MESURE DES COMPÉTENCES

Méthodes psychométriques utilisées dans le cadre des évaluations des élèves

Thierry Rocher
MENESR-DEPP, bureau de l'évaluation des élèves

Cet article présente les méthodes psychométriques qui sont généralement employées dans les programmes d'évaluations standardisées des compétences des élèves, au niveau national et au niveau international. Nous proposons un panorama de ces méthodes, de façon pédagogique, mais également technique. Leurs fondements théoriques ainsi que leurs hypothèses sous-jacentes sont présentés. Nous montrons leur intérêt d'un point de vue pratique, mais également leurs limites. Enfin, une description des analyses psychométriques réalisées dans le cadre d'une évaluation du cycle Cedre est proposée.

Les programmes d'évaluations standardisées réalisés à 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 [Salines et Vrignaud, 2001 ; Bottani et Vrignaud, 2005 ; Trosseille et Rocher, dans ce numéro, p. 15]. D'un point de vue méthodologique, elles reposent sur des échantillons représentatifs [Garcia, Le Cam et Rocher, dans ce numéro, p. 101] et suivent des procédures standardisées afin de limiter l'erreur de mesure à tous les niveaux (passation, correction, etc.). Ces évaluations se distinguent d'autres enquêtes notamment à travers l'emploi d'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. Ces méthodes restent relativement méconnues

en France. Largement employées dans les évaluations nationales ou internationales, elles sont peu diffusées, que ce soit dans le monde académique, le monde éducatif ou encore celui de la statistique publique. Cet article a pour objectif de dresser un panorama des méthodes psychométriques employées dans les programmes d'évaluations standardisées et de présenter de manière pédagogique leurs fondements théoriques et leurs aspects pratiques. Nous présentons tout d'abord le cadre conceptuel de la mesure des compétences des élèves, qui consiste à considérer que les performances observées aux items d'une évaluation sont les manifestations d'une variable latente, non observable directement. Après avoir introduit quelques éléments descriptifs, nous présentons les modélisations habituellement employées, à savoir les modèles de réponse à l'item. Nous montrons l'intérêt de ces modèles, à la fois sur le plan théorique et sur le plan pratique, et nous étudions les hypothèses fondamentales sur lesquelles ils reposent. Enfin, nous décrivons le déroulement des analyses psychométriques qui sont réalisées dans le cadre d'une évaluation Cedre (cycle des évaluations disciplinaires réalisées sur échantillons).

CADRE GÉNÉRAL

Mesurer une variable latente

Les programmes d'évaluation des acquis des élèves, tels que PISA ou Cedre, se situent au carrefour de deux traditions méthodologiques : celle de la psychométrie, pour ce qui relève de la mesure de dimensions psychologiques, en l'occurrence des acquis cognitifs ; celle des enquêtes statistiques pour ce qui a trait aux procédures de recueil des données.

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 »*. Ainsi, le terme de « construit » est 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-réinscription 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. Notons que cette approche n'est pas propre au domaine de la cognition. On retrouve ce type de variable en économie avec par exemple la notion de propension, en sciences politiques avec la notion d'opinion ou encore en médecine avec la notion de qualité de vie [voir par exemple : Falissard, 2008].
Envisager les résultats à une évaluation comme résultant d'un processus de mesure d'une variable latente ne s'impose pas de lui-même. En effet, il est tout à fait possible de considérer uniquement le nombre de points obtenus à un test et de ne pas donner plus de significations à cette statistique qu'un score observé à un test. Mais cette démarche est très fruste d'un point de vue théorique et trouve vite des limites en pratique, notamment en termes de comparabilité entre différentes populations ou entre différentes épreuves. Le cadre conceptuel de la mesure d'une variable latente est plus adapté à la problématique de l'évaluation des acquis des élèves, comme nous le verrons dans cet article.

Un exemple introductif

Avant d'entrer dans des considérations plus techniques, nous présentons tout d'abord un exemple d'application qui a pour seul objectif d'illustrer de façon pédagogique les grandes notions de psychométrie.
Cet exemple porte sur la taille des individus. 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.
Ce cas d'école est depuis longtemps utilisé aux Pays-Bas dans les cours de psychométrie : Glas [2008] en donne quelques illustrations. 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. À 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[1]. Nous avons calculé un score de façon très simple à partir des 24 questions en attribuant 1 point pour chacune d'entre elles, en fonction de la modalité associée à une taille plus élevée : par exemple, les individus obtiennent un 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, 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 [voir par exemple Newton et Shaw, 2014 ; en français, Laveault et Grégoire, 2002]. 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.
Au-delà de la validité, 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 *« À deux sous un parapluie, c'est souvent moi qui le tiens »*, 89 % des hommes répondent oui contre

1. Il ne s'agit donc pas de la taille exacte mais de la taille déclarée, ce qui peut introduire un décalage, par le jeu des arrondis que les personnes font naturellement concernant leur taille : par exemple, on observe certaines concentrations, autour de 165 cm, mais peu de valeurs telles que 163 cm… Nous supposerons cependant ici que la taille est déclarée sans erreur.

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[2]. L'étude des FDI est fondamentale en matière de comparaison temporelle ou internationale des acquis des élèves. Nous revenons plus en détail sur cette notion par la suite.

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, présenté plus loin, est un indicateur de fidélité du test. En l'occurrence, pour le questionnaire sur la taille, il a pour valeur 0,80, ce qui est 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 de l'article de Glas [2008] : « 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 aux pieds 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

2. Une autre question présente un fonctionnement différentiel du même ordre : « Au supermarché, je dois souvent demander de l'aide pour attraper des produits en haut des gondoles ». Aucun homme ne répond oui à cette question, alors qu'un tiers des femmes répond positivement, en lien avec leur taille… Nous laissons ici au lecteur le soin de formuler sa propre interprétation.

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 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.

APPROCHE CLASSIQUE

Dans un premier temps, nous posons quelques notations et nous présentons les principales statistiques descriptives utilisées pour décrire un test, issues de la « théorie classique des tests » que nous évoquons rapidement.

Réussite et score

On note n le nombre d'élèves ayant passé une évaluation composée de J items. On note Y_i^j la réponse de l'élève i (i = 1, ..., n) à l'item j (j = 1, ..., J). Dans notre cas, les items sont dichotomiques, c'est-à-dire qu'ils ne prennent que deux modalités (la réussite ou l'échec) :

$$Y_i^j = \begin{cases} 1 & \text{si l'élève } i \text{ réussit l'item } j \\ 0 & \text{si l'élève } i \text{ échoue à l'item } j \end{cases} \quad (1)$$

Le taux de réussite à l'item j est la proportion d'élèves ayant réussi l'item j. Il est noté p_j :

$$p_j = \frac{1}{n}\sum_{i=1}^{n} Y_i^j \quad (2)$$

Le taux de réussite d'un item renvoie à son niveau de difficulté. C'est certainement la caractéristique la plus importante, qui permet de construire un test de niveau adapté à l'objectif de l'évaluation, en s'assurant que les différents niveaux de difficulté sont balayés.

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.

