Nous présentons le modèle le plus simple, proposé par le mathématicien danois George Rasch en 1960 :

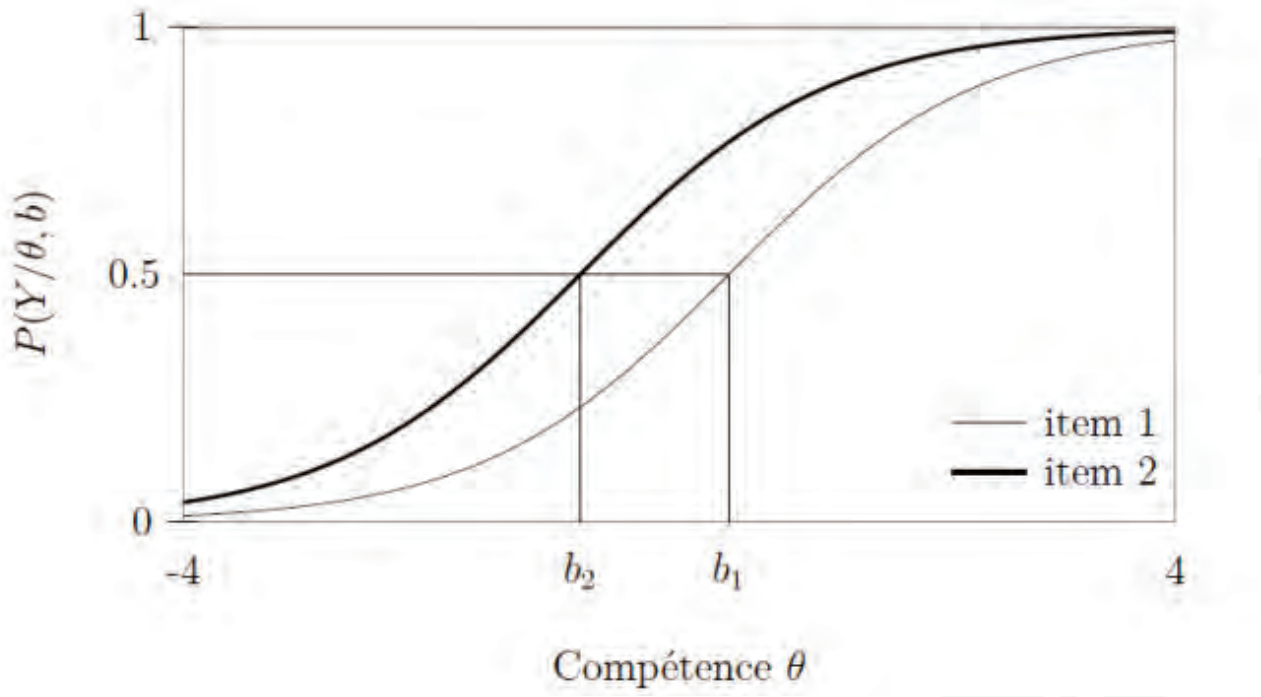
$$P_{ij} = P(Y_{ij} = 1 / \theta_i, b_j) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}}$$

i.e. la probabilité Pij que l'élève i réussisse l'item j est une fonction sigmoïde[4] du niveau de compétence θi de l'élève i et du niveau de difficulté bj de l'item j.
La fonction sigmoïde étant une fonction croissante, il ressort que la probabilité de réussite augmente lorsque le niveau de compétence de l'élève augmente et diminue lorsque le niveau de difficulté de l'item augmente, ce qui traduit à l'évidence les relations attendues entre réussite, difficulté et niveau de compétence. L'intérêt de ce type de modélisation, et ce qui explique son succès, c'est de séparer deux concepts-clé, à savoir la difficulté de l'item et le niveau de compétence de l'élève.

3. C'est le cas par exemple des échelles de température. S'il fait 20 °C à Paris, 30 °C à Grenoble et 40 °C à Rome, l'écart de température entre Rome et Paris est deux fois plus grand que celui entre Grenoble et Paris. C'est également vrai en Fahrenheit, après transformation linéaire. En revanche, on ne peut pas dire qu'il fait deux fois plus chaud à Rome qu'à Paris, cela dépend de l'échelle utilisée. Seules les échelles dites de rapport (poids, taille, revenu, etc.) permettent des comparaisons de rapports.

4. La fonction sigmoïde est définie par : $\forall x, f(x) = \frac{e^x}{1+e^x}$, à valeur dans]0, 1[. Cette fonction (par ailleurs classiquement utilisée pour la régression logistique) a l'avantage d'être très proche de la fonction de répartition de la loi normale, tout en étant plus facile à manipuler dans les procédures d'estimation.

