LA MESURE DES COMPÉTENCES : LES LOGIQUES CONTRADICTOIRES DES ÉVALUATIONS INTERNATIONALES

X. D'HAULTFOEUILLE [*], *F. MURAT* [**] *et T. ROCHER* [***]

[*] INSEE, Unité "Méthodes Statistiques"
[**] INSEE, Division "Etudes Sociales"
[***] Ministère de l'Education Nationale, DPD

Le besoin d'évaluer le fonctionnement du système éducatif sur des bases objectives est devenu particulièrement fort, à la fois chez les acteurs de ce système (décideurs politiques, professeurs, etc.) et chez les « consommateurs » (les élèves et leurs parents). Depuis 1992, l'OCDE publie chaque année pour répondre à cette demande une série d'indicateurs sur l'école[1]. Cette publication propose des informations sur les moyens mis en œuvre (part du PIB consacrée à l'éducation, formation des enseignants, taille des classes...), ainsi que des chiffres relatifs aux résultats obtenus : taux d'accès en fin d'études secondaires, efficacité des diplômes sur le marché du travail mais aussi des indicateurs concernant les compétences des populations scolaires ou ayant fini leurs études. La place grandissante de ces derniers indicateurs incitent à s'interroger sur leur valeur.

Dans un premier temps, on présentera les objectifs des évaluations de compétences, c'est-à-dire comment elles permettent d'éclairer le débat sur le fonctionnement du système éducatif. On s'intéressera plus particulièrement au cas des enquêtes internationales. Nous donnerons ensuite un aperçu des théories statistiques inhérentes à ces domaines. Les deux dernières parties seront consacrées aux problèmes concernant l'élaboration de la mesure et la comparabilité des résultats d'un pays à l'autre. Ces difficultés affectent les niveaux moyens aussi bien que les inégalités au sein de chaque population.

[1] Cette publication annuelle a pour titre « Regards sur l'Education ».

1. Evaluation des compétences : définition et utilité

1.1. Evaluation individuelle et évaluation statistique

A moins d'avoir effectué sa scolarité en marge du système scolaire habituel, il est probable que le lecteur aura déjà répondu, plus ou moins directement, à plusieurs centaines de tests d'évaluation de ses compétences. On peut donner quelques exemples : la dictée de CE2, le contrôle de math de 3ème, un bac « blanc », le vrai bac, un concours de la fonction publique, les tests psychotechniques des « 3 jours », la note administrative pour les attachés INSEE... Toutes ces formes d'évaluations ont des objectifs bien différents mais au moins un point commun : elles ont une implication essentiellement individuelle. C'est l'individu interrogé qui a intérêt à réussir l'évaluation, pour obtenir un diplôme ou une place sur le marché du travail ou quelques mois d'avancement dans son échelle de salaires.

Les évaluations *statistiques* dont nous allons parler ont une visée bien distincte : c'est le système éducatif dans son ensemble qui est évalué au travers d'un échantillon d'individus. Ceux-ci n'ont généralement rien à gagner[2]. Il s'agit d'étudier le niveau de la population *dans son ensemble* (c'est-à-dire en moyenne ou en considérant plus finement la distribution des résultats, en particulier pour les groupes les plus en difficulté). Ce changement de perspective explique bien des particularités des évaluations statistiques de compétences[3].

De plus, ces évaluations sont souvent accompagnées d'un questionnement statistique plus « classique ». En même temps qu'on évalue un élève, on va recueillir une somme d'informations relatives à ses caractéristiques sociodémographiques, ses opinions sur l'école, celles de ces parents, les opinions et pratiques pédagogiques de ses professeurs, les caractéristiques de l'établissement qui l'accueille... Dans le cas où l'échantillon est suivi sur plusieurs années, on s'intéresse aussi à son parcours scolaire, aux différents établissements qu'il a fréquentés. On le voit, il existe une quantité formidable de données qui pourront servir dans certains cas à expliquer les

---

[2] On verra d'ailleurs que cela peut avoir un certain effet sur leur motivation.

[3] Il ne faut cependant pas établir une frontière trop rigide entre les deux types d'évaluations. Certaines peuvent remplir, de façon plus ou moins satisfaisante, les deux objectifs. L'obtention du baccalauréat n'est pas qu'un événement heureux pour le candidat : un taux élevé de réussite est aussi un motif de satisfaction pour l'école. De même les évaluations nationales en CE2 et 6ème ont des objectifs prioritairement pédagogiques (il s'agit d'aider chaque professeur à repérer les difficultés rencontrées par les élèves) mais sont aussi utilisées pour donner une image de l'état des compétences à ces deux niveaux.

compétences de l'individu et dans d'autres cas seront elles-mêmes éclairées par la connaissance de ces compétences.

La constitution d'une évaluation de compétence fait appel à un processus un peu différent de celui d'une enquête ordinaire :

- Tout d'abord, un groupe de travail composé de professeurs et de pédagogues constitue un protocole d'évaluation à partir d'un tableau de compétences. Il est alors important de bien définir la ou les compétences que l'on veut évaluer (cf. 1.4). Un soin particulier est apporté à ce que chaque compétence jugée importante soit évaluée par un ou plusieurs exercices. Un autre aspect qui fait différer sensiblement ces protocoles des évaluations ordinaires est la recherche d'une standardisation la plus poussée possible. Tout cahier d'exercices doit pouvoir être corrigé de la même façon par n'importe quel correcteur (qu'on suppose tout au moins spécialiste du domaine évalué). Cela explique peut-être pourquoi les protocoles comportent souvent des QCM et pourquoi certains domaines sont rarement concernés (expression écrite et a fortiori orale).

- Vient ensuite une phase plus habituelle de recueil de l'information, passant par la constitution d'un échantillon (en utilisant souvent le degré « école » ou « classe ») l'envoi des questionnaires, leur passation, la correction (souvent par les professeurs des élèves interrogés), le retour des questionnaires, leur saisie (parfois par lecture optique) et les premiers traitements statistiques indispensables (« nettoyage » des fichiers, repondération...). Les enquêtes à domicile présentent, quant à elles, des difficultés particulières.

- C'est alors la phase la plus spécifique de ce type d'enquête : l'élaboration d'indicateurs de résultats. On a recours à des modélisations assez complexes, que l'on présentera dans les parties suivantes.

- Enfin, une fois les indicateurs de résultats construits, il ne reste plus qu'à les étudier dans des directions variées (voir chapitre 1.2.)

Cette présentation sommaire de la façon dont on élabore une évaluation de compétences met en évidence la diversité des acteurs. Parmi les plus importants, on peut citer : les pédagogues à l'origine des protocoles, les psychométriciens élaborant la mesure des compétences et les statisticiens étudiant les résultats.

1.2. Utilité des évaluations de compétences

L'analyse de ces évaluations peut se faire sous des angles très divers. Nous allons présenter différentes voies possibles en nous attardant davantage sur celles qui seront abordées dans la suite.

1) **Analyses pédagogiques** : même si nous n'insisterons pas sur cet aspect, il faut se souvenir que les résultats publiés ont peu de sens si l'on ne revient pas au protocole d'évaluation. On ne peut se contenter d'une moyenne en affirmant qu'on mesure la « lecture » ou les « mathématiques ». L'analyse par sous-domaines, voire par exercices doit être menée, généralement par les personnes à l'origine des questionnaires, qui pourront à la lumière de leurs attentes permettre de dire si les résultats sont satisfaisants ou non. Ce type d'études peut conduire les décideurs à changer les programmes, pour que soit mis l'accent sur un domaine qui est apparu peu maîtrisé.

2) **Analyses statistiques** : on se contente ici d'un indicateur très synthétique de résultats pour étudier les liens qu'il entretient avec d'autres variables. Cette analyse peut se faire à deux niveaux :

a) **Niveau individuel** : on étudie de façon plus ou moins fine, dans le cadre d'un échantillon donné, les écarts de résultats entre individus. Cela peut consister à calculer les différences de scores moyens entre plusieurs groupes (écarts entre filles et garçons, entre ouvriers et cadres...). Pour aller plus loin, on a souvent recours aux techniques économétriques, pour isoler les effets propres de chaque caractéristique. D'autres pistes d'études à l'échelle individuelle peuvent être envisagées : lien entre compétence et parcours scolaire, mesure de la progression en procédant à deux évaluations, au début et en fin d'année, études docimologiques (comparaison des résultats à une évaluation standardisée et des notes données par le professeur)... Les évaluations de compétence permettent d'éclairer d'une façon originale le débat sur le fonctionnement du système éducatif.

   Cependant, ce débat s'est déjà beaucoup développé sans avoir recours aux comparaisons de compétences. On étudie souvent les inégalités devant l'école à l'aide des différences de parcours scolaires (obtention de tel ou tel diplôme, redoublement, etc.). La netteté des résultats généralement dégagés conduit à se demander ce qu'apporte l'usage des comparaisons de compétences.

   En fait, la plupart du temps, on peut regretter que les indicateurs de parcours scolaire donnent une vision trop schématique de la réussite ou de l'échec : la population ne se ramène pas à deux groupes, les « bons » et les « mauvais », ceux qui sont en retard et les autres. De même, dans une perspective diachronique, on évoque souvent la dévaluation des diplômes ce qui remet en cause l'usage de cette donnée comme mesure absolue de la

réussite scolaire. Enfin, la question se pose de façon encore plus cruciale quand on s'intéresse aux élèves des écoles primaires car pour mettre en évidence des inégalités sociales, encore faut-il mesurer des écarts ! Or, au début de l'école primaire, il n'y a pas de diplôme, pas encore de retard significatif et pourtant il serait audacieux d'affirmer qu'il n'existe encore aucun écart entre enfants. C'est là sans doute la grande utilité des évaluations de compétences que de permettre d'étudier les écarts sociaux alors qu'ils ne sont pas encore matérialisés en divergences de parcours scolaires. Il va sans dire que ces considérations ne remettent pas en cause les résultats obtenus jusqu'à présent en utilisant l'accès à tel diplôme ou telle formation comme indicateur de réussite scolaire. L'ampleur des écarts qu'ils permettent de mettre en évidence, malgré leur caractère un peu schématique, justifie leur usage. De plus, ces variables ont une signification claire et concrète (on peut savoir parfaitement qui a son bac ou non), ce qui, on le verra, n'est pas forcément le cas des mesures psychométriques.

b) **Niveau agrégé** : on ne retient alors qu'une valeur très synthétique calculée sur l'ensemble de l'échantillon (la moyenne, l'écart-type ou l'écart entre deux groupes caractéristiques) et on compare cette valeur à celle que l'on trouve sur un autre population à une évaluation identique. On peut ainsi envisager différents types de comparaisons :

   i) ***Comparaisons temporelles*** : on compare l'indicateur avec la valeur obtenue autrefois, pour déterminer si la situation s'est améliorée ou pas.

   ii) ***Comparaisons internationales*** : on compare l'indicateur avec ce que l'on trouve dans d'autres pays pour situer le système éducatif français et chercher des pistes auprès des pays les meilleurs. Comme dans le cas des comparaisons temporelles d'ailleurs, les analyses peuvent se ramener à deux grandes questions :

      (1) *Etude des moyennes.* Le niveau est-il bon ? la moyenne est-elle élevée ?

      (2) *Etude des écarts-type.* Les inégalités sont-elles importantes ? L'écart-type est-il élevé ou l'écart entre enfants de cadres et enfants d'ouvriers est-il élevé ?

Ce sont ces deux questions et surtout la première qui vont nous préoccuper ici, dans le cadre particulier des enquêtes internationales.

1.3. Les enquêtes internationales

Les récentes enquêtes internationales à grande échelle concernent généralement les mathématiques, les sciences ou la lecture. Le processus d'élaboration des questionnaires et de traitement des données est sensiblement différent de ce qui se fait en France (par exemple à la Direction de la Programmation et du Développement) sur deux points importants : d'une part, les concepteurs des évaluations ont souvent recours à la technique des cahiers tournants (l'ensemble des items conçus, généralement assez large, est découpé en parties non disjointes d'une quarantaine de questions ; chaque individu ne passera qu'un seul de ces sous-ensembles) ; d'autre part, les responsables de ces enquêtes font souvent appel à des techniques spécifiques de traitement des données. Enfin, le souci d'établir des comparaisons équitables impose une expertise poussée des procédures d'échantillonnage et de calcul des résultats.