Le score observé à l'évaluation pour l'élève i, noté S_i, correspond au nombre d'items réussis par l'individu i :

$$S_i = \sum_{j=1}^{J} Y_i^j \qquad (3)$$

La théorie classique des tests a précisément pour objet d'étude le score S_i obtenu par un élève à un test. Elle postule notamment que ce score observé résulte de la somme d'un score « vrai » inobservé et d'une erreur de mesure. Un certain nombre d'hypothèses portent alors sur le terme d'erreur [pour plus d'informations, voir par exemple Laveault et Grégoire, 2002].

Fidélité

Dans le cadre de la théorie classique des tests, la fidélité (*reliability*) est définie comme la corrélation entre le score observé et le score vrai : le test est fidèle, lorsque l'erreur de mesure est réduite. Une manière d'estimer cette erreur de mesure consiste par exemple à calculer les corrélations entre les différents sous-scores possibles : plus ces corrélations sont élevées, plus le test est dit fidèle[4].
Le coefficient α de Cronbach est un indice destiné à mesurer la fidélité de l'épreuve. Il est compris entre 0 et 1. Sa version « standardisée » s'écrit :

$$\alpha = \frac{J\,\overline{r}}{1 + (J-1)\overline{r}} \qquad (4)$$

où $\overline{r}$ est la moyenne des corrélations inter-items.
De ce point de vue, cet indicateur renseigne sur la consistance interne du test. En pratique, une valeur supérieure à 0,8 témoigne d'une bonne fidélité[5].

Indices de discrimination

Des indices importants concernent le pouvoir discriminant des items. Nous présentons ici l'indice « r-bis point » ou coefficient point-bisérial qui est le coefficient de corrélation linéaire entre la variable indicatrice de réussite à l'item Y^j et le score S. Appelé également « corrélation item-test », il indique dans quelle mesure l'item s'inscrit dans la dimension générale. Une autre manière de l'envisager consiste à le formuler en fonction de la différence de performance constatée entre les élèves qui réussissent l'item et ceux qui échouent. En effet, on peut montrer que :

$$r_{bis-point}(j) = corr(Y^j, S) = \frac{\overline{S}_{(j1)} - \overline{S}_{(j0)}}{\sigma_S}\sqrt{p_j(1-p_j)} \qquad (5)$$

où $\overline{S}_{(j1)}$ est le score moyen sur l'ensemble de l'évaluation des élèves ayant réussi l'item j, $\overline{S}_{(j0)}$ celui des élèves ayant échoué à l'item et σ_S est l'écart-type des scores.

4. Notons au passage que la naissance des analyses factorielles est en lien avec ce sujet : Charles Spearman cherchait précisément à dégager un facteur général à partir de l'analyse des corrélations entre des scores obtenus à différents tests.
5. La littérature indique plutôt un seuil de 0,70 [Peterson, 1994]. Cependant, comme le montre la formule ci-dessus, le coefficient α est lié au nombre d'items, qui est important dans les évaluations conduites par la DEPP afin de couvrir les nombreux éléments des programmes scolaires. Des facteurs de correction existent néanmoins et permettent de comparer des tests de longueurs différentes.

C'est donc bien un indice de discrimination, entre les élèves qui réussissent et ceux qui échouent à l'item. En pratique, on préfère s'appuyer sur les $r_{bis-point}$ corrigés, c'est à dire calculés par rapport au score à l'évaluation privée de l'item considéré. Une valeur inférieure à 0,2 indique un item peu discriminant [Laveault et Grégoire, 2002].

MODÈLES DE RÉPONSE À L'ITEM (MRI)

Dans la pratique, l'approche classique comporte certaines limites. En se concentrant sur l'analyse du score observé, c'est-à-dire du nombre de bonnes réponses aux items d'un test donné, les résultats dépendent de l'ensemble des items considérés. L'approche classique permet donc difficilement 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. Nous présentons quelques-uns de ces modèles.

Présentation générale

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. De manière très générale, les MRI peuvent être présentés de la manière suivante :

$$P(Y = k|\boldsymbol{\theta}, \boldsymbol{\xi}) = F(\boldsymbol{\theta}, \boldsymbol{\xi}, k) \qquad (6)$$

La probabilité qu'un élève donne la réponse k à l'item Y dépend de caractéristiques $\boldsymbol{\theta}$ concernant l'élève et de caractéristiques $\boldsymbol{\xi}$ concernant l'item Y. La fonction F est typiquement une fonction de répartition, à valeur dans]0, 1[.

En comparaison de la théorie classique des tests, ces modèles ont l'avantage de séparer ce qui relève des élèves de ce qui relève des items, la réponse résultant d'une interaction entre ces deux composantes. Les MRI ont un intérêt pratique pour la construction de tests et que nous détaillons par la suite : si le modèle est bien spécifié sur un échantillon donné, les paramètres des items – en particulier leurs difficultés – peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves – en particulier, leur niveau de compétence. Les modèles de réponse à l'item ont donné lieu à une littérature extrêmement fournie. Le lecteur intéressé est invité à consulter, par exemple, Embretson et Reise [2000] ou bien, en français, Bertrand et Blais [2004].

Notre attention va se concentrer sur le cas où $\boldsymbol{\theta}$ est un scalaire (un nombre réel), c'est-à-dire que le MRI est dit unidimensionnel. En outre, nous nous restreignons ici au cas d'items dichotomiques ($k \in \{0,1\}$). Des extensions existent, mais leur présentation sort du cadre de cet article.

Modèle de Rasch (1PL)

Proposé par Rasch [1960], le modèle le plus simple, appelé aussi MRI « à un paramètre » (1PL pour *One-Parameter Logistic*) s'écrit de la manière suivante :

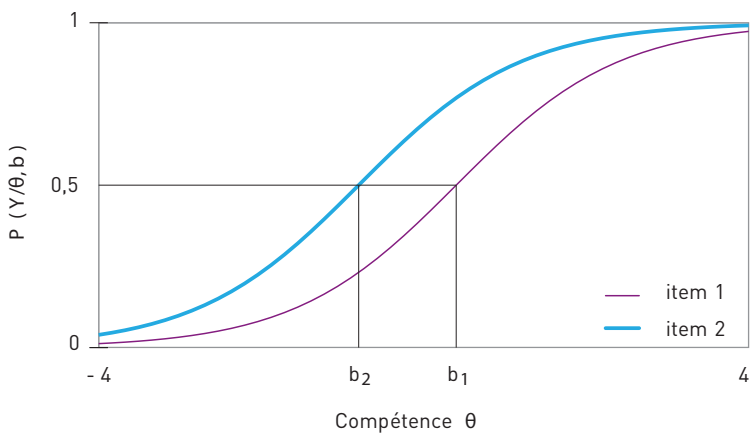
$$P_{ij} = P(Y_i^j = 1|\theta_i, b_j) = \frac{e^{\theta_i - b_j}}{1 + e^{\theta_i - b_j}} \quad (7)$$

i.e. la probabilité P_{ij} que l'élève i réussisse l'item j est une fonction sigmoïde[6] du niveau de compétence de l'élève i et du niveau de difficulté b_j 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.

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 ($\theta_i - b_j$). 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 b_j 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**.

▸ Figure 1 Modèle de réponse à l'item – 1 paramètre



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.