Autre avantage : le niveau de compétence des élèves et la difficulté des items sont placés sur la même échelle, par le simple fait de la soustraction (θi −bj). Cette propriété permet d'interpréter le niveau de difficulté des items par rapprochement avec le continuum de compétence. Ainsi, les élèves situés à un niveau de compétence égal à bj auront 50 % de chances de réussir l'item, ce que traduit visuellement la représentation des courbes caractéristiques des items (CCI) selon ce modèle (figure 1).



Note de lecture : la probabilité de réussir l'item (en ordonnées) dépend du niveau de compétence (en abscisse). Par définition, le paramètre de difficulté d'un item correspond au niveau de compétence ayant 50 % de chances de réussir l'item. Ainsi, l'item 1 en trait fin est plus difficile que l'item 2 en trait plein. La probabilité de le réussir est plus élevée quel que soit le niveau de compétence.

Figure 1. Modèle de réponse à l'item -1 paramètre

3 Assurer la comparabilité

Les MRI sont très utiles dès lors qu'il s'agit de comparer les niveaux de compétence de différents groupes d'élèves. Par exemple, dans le cadre de comparaisons temporelles, la reprise à l'identique de l'ensemble des items passés lors de la précédente enquête n'est pas forcément pertinente, au regard de l'évolution des programmes scolaires, des pratiques, de l'environnement, etc. Certains items doivent être retirés, d'autres ajoutés. Par conséquent, les élèves des deux cohortes passent une épreuve en partie différente. Dès lors, comment assurer la comparabilité des résultats?

Cette problématique renvoie à la notion d'ajustement des métriques ou de parallélisation des épreuves (en anglais : *equating*). Il s'agit de positionner sur la même échelle de compétence les élèves de différentes cohortes, à partir de leurs résultats observés à des évaluations différentes. De nombreuses techniques existent et sont couramment employées dans les programmes d'évaluations standardisées. Typiquement, les comparaisons sont établies à partir d'items

communs, repris à l'identique d'un moment de mesure à l'autre. Les modèles de réponse à l'item fournissent alors un cadre approprié, dans la mesure où ils distinguent les paramètres des items, qui sont considérés comme fixes, des paramètres des élèves, considérés comme variables.

Plusieurs stratégies d'estimation sont possibles. La première vise à estimer les paramètres des items – la difficulté βj et les discriminations aj pour un MRI à deux paramètres – à partir des données de la première cohorte, en fixant la moyenne et l'écart-type des niveaux de compétence θi, par exemple à 0 et à 1 respectivement. Les valeurs des paramètres des items communs sont considérées comme fixes et elles sont utilisées pour estimer les θi de la deuxième cohorte.

Une autre possibilité, appelée « estimation concourante », consiste à envisage toutes les données de manière simultanée en autorisant des différences de niveau de compétence entre groupes. Les réponses des élèves aux items qu'ils n'ont pas vus sont traitées comme des valeurs manquantes par l'algorithme d'estimation (cf. DEPP, 2015 pour plus de détails).

4 Hypothèses

4.1 L'hypothèse d'unidimensionnalité

L'uni-dimensionnalité est une hypothèse fondamentale des modèles présentés précédemment. Seul le niveau de compétence θ explique la réussite à un item de difficulté et de discrimination données. Le respect de cette hypothèse est une condition préalable à la mise en œuvre de ces modèles. Si d'autres facteurs entrent en ligne de compte dans la probabilité de réussite aux items – par exemple une compétence différente de celle visée –, l'hypothèse d'uni-dimensionnalité doit être rejetée et le modèle ne peut être appliqué.

Bien que fondamentale, cette hypothèse est rarement testée statistiquement. Pour cause, la notion d'uni-dimensionnalité a longtemps souffert d'une absence de définition formelle. Ainsi, une quantité impressionnante d'indices ont été mis au point et visent à évaluer l'importance d'une dimension principale. Mais la plupart d'entre eux souffrent d'un manque de fondement théorique ainsi que de faiblesses techniques. Il existe cependant une définition plus formelle de l'uni-dimensionnalité, à partir de la notion d'indépendance locale, c'est-à-dire l'indépendance des réussites entre deux items, conditionnellement à la dimension visée. En effet, là encore, si une corrélation est constatée entre items, après avoir contrôlé du niveau à l'ensemble du test, c'est qu'une deuxième dimension est intervenue dans la réussite à ces deux items.