Dans la suite, nous allons tirer la plupart de nos exemples des deux dernières enquêtes effectuées par les organismes internationaux. Il convient donc de les présenter rapidement.

1.3.1. TIMSS (Third International Mathematics and Science Study)

Plus de quarante pays ont participé à cette enquête sur les mathématiques et les sciences, organisée par l'IEA (International Association for the Evaluation of Educational Achievement) pendant l'année scolaire 1994/1995. Trois populations différentes étaient visées : la population 1 regroupant les élèves de 9 ans (le plus souvent en $3^{ème}$ et $4^{ème}$ années de scolarité), la population 2 les élèves de 13 ans (en $7^{ème}$ et $8^{ème}$ années) et la population 3 les élèves de fin d'études secondaires.

Un test différent était administré à chacune des trois populations. L'évaluation portait sur plusieurs sous-domaines comme la géométrie ou les probabilités pour les mathématiques ou bien la physique ou la chimie pour les sciences. Cependant, les concepteurs n'ont conservé que deux échelles : une pour les mathématiques et l'autre pour les sciences, afin de publier un classement général pour chacun de ces deux domaines (même si des résultats par sous-domaines, voire par items, sont aussi présentés).

1.3.2. IALS (International Adult Literacy Survey)

Cette enquête a eu lieu la première fois en 1994. Elle concernait les individus âgés de 16 à 65 ans de huit pays. En 1997, on comptait au total environ vingt-cinq pays de l'OCDE ayant participé aux différentes vagues de l'enquête. L'Office national de statistique canadien (Statistics Canada) et l'institut privé américain ETS (Educationnal Testing Service) ont mené cette étude en collaboration avec l'OCDE.

Le test, inspiré de l'enquête américaine NALS (National Adult Literacy Survey), visait à mesurer les capacités de lecture et de compréhension de documents de la vie courante. Partant du concept de « littératie », l'évaluation concernait trois domaines : la compréhension de textes suivis, la compréhension de textes schématiques et la compréhension de textes quantitatifs. Pour chacun de ces domaines, les concepteurs ont défini cinq niveaux de compétences. En général, les publications donnent pour chaque domaine la répartition des personnes interrogées, selon ces niveaux. Il n'y a pas d'analyse pédagogique plus détaillée.

Il est important de souligner qu'il s'agissait là, contrairement à TIMSS, d'une enquête portant sur une population d'adultes interrogés à leur domicile. Ce public, ayant pour une part quitté le système scolaire depuis longtemps, se montre sans doute moins réceptif à un questionnement sur leurs compétences. De plus, il va de soi que l'on ne cherche pas à mesurer des compétences scolaires, mais alors, que veut-on évaluer ?

1.4. Les compétences en question

L'évaluation statistique de compétences amène à s'interroger de façon marquée sur ce que l'on veut mesurer. Les intitulés des enquêtes font référence aux mathématiques, aux sciences ou à la littératie. Les responsables se proposent de donner une image des compétences d'une population dans ces domaines. Ceci amène plusieurs questions. On peut tout d'abord se demander si la compétence en mathématiques ou en lecture existe ou si au contraire on ne mesure que des performances à un test (performances que l'on juge plus ou moins représentatives des tâches que l'individu peut être amené à remplir). Et même si l'on accepte, au vu de l'analyse statistique, l'existence de facteurs latents, rien n'assure qu'on puisse les réduire à un seul. Dans le cas des mathématiques, on verra que l'on peut distinguer compétences en algèbre et en géométrie. La question est de savoir jusqu'à quel niveau de détail on doit aller et s'il faut supposer une dimension différente pour chaque processus mental (il y aurait une dimension pour l'addition, une pour la multiplication, une pour le repérage d'un mot, une pour le repérage d'une lettre, etc.). Cette préoccupation est extrêmement importante dans le domaine scolaire, où du fait de l'étendu des programmes, on pose souvent des exercices sur des sujets très variés (elle est beaucoup moins forte dans le cas des tests de QI, où les items recourent à des raisonnements très proches voire quasi identiques).

On peut aussi s'interroger sur le caractère universel de la dimension mesurée. La lecture en français et la lecture en japonais relèvent-elles de la même compétence ? Même une discipline comme les « sciences » recouvre des définitions et des contenus d'enseignement très différents d'un pays à l'autre.

Ce problème est encore plus délicat lorsqu'on travaille sur des populations adultes (cas de l'enquête IALS). On peut s'intéresser aux compétences qui sont utiles sur le

marché du travail mais on se heurte à une nouvelle difficulté : est-ce qu'un jeune homme ou une personne mûre ont besoin des mêmes compétences pour trouver un emploi ? Est-ce que ce sont les mêmes compétences qui facilitent l'insertion professionnelle pour un homme ou une femme, en France et au Chili ?

Un dernier point mérite d'être évoqué concernant les difficultés de mesure. Il peut paraître plus facile de repérer les individus qui ne maîtrisent pas les compétences de base pour effectuer convenablement les actes de leur vie quotidienne que de distinguer chez les personnes n'ayant pas de difficultés, celles qui sont très compétentes de celles qui le sont un peu moins. En particulier, on peut penser que la hiérarchie obtenue sera très sensible au domaine évalué. Pour reprendre la problématique du paragraphe précédent, une fois que l'on a vu que tel individu avait les bases en lecture et en calcul pour s'insérer convenablement dans la société, faut-il voir s'il est Shakespeare ou cordonnier ?

Ce problème concernant la dimensionnalité du domaine d'intérêt est fondamental car la constitution du questionnaire et les traitements statistiques à effectuer en dépendent. Cependant, le discuter en détail conduirait à aborder les théories psychologiques, sociologiques et philosophiques, autant que techniques. Nous nous placerons donc par la suite dans le cadre dominant des évaluations internationales, qui suppose l'unidimensionnalité du domaine évalué, en évoquant parfois les quelques alternatives qui sont proposées.

2. Eléments de psychométrie

Avant de commenter les résultats issus des évaluations internationales, il paraît souhaitable de présenter leurs fondements théoriques. On insistera plus particulièrement sur les Modèles de Réponse à l'Item (MRI), dont l'emploi est largement répandu dans les organismes internationaux.

2.1. Présentation rapide et historique

La tentation de mettre l'âme humaine en chiffre est très ancienne. Au siècle dernier, on a cherché dans les caractéristiques du cerveau, en particulier son volume, l'origine de l'intelligence, avec souvent l'intention plus ou moins explicite de justifier les inégalités sociales, raciales et sexuelles. Dans un autre état d'esprit, les premiers tests d'intelligence élaborés par Binet étaient destinés à repérer les enfants en difficultés pour leur apporter l'aide nécessaire. Ils sont toutefois assez rapidement devenus des outils au service d'une vision plus conservatrice, voire eugéniste, de la société. Pour cette raison et à cause d'un certain manque de fondement du point vue de la théorie psychologique, ils ont été assez contestés avant de retrouver un certain crédit.

Du point de vue technique, l'étude des réponses aux tests d'intelligence a donné lieu à une théorie assez développée, partant le plus souvent d'une approche probabiliste. Les outils d'analyses de données ont ainsi vu le jour, en grande partie pour alimenter le débat sur la structure factorielle de l'intelligence (un seul facteur, appelé *g*, ou plusieurs comme celui d'intelligence verbale, etc.). Dans le domaine des évaluations scolaires, la dernière décennie a vu le développement rapide des Modèles de Réponse à L'Item (MRI) très souvent employés dans le cadre des comparaisons internationales.

Le petit exposé qui va suivre ne prétend pas être un cours complet et structuré de psychométrie mais donne quelques éléments indispensables à la compréhension des résultats publiés dans le domaine de l'évaluation des compétences.

2.2. La théorie de la mesure

2.2.1. Les différents « niveaux » de mesure

Les manuels de psychométrie commencent généralement par des considérations assez détaillées sur la nature de la mesure. Les psychométriciens ne se contentent pas de la distinction entre variables qualitatives et variables quantitatives mais ont élaboré une théorie de la mesure beaucoup plus formalisée.

On considérera comme mesure un ensemble de symboles, généralement chiffrés, représentant des faits empiriques. Les différents types de mesure doivent respecter quelques propriétés de base (définition d'une relation d'identité symétrique et

transitive) et se distinguent entre elles en fonction des propriétés supplémentaires qu'elles possèdent. On les classe ainsi en différents « niveaux », que l'on peut aussi définir en donnant les transformations applicables sans perte d'information. Ces notions permettent ensuite de définir les statistiques appropriées. Quatre échelles principales sont généralement présentées :

- l'échelle ***nominale***, où la mesure est un simple ensemble d'étiquettes. Toute bijection de l'ensemble d'étiquettes dans un autre donne une mesure équivalente (en revanche, une transformation agrégeant certains cas donne une mesure non équivalente). Exemples : la nomenclature de professions INSEE.

- l'échelle ***ordinale***, qui impose l'existence d'une relation d'ordre. On obtient une mesure équivalente si la mesure finale respecte une relation d'ordre et que deux objets classés dans un certain ordre pour la première mesure le sont pour la seconde (la transformation doit donc être monotone stricte). Exemples : les classes d'âge, les niveaux scolaires, les classements sportifs...

- l'échelle ***d'intervalle***, où l'on arrive à une notion plus proche de ce que l'on entend par mesure, c'est-à-dire une donnée chiffrée sur laquelle on peut effectuer des opérations arithmétiques. La propriété fondamentale de ce niveau de mesure est que les écarts entre les valeurs ont une signification. Un écart de 10 unités a la même importance si le niveau est 20, -350 ou 2000 : il est 2 fois plus grand qu'un écart de 5. Ce rapport entre les écarts doit se retrouver dans toutes les mesures équivalentes ce qui implique que l'on n'obtient une mesure équivalente que par une transformation affine sur les nombres. Exemples : les températures °C ou °F, la position sur un axe de coordonnées, la plupart des variables psychologiques...

- l'échelle de ***rapport***, où non seulement les écarts, mais aussi les rapports, doivent être comparables. Cela impose l'existence d'un 0 absolu. Seule la multiplication par une constante permet de passer du mètre à l'inch, de la livre au kilo, de la seconde à l'année, de l'euro au dollar.

2.2.2. Le choix des indicateurs

Les précédentes distinctions ne sont pas simplement d'ordre théorique mais ont des implications pratiques fortes : il est indispensable de s'interroger sur la nature de la mesure quand on souhaite construire des indicateurs de résultats.

Il y a plusieurs façons de voir si la mesure sur laquelle on travaille autorise ou non une statistique. On peut par exemple s'intéresser à la façon dont la statistique est obtenue et l'interdire si elle utilise des opérations non autorisées : par exemple, en toute rigueur, on ne peut pas calculer de moyenne sur des données ordinales puisque l'addition n'est pas possible ; on ne peut pas non plus calculer de rapport interdéciles sur des échelles d'intervalle.

Une autre méthode consiste à déterminer si l'indicateur donne les mêmes résultats sur des échelles équivalentes, obtenues par les transformations possibles. Nous allons prendre deux exemples : celui de l'écart-type et du coefficient de variation. Ces indicateurs sont-ils valides sur les échelles d'intervalles et sur les échelles de rapport ?

On passe d'une échelle d'intervalle E_{int} de mesure M_{int} à une échelle d'intervalle équivalente E_{int}' de mesure M_{int}' par une transformation linéaire $M_{int}'=a\times M_{int}+b$ (avec a différent de 0, de préférence strictement positif). Pour une échelle de rapport E_{rap} de mesure M_{rap}, la transformation est $M_{rap}'=a\times M_{rap}$ (avec a différent de 0).

Voyons si l'interprétation de deux écarts-type est affectée par le changement de mesure. Si deux populations ont sur l'échelle E_{int} des écarts-type s_1 et s_2 tels que $s_1=K\times s_2$ alors sur l'échelle E_{int}', $s_1'= a\times s_1'$, $s_2'= a\times s_2'$ et l'on retrouve $s_1'=K\times s_2'$. On voit immédiatement que les transformations permises pour les échelles de rapports respectent aussi cette propriété. L'ordre des dispersions pour l'une des mesure est le même pour l'autre.