6. La fonction sigmoïde est définie par : $\forall x, f(x) = \frac{e^x}{1 + e^x}$, à valeur dans]0, 1[.

Modèle à deux paramètres (2PL)

Birnbaum [1968] a proposé d'introduire un deuxième paramètre, dit de discrimination :

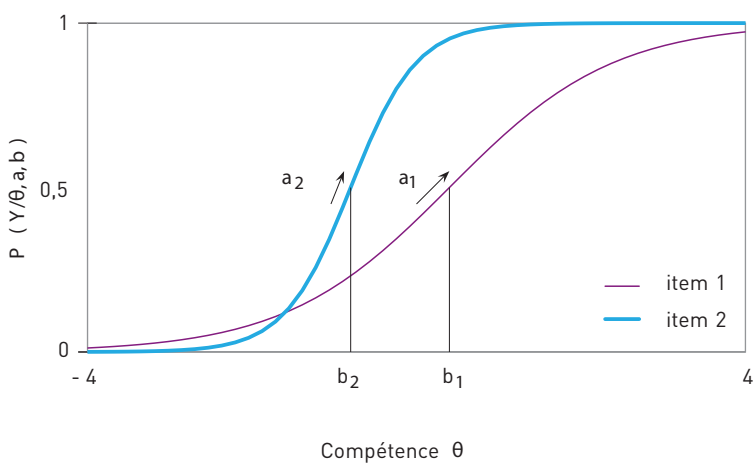
$$P_{ij} = P(Y_i^j = 1|\theta_i, a_j, b_j) = \frac{e^{1,7a_j(\theta_i - b_j)}}{1 + e^{1,7a_j(\theta_i - b_j)}} \quad (8)$$

où a_j ($a_j > 0$) représente la pente au point d'inflexion de la courbe caractéristique de l'item j qui varie d'un item à l'autre et la constante 1,7 est introduite pour rapprocher la fonction sigmoïde de la fonction de répartition de la loi normale ▸ **Figure 2**.

Pour un item très discriminant, la probabilité de le réussir sera très faible en deçà d'un certain niveau de compétence et très élevée au-delà de ce même niveau. Ainsi, une faible différence de niveau de compétence peut conduire à des probabilités de réussite très différentes. C'est le cas de l'item 2 sur la figure 2. De son côté, un item peu discriminant pourra conduire à de faibles différences de probabilité de réussite, pour un écart de niveau de compétence important.

Cet indice de discrimination peut ainsi être interprété en termes de quantité d'information portée par l'item. Nous ne développons pas cette notion statistique ici, mais l'idée est la suivante : un élève qui réussit un item très discriminant se situe très certainement au-dessus du niveau de difficulté de l'item sur l'échelle de compétence, alors que pour un item de discrimination faible, l'incertitude est plus grande quant à la position de l'élève sur l'échelle. De ce point de vue, un item discriminant apporte de l'information.

Figure 2 **Modèle de réponse à l'item – 2 paramètres**



Note de lecture : la probabilité de réussir l'item (en ordonnées) dépend du niveau de compétence (en abscisse). L'item 1 en trait fin est plus difficile que l'item 2 en trait plein ($b_1 > b_2$), et il est moins discriminant ($a_1 < a_2$).

Par rapport au cadre du modèle de Rasch, l'estimation des paramètres est plus complexe (voir annexe p. 57). Mais au-delà des aspects techniques, certaines propriétés ne sont plus valables dans le cas du modèle à deux paramètres. C'est le cas de la propriété dite d'« objectivité spécifique », qui pourrait se résumer au fait que dans le modèle de Rasch, la probabilité de réussir un item d'un certain niveau de difficulté est toujours inférieure à la probabilité de réussir un item plus difficile. Mais dans le cas du modèle à deux paramètres, les courbes caractéristiques peuvent se croiser : un item peut alors apparaître plus facile ou plus difficile selon le niveau de compétence considéré. Selon certains auteurs, la propriété d'objectivité spécifique conférerait à l'opération de mesure en sciences sociales des propriétés équivalentes à celles prévalant en sciences physiques [Andrich, 2004]. Les tenants de cette vision sont donc partisans de construire l'instrument de mesure en fonction des propriétés du modèle de Rasch : les caractéristiques du test doivent satisfaire les exigences du modèle, d'où l'appellation de théorie de réponse à l'item (*Item Response Theory*)[7]. Cependant, en pratique, l'égalité des discriminations imposée par le modèle de Rasch est une contrainte très exigeante : elle revient à éliminer de nombreux items après avoir estimé leur adéquation au modèle. En effet, la prise en compte de la discrimination permet de mieux modéliser le fonctionnement des items et revient finalement à donner un poids plus important aux items les plus discriminants.

Notons enfin que les modèles présentés ne sont pas identifiables (voir annexe p. 57). Il est nécessaire de fixer des valeurs arbitraires concernant la moyenne et l'écart-type des θ. Ainsi que nous l'avons évoqué dans l'exemple introductif, le continuum θ peut s'apparenter à une échelle de température sur laquelle il est possible d'opérer des transformations.

Autres MRI

Ces modèles admettent de nombreuses variantes qui sortent du cadre de cet article. Ainsi, une extension « naturelle » des modèles présentés précédemment vers une représentation multidimensionnelle consiste à supposer que θ n'est plus une seule variable latente, mais un vecteur de dimension D. Des distinctions concernent alors la nature compensatoire ou non des dimensions.
Une autre approche assez courante consiste à introduire un troisième paramètre, dit de pseudo-chances (*guessing*). Il s'agit d'une asymptote horizontale non nulle à la courbe caractéristique de l'item : la probabilité de réussite pour les faibles niveaux de compétence ne tend plus vers 0, mais vers une certaine valeur positive, qui dépend de l'item, et qui représente la chance de réussite « au hasard ».
Enfin, nous ne développons pas le cas d'items polytomiques mais des extensions existent selon que l'on considère que les réponses possibles sont un nombre de points attribués du fait de la maîtrise de différents aspects – *partial credit models* – ou bien que l'on considère que les réponses sont hiérarchisées en niveaux plus ou moins corrects imbriqués les uns dans les autres – *graded response models*.

7. C'est d'ailleurs la terminologie la plus répandue dans la littérature en langue anglaise, bien qu'elle soit critiquée, notamment par Goldstein [1989] qui considère qu'il ne s'agit pas d'une théorie mais bien d'une modélisation.

APPLICATIONS

Les MRI ont de nombreux avantages pratiques. Nous donnons un aperçu de quelques applications montrant l'intérêt d'avoir recours à ces modèles. Bien entendu, leur mise en œuvre est soumise au respect de certaines hypothèses que nous décrivons dans la section suivante.

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 β_j et les discriminations a_j 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 θ_i, 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 θ_i de la deuxième cohorte.
Une autre possibilité consiste à estimer les paramètres des items sur chacun des groupes pris séparément. Les paramètres des items communs sont alors « alignés » de manière à en déduire les différences de compétences entre groupes. En effet, dans le cas du modèle à deux paramètres par exemple, les modifications

$$\theta^* = A\theta + B,\ b_j^* = Ab_j + B \text{ et } a_j^* = a_j \ / \ A \qquad (9)$$

ne modifient pas la probabilité de réussite. L'*equating* consiste alors à déterminer les coefficients A et B tels que les paramètres des items communs aux deux évaluations soient proches, selon qu'ils sont estimés sur un groupe ou sur un autre. Puis, l'ajustement des métriques se déduit en appliquant $\theta^* = A\theta + B$, c'est-à-dire une transformation linéaire, de la même manière que l'on passe des degrés Celsius aux degrés Fahrenheit. De nombreuses méthodes ont été proposées pour estimer A et B [Kolen et Brennan, 2004][8].