Notons que l'uni-dimensionnalité stricte n'existe probablement pas. Les processus mis en œuvre pour réussir un ensemble d'items sont complexes et varient selon les sujets et les contextes. Dès lors, il est difficilement concevable que ces processus se réduisent rigoureusement à une seule et même dimension. C'est pourquoi, en pratique, évaluer l'uni- dimensionnalité revient en fait à évaluer l'existence d'une dimension dominante, à l'aide par exemple d'analyses factorielles exploratoires, en facteurs communs et spécifiques.

4.2 Les fonctionnements différentiels d'items

Nous l'avons évoqué avec le questionnaire sur la taille : un fonctionnement différentiel d'item (FDI) apparaît entre des groupes d'individus dès lors qu'à niveau égal sur la variable latente mesurée, la probabilité de réussir un item donné n'est pas la même selon le groupe considéré. La question des FDI est importante car elle renvoie à la notion d'équité entre les groupes : un test ne doit pas risquer de favoriser un groupe par rapport à un autre. Ainsi, aux Etats-Unis, quantité de tests sont passés au crible dans le but de déterminer la présence d'éventuels biais

d'items (« Male/Female », « Black/White », ...) surtout si les résultats ont des conséquences sur le devenir des individus, comme pour les tests de sélection d'entrée à l'Université, les tests de recrutement, etc. Les évaluations standardisées à grande échelle sont également concernées, en particulier les évaluations internationales qui doivent assurer la comparabilité des difficultés des items d'un pays à l'autre. C'est en effet l'hypothèse forte qui est faite dans le cadre des évaluations internationales : l'opération de traduction ne modifie pas la difficulté de l'item. Or, des analyses montrent que la hiérarchie de difficulté des questions posées est à peu près conservée pour des pays partageant la même langue, mais qu'elle peut être bouleversée entre deux pays ne parlant pas la même langue.

Une définition formelle du FDI peut s'envisager à travers la propriété d'invariance conditionnelle : à niveau égal sur la compétence visée, la probabilité de réussir un item donné est la même quel que soit le groupe de sujets considéré. En réalité, deux conditions sont nécessaires et suffisantes pour qu'un FDI se manifeste : l'item est sensible à une seconde dimension distincte de la dimension principale visée par le test et les groupes se différencient sur cette seconde dimension conditionnellement à la dimension principale. En guise d'illustration, considérons un item, dans une épreuve de mathématiques, qui nécessite la lecture d'un texte. Cet item est donc sensible à une dimension parasite. En outre, les filles ont de meilleures performances en lecture, et ce à niveau égal en mathématiques. L'item est fortement susceptible de présenter un fonctionnement différentiel selon le genre. Ce simple exemple permet d'entrevoir le lien entre dimensionnalité et fonctionnement différentiel, lien qui peut être formellement démontré et qui doit conduire à envisager les FDI de manière plus large que des indicateurs de biais.

5 Perspectives

Les principes méthodologiques présentés ici sont aujourd'hui prédominants dans le domaine des évaluations standardisées. Ce type d'approche comporte cependant des limites. Par exemple, l'hypothèse d'unidimensionnalité est évidemment contestable lorsqu'on sait la multiplicité des compétences mises en jeu lors de la résolution d'une tâche. Des modélisations permettent cependant de prendre en compte la multidimensionnalité, tels que les modèles dits de classification diagnostique qui permettent d'établir des profils d'élèves à partir de leurs réponses et d'une analyse a priori des items selon un cadre théorique autorisant une structure complexe (chaque item est relié à un ensemble d'attributs que les élèves sont censés maîtriser pour réussir l'item).

Du point de vue des perspectives, notons enfin que l'avènement du numérique dans le domaine des évaluations standardisées amènera sans doute progressivement à reconsidérer les modélisations en cours, afin d'intégrer les « traces » laissées par les élèves lors de leur activité pendant l'évaluation.

Références

Arzoumanian, P. & Dalibard, E. (2015). CEDRE 2014 - Mathématiques en fin de collège : une augmentation importante du pourcentage d'élèves de faible niveau, Note d'information, n°19, mai 2015.

Baudelot, C. & Establet, R. (2009). L'élitisme républicain - L'école française à l'épreuve des comparaisons internationales. Paris : Seuil, la République des idées.

DEPP (2015). CEDRE - Rapport technique Sciences expérimentales 2013 Collège, Paris : DEPP. (en ligne : http://educ.gouv.fr/c81218).

Gould, S. J. (1997). La mal-mesure de l'homme, Paris : Odile Jacob.

Trosseille, B. & Rocher, T. (2015). Les évaluations standardisées des élèves. Perspective historique, Education et formations, n°86-87. (en ligne : http://educ.gouv.fr/c88746).