En ce qui concerne le coefficient de variation commençons par les échelles de rapports. Si deux populations ont sur l'échelle E_{rap} des coefficients de variation $CV_1=s_1/m_1$ et $CV_2=s_2/m_2$, en passant à l'échelle équivalente E_{rap}' on obtient $CV_1'=s_1'/m_1'=(a\times s_1)/(a\times m_1)=s_1/m_1=CV_1$. Le coefficient de variation n'est pas affecté par un changement d'échelle (c'est ce qui explique sa popularité). Les interprétations sont donc équivalentes d'une échelle à l'autre.

En revanche si l'on travaille sur des échelles d'intervalles, la transformation est plus compliquée et donne comme équivalent $CV_1'=s_1'/m_1'=(a\times s_1)/(a\times m_1+b)$ et $CV_2'=s_2'/m_2'=(a\times s_2)/(a\times m_2+b)$. On ne voit pas clairement l'équivalence des deux échelles. Plutôt que d'effectuer un travail d'analyse sur la fonction $f(x,y)=(a\times x)/(a\times y+b)$, nous allons prendre un petit exemple. Soient deux séries de température en °F qui donnent pour l'une : une moyenne de 50°F et un écart-type de 5°F et pour l'autre une moyenne de 80°F et un écart-type de 10°F. Les coefficients de variation sont 0,1 et 0,125 : la deuxième série paraît pour cet indicateur plus dispersée. Si nous convertissons toutes ces températures en °C (°C=5/9 °F-160/9) les moyennes deviennent respectivement 10 °C et 26,67 °C, les écarts-type 2,78 °C et 5,56 °C et les coefficients de variations 2,78 et 2,08 : c'est la première série qui paraît alors la plus dispersée ! Il paraît donc difficile de donner un jugement fiable à partir de cet indicateur. On note en revanche que l'examen des écarts-type sont concordants. Cela ne signifie pas que cet indicateur soit vraiment pertinent mais il respecte au moins la clause de base de tout jugement scientifique, la non-contradiction.

2.3. *Indicateurs classiques*

Les indicateurs les plus simples que l'on puisse construire à partir d'une épreuve composée de N questions proposées à un échantillon de K individus sont les suivants :

- On mesure le niveau de l'individu par un *score*, la proportion de bonnes réponses, c'est-à-dire le nombre de bonnes réponses divisé par le nombre total de questions.

- On mesure la difficulté d'un item par la proportion d'individus qui le réussissent.

- Pour les items, on définit généralement une autre caractéristique : sa discrimination. Il s'agit de la corrélation, calculée sur l'ensemble des individus, entre la réussite à l'item et le score calculé sur la totalité des items (ou très souvent en enlevant l'item en question). Plusieurs indicateurs peuvent être proposés. Le plus simple mais pas forcément le meilleur est tout simplement le coefficient de corrélation linéaire entre le score et l'indicatrice de réussite à l'item. La discrimination indique dans quelle mesure la réussite à tel item permet de prédire un score élevé. Dans l'absolu, les items les plus discriminants sont les meilleurs (car ils reflètent bien la dimension évaluée par l'ensemble de l'épreuve).

Il convient d'ajouter à cela quelques notions permettant de juger de la qualité globale d'une épreuve. On a généralement recours à deux types d'indicateurs :

- Indicateurs de fidélité : le calcul d'un score revient donc dans le cas le plus simple à faire la somme des indicatrices de réussite aux items. Comme dans tous les cas où l'on additionne des données, il faut vérifier qu'aucune carotte ne s'est glissée parmi les choux. En d'autres termes, la sommation n'est pertinente que si tous les items mesurent la même chose. Ceci se teste de plusieurs façons qui reviennent toutes à étudier les corrélations entre items. Si les corrélations entre items sont élevés, quand réussir à l'un accroît significativement les chances de réussir à un autre, on considère que l'épreuve est *fidèle* ou qu'elle a une bonne *consistance interne*. On verra plus loin les méthodes utilisées pour évaluer cette *fidélité* quand on cherchera à évaluer l'unidimensionnalité de l'épreuve.

- Indicateurs de validité : pour s'assurer que l'on a une bonne épreuve, une autre méthode consiste à comparer les résultats qu'elle donne soit avec une donnée connue et censée être liée à la compétence (par exemple, le redoublement dans le cas d'une évaluation scolaire) soit avec une autre épreuve faisant autorité. Les indicateurs de validité donnent donc une idée de la valeur prédictive de l'épreuve.

Tous ces indicateurs ont le mérite de la simplicité mais ils ont quelques inconvénients. Tout d'abord, ils ont les défauts de tous les pourcentages, c'est-à-dire que leur « comportement sur les bords » devient particulier. Une population ayant un score moyen élevé, proche de 100 % (ou faible et proche de 0) risque de présenter des écarts entre individus assez faibles par rapport à une population dont le taux de réussite est proche de 50 %. Or ce résultat sera certainement inversé si l'on change d'épreuve, en en donnant une plus difficile, pour faire en sorte que le taux de réussite de la première population s'approche de 50 %, alors que celui de la deuxième tende vers 0. Pour reprendre les termes de la théorie de la mesure que nous avons développée, il est à craindre que les scores en % ne suivent ni une échelle de rapport, ni une échelle d'intervalle, quand on s'écarte trop de 50 %[4].

Les indicateurs posent d'autres problèmes quand on veut les comparer entre eux. Il n'est pas possible de dire que telle population est moins bonne que telle autre, parce qu'elle a un score moyen inférieur. Il faut que chacune des populations ait été soumise une évaluation identique. Or dans bien des cas, c'est une contrainte très forte.

---

[4] Ce résultat ne semblera pas singulier à ceux qui ont suivi dans la Revue Française de Sociologie le débat sur la comparaison des taux d'accès. Si le taux d'accès des enfants d'ouvriers passe de 1 % à 10 % tandis que celui des enfants de cadres passe de 39 % à 51 %, peut-on dire que les inégalités ont diminué ? Doit-on faire la différence des taux, leur rapport, étudier les taux de variation ?
Là aussi, on voit que toutes ces questions amènent à se demander quel est le niveau de mesure du taux d'accès. Est-ce une mesure d'intervalle ou de rapport ? En fait, les chercheurs semblent considérer qu'elle n'est véritablement ni l'une ni l'autre mais simplement ordinale. Son échelle naturelle serait logistique : il faudrait donc appliquer une transformation du type log(p/(1-p)) pour aboutir à une échelle d'intervalle. Cependant, on s'accorde aussi à dire que si les pourcentages restent dans des valeurs moyennes (entre 25 et 75 %) on peut les considérer comme relevant directement d'une mesure d'intervalle. La conclusion pratique que l'on peut tirer de l'examen des articles abordant ce problème est que les indicateurs logistiques sont sans doute les plus justes mais qu'ils doivent surtout servir à vérifier les commentaires portant sur des indicateurs plus simples. La lisibilité ne doit pas être sacrifiée à un trop grand souci de rigueur. De plus, même si l'indicateur logistique conduit à dire qu'un écart de taux d'accès au baccalauréat de 92 % à 99 % est statistiquement énorme, en pratique, c'est une inégalité relativement peu apparente, même si cela peut suggérer que le problème des inégalités n'est pas résolu et risque de réapparaître à un niveau plus sélectif du système éducatif.

3. Les modèles de réponse à l'item

3.1. Présentation

Les modèles de réponse à l'item (MRI ou IRT en anglais) proposent d'expliquer la réussite à un item en fonction de la compétence de l'individu et de l'item lui-même. La probabilité pour un individu j de répondre à la question i sera notée :

$$\Pr(x_{ij} = 1) = F(\theta_j, \beta_i)$$

où x_{ij} est la réponse de j à la question i, θ_j la compétence de l'individu j, β_i le vecteur caractéristique de l'item i et F la fonction de lien[5].

De tels modèles présentent l'avantage de séparer les concepts, puisque compétences individuelles et caractéristiques des items sont définies de façon totalement indépendante[6]. On verra dans la partie 3.2. les hypothèses sous-jacentes à ce modèle.

On remarque une certaine analogie avec les modèles économétriques de panels : les différentes questions posées à un individu peuvent être assimilées à des prises d'informations successives. La compétence θ_j correspondrait à « l'effet fixe » ou effet individuel dans ces modèles.

3.1.1. Le modèle de Rasch

Il existe plusieurs MRI suivant le nombre de paramètres que l'on utilise pour décrire l'item. Historiquement, le premier est le modèle à un paramètre, ou modèle de Rasch (1960) :

$$\Pr(x_{ij} = 1 | \theta_j, b_i) = \frac{1}{1 + e^{(b_i - \theta_j)}}$$

où b_i est la difficulté de l'item i.

---

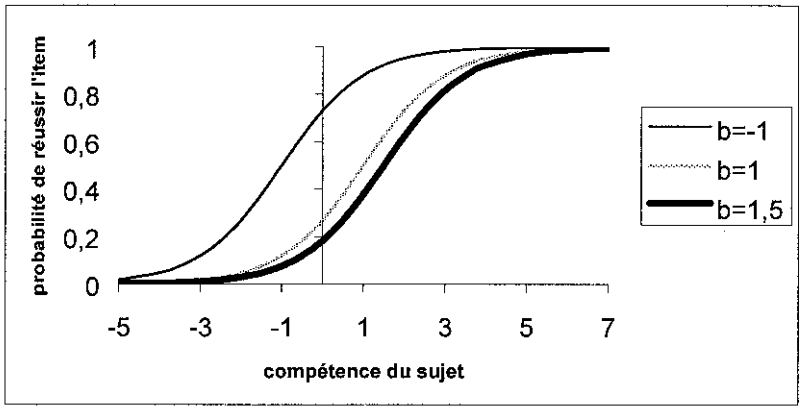
[5] On se limitera ici à la fonction logistique mais d'autres choix sont possibles.

[6] Sur ce point, les psychométriciens tombent parfois dans l'ambiguïté : certains transforment l'hypothèse (i.e. la relation proposée par le modèle et ses implications sur les propriétés requises par les données) en fait imposé par la nature du modèle. Par exemple, certains auteurs affirment que les paramètres de difficulté peuvent être calibrés sur n'importe quel échantillon puisqu'ils ne dépendent pas du groupe d'individus sur lesquels ils sont calculés. Cet avantage existe mais il est théorique : il faut avant tout qu'il y ait adéquation entre le modèle et les données.

Une même échelle latente s'applique donc aux individus et aux questions, ce qui fait tout l'intérêt du modèle. $(b_i - \theta_j)$ représente la différence de niveaux entre les deux : si $(b_i-\theta_j)<0$, la probabilité de réussite sera supérieure à 0.5, inférieure sinon.

Ce modèle fait l'hypothèse que la dépendance entre la probabilité de réussir un item et la compétence des individus peut être représentée comme sur le graphique 1. Les courbes des 3 items pris en exemple sont toutes parallèles. La difficultés de l'item correspond au point d'inflexion de la courbe. On voit que c'est aussi la valeur de la compétence des individus qui ont autant de chance de réussir l'item que de le rater.

Graphique 1 : modèle à un paramètre (ou modèle de Rasch)



3.1.2. Le modèle à deux paramètres

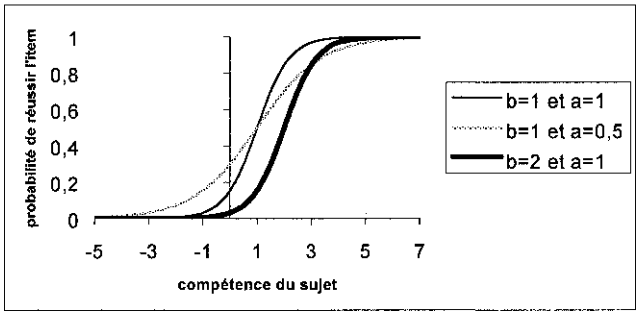
Pour affiner le modèle, on peut ajouter une caractéristique à l'item :

$$\Pr(x_{ij}=1|\theta_j,a_i,b_i)=\frac{1}{1+e^{-D\times a_i(\theta_j-b_i)}}$$

où a_i est le coefficient de discrimination de la question i et D un facteur d'échelonnement permettant le passage à la fonction de lien ogive normale (constante égale à 1,7).