8. Par exemple, une méthode très simple, dite *mean/mean*, consiste tout simplement à remplacer les a et les a_j^* par leurs moyennes respectives $\overline{a}$ et $\overline{a^*}$ pour calculer A, puis à remplacer les b_j et les b_j^* par leurs moyennes respectives $\overline{b}$ et $\overline{b^*}$ pour calculer B. De leur côté, Stocking et Lord [1983] ont proposé une procédure plus complexe qui consiste à minimiser une fonction de perte pour trouver A et B.

Une perspective différente, appelée « estimation concourante », 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 (voir annexe). C'est la stratégie qui est retenue dans Cedre, dans la mesure où elle conduit aux résultats les plus stables, ainsi que le rapporte le plus souvent la littérature scientifique sur le sujet.
Dans tous les cas, l'hypothèse est faite que les items communs « fonctionnent » de la même manière, quel que soit le groupe d'élèves considéré. Cela signifie que leur difficulté n'est pas altérée d'un groupe d'élèves à l'autre. Cette hypothèse est fondamentale et renvoie à la notion de fonctionnements différentiels d'items que nous développons plus loin.

Quelques variantes

Nous avons présenté le principe de l'*equating* entre deux groupes d'élèves à partir d'items communs. La même démarche peut être employée pour estimer des niveaux de compétence comparables, pour les mêmes élèves, à différents moments de mesure.
Par exemple, dans le cadre de l'évaluation de l'enseignement intégré de science et technologie (EIST), les élèves ont été suivis de la sixième à la troisième et ont passé cinq évaluations [Le Cam et Cosnefroy, dans ce numéro, p. 283]. Deux évaluations successives comportent des items communs, ce qui a permis de calculer des scores sur une échelle commune aux cinq temps de mesure, facilitant ainsi l'analyse des progressions des élèves au cours du collège. La même méthode a été appliquée aux données du panel d'élèves de sixième pour suivre l'évolution des acquis des élèves entre la sixième et la troisième, à partir d'items repris entre les deux évaluations [Ben Ali et Vourc'h, dans ce numéro, p. 211].

Une autre application du principe d'*equating* consiste à ajuster les métriques *via* les élèves et non plus *via* les items. Par exemple, pour la reprise de l'évaluation Lire, écrire, compter en CM2 [Rocher, 2008], l'épreuve de calcul de 2007 a repris des items de l'évaluation de 1987, mais également des items d'une autre évaluation sur échantillon, datant de 1999. Il n'y avait aucun item commun entre l'épreuve de 1987 et celle de 1999, mais grâce à la reprise des deux épreuves en 2007, il est possible d'estimer les paramètres de l'ensemble des items et d'en déduire les niveaux de compétences positionnés sur la même échelle, quelle que soit l'année considérée.
Un autre dispositif, courant en matière de tests de langues, consiste à estimer au fur et à mesure les paramètres des items, afin de constituer une banque d'items dans laquelle il sera possible de piocher pour proposer aux candidats des épreuves différentes, selon les moments, selon les pays, afin d'éviter les risques d'exposition et de triches tout en garantissant l'établissement de scores comparables quelle que soit l'épreuve passée. Cette mécanique repose parfois sur la gestion de « flux » d'items. Par exemple, un candidat passe une épreuve dont une partie est composée d'items sélectionnés dans une banque dont on connaît les paramètres et l'autre partie est constituée d'items non calibrés, qui ne sont pas pris en compte dans le calcul du score pour ce candidat, mais les données recueillies serviront à estimer les paramètres de ces nouveaux items.

Ce même principe – dit de pré-test/post-test – a été appliqué pour les évaluations nationales exhaustives de CE1 et de CM2 ayant eu lieu entre 2009 et 2012 afin de comparer l'évolution des scores d'une année sur l'autre, alors que les épreuves étaient entièrement renouvelées chaque année, afin d'éviter les risques de bachotage[9].

Évaluations adaptatives

Un cadre d'application très important des MRI est celui des évaluations dites « adaptatives ». Le principe est le suivant : chaque élève passe une première épreuve ; s'il échoue, une épreuve plus facile lui est proposée ; s'il réussit, il passera une épreuve plus difficile. Ce processus itératif conduit à une estimation plus précise et plus rapide du niveau de compétence de chaque élève. En outre, proposer aux élèves des items de difficulté adaptée à leur niveau peut apparaître comme un levier pour favoriser la motivation des élèves par rapport à la situation d'évaluation [Keskpaik et Rocher, dans ce numéro, p. 119]. Avec le développement de l'informatique, cette procédure s'est répandue dans le domaine de l'évaluation [Wainer, 2000]. À chaque item, selon la réponse de l'élève, son niveau de compétence est réestimé et l'ordinateur propose un nouvel item dont la difficulté correspond à ce niveau. En positionnant sur la même échelle les paramètres de difficulté des items et les niveaux de compétences des élèves, les modèles de réponse à l'item sont particulièrement prisés dans le domaine des tests adaptatifs.

La principale contrainte de ce type de procédure est qu'il est nécessaire d'avoir estimé au préalable le niveau de difficulté d'un grand nombre d'items[10]. Cela suppose que chaque item ait été passé par un échantillon représentatif de la population visée, que sa difficulté ait été estimée et enregistrée dans une banque d'items, dans laquelle il sera possible de choisir le plus approprié lors de la procédure de test adaptatif. La constitution d'une telle banque implique un coût financier très important, qui limite la mise en pratique des tests adaptatifs[11].

Il existe d'autres stratégies d'adaptation, moins exigeantes. C'est le cas par exemple des procédures d'orientation (*multi-stage testing*) utilisées dans les enquêtes auprès des adultes IVQ et Piaac [Murat et Rocher, dans ce numéro, p. 83]. L'adaptation des items n'est pas faite individuellement mais pour des groupes de sujets déterminés en fonction de leurs résultats à un test d'orientation. Cette procédure est moins contraignante en pratique. Le recours à l'ordinateur n'est pas requis. Elle a l'avantage de pouvoir être appliquée pour une passation collective de tests papier-crayon, comme ce fut le cas par exemple avec les anciens tests de la Journée d'appel de

9. Pour cette approche, plusieurs approches ont été mises en concurrence, dont les modèles de réponse à l'item. D'ailleurs, après analyse, et pour des raisons pratiques, ce ne sont finalement pas les modèles de réponse à l'item qui ont été retenus mais une approche non paramétrique [Rocher, 2011]. En effet, les comparaisons de résultats entre les années pouvaient être réalisées directement après la passation, dans ces écoles, sur la base des scores observés (nombre de bonnes réponses). L'approche non paramétrique a ainsi permis d'établir des règles simples de passage entre les scores, permettant ainsi à chaque école d'assurer la comparabilité temporelle des résultats. Cela montre que les MRI, bien que très adaptés à ces problématiques, ne sont pas nécessairement incontournables et que d'autres méthodes sont envisageables, selon les contraintes des évaluations.

10. En faisant l'hypothèse sans doute assez forte que le niveau de difficulté de l'item existe indépendamment du test dans lequel il se situe.