Si a_i est grand, $(\theta - b)$ sera « amplifié » : une personne ayant une compétence plus forte que la difficulté de la question réussira très probablement la question tandis qu'une personne de compétence moindre n'aura que très peu de chances d'y répondre avec succès. A l'extrême inverse, si ce coefficient est nul, la probabilité de réussite sera constante quelle que soit la compétence de l'individu. C'est pourquoi a_i est appelé coefficient de discrimination. Le graphique 2 présente les courbes caractéristiques théoriques de quelques items. On voit que contrairement au cas précédent, elles ne sont plus parallèles et peuvent se croiser.

Graphique 2 : modèle à deux paramètres



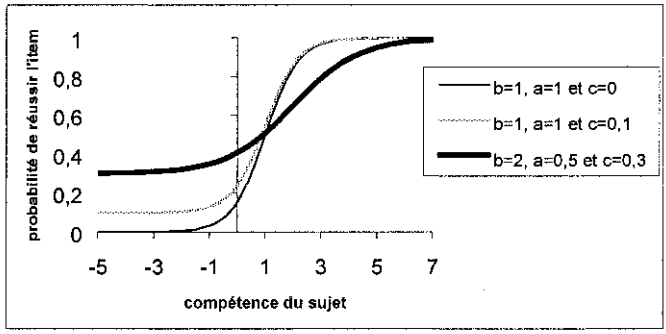
3.1.3. Le modèle à trois paramètres

Certains items peuvent être réussis au hasard, comme les questions à choix multiples. L'ajout d'un facteur de chance c_i peut dans ce cas s'avérer pertinent. Le modèle devient alors

$$\Pr(x_{ij}=1|\theta_j,a_i,b_i,c_i)=c_i+(1-c_i)\frac{1}{1+e^{-D\times a_i(\theta_j-b_i)}}$$

Ainsi, on considère que la probabilité de réussir l'item est supérieure à c_i, quelle que soit la compétence de l'individu. Si par exemple la réponse à la question est oui ou non, on pourra poser c_i=0,5 puisque même si on ne connaît absolument pas la réponse, on a une chance sur deux de répondre correctement. Mais l'ajout de ce paramètre n'est pas toujours si simple car si la question posée est une question « piège », il se peut très bien que l'on observe des taux de réussite inférieurs à 50 %, en contradiction avec le modèle.

Graphique 3 : modèle à trois paramètres



3.2. Hypothèses des modèles

En théorie, les modèles de réponse à l'item présentent des avantages très intéressants pour les constructeurs de tests ou d'évaluations. Cependant, ils reposent sur des hypothèses fortes, qu'il est nécessaire de vérifier avant toute utilisation.

3.2.1. L'hypothèse d'unidimensionnalité

Les performances des individus aux items ne sont mesurées que par un seul paramètre[7]. On postule donc que la réussite aux items est conditionnée par un seul et même facteur. Dans la pratique, cette hypothèse n'est évidemment pas totalement remplie vu la complexité des processus mentaux permettant la réussite aux items. De plus, il n'existe pas de méthode parfaite pour valider cette hypothèse car la dimensionnalité des items est un concept difficilement mesurable.

Deux approches ont cours. La première vise à mesurer la cohérence globale : on aura alors recours aux analyses factorielles ou aux α de Cronbach. La deuxième analyse le fonctionnement différentiel des items.

Les analyses factorielles

En général, on considère qu'une évaluation est unidimensionnelle si l'ensemble des items s'ajuste convenablement au premier facteur d'une analyse factorielle. De manière pratique, on compare l'inertie expliquée par le premier facteur à celle expliquée par le second. On considère en général que l'hypothèse est vérifiée lorsque le rapport des deux est supérieur à 4.

Le coefficient α de Cronbach

Ce coefficient est employé pour évaluer la consistance interne de l'épreuve. Dans le cas où tous les items sont dichotomiques, il est défini de la manière suivante[8] :

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} p_i(1-p_i)}{\sigma^2}\right)$$

où p_i est la proportion de réussite à l'item i

σ^2 est la variance de l'épreuve (c'est-à-dire la variance des proportions de bonnes réponses obtenues par individu).

---

[7] Cela ne signifie pas que l'évaluation entière repose sur une seule échelle. Dans IALS, par exemple, 3 axes de compétence sont définis. On ne s'intéressera donc ici qu'aux items mesurant la même dimension.

[8] De manière plus générale, le α est défini comme la moyenne des corrélations entre toutes les sous-épreuves que l'on peut construire à partir des N items.

A partir d'une valeur de 0,8, on considère que l'hypothèse est acceptable. Mais plusieurs mises en garde peuvent être faites (Cortina, 1993) : par exemple, le degré de corrélation entre les items ne mesure pas forcément l'homogénéité de l'épreuve, c'est une condition nécessaire mais pas suffisante. Ce coefficient dépend de l'écart-type et une forte valeur de α peut provenir d'une forte disparité de niveaux entre les individus de l'échantillon et non pas de l'homogénéité des items. De plus, il augmente sensiblement avec le nombre d'items de l'épreuve.

Le fonctionnement différentiel des items (FDI)

- La statistique de Mantel-Haenszel.

On cherche ici à repérer les items dont la réussite ne seraient pas uniquement conditionnée par le score de l'individu. On teste ainsi l'influence, à compétence constante, des facteurs comme le sexe, l'âge, les diplômes sur les réponses à une question donnée. Pour cela, on peut utiliser la statistique de Mantel - Haenszel. On considère deux groupes (par exemple les hommes et les femmes), qu'on découpe en J classes de compétences équivalentes. Le tableau ci-dessous présente les effectifs de chaque catégorie :

		Réponse à l'item		
		correcte	incorrecte	Total
Niveau j	Groupe 1	A_j	B_j	n_{1j}
	Groupe 2	C_j	D_j	n_{2j}
	Total	m_{1j}	m_{0j}	T_j

La statistique de Mantel-Haenszel est donnée par :

$$MH = \frac{\left(\left| \sum_{j=1}^{J} (A_j - E(A_j)) \right| - \frac{1}{2} \right)^2}{\sum_{j=1}^{J} Var(A_j)}$$

où $E(A_j) = \frac{n_{1j} m_{1j}}{T_j}$ et $Var(A_j) = \frac{n_{1j} n_{2j} m_{1j} m_{2j}}{T_j^2 (T_j - 1)}$

Sous l'hypothèse H_0 : A/B = C/D (la réussite à l'item est indépendante du groupe considéré), on a approximativement : MH ~ χ_1^2

- Le biais statistique.

Dans le même ordre d'idée, une autre façon de procéder est d'utiliser l'indice suivant, proposé par Lord :

$$d_j = \frac{\Delta b_j}{\sqrt{S_{j,1}^2 + S_{j,2}^2}}$$

où $\Delta b_j = b_{j,1} - b_{j,2}$ est la différence des estimations de la difficulté de l'item j pour les pays 1 et 2.

$S_{j,k}$ est l'erreur type de l'estimation de $b_{j,k}$

d_j est appelé usuellement SIB (pour Statistical Index of Bias).

A priori, les deux populations ont des compétences moyennes différents. Afin de supprimer cet effet, on impose que la moyenne des b_j soit nulle pour chacun des deux groupes.

On teste l'hypothèse H_0 : d=0 conte H_1 : $d \neq 0$.

Sous H_0, $d \sim N(0,1)$. Donc, pour un seuil de 5%, si |d |>1.96, la question présente un fonctionnement différentiel significatif.

Notons que ces deux statistiques ont le défaut de dépendre de la taille de la population. Si le nombre d'individus est trop faible, H_0 est presque systématiquement acceptée. En d'autres termes, on risque d'affirmer que les items correspondent bien au modèle simplement parce que le nombre d'individus ne permet pas de mettre en évidence un écart statistique. Ce risque est particulièrement important quand on travaille sur des échantillons de taille modeste et en tenant compte de l'usage des cahiers tournants (chaque item n'est passé que par une partie de la population).

3.2.2. L'hypothèse d'indépendance locale

Cette hypothèse signifie qu'à un niveau de compétence donné, il n'existe pas de corrélation entre les réponses aux items. Elle rejoint le postulat d'unidimensionnalité dans la mesure où seule la compétence des individus explique la réussite aux items. Par exemple, une question intitulée comme ceci : « A partir des résultats de la question précédente... » violerait le principe d'indépendance locale. En pratique, pour vérifier cette hypothèse, on utilise le coefficient ϕ^2.

Lorsque les items sont dichotomiques, il s'agit du coefficient de corrélation linéaire R^2 entre les variables indicatrices de deux items :

$$\phi^2 = \frac{(p_{11}p_{00} - p_{10}p_{01})^2}{p_1(1-p_1)p_2(1-p_2)}$$

où p_{ij} est la proportion de réponses (i à l'item 1, j à l'item 2)

p_1 est la proportion de bonnes réponses à l'item 1

p_2 est la proportion de bonnes réponses à l'item 2

En pratique, on calcule la moyenne de tous les ϕ obtenus sur l'ensemble d'un questionnaire, pour un groupe d'individus de niveaux homogènes. Si les résultats sont proches de 0, l'hypothèse d'indépendance locale est validée.

3.3. Estimation

3.3.1. Estimation des paramètres d'items

Les modèles MRI ont ceci de très particulier qu'ils s'appuient sur des variables explicatives inobservables a priori. En effet, même si les items sont élaborés par des spécialistes (professeurs, psychométriciens) suivant certains critères de difficulté (comme la lisibilité du document, le nombre de distracteurs, la quantité d'informations à relever...), les paramètres des items ne sont pas déterminés au départ mais estimés conjointement aux scores des individus.

On se place tout d'abord dans le cas où tout le monde a répondu aux mêmes items. Soit $X=[x_{ij}]_{(1\leq i\leq N,\ 1\leq j\leq K)}$ la matrice de réponses des K individus aux N questions. Sous l'hypothèse d'indépendance locale, la fonction de vraisemblance s'écrit[9] :

$$L(X|\theta,\beta) = \prod_{j=1}^{K}\prod_{i=1}^{N} P_{ij}^{\ x_{ij}}(1-P_{ij})^{1-x_{ij}}$$

où θ est le vecteur de scores de l'ensemble des individus

β est le vecteur des paramètres des items

$P_{ij}=P(x_{ij}=1 \mid \theta, \alpha)$

---

[9] L'autre hypothèse est bien sûr qu'à θ constant, P_{ij} est constant quel que soit le pays d'origine.

Soit f la densité de θ considérée comme une variable aléatoire continue. On a donc :

$$L(X,\theta|\beta) = \prod_{j=1}^{K}\prod_{i=1}^{N} P_{ij}^{\,x_{ij}}(1-P_{ij})^{1-x_{ij}}\ f(\theta)$$

Et par conséquent :

$$L(X|\beta) = \int_{\theta}\prod_{j=1}^{K}\prod_{i=1}^{N} P_{ij}^{\,x_{ij}}(1-P_{ij})^{1-x_{ij}}\ f(\theta)\,d\theta$$

A ce stade, la solution la plus simple est de considérer f comme connue. On peut postuler, par exemple, que θ suit une loi normale centrée réduite. On estime alors directement β par maximum de vraisemblance.

3.3.2. Les cahiers tournants

Jusqu'à présent , on a considéré que tous les individus répondaient à un seul questionnaire. En réalité, dans les évaluations internationales, on utilise le principe des cahiers tournants. Ainsi, dans IALS par exemple, les 114 items de départs sont répartis en sept blocs et ces blocs sont regroupés par trois dans sept questionnaires : le premier questionnaire, ainsi, est formé des blocs 1, 2 et 4 et le deuxième des blocs 2, 3 et 5. On a donc la possibilité de faire passer un plus grand nombre d'items sans alourdir le temps de passation, et d'obtenir plus de résultats item par item.

Par ailleurs, l'influence d'une question inadéquate est réduite, puisque, dans notre exemple, 3/7 de la population seulement y répond. En contrepartie, il faut s'assurer que les items des différents blocs sont sur une échelle commune, ne serait-ce que pour que les individus soient également sur une échelle commune. Par exemple, une simple proportion de bonnes réponses ne conviendrait pas pour de tels questionnaires car si l'un des blocs était plus difficile que les autres, les scores obtenus seraient inférieurs à la moyenne sans pour autant refléter une faiblesse générale de cette sous-population. Le modèle MRI s'adapte quant à lui relativement bien à cette contrainte dans la mesure où il tient déjà compte des caractéristiques des questions. Un seul problème subsiste en fait : l'indétermination du modèle. En effet, en posant, pour le modèle à deux paramètres :

$$a_i^* = a_i / k_2,\ b_i^* = k_1 + k_2 b_i \ \text{ et } \ \theta_j^* = k_1 + k_2\theta_j$$

on obtient : $P(x_{ij}=1 \mid \theta_j, a_i, b_i) = P(x_{ij}=1 \mid \theta_j^*, a_i^*, b_i^*)$

Ainsi, deux estimations successives de a et b peuvent donner des résultats différents. Il existe plusieurs techniques pour relier les paramètres entre eux, suivant le type de méthode de cahiers tournants utilisés. Si au moins un groupe d'items est commun (cas de IALS), on peut par exemple effectuer une estimation simultanée de tous les items et standardiser ensuite soit les b_i, soit les θ_j obtenus.