11. Autre difficulté, il faut aussi que la réponse du sujet soit corrigée immédiatement, ce qui rend difficile le recours à un codage manuel et impose une procédure d'estimation des compétences intégrée à l'outil de collecte.

préparation à la défense [Rivière, De La Haye *et alii*, 2010]. Elle ne nécessite pas d'estimer au préalable la difficulté des items et donne potentiellement des résultats plus précis que ceux obtenus par un seul test, dans le cas où les niveaux de compétence sont très dispersés (cf. une application aux données d'IVQ : Murat et Rocher, 2009]. Au-delà des aspects pratiques, cette procédure se justifie également sur le plan théorique. Les dimensions cognitives intéressantes à évaluer ne sont pas forcément les mêmes selon les niveaux de compétences. Pour les personnes en difficulté face à l'écrit, il peut être intéressant d'insister sur les processus de bas niveaux comme le décodage des mots, alors que pour les autres personnes, différents aspects de la compréhension pourront être plus finement évalués. Ainsi, ce n'est pas seulement la difficulté du test qui est adaptée, mais la nature même de ce qu'il est censé mesurer.

Cahiers tournants

Nous présentons un autre cas pratique d'utilisation des MRI avec la méthode dite des « cahiers tournants ». Cette méthode est utilisée pour évaluer un nombre important d'items sans allonger le temps de passation. Elle consiste à répartir les items dans des cahiers différents qui comportent des items communs. Cette répartition doit répondre à certaines contraintes[12].

Par exemple, pour l'évaluation Cedre en sciences expérimentales de 2013 en troisième, l'équivalent de six heures et demie d'évaluation ont été créées. En effet, Cedre a pour objectif d'évaluer les acquis des élèves au regard des programmes scolaires. L'« univers » des items est donc très large. Le matériel a été réparti dans 13 blocs d'une demi-heure chacun. Ces blocs ont été ensuite répartis dans 13 cahiers différents, chaque cahier contenant 4 blocs. Ainsi, les élèves sont soumis à deux heures d'évaluation, ce qui est raisonnable.

La manière d'agencer les 13 blocs dans les 13 cahiers « tournants » répond à plusieurs contraintes :

– chaque bloc se retrouve le même nombre de fois au total, afin d'équilibrer le « poids » de chaque bloc ;

– chaque association de blocs (chaque paire) se trouve au moins une fois dans un cahier, afin de pouvoir calculer toutes les corrélations inter-items ;

– un bloc se retrouve à chacune des dispositions possibles : le bloc 1 apparaît en première position dans un des cahiers, en deuxième position dans un autre cahier, etc.

Le **tableau 1 p. 52** donne la répartition des blocs dans les cahiers, pour l'évaluation Cedre de troisième en sciences expérimentales en 2013. Le plan de rotation respecte les principes énoncés ci-dessus. Par ailleurs, cette évaluation est composée pour près de la moitié de blocs d'items repris de l'évaluation de 2007 afin d'établir des comparaisons. Les procédures d'estimation des MRI permettent facilement de gérer les valeurs manquantes aléatoires induites par la méthode des cahiers tournants. En outre, l'objectif est bien de rendre compte de la distribution des niveaux de compétences de manière globale, et non pas de manière individuelle, pour chaque élève, qui n'a pas passé les mêmes items que son voisin.

12. Cette méthode est en réalité une adaptation de procédures d'analyse de variance dans le cas de plans d'expérience incomplets [Cochran et Cox, 1950].

Tableau 1 Répartition des blocs dans les cahiers pour l'évaluation Cedre sciences expérimentales 2013

Cahiers	Séquence 1	Séquence 2	Séquence 3	Séquence 4
1	SVT 1*	SVT 3	SVT 4*	PHY B*
2	SVT 2	SVT 4*	SVT 5	PHY C
3	SVT 3	SVT 5	SVT 6*	PHY D*
4	SVT 4*	SVT 6*	PHY A	PHY E
5	SVT 5	PHY A	PHY B*	PHY F*
6	SVT 6*	PHY B*	PHY C	MIX*
7	PHY A	PHY C	PHY D*	SVT 1*
8	PHY B*	PHY D*	PHY E	SVT 2
9	PHY C	PHY E	PHY F*	SVT 3
10	PHY D*	PHY F*	MIX*	SVT 4*
11	PHY E	MIX*	SVT 1*	SVT 5
12	PHY F*	SVT 1*	SVT 2	SVT 6*
13	MIX*	SVT 2	SVT 3	PHY A

Note de lecture : le cahier 1 est composé de quatre blocs : SVT 1*, SVT 3, SVT 4* et PHY B*.
Les blocs étoilés sont les blocs repris de 2007.

HYPOTHÈSES

L'hypothèse d'unidimensionnalité

L'unidimensionnalité 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'unidimensionnalité doit être rejetée et le modèle ne peut être appliqué.

Bien que fondamentale, cette hypothèse est rarement testée statistiquement. Pour cause, la notion d'unidimensionnalité 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 [Hattie, 1985]. Il faut attendre Stout [1987] pour poser une définition plus formelle de l'unidimensionnalité, 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'unidimensionnalité stricte n'existe probablement pas. Les processus mis en œuvre pour réussir un ensemble d'items sont complexes et varient selon les élèves et les contextes. Dès lors, il est difficilement concevable que ces processus se réduisent rigoureusement à une seule et même dimension [Goldstein, 1980]. C'est pourquoi, en pratique, évaluer l'unidimensionnalité revient en fait à évaluer l'existence d'une dimension dominante [Blais et Laurier, 1997][13].

13. Cela rejoint la démarche en analyse factorielle exploratoire qui consiste à comparer les valeurs propres des différents facteurs. D'ailleurs, les MRI peuvent être vus comme des analyses en facteurs communs [Rocher, 2013].

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 États-Unis, quantité de tests sont passés au crible dans le but de déterminer la présence d'éventuels biais d'items (*« Male/Female »*, *« Black/White »,* etc.) 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 [Vrignaud, 2002]. 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 [Rocher, 2003].

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. Formellement, un fonctionnement différentiel se traduit donc par :

$$P(Y|Z,G) \neq P(Y|Z) \tag{10}$$

où Y est le résultat d'une mesure de la compétence visée, typiquement la réponse à un item ; Z est un indicateur du niveau de compétence des sujets ; G est un indicateur de groupes de sujets.

La probabilité de réussite, conditionnellement au niveau mesuré, est identique pour tous les groupes de sujets. 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 [Rocher, 2013] et qui doit conduire à envisager les FDI de manière plus large que des indicateurs de biais.

Ainsi, une analyse de FDI qui intègre des éléments d'interprétation apporte des renseignements précieux au chercheur qui s'interroge sur les différences entre groupes de sujets, sur la dimensionnalité ou sur le caractère universel de certains concepts [Vrignaud, 2002]. Les biais ne sont alors plus envisagés comme des nuisances dans le processus de mesure, mais comme des éléments explicatifs, au service d'une démarche heuristique.