3.3.3. Estimation des scores : la méthode des valeurs plausibles

Une fois déterminé β, on peut maximiser en θ la fonction de vraisemblance $L(X|\theta, \beta)$. Cette solution est la plus simple mais possède l'inconvénient de donner des résultats infinis si la personne a répondu juste ou faux à toutes les questions. On peut contourner cet obstacle en utilisant une estimation bayésienne. Si la moyenne des estimateurs est sans biais par rapport à la moyenne de f, l'écart-type des estimateurs est en général plus petit que celui de f, ce qui peut poser des problèmes de comparaison de populations.

Une autre solution souvent retenue est la méthode des valeurs plausibles. L'objectif est de fournir des estimations optimales à l'échelle des populations et non des individus. Cela permet également d'obtenir des résultats continus et non discrets (avec la méthode de maximum de vraisemblance, par exemple, on obtient un nombre fini de θ différents). Pour ce faire, on tire au hasard une valeur de θ dans sa distribution a posteriori (c'est-à-dire la distribution de θ étant donné les résultats aux items). On s'appuie sur la relation bayésienne suivante :

$$P(\theta \mid x) = C\, P(x \mid \theta)\, P(\theta) \qquad (1)$$

où C est une constante.

Les paramètres des items étant déterminés, $P(x \mid \theta)$ est connu. Reste a fixer la distribution a priori. On considère par exemple : $\theta \sim N(0, 1)$[10].
De par la complexité de la vraisemblance, on peut supposer également que la distribution a posteriori est normale, et on calcule son espérance et sa variance à partir de (1). On tire ensuite un θ dans cette distribution. Très clairement, l'estimateur obtenu n'est pas optimal pour les individus puisqu'on a une probabilité non nulle de tirer un θ très faible pour un individu ayant réussi entièrement le questionnaire, et inversement.

---

[10] On pourra toujours contester cette hypothèse de normalité. La symétrie des compétences, par exemple, semble hasardeuse : peut-on dire qu'il y a autant de Proust que d'illettrés ? Souvent, on préfère se limiter à des populations a priori plus homogènes (et, on l'espère, davantage « normales »). Cela revient à introduire des caractéristiques socio-démographiques dans l'équation (1). Ce raffinement complique singulièrement les procédures d'estimation (NCES, [98]).

3.4. Relation avec la proportion de bonnes réponses

La proportion de bonnes réponses constitue sans doute l'estimateur le plus naturel de réussite d'un individu à une évaluation. On va chercher à savoir en quoi la compétence estimée par le modèle de réponse à l'item est « supérieure » à cette statistique élémentaire. Pour comprendre le lien entre ces deux approches a priori divergentes, analysons le modèle suivant :

$$P(\theta) = Pr(x_i = 1 | \theta) = \frac{1}{1 + e^{-\theta}}$$

Il s'agit en quelque sorte d'un modèle MRI à 0 paramètre puisqu'on ne prend pas en compte les caractéristiques de l'item (on considère que $\forall i\ b_i = 0$). On détermine θ à l'aide de la fonction de vraisemblance suivante :

$$L(X | \theta) = \prod_{i=1}^{N} P(\theta)^{(x_{ij}=1)} (1 - P(\theta))^{1-(x_{ij}=1)}$$

θ vérifie : $[\ \ln(X \mid \theta)\ ]' = 0$
C'est-à-dire :

$$\sum_{i=1}^{N} (x_{ij} = 1) \frac{P'(\theta)}{P(\theta)} - [1 - (x_{ij} = 1)] \frac{P'(\theta)}{1 - P(\theta)} = 0$$

Or $P'(\theta) = P(\theta)(1 - P(\theta))$ donc

$$P(\theta) = \frac{1}{N} \sum_{i=1}^{N} (x_i = 1)$$

En considérant p la proportion de réponses correctes, on obtient ainsi :

$$\theta = \ln\left(\frac{p}{1-p}\right)$$

Dans la suite on notera $\ln\left(\frac{p}{1-p}\right) = p_{log}$

Ainsi, à une transformation logistique près, on peut affirmer que la proportion de bonnes réponses est la compétence obtenue pour un modèle MRI à 0 paramètre.
La question initiale se pose donc dans les termes suivants : l'apport des caractéristiques d'item modifie-t-il sensiblement les compétences des individus ?
Prenons tout d'abord le cas du modèle à 1 paramètre.

Supposons qu'on ait calculé θ par maximum de vraisemblance. θ maximise donc ln(L) où L est la vraisemblance du modèle. Par conséquent

$$[\ln(L)]'(\theta) = 0$$

Soit

$$\sum_{i=1}^{N} (x_i = 1) \frac{P'_i(\theta)}{P_i(\theta)} - [1 - (x_i = 1)] \frac{P'_i(\theta)}{1 - P_i(\theta)} = 0$$

Or pour un modèle à un paramètre : $P'_i(\theta) = P_i(\theta)(1 - P_i(\theta))$. On obtient donc

$$\frac{1}{N}\sum_{i=1}^{N}(x_i = 1) = \frac{1}{N}\sum_{i=1}^{N}P_i(\theta)$$

En d'autres termes, la moyenne des probabilité de réussite du modèle est égale à la proportion p d'items réussis.

Si l'on remplace $P_i(\theta)$ par son expression, on obtient

$$p = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{1+e^{b_i-\theta}}$$

Il existe donc une relation fonctionnelle entre p et θ, et donc également entre p_{log} et θ. De plus, p est une fonction continue strictement croissante de θ. Cela signifie qu'à un p donné on associe un unique θ. Donc, pour une épreuve de N items sans cahiers tournants, comme il n'y a que N-1 valeurs possibles de p (on ne tient pas compte des extrêmes qui sont des cas particuliers), il n'y aura également que N-1 valeurs possibles de θ. L'utilisation des valeurs plausibles dans le cas d'un modèle à 1 paramètre est donc particulièrement pertinente si l'on souhaite obtenir un score continu.

Par ailleurs, comme $e^{b_i-\theta} > 0$, on a, par convexité de $x \to \frac{1}{1+x}$:

$$p \geq \frac{1}{1+e^{-\theta}E_1}$$

où $E_1 = \frac{1}{N}\sum_{i=1}^{N} e^{b_i}$

et, en considérant b_{min} le minimum des b_i :

$$\frac{1}{1+e^{b_{min}-\theta}} \geq \frac{1}{1+e^{b_i-\theta}}$$

Donc :

$$\frac{1}{1+e^{b_{min}-\theta}} \geq p \geq \frac{1}{1+e^{-\theta}E_1}$$

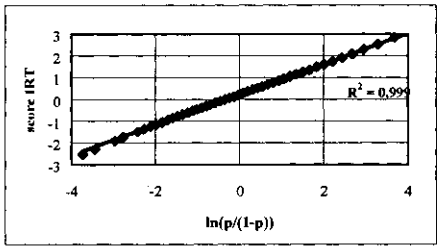
Comme $x \to \ln(x/(1-x))$ est une fonction croissante de x, on peut encadrer p_{log} :

$$\theta - b_{min} \geq p_{log} \geq \theta - \ln(E_1)$$

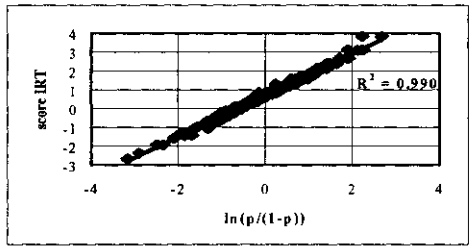
Ce résultat théorique se retrouve en pratique de manière encore plus spectaculaire. Sur les données françaises de IALS, on obtient ainsi R^2=0.96 entre p_{log} et θ. De plus, on constate que les divergences sont essentiellement dues à l'imprécision des estimateurs de θ dans le cas où les individus ont répondu à très peu d'items. Lorsqu'on retire les personnes ayant répondu à moins de 20 items (l'épreuve en compte en moyenne 50), le R^2 atteint 0.99. Et si l'on prend en compte seulement les individus ayant passé entièrement un même cahier, on obtient des R^2 de l'ordre de 0.999.

Le modèle à 2 paramètres possède l'avantage d'accorder plus ou moins de poids aux items dans l'estimation du θ. En théorie, les items qui fonctionnent mal interviendront peu tandis que ceux qui discriminent parfaitement la population auront une importance cruciale dans l'estimation du θ. Cela sous-entend également que contrairement au modèle à 1 paramètre, plusieurs θ sont possibles pour un p donné. En pratique, les corrélations entre θ et les proportions de bonnes réponses sont encore très élevées, puisqu'on obtient R^2=0.95 sur l'ensemble des individus. Si l'on calcule cet indicateur cahier par cahier, les R^2 avoisinent 0.99.

Modèle à 1 paramètre



Modèle à 2 paramètres



D'une manière analogue, on constate que l'estimateur intuitif de la difficulté des items, c'est-à-dire la proportion d'individus réussissant l'item, est très proche des b_j. Toujours pour les données françaises de IALS, on obtient R^2=0.96 avec un modèle à 1 paramètre.

Etant donné, d'une part, la complexité des estimations, et d'autre part, la similitude des résultats avec les estimateurs intuitifs, on peut donc s'interroger sur la pertinence de la mise en œuvre d'un modèle MRI. Un argument est souvent invoqué : l'avantage qu'il procure pour résoudre le problème des cahiers tournants. Cela n'est pas faux. Mais, dans une épreuve comme IALS où les items ne sont pas étudiés un par un, rien ne démontre que les cahiers tournants sont indispensables.

3.5. Les limites du modèle

3.5.1. La question des extrêmes

Comment peut-on comprendre le modèle de réponse à l'item ? Partons d'un modèle avec seuil :

$$\Pr(x_{ij}=1)=\begin{cases}1 \text{ si } (\theta_j - b_i) > s \\ 0 \text{ sinon}\end{cases}$$

On peut maintenant supposer qu'à chaque question, la compétence « vraie » de l'individu est perturbée par un aléa imputable à la question, aux conditions extérieures... et ne peut s'observer directement. En posant donc $\theta_{j\ obs}=\theta_j + \varepsilon_{ij}$, on obtient, si G est la fonction de distribution de $-\varepsilon_{ij}$:

$$\Pr(x_{ij}=1)=\Pr(-\varepsilon ij<(\theta_j - b_i) - s)=G((\theta_j - b_i) - s)$$

Si l'on suppose que les résidus sont gaussiens d'écart-type constant, on obtient un modèle MRI à un paramètre. Et si on suppose que l'écart-type du résidu varie en fonction des questions, on obtient un modèle MRI à deux paramètres. L'écart-type du résidu correspond en fait au coefficient de discrimination a_i de la question : si celui-ci est très faible, l'écart-type du résidu est tellement important qu'il empêche de distinguer la compétence individuelle. Si a_i est très fort, l'écart-type est quasi-nul et on retrouve pratiquement notre modèle de seuil initial.

Un point paraît discutable : la compétence « vraie » peut être modifiée à la fois négativement par des facteurs aléatoires de perturbation, ce qui semble logique - distraction par des éléments extérieurs, manque de motivation, lassitude face au questionnaire etc. - mais aussi positivement, ce qui est nettement moins évident. On peut à la rigueur imaginer que l'attrait d'un individu pour le domaine abordé par la question favorise ses chances de réussite. Que penser, en revanche, d'un élève de $6^{ème}$ qui aurait une probabilité non nulle de répondre à une question posée à Polytechnique ? Cette situation n'est pas si caricaturale que cela puisque le modèle MRI stipule justement que toutes les questions se valent et que seule compte la probabilité de réussite : réussir une question simple avec une probabilité de 0.8 équivaudrait à réussir une question difficile avec une probabilité de 0.07, par exemple. Partant de là, les concepteur de IALS n'ont construit qu'un très faible nombre d'items élémentaires (4 ou 5 suivant le questionnaire), et ont donc évalué en grande partie la compétence des individus à partir d'items plus ardus. Il semble pourtant hasardeux de considérer comme illettrée une personne qui n'aurait réussi « que » 7 % des questions les plus difficiles, puisqu'a priori un illettré serait totalement incapable de répondre à ces questions. Les limites du modèle sont très claires ici.

De plus, la façon de présenter les résultats par les responsables de l'OCDE est parfois malhonnête. En effet, ils définissent une population de niveau 1 d'un point de vue statistique : il s'agit des individus qui auraient, suivant le modèle, 80 % de chance de réussir les items de ce niveau et moins pour les items de difficulté supérieure. Cependant, ils la présentent également comme l'ensemble des individus « ayant un niveau de compétences très faible ; par exemple, la personne peut être incapable de déterminer correctement la dose d'un médicament à administrer à un enfant d'après le mode d'emploi indiqué sur l'emballage ». On note déjà un glissement sémantique puisque d'un côté ces individus réussiraient régulièrement les items de base (et seulement ceux-là) tandis que la deuxième formulation semble suggérer qu'ils ne sont mêmes pas capables de résoudre les problèmes les plus simples. Cette dernière formulation a sans doute contribué à ce que l'on considère cette population comme illettrée. Or, selon le calcul de l'OCDE, 40 % des Français appartenaient à cette catégorie, chiffre largement supérieur aux estimations les plus pessimistes du taux d'illettrisme. D'ailleurs, on constate que 94 % des individus ont répondu correctement à la question donné en exemple dans la définition de l'OCDE (lecture d'une notice d'un médicament). Est-ce vraiment compatible avec les 40 % obtenus par ailleurs ?

3.5.2. La même justice pour tous ?

Rappelons qu'une des hypothèses du modèle est que les items fonctionnent identiquement dans tous les pays. C'est une des conséquences du postulat d'unidimensionnalité. La statistique de Mantel-Haenszel ou le SIB permet de tester cette hypothèse (voir paragraphe 3.2.1.). L'existence de nombreux fonctionnements différentiels viole le principe d'unidimensionnalié, et remet totalement en cause la possibilité de classements internationaux, qui constitue pourtant l'enjeu politique majeur de ce type d'enquête. Nous allons maintenant étudier ce problème dans un cas particulier.

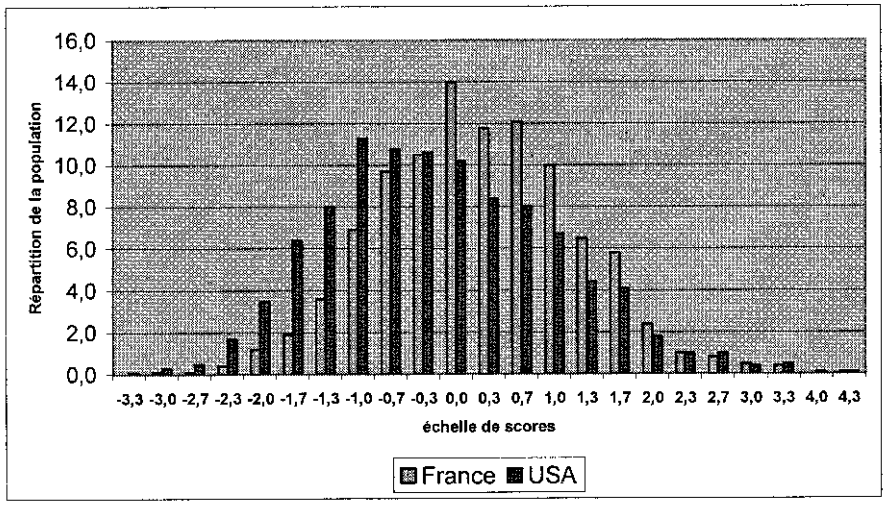
4. Les biais d'items et l'étude du niveau moyen

4.1. France - Etats-Unis : le match était-il truqué ?

Nous allons procéder maintenant à une mini-enquête internationale, pour montrer dans les grandes lignes, les traitements généralement effectués. L'exemple que nous allons présenter ici est issu de l'enquête TIMSS menée en 1995 au niveau de la cinquième et de la quatrième. On se restreindra à l'étude de l'épreuve de mathématiques dans deux pays : la France et les Etats-Unis. Environ 6000 élèves ont répondu à l'épreuve en France et plus de 10 000 aux Etats-Unis. On a déterminé les scores des élèves et les paramètres des items grâce au logiciel CONQUEST, à partir d'un modèle de Rasch, qui a également été utilisé par les responsables de TIMSS pour établir les résultats publiés.

Le logiciel propose différentes façons d'attribuer un score à chaque élève. Nous avons retenu la méthode la plus simple, dans le cadre des MRI, celle utilisant l'estimation du maximum de vraisemblance, qui donne un score très proche du logit du taux de réussite, voire de ce taux lui-même. Le graphique 4 présente la distribution des scores aux Etats-Unis et en France. Les élèves français apparaissent sensiblement meilleurs que leurs camarades américains.

Graphique 4 : Répartition des scores aux Etats-Unis et en France



Nous allons surtout nous intéresser maintenant aux paramètres des items. Le logiciel donne pour chaque item sa difficulté et son « fit », accompagné de l'erreur de mesure

(voir tableau 1). Le « fit » mesure l'écart entre la discrimination de l'item et la discrimination moyenne : si trop de « fits » sont différents de 1, il est souhaitable d'utiliser un modèle à 2 paramètres. L'indétermination linéaire affectant les difficultés est résolue en fixant la moyenne à 0.

Tableau 1 : extrait des paramètres d'items calculés par CONQUEST

```
ConQuest: Generalised Item Response Modelling Software
Tue Nov 07 15:39:57
TABLES OF RESPONSE MODEL PARAMETER ESTIMATES
TERM 1: item
```

VARIABLES			UNWGHTED FIT		WGHTED FIT	
item	ESTIMATE	ERROR	MNSQ	T	MNSQ	T
1 RMA01	0.052	0.016	0.83	-9.2	0.86	-16.6
2 RMA02	-1.057	0.017	1.03	1.5	1.01	0.9
3 RMA03	0.060	0.016	0.99	-0.5	0.99	-0.7
4 RMA04	0.460	0.016	1.03	1.7	1.00	-0.5
5 RMA05	-0.543	0.016	1.11	5.3	1.09	9.3
6 RMA06	-1.270	0.017	1.12	5.8	1.03	3.2
7 RMB07	-0.710	0.021	1.02	1.0	0.99	-0.9
8 RMB08	-0.647	0.021	1.15	5.6	1.10	7.2
9 RMB09	-0.264	0.021	1.09	3.4	1.05	3.7
(...)						
151 RMU01B	1.566	0.027	0.90	-3.5	0.96	-2.0
152 RMU02A	3.267	0.033	1.10	3.2	1.06	2.5
153 RMU02B	2.651	0.031	0.98	-0.7	1.01	0.6
154 RMV02	1.915	0.028	1.43	12.4	1.28	11.6
155 RMV03	0.282	0.026	0.95	-1.7	0.95	-2.6
156 RMV04	1.259*					

```
An asterisk next to a parameter estimate indicates that it is
constrained
Separation Reliability = 1.000
Chi-square test of parameter equality = 271886.991,
df = 155, Sig Level = 0.000
```

Les erreurs de mesure données ici pourraient permettre de tester des « fonctionnements différentiels ». Il suffirait de faire tourner le modèle dans chaque pays et de comparer pour chaque item les deux paramètres de difficulté obtenus, en tenant compte de l'erreur de mesure. Des tests montreraient si les paramètres de difficultés sont significativement différents. En fait, le logiciel CONQUEST intègre cette possibilité et le tableau 2 donne pour chaque item, ce qu'il faut ajouter au paramètre de difficulté calculé sur les deux pays, pour obtenir celui spécifique à la France. Si ce coefficient est significativement différent de 0 (on utilise l'erreur de mesure pour le tester) on peut dire que l'item est plus difficile en France qu'aux Etats-Unis (si son signe est positif) ou moins difficile (si le signe est négatif).

Tableau 2 : extrait des paramètres concernant les fonctionnements différentiels

VARIABLES				UNWGHTED FIT		WGHTED FIT	
item	pays	ESTIMATE	ERROR	MNSQ	T	MNSQ	T
1 RMA01	1 fra	-0.186	0.016	0.83	-9.2	0.86	-16.6
2 RMA02	1 fra	-0.092	0.017	1.03	1.5	1.01	0.9
3 RMA03	1 fra	0.257	0.016	0.99	-0.5	0.99	-0.7
4 RMA04	1 fra	0.062	0.016	1.03	1.7	1.00	-0.5
5 RMA05	1 fra	-0.413	0.016	1.11	5.3	1.09	9.3
6 RMA06	1 fra	-0.039	0.017	1.12	5.8	1.03	3.2
7 RMB07	1 fra	0.124	0.021	1.02	1.0	0.99	-0.9
8 RMB08	1 fra	0.265	0.021	1.15	5.6	1.10	7.2
(...)							
150 RMU01A	1 fra	0.784	0.027	0.89	-3.7	0.97	-2.1
151 RMU01B	1 fra	0.985	0.027	0.90	-3.5	0.96	-2.0
152 RMU02A	1 fra	-0.173	0.033	1.10	3.2	1.06	2.5
153 RMU02B	1 fra	-0.721	0.031	0.98	-0.7	1.01	0.6
154 RMV02	1 fra	0.473	0.028	1.43	12.4	1.28	11.6
155 RMV03	1 fra	0.181	0.026	0.95	-1.7	0.95	-2.6
156 RMV04	1 fra	-0.329*					

An asterisk next to a parameter estimate indicates that it is constrained

Separation Reliability = 0.997 Chi-square test of parameter equality = 42360.449, df = 155, Sig Level = 0.000

On le voit, la plupart des items présentent dans cette analyse un fonctionnement différentiel. Pour éclairer cela, dans le cas de trois items (A05, B08 et U01B), il est sans doute intéressant d'utiliser la technique rudimentaire du Khi² en classant la population totale suivant les quartiles de score et en regardant pour chaque quartile, le taux de réussite par pays à l'item étudié (on neutralise ainsi le fait qu'une des populations a des moins bons résultats). Normalement, en l'absence de biais, les taux de réussite devraient être identiques (les élèves « médiocres » américains devraient avoir autant de chance de réussir l'item que les élèves « médiocres » français). On constate au contraire qu'ils sont assez différents (voir tableau 3). Pour le premier item A05, on note que l'écart entre la France et les Etats-Unis est considérable : on tourne autour de 20 % par quartile et l'écart est encore plus fort pour l'ensemble de la population, du fait d'un effet de structure (les jeunes français sont en moyenne meilleurs donc...). L'item U01B donne des résultats opposés : là, que ce soit par quartile ou sur l'ensemble de la population, c'est en faveur des jeunes américains que l'on observe un écart !

Manifestement la réussite à chacun de ces deux items doit faire appel à des éléments extérieurs à la pure dimension mathématique, sinon les conclusions sur le niveau des français et des américains risquent d'être contradictoires suivant qu'on prend l'un ou l'autre. L'item B08 est intéressant en ce qu'il montre un biais alors que l'analyse trop précipitée des taux de réussite l'aurait laissé passer. En effet, les taux de réussite sur l'ensemble de la population paraissent équivalents en France et aux Etats-Unis

(67,5 %) mais par quartile, on note un léger mais significatif avantage pour les jeunes américains et seul l'effet de structure explique l'équivalence sur la totalité des élèves. En d'autres termes, trouver des réussite équivalentes alors que les populations ont des niveaux a priori différents est le signe d'un biais. En fait, cela conduit à utiliser une méthode simplifiée pour tester les biais : on compare l'écart entre France et Etats-Unis sur un item donné avec celui que l'on trouve sur l'ensemble des items.

Tableau 3 : Taux de réussite par pays et par quartiles de score à trois items.