En pratique, de très nombreuses méthodes ont été proposées afin d'identifier les FDI. Ces méthodes ont chacune des avantages en matière d'investigation des différents

éléments pouvant conduire à l'apparition de ces FDI [Rocher, 2013]. Dans le cas des évaluations standardisées menées à la DEPP, il s'agit avant tout d'identifier les fonctionnements différentiels pouvant apparaître entre deux moments de mesure, s'agissant des items repris à l'identique. Dans ce cas, les différentes méthodes d'identification donnent des résultats relativement proches. Une stratégie très simple, employée dans Cedre, consiste donc à comparer les paramètres de difficulté des items repris, estimés de façon séparée pour les deux années. Si la difficulté d'un item a évolué, comparativement aux autres items, c'est le signe d'un fonctionnement différentiel, qui peut être lié par exemple à un changement de programmes ou de pratiques, comme nous le montrons dans l'illustration présentée dans la section suivante.

MÉTHODOLOGIE SUIVIE POUR LES ÉVALUATIONS CEDRE

En écho aux éléments théoriques exposés, nous présentons concrètement dans cette dernière partie la méthodologie suivie, en matière d'analyse psychométrique, par les évaluations Cedre. Cedre a pour objet de mesurer les acquis des élèves au regard des programmes scolaires, à partir d'évaluations réalisées par des échantillons représentatifs d'élèves, en CM2 et en troisième [voir Trosseille et Rocher, dans ce numéro, p. 15]. Chaque année, une discipline différente est évaluée et des comparaisons sont effectuées tous les cinq ou six ans.
L'exemple retenu est celui de l'évaluation des compétences des élèves de troisième en sciences expérimentales qui a établi une comparaison à six ans d'intervalle, entre 2007 et 2013. Les grandes lignes de la méthodologie employée sur les aspects psychométriques sont présentées. Pour plus de détails, le lecteur est invité à consulter le rapport technique disponible sur Internet [Bret, Garcia *et alii*, 2015].

Le matériel d'évaluation

En 2007, les élèves avaient passé 207 items au total dont 103 ont été repris pour l'évaluation de 2013 et 104 non repris. Cette sélection repose sur des critères statistiques ainsi que pédagogiques. En particulier, des items peuvent ne pas être retenus pour des raisons liées à l'évolution des programmes ou des pratiques.
En 2012, lors de l'expérimentation[14], 106 items ont été testés sur un échantillon d'environ 3 500 élèves. Après analyse, 72 items ont été retenus pour l'évaluation de 2013. Cette sélection repose principalement sur l'examen de statistiques descriptives concernant les items tels que la répartition des réponses données, le taux de réussite, le taux de non-réponse, le pouvoir discriminant (le « r-bis point »). Une vérification est faite quant à la précision des items sélectionnés selon le niveau de compétence[15], afin de s'assurer que le continuum est bien couvert.
Au final, en 2013, les élèves ont passé 175 items, dont 103 étaient des items repris de 2007 et 72 des items nouveaux. Ces items ont été répartis en 13 blocs, ventilés dans 13 cahiers selon le schéma présenté dans le tableau 1 : 7 blocs ont été repris à l'identique

14. Chaque évaluation est précédée d'une phase expérimentale l'année n-1.
15. D'un point de vue technique, la précision d'un item est l'inverse de la racine carrée de l'information de Fisher.

de l'évaluation de 2007 et 6 blocs nouveaux ont été intégrés en 2013.
Notons enfin que sur les 72 nouveaux items introduits en 2013, 32 items sont des questions ouvertes appelant une réponse rédigée et nécessitant la mise en œuvre de procédures standardisées de correction (supervision, corrections multiples, etc.). Chacun des deux formats de questions – QCM et questions ouvertes – présentent des avantages et des inconvénients : les premières forcent les choix de réponse mais garantissent l'objectivité du codage, tandis que les secondes permettent l'authenticité des réponses mais leur correction nécessite d'être très contrôlée [Vrignaud, 2003].

Les étapes

Les principales étapes de l'analyse psychométrique sont les suivantes :

1. Analyse « classique » des items
2. Étude de la dimensionnalité
3. Détection des fonctionnements différentiels d'items (avec le cycle précédent)
4. Étude de la qualité d'ajustement des items au modèle de réponse à l'item (MRI)
5. Application du MRI
6. *Equating* : ancrage avec le cycle précédent pour assurer la comparabilité des scores.

Suite à l'analyse « classique » menée sur l'ensemble des élèves (de 2007 et de 2013), 33 items ont été supprimés pour cause de mauvaise discrimination (r-bis < 0,2) : 19 items de 2007, 13 items communs et 1 de 2013. Il apparaît que cette suppression concerne pour l'essentiel des items construits en 2007, ce qui renvoie en effet à des niveaux de discrimination moins robustes pour cette évaluation, déjà observés en 2007 mais au-dessus du seuil de 0,3 à l'époque. En revanche, nous pouvons observer que l'expérimentation de 2012 a bien joué son rôle puisqu'un seul item présente une mauvaise discrimination. Au final, les analyses portent donc sur une évaluation composée de 85 items de 2007 non repris en 2013, de 90 items de 2007 repris en 2013 et de 71 items nouveaux en 2013.
L'étude dimensionnelle a montré une forte unidimensionnalité. Ainsi, sur les items passés en 2013, l'analyse factorielle des items sur la base des coefficients de corrélations tétrachoriques[16] a révélé une première valeur propre de 32,9 contre 3,6 pour la deuxième, ce qui témoigne de la présence d'une dimension principale prépondérante. En particulier, les items repris de 2007 et les items nouveaux de 2013 peuvent être considérés comme relevant d'une même dimension.
L'analyse des FDI a permis de détecter 5 items (la règle retenue est celle d'un écart de paramètres de difficulté β d'au moins 0,5) : 3 items en faveur de 2007, 2 items en faveur de 2013. Tous ces items sont des items de physique-chimie. Ils ont été éliminés des calculs. L'évolution des programmes est susceptible de produire des FDI. Ainsi, les 3 items présentant un FDI en défaveur des élèves de 2013 sont des items de physique-chimie portant sur la combustion. Or, par le biais de changements de programmes, il se trouve que la combustion n'est plus abordée en troisième. Si

16. Le coefficient de corrélation tétrachorique entre deux items est le coefficient de corrélation estimé entre les deux variables normales latentes qui conditionnent la réussite à chacun des items. Il est moins sensible aux effets seuil et plafond que le coefficient de corrélation linéaire, ou Φ , dans le cas d'items dichotomiques [Rocher, 1999].

ce type d'analyse peut souvent se révéler pertinent[17], il arrive qu'aucune explication ne soit trouvée à l'apparition de FDI.

Le calcul des scores

L'estimation des paramètres des items et des scores a été réalisée sur l'ensemble des élèves des deux années 2007 et 2013. Un modèle de réponse à l'item à deux paramètres a été employé. Ce choix se justifie par la variabilité des items en matière de pouvoir discriminant. Le modèle présente de bons critères d'ajustement aux données. D'ailleurs, les items présentent tous un indice dit de « FIT » acceptable, c'est-à-dire que leurs paramètres estimés permettent de rendre compte correctement des données.

Les scores estimés sont alors standardisés de sorte que les élèves de 2007 aient une moyenne de 250 et un écart-type de 50. Puis, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en deçà duquel se situent 15 % des élèves (groupes 0 et 1), nous déterminons le score-seuil au-delà duquel se situent 10 % des élèves (groupe 5). Entre ces deux niveaux, l'échelle a été scindée en trois parties d'amplitudes de scores égales correspondant à trois groupes intermédiaires. Ces choix sont arbitraires et ont pour objectif de décrire plus précisément le continuum de compétence.