	A05		B08		U01B	
	France	USA	France	USA	France	USA
Premier quartile	52,2%	33,7%	49,5%	50,4%	2,1%	9,4%
Deuxième quartile	71,2%	46,5%	63,9%	65,9%	7,5%	26,5%
Troisième quartile	83,6%	57,4%	71,7%	73,7%	14,3%	52,3%
Quatrième quartile	94,0%	73,0%	79,6%	82,9%	34,5%	75,0%
Ensemble	75,7%	52,6%	67,5%	67,6%	13,2%	41,9%

En reprenant les données du tableau 2, on note de plus que les items se voyaient attribuer respectivement un coefficient de fonctionnement différentiel de –0,413 ; 0,265 et 0,985, tous significativement différents de 0 ce qui confirme l'existence de biais.

Cependant, cette exposé paraîtra un peu sec si l'on continue à traiter les items comme de simples numéros. C'est pourquoi nous allons présenter quelques exemples selon le type de biais rencontré. Par souci de clarté, on se contentera de donner les taux de réussite car les biais présentés sont généralement suffisamment forts pour apparaître ainsi.

4.2. Explications des biais statistiques

Il est ainsi possible de mettre en évidence un assez grand nombre d'items biaisés, qui semblent inégalement difficiles selon le pays. Les raisons qui peuvent expliquer un tel phénomène sont nombreuses et nous allons en présenter quelques-unes.

4.2.1. Les problèmes de traduction

Les évaluations internationales sont proposées dans la langue de chaque pays participant, ce qui est une source de biais importante. Sans parler des erreurs qui peuvent se glisser, il est de nombreux exemples où une connaissance sommaire de l'anglais et du français permet de voir que même si les traductions sont exactes, les exercices ne sont pas de difficultés équivalentes.

Ces problèmes de traduction semblent avoir été particulièrement nets dans le cas de l'enquête IALS. Ainsi, des différences importantes entre les textes français et américains ont subsisté, que l'on peut regrouper en trois catégories[11] :

- les erreurs de traduction.

- l'absence de répétition des termes en français

- imprécisions des termes français par rapport à l'anglais.

Les erreurs de traduction regroupe ici tous les faux-sens commis lors de la traduction. Ces faux-sens rendent la question plus difficile, voire erronée. Par exemple, l'intitulé d'une question était le suivant : « Sur la carte météo, entourez l'endroit où de fortes précipitations sont prévues dans la semaine. » Or les cartes météo donnaient seulement les prévisions pour les journées de vendredi à dimanche. Et l'on s'est aperçu que bon nombre de personnes essayait de déduire, à partir des cartes du week-end, le temps qu'il ferait la semaine suivante. En américain, la question ne prêtait pas à confusion puisqu'elle portait sur la météo du week-end. Conséquence immédiate : le taux de réussite en France est 26 points en dessous celui des Etats-Unis, alors que l'écart n'est que de 12 points sur l'ensemble des items. On trouvera dans l'article cité en note bien d'autres exemples de ce type.

Dans le cas de TIMSS, on repère quelques cas où sans qu'on puisse parler d'erreurs, les exercices ne sont pas équivalents en français et en anglais.

Ainsi les items J15 et P09 portaient sur des triangles homothétiques. Sans entrer dans le détail de ces deux questions, il est intéressant de noter que l'expression anglaise « the triangles are similare » est sans doute moins parlante que sa traduction, tout à fait exacte d'ailleurs, « le premier triangle est un agrandissement de l'autre » ce qui peut expliquer le très net écart en faveur de la France observé sur ces deux questions (79 % contre 66 % de réussite pour J15, 58% contre 28 % pour P09).

Parfois, un petit changement qui peut paraître anodin bouleverse la difficulté de la question. Ainsi, 68 % des élèves américains peuvent trouver l'équivalent de y^3 parmi cinq propositions, quand on donne comme bonne réponse y×y×y. Dans le questionnaire français, on a préféré remplacer la bonne réponse par une autre expression équivalente (y.y.y) avec des conséquences importantes : seuls 32 % des français reconnaissent l'égalité soit deux fois moins que les américains. A titre de comparaison, une question portant sur l'égalité m+m+m+m=4m est réussie par 60 % des élèves français contre 43 % des américains ! On peut donc penser que c'est l'usage d'une formule plus « scientifique » et moins familière qui a perturbé les jeunes français. Le problème aurait pu dans ce cas être évité mais dans bien d'autres, on est complètement désarmé quand les notations ou les formulations habituellement

[11] cf. F. Guérin Pacé et A. Blum, in « L'illusion comparative », 1999.

utilisées dans les deux langues sont très différentes, sans être a priori de même difficulté : soit on prend dans chaque langue la bonne expression et les élèves où elle est plus complexe sont défavorisés ; soit on harmonise les deux expressions mais cela conduit à proposer dans un pays une expression qui n'est pas celle habituellement utilisée.

De façon générale, il est indispensable d'accorder une place importante à la traduction des questionnaires, sans hésiter à avoir recours à la technique de rétrotraduction (Cette méthode consiste à traduire en retour vers la langue de départ, et à confronter le texte obtenu avec l'original).

4.2.2. Les biais culturels

On parle de biais culturel quand certaines catégories d'individus sont favorisées (ou parfois défavorisées, on va le voir) par leur connaissance a priori du support sur lequel portent les questions. Pour prendre un exemple caricatural, une question portant sur la Révolution française risque de favoriser les élèves français, tandis que ce seront les élèves américains tireront avantage d'une question liée à la guerre de Sécession. Il faut noter que la familiarité avec le support peut parfois causer des déboires et jouer dans le mauvais sens. Ainsi, dans IALS, une question consistait à identifier les comédies à partir des critiques de 4 films. Or, on observe en France que de nombreux enquêtés ont donné comme réponse un film dont la description ne l'apparente pourtant pas à une comédie. La seule explication est la présence dans ce film de l'acteur Michel Blanc, bien connu des Français pour ses rôles dans de nombreuses comédies, mais peu célèbre à l'étranger. Ici, la connaissance a priori des individus a joué en leur défaveur.

Les exercices de TIMSS, de forme assez scolaire, ne semblent pas tomber souvent sous le coup des problèmes culturels. On peut cependant évoquer un facteur qui a pu jouer (mais son influence semble faible) : l'usage systématique du système métrique. Ainsi, une question demandait quelle était l'unité la plus adaptée pour mesurer la masse d'un œuf : le centimètre, le millilitre, le gramme ou le kilogramme. Cet exercice est réussi à 88 % en France contre 68 % aux Etats-Unis. Il n'est pas impossible que la plus grande familiarité des élèves français avec ces unités ait joué en leur faveur.

4.2.3. Les différences de programmes

Dans le cas des évaluations scolaires, l'une des sources les plus importantes de biais (mais ici la connotation péjorative de ce terme va paraître particulièrement inadéquate) est l'influence des programmes des différents pays. Pour rester dans un cadre simple, si dans tel pays, l'accent est particulièrement mis sur un certain domaine de la discipline évaluée, on risque d'observer que tous les items relevant de ce domaine seront biaisés en faveur du pays, tandis que beaucoup des autres items seront biaisés dans l'autre sens.

Nous n'avons malheureusement pas la place ici de présenter, dans le cadre de notre comparaison France/Etats-Unis, les items où un tel type d'explications paraît pertinent, c'est pourquoi nous nous contenterons d'une présentation agrégée. En effet, il est apparu à l'examen que de nombreux items « favorisant » la France relevaient de la géométrie, tandis que les items d'algèbre étaient « plus difficiles » dans notre pays. Le tableau 4 permet de vérifier ce résultat en présentant les scores moyens (en %) pour les différents domaines des mathématiques.

Tableau 4 : taux de réussite de le France et des Etats-Unis, par niveau, pour les différents domaines de mathématiques

Niveau	Pays	Total	Fraction et sens des nombres	Géométrie	Algèbre	Représentation des données, probabilités	Mesure	Proportionnalité
Nombre d'items		*151*	*51*	*23*	*27*	*21*	*18*	*11*
4ème	France	61	64	66	54	71	57	49
	Etats-Unis	53	59	48	51	65	40	42
5ème	France	51	53	58	39	63	49	41
	Etats-Unis	48	54	44	44	60	36	38

Considérons l'exemple de la cinquième. Sur les 151 items que comprend l'épreuve, la France se situe un peu au-dessus des Etats-Unis (3 points exactement ce qui, du fait de l'imprécision due à l'échantillonnage, n'est pas significatif) mais la situation est très différente suivant les domaines : en géométrie, les élèves français distancent nettement leurs camarades américains (58 % de réussite aux 23 items de ce type contre 44 %) alors qu'en algèbre ils sont sensiblement en retrait (39 % de réussite aux 21 items contre 44 % aux Etats-Unis). On pourrait encore affiner l'analyse : parmi les items du quatrième domaine, les français ont peu brillé dans les questions de probabilité. Par exemple, si on leur dit qu'un vase contient 9 jetons numérotés de 1 à 9, ils sont 61 % (contre 70 % aux Etats-Unis) à affirmer qu'on a 4 chances sur 9 de tirer un jeton pair (beaucoup optent pour 1 chance sur 2).

Il est à peine besoin de souligner l'intérêt de ce genre d'analyses. Si l'on souhaite que les évaluations internationales ne se résument pas à un simple tableau d'honneur mais servent à définir une politique éducative, l'examen des résultats domaine par domaine, voire item par item, par les spécialistes des disciplines évaluées, permettra d'infléchir la teneur des programmes, en montrant les points où il convient de faire un effort particulier.

4.2.4. Les relations différentes au questionnaire

Nous achèverons cette présentation rapide des causes possibles de biais d'item par des considérations sur le rapport au questionnaire. La forme générale des questions peut influencer les réponses, si elle paraît aux élèves trop déroutante. Mais surtout, nous avons pu constater que les réactions des élèves face aux efforts que nécessite la réponse, donnent lieu à des stratégies très différentes en France et aux Etats-Unis. Nous détaillons par la suite deux exemples pour mettre en évidence ce phénomène mais des analyses plus agrégées permettraient sans aucun doute de confirmer ces résultats.

Deux exercices portaient sur la manipulation des expressions du premier degré. Nous ne nous attarderons pas sur leur différence en terme de contenu pédagogique. Les différences de forme paraissent aussi importantes. L'item L16 demandait : « trouver x tel que 10x-15=5x+20 ». L'item N13, lui aussi sous forme de question ouverte, demandait « si x=2, combien vaut (7x+4)/(5x-4) ? ».

Tout d'abord, il est intéressant de noter une différence entre la France et les Etats-Unis concernant les non-réponses. Pour la question L16, 32 % des français n'ont rien répondu, contre seulement 18 % des américains. Sachant que les non-réponses sont souvent interprétées comme des réponses fausses, cela serait le signe que la question est particulièrement difficile pour les français, et si difficile qu'ils n'essayent même pas de répondre. Or elle est réussie à 32 % en France soit 8 points de plus qu'aux Etats-Unis ! Notons qu'on retrouverait sur d'autres items de l'épreuve (sous forme de questions ouvertes essentiellement) cette « contradiction » entre taux de réussite et taux de non-réponses. Il semble que les élèves français ne fassent pas de difficulté pour vérifier la concordance de deux propositions et cocher la case correspondant à la bonne réponse mais nombre d'entre eux rechignent à s'investir dans un problème qui semble demander plus d'effort et surtout, exigence jugée souvent inadmissible, où il faut écrire sa réponse ! On peut voir là une certaine démotivation face à une évaluation qui paraît sans enjeu. D'un point vue technique, cela amène à se poser des questions sur le traitement des non-réponses[12].

---

[12] Dans de nombreuses enquêtes, les non-réponses se voient attribuer une signification différentes selon leur emplacement. Quand on trouve à partir d'un certain exercice un nombre important de non-réponses qui se suivent jusqu'à la fin du questionnaire, on considère que, manque de temps ou de motivation, la personne s'est arrêtée là et que puisqu'elle n'a pas répondu aux dernières questions, on ne les fait pas entrer dans le calcul du score (les modèles MRI permettent de faire ceci en tenant compte de la difficulté des questions non faites). En revanche, toute non-réponse « interne », c'est-à-dire suivie par au moins une réponse, bonne ou non, est considérée comme une mauvaise réponse. Alain Blum et France Guérin-Pace ont montré dans l'article déjà cité qu'une telle façon de voir était dangereuse et qu'il y avait bien des cas où la personne interrogée « sautait » des exercices non parce qu'ils

Le cas de l'item N13 est différent. Les taux de non-réponses sont comparables (17 % en France contre 13 % aux Etats-Unis). Ce qui est intéressant ici, c'est de distinguer les bonnes réponses. Un peu moins de la moitié des élèves en France comme aux Etats-Unis réussissent l'item (il serait donc biaisé, d'après ce que nous avons vu, en faveur des américains) mais surtout les bonnes réponses ne sont pas les mêmes dans les deux pays. Aux Etats-Unis, 42 % répondent « 3 » et 6 % répondent « 18/6 ». En France, ils sont 33 % seulement à répondre « 3 » et 15 % à s'être contentés de donner la fraction sans la réduire complètement. On peut y voir là encore, plus que des écarts de compétences, un différentiel de motivation face au questionnement.