En effet, les modèles de réponse à l'item ont l'avantage de positionner sur la même échelle les scores des élèves et les difficultés des items. Ainsi, chaque item est associé à un des six groupes, en fonction des probabilités estimées de réussite selon les groupes. Un item est dit « maîtrisé » par un groupe dès lors que l'élève ayant le score le plus faible du groupe a au moins 50 % de chance de réussir l'item. Les élèves du groupe ont alors plus de 50 % de chance de réussir cet item.

À partir de cette correspondance entre les items et les groupes, une description qualitative et synthétique des compétences maîtrisées par les élèves des différents groupes est proposée. Ces principaux résultats sont présentés dans une *Note d'information* [Bret, Garcia, Roussel, 2014].

Perspectives

Les principes méthodologiques présentés sont aujourd'hui prédominants dans le domaine des évaluations standardisées. Ce type d'approche comporte cependant des limites. Par exemple, les modèles de réponse à l'item sont des outils puissants, d'un point de vue pratique, mais ils reposent sur des hypothèses fortes. En particulier, l'hypothèse d'unidimensionnalité est évidemment contestable lorsqu'on sait la multiplicité des compétences mises en jeu lors de la résolution d'une tâche.

Comme nous l'avons évoqué, des modélisations permettent de prendre en compte la multidimensionnalité, mais le plus souvent ces modèles sont multi-unidimensionnels, chaque item se rapportant à une seule dimension. C'est cette structure simple qui est le plus souvent considérée alors que c'est sans doute une structure complexe

17. Un autre exemple tiré de Cedre histoire-géographique et éducation civique en troisième entre 2006 et 2012 : les items proposés ayant trait à la connaissance des règles électorales ont présenté des FDI en faveur des élèves de 2012 par rapport aux élèves de 2006. En effet, l'évaluation de 2012 s'est déroulée au mois de mai, en pleine période d'élections.

qui prévaut. Des modèles existent et prennent en considération ces aspects : 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.

Annexe – Procédures d'estimation des MRI

D'un point de vue statistique, les modèles de réponse à l'item peuvent être formulés de manière plus générale comme des analyses factorielles d'items ou encore comme des modèles multiniveaux, avec les items comme effets fixes et le niveau de compétence comme effet aléatoire [Goldstein, Bonnet et Rocher, 2007 ; Rocher, 2013]. Plusieurs méthodes ont cependant été spécifiquement développées pour estimer les paramètres du modèle. Baker et Kim [2004] les décrivent de manière précise. L'estimation est généralement conduite en deux temps : l'estimation des paramètres des items puis l'estimation des θ en considérant les paramètres des items comme fixes. Nous donnons des éléments concernant quelques méthodes, qui reposent sur la maximisation de vraisemblance, pour le modèle de Rasch et pour le modèle à deux paramètres.
Nous reprenons les notations des équations (7) et (8) p. 45 et 46 qui formulent la probabilité P_{ij} d'un élève i de répondre correctement à un item j, respectivement dans le cadre d'un modèle de Rasch et dans le cadre d'un modèle de réponse à l'item à deux paramètres, pour un item dichotomique.

Notons tout d'abord que les modèles présentés ne sont pas identifiables. Par exemple, dans le modèle à deux paramètres, les transformations $\theta_i^* = A\theta_i + B$, $b_j^* = Ab_j + B$ et $a_j^* = a_j / A$ avec A et B deux constantes ($A > 0$), conduisent aux mêmes valeurs des probabilités. Généralement, l'indétermination est levée en standardisant la distribution des θ (moyenne de 0 et écart-type de 1), ou bien dans le cadre du modèle de Rasch en fixant leur difficulté moyenne des items à 0.
Sous l'hypothèse d'indépendance locale des items, la fonction de vraisemblance s'écrit :

$$L(\mathrm{y}, \boldsymbol{\xi}, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{J} P_{ij}^{y_{ij}} [1 - P_{ij}]^{1-y_{ij}} \qquad (11)$$

où y est le vecteur des réponses aux items (*pattern*), $\boldsymbol{\xi}$ est le vecteur des paramètres des items.

Modèle de Rasch

Pour estimer les paramètres du modèle de Rasch, la procédure CML (*Conditional Maximum Likelihood*) peut être employée. Cette procédure consiste à conditionner la vraisemblance par le score observé à l'ensemble des items S_i (nombre de bonnes réponses), qui entretient une relation bijective avec θ_i. En effet, l'intérêt du modèle de Rasch, en matière d'estimation, résulte de ce qu'il définit un modèle exponentiel, au sens statistique du terme. Or, un modèle exponentiel admet une statistique exhaustive[18]. En l'occurrence, le score S_i est une statistique exhaustive pour le modèle de Rasch.

La connaissance d'une statistique exhaustive simplifie grandement la procédure d'estimation des paramètres : conditionner les pattern y_i par S_i permet en effet d'obtenir une vraisemblance indépendante de θ. La densité conditionnelle se calcule alors en utilisant le fait que S_i suit une loi multinomiale. Il s'agit alors d'un problème classique de maximisation de vraisemblance pour estimer les paramètres b_j.

18. Une statistique S(X), fonction de la variable (ou du vecteur) aléatoire X, est dite exhaustive si la loi de X conditionnellement à S(X) est indépendante des paramètres d'estimation.

Modèle à deux paramètres

Dans le cadre d'un modèle MRI à deux paramètres, la propriété d'exhaustivité du score observé n'est plus satisfaite. La procédure CML ne peut être appliquée. D'autres techniques d'estimation, plus coûteuses d'un point de vue algorithmique, doivent être employées.

Une première approche « naturelle » consisterait à annuler les dérivées de L par rapport aux paramètres du modèle, puis résoudre un système de $2J + n$ équations, par exemple avec une méthode itérative de type Newton-Raphson. Cette procédure appelée JML pour *Joint Maximum Likelihood* conduit cependant à des estimateurs biaisés. En effet, le nombre de paramètres augmente avec le nombre d'observations – un θ_i pour chaque observation –, ce qui ne correspond pas au cadre habituel des résultats sur les estimateurs sans biais convergents.

La procédure de maximisation de la vraisemblance marginale MML (*Marginal Maximum Likelihood*) permet de lever cette difficulté.

Estimation des paramètres des items (procédure MML)

La procédure MML consiste à estimer les paramètres des items en supposant que les paramètres des individus sont issus d'une distribution fixée *a priori* (le plus souvent normale). La maximisation de vraisemblance est *marginale* dans le sens où les paramètres concernant les individus n'apparaissent plus dans la formule de vraisemblance.

Si θ est considérée comme une variable aléatoire de distribution connue, la probabilité inconditionnelle d'observer un *pattern* y_i donné peut s'écrire :

$$P(\mathrm{y} = \mathrm{y}_i) = \int_{-\infty}^{+\infty} P(\mathrm{y} = \mathrm{y}_i|\theta_i) g(\theta_i) d\theta_i \qquad (12)$$

avec g la densité de θ.

L'objectif est alors de maximiser la fonction de vraisemblance :

$$L = \prod_{i=1}^{n} P(\mathrm{y} = \mathrm{y}_i) \qquad (13)$$

Cependant, l'annulation des dérivées de L par rapport aux a_j et aux b_j conduit à résoudre un système d'équations relativement complexe et à procéder à des calculs d'intégrales qui peuvent s'avérer très coûteux en termes de temps de calcul.

La résolution de ces équations est classiquement réalisée grâce à l'algorithme EM (*Expectation-Maximization*) impliquant des approximations d'intégrales par points de quadrature. L'algorithme EM est théoriquement adapté dans le cas de valeurs manquantes. Le principe général est de calculer l'espérance conditionnelle de la vraisemblance des données complètes (incluant les valeurs manquantes) avec les valeurs des paramètres estimées à l'étape précédente, puis de maximiser cette espérance conditionnelle pour trouver les nouvelles valeurs des paramètres. Le calcul de l'espérance conditionnelle nécessite cependant de connaître (ou de supposer) la loi jointe des données complètes. Une version modifiée de l'algorithme considère dans notre cas le paramètre θ lui-même comme une donnée manquante.

En outre, ce cadre d'estimation permet aisément de traiter des valeurs manquantes structurelles, par exemple dans le cas de cahiers tournants ou bien dans le cas de reprise partielle d'une évaluation.

Une fois les paramètres des items estimés, ils sont considérés comme fixes et il est possible d'estimer les θ_i, par exemple *via* la maximisation de la vraisemblance donnée par l'équation (10) p. 53. Les enquêtes internationales proposent quant à elles une méthode d'imputation multiple pour estimer les θ_i. Elles fournissent pour chaque élève un jeu de « valeurs plausibles », calculées selon une logique bayésienne, c'est-à-dire qui tient compte de l'information disponible par ailleurs, en l'occurrence celle des questionnaires de contexte [OCDE, 2012].

Pour finir, notons que les logiciels disponibles pour mener à bien ces calculs sont majoritairement des logiciels commerciaux dont le fonctionnement exact n'est pas très explicite. C'est pourquoi nous implémentons actuellement en interne ces procédures avec le logiciel libre R.

BIBLIOGRAPHIE

Andrich D., 2004, "Controversy and the Rasch model", *Medical Care*, vol. 42, No. 1, p. 1-7.

Baker F. B., Kim S.-H., 2004, *Item response theory – Parameter estimation techniques,* 2e ed., New York, Marcel Dekker.

Bertrand R., Blais J.-G., 2004, *Modèles de mesure – L'apport de la théorie des réponses aux items*, Sainte-Foy, Presses de l'Université du Québec.

Bret A., Garcia É., Rocher T., Roussel L., Vourc'h R., 2015, *Cedre, 2013 – Cycle des évaluations disciplinaires réalisées sur échantillons*, Rapport technique, MENESR-DEPP. Consultable en ligne : www.education.gouv.fr/methodologie-cedre.html

Bret A., Garcia E., Roussel L., 2014, « Cedre 2013 – Sciences en fin de collège : stabilité des acquis des élèves depuis six ans », *Note d'information*, n° 14-28, MENESR-DEPP.

Cochran W. G., Cox G. M., 1950, *Experimental designs,* New York, John Wiley and Sons.

Embretson S. E., Reise, S. P., 2000, *Item response theory for psychologists*, New Jersey, Lawrence Erlbaum Associates inc. publishers.

Falissard B., 2008, *Mesurer la subjectivité en santé – Perspective méthodologique et statistique,* 2e éd., Issy-les-Moulineaux, Elsevier-Masson.

Glas C. A. W., 2008, "Item response theory in educational assessment and evaluation", *Mesure et Evaluation en Education*, vol. 31, No. 2, p. 19-34.

Goldstein H., 1980, "Dimensionality, bias, independance and measurement scale problems in latent trait score models", *British Journal of Mathematical and Statistical Psychology*, vol. 33, No. 2, p. 234-246.

Goldstein H., Bonnet, G., Rocher T., 2007, "Multilevel structural equation models for the analysis of comparative data on educational performance", *Journal of Educational and Behavioral Statistics*, vol. 32, No. 3, p. 252-286.

Goldstein H., Wood R., 1989, "Five decades of item response modelling", *The British Journal of Mathematical and Statistical Psychology*, vol. 42, No. 2, p. 139-167.

Hattie J., 1985, "Methodology review : assessing unidimensionality of tests and items", *Applied Psychological Measurement*, vol. 9, No. 2, p. 139-164.

Kolen M. J., Brennan R. L., 2004, *Test equating, scaling and linking,* New York, Springer.

Laveault D., Grégoire J., 2002, *Introduction aux théories des tests en psychologie et en sciences de l'éducation,* Bruxelles, De Boeck.

MURAT F., ROCHER T., 2009, « Création d'un score global dans le cadre d'une épreuve adaptative », *Économie et Statistique*, n° 424-425, Insee, p. 149-178.

NEWTON P., SHAW S., 2014, *Validity in Educational and Psychological Assessment,* London, Sage Publications Ltd.

OCDE, 2012, *PISA 2009 – Technical Report*, Paris, OCDE.

PETERSON R. A., 1994, "Cronbach's Alpha Coefficient : A Meta-Analysis", *Journal of Consumer Research*, No. 21, p. 381-391.

RIVIÈRE J.-P., DE LA HAYE F., GOMBERT J.-E., ROCHER T., 2010, « Les jeunes Français face à la lecture : nouvelles pistes méthodologiques pour l'évaluation massive des performances cognitives », *Revue Française de Linguistique Appliquée*, n° 15, p. 121-144.

ROCHER T., 2013, *Mesure des compétences : les méthodes se valent-elles ? Questions de psychométrie dans le cadre de l'évaluation de la compréhension de l'écrit*, thèse de doctorat, Université Paris Ouest Nanterre La Défense.

ROCHER T., 2011, « Ajustement des évaluations nationales de CM2 (janvier 2009 - janvier 2010) », *Document de travail, série « Méthodes »*, n° 2011-M04, MEN-DEPP.

ROCHER T., 2008, « Lire, écrire, compter : les performances des élèves de CM2 à vingt ans d'intervalle (1987-2007) », *Note d'information*, n° 08.38, MEN-DEPP.

ROCHER T., 2003, « La méthodologie des évaluations internationales de compétences », *Psychologie* et *Psychométrie*, vol. 24, n° 2-3, p. 117-146.

STOCKING M. L., LORD, F. M., 1983, "Developing a common metric in item response theory", *Applied Psychological Measurement*, vol. 7, No. 2, p. 201-210.

STOUT W., 1987, "A non parametric approach for assessing latent trait unidimensionality", *Psychometrika*, vol. 52, No. 4, p. 293-325.

VRIGNAUD P., 2008, « La mesure de la littératie dans PISA : la méthodologie est la réponse, mais quelle était la question ? », *Éducation & formations,* n° 78, MEN-DEPP, p. 69-84.

VRIGNAUD P., 2003, « Objectivité et authenticité dans l'évaluation – Avantages et inconvénients des questions à choix multiples et des questions à réponses complexes : importance du format de réponse pour l'évaluation des compétences verbales », *Psychologie et Psychométrie*, vol. 24, n° 2-3, p. 147-188.

VRIGNAUD P., 2002, « Les biais de mesure : savoir les identifier pour y remédier », *Bulletin de Psychologie*, vol. 55, n° 6, p. 625-634.