4.2.5. Que faire des items biaisés ?

Les sources de biais sont nombreuses mais elles ne sont pas de même nature. Les erreurs de traduction ou de présentation peuvent être corrigées. En revanche, la recherche d'équivalents dans toutes les langues ou les différences de motivation des individus sont beaucoup plus difficiles à prendre en compte. Enfin, les différences de réussite dues aux programmes scolaires ne semblent pas devoir être effacées, compte tenu de leur intérêt pédagogique.

En fait, face au biais, deux attitudes extrêmes sont possibles. Les intégristes de la mesure unidimensionnelle, généralement dévots des MRI, considèrent que les items biaisés détériore la qualité de la mesure, qu'il faut les éliminer quand par négligence, ils subsistent. En effet, d'un point de vue théorique, les items n'ont aucun intérêt pris individuellement, ils ne sont que des intermédiaires servant à l'élaboration de la dimension évaluée, unique et unidimensionnelle. On peut imaginer une attitude plus (trop ?) souple qui consisterait à construire une épreuve censée mesurer au mieux ce que l'on veut et tout ce que l'on veut, à analyser les résultats pour déterminer a posteriori si l'on ne pourrait pas en tirer un ou plusieurs facteurs synthétiques dominants. On pourrait alors reprocher à une telle approche d'être trop empirique. Il est en effet préférable d'avoir une théorie bien définie avant d'aborder la constitution d'un protocole et l'analyse des résultats. Mais, plutôt que de rejeter systématiquement les items biaisés, il est plus intéressant de profiter de leur éclairage sur les différences de réussite entre les pays.

---

lui semblaient hors de portée mais par manque d'intérêt. Ce que nous venons d'exposer rejoint cette analyse.

4.3. Généralisation aux comparaisons multiples

Il semble intéressant de tenter de généraliser les analyses menées ci-dessus au cas des comparaisons entre plus de deux pays. Comment dans ce cas-là mettre en évidence des biais ? Pour cela, on peut étudier les profils de réussite des pays, c'est-à-dire pour chaque pays la hiérarchie des items selon leur pourcentage de bonnes réponses. Théoriquement, si on respecte l'hypothèse d'unidimensionnalité, tous les pays devraient avoir le même profil de réussite (les items ayant la même difficulté dans chaque pays, ils doivent se retrouver dans le même ordre en terme de réussite[13]). Si on observe des variations de profils de réussite d'un pays à l'autre, on peut supposer que ces variations sont liées à des différences culturelles et ne sont pas le simple fruit d'un effet aléatoire. Pour le vérifier, nous avons procédé à une classification hiérarchique des pays selon leurs profils de réussite aux items de mathématiques de l'enquête TIMSS pour la population 2. Cette méthode est reprise d'une analyse similaire menée par F. Guérin-Pace et A. Blum sur les données de IALS (1999). D'ailleurs, on aboutit ici au même constat : la proximité des profils de réussite des pays coïncide en général avec leur proximité géographique, culturelle ou linguistique. D'où la formation de groupes de pays à peu près cohérents selon ces critères (voir graphique 5).

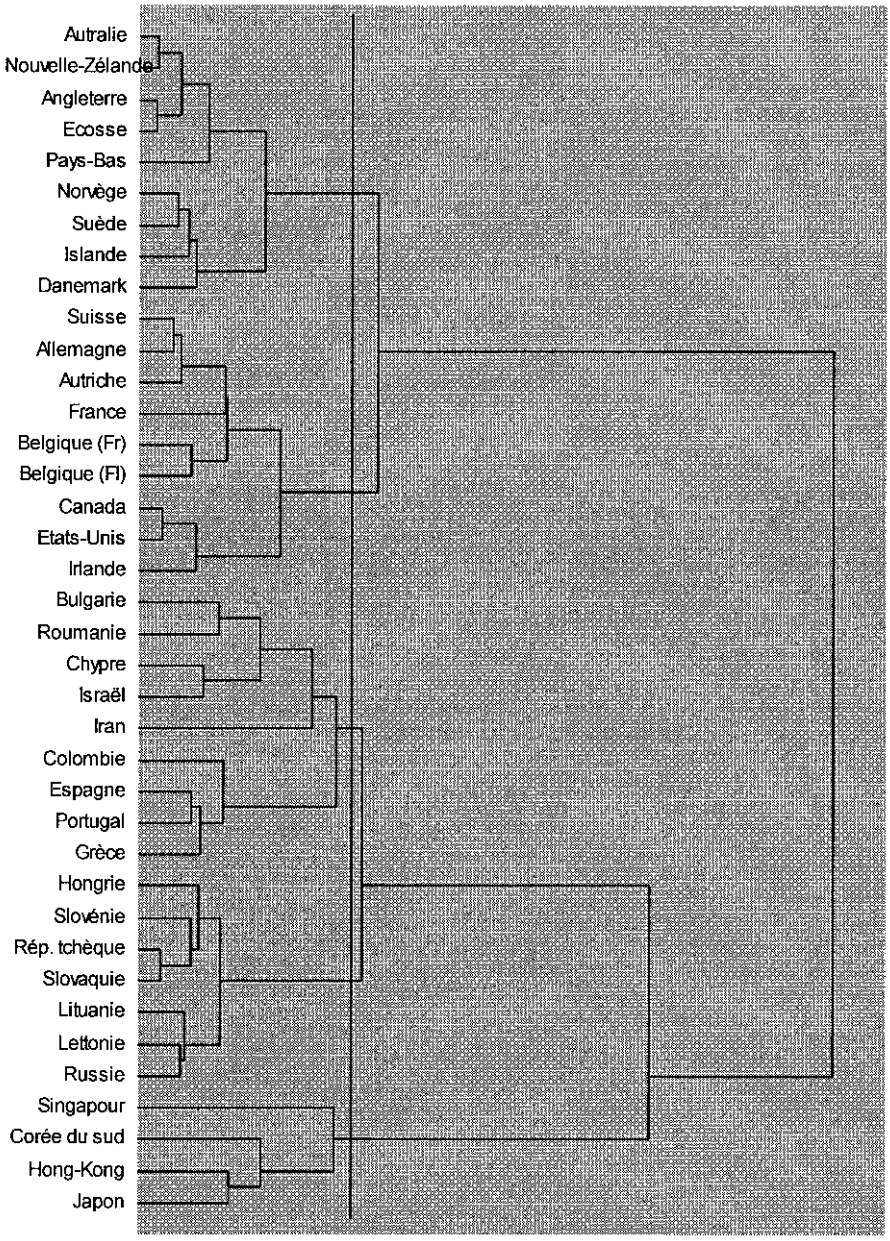
Les quatre pays asiatiques (Singapour, la Corée du Sud, Hong-Kong et le Japon) forment un groupe homogène. Une grande classe rassemble les pays d'Europe occidentale (sauf l'Espagne et le Portugal) ainsi que les Etats-Unis, le Canada, l'Australie et la Nouvelle-Zélande. De ce groupe se dégagent plusieurs sous-groupes intéressants comme par exemple les pays d'Europe du Nord ou les deux groupes de pays anglophones. Les pays d'Europe de l'Est se retrouvent pratiquement tous dans le même grand groupe (à part la Bulgarie et la Roumanie). Le dernier regroupement est certainement le plus hétérogène : même si l'on observe des proximités géographiques ou culturelles (comme l'Espagne et le Portugal, par exemple), il rassemble des pays aussi divers que Chypre, l'Iran ou la Bulgarie.

Pour illustrer l'arbre de classification, nous avons représenté graphiquement les croisements des pourcentages de bonnes réponses de l'Angleterre et de l'Australie d'une part, et de l'Angleterre et du Japon d'autre part (graphiques 6 et 7).

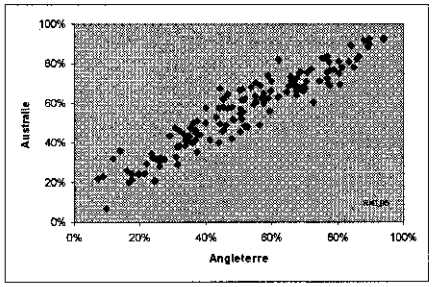
---

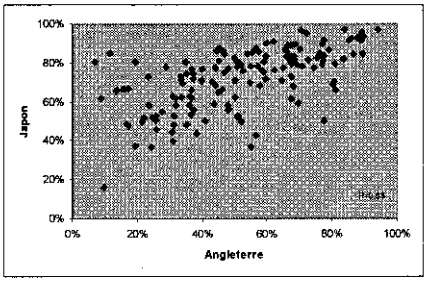
[13] Ce résultat est certain pour les modèles à 1 paramètre (cas de TIMSS) où, on l'a vu, le taux de réussite est équivalent au paramètre de difficulté donné par le modèle. La situation est plus complexe quand on utilise un modèle à 2 paramètres.

Graphique 5 : Arbre de classification des pays selon le profil de réussite



Graphiques 6 et 7 : comparaison des profils de réussite de l'Angleterre avec ceux de l'Australie et du Japon





Chaque point représente un item dont l'abscisse est le pourcentage de bonnes réponses observé en Angleterre et l'ordonnée le pourcentage de bonnes réponses au Japon ou en Australie. Il ressort assez clairement de ces représentations que l'échelle de réussite des items de l'Angleterre est très proche de celle de l'Australie mais assez éloignée de celle du Japon. D'ailleurs, la corrélation des rangs s'élève à 0.95 dans le premier cas contre seulement 0.63 dans le second.

4.4. Stabilité des classements des moyennes

A partir de la classification précédente, on a établi une partition des pays en cinq classes. Pour chacune de ces classes, on a déterminé les items les plus caractéristiques, c'est-à-dire les items particulièrement réussis ou échoués par ces populations.

Le classement des pays se modifie sensiblement selon qu'on le calcule sur les points faibles ou sur les points forts d'une classe particulière. Par exemple, nous avons classé les pays selon le taux de réussite moyen obtenu aux 20 items constituant les points faibles du groupe des pays asiatiques. On constate que, par rapport au classement établi sur l'ensemble des items, ces pays voient leurs positions considérablement diminuées. Ainsi Singapour passe de la 1ère à la 4ème place, le Japon de la 2nde à la 14ème, la Corée du Sud de la 3ème à la 19ème et Hong-Kong de la 4ème à la 24ème. Une autre pondération des items peut donc amener à un classement final très différent.

Ce dernier résultat conduit à s'interroger sur la robustesse des classements publiés par l'OCDE. Les imprécisions du processus d'élaboration de la mesure impliquent une certaine instabilité dans les palmarès des niveaux moyens. Il faut cependant reconnaître que la simulation du paragraphe précédent n'est pas complètement satisfaisante sur ce point. En effet, on s'est placé là dans le pire des cas pour les pays asiatiques, sur *la* sous-épreuve parmi les millions que l'on pourrait constituer à partir

des items passés qui donne un résultat très divergent du résultat publié. Il pourrait être intéressant de procéder à ce genre de simulation sans a priori, en constituant à partir d'un ensemble d'items, toutes les sous-épreuves possibles (voir pistes de recherche). Ces analyses ont cependant l'intérêt de montrer qu'on observe parfois de véritables bouleversements de classements et surtout, il est assez significatif que l'épreuve qui défavorise Honk-Kong soit aussi celle, à peu de choses près, qui défavorise le Japon ou Singapour.

Ceci nous amène à reprendre la question des biais d'items, en envisageant leur influence sur les classements des niveaux moyens des pays. Il est toujours possible de contester les résultats d'une évaluation en remettant en cause la composition des épreuves, en arguant que la dimension ne serait pas tout à fait la même ainsi si une équipe d'un pays particulier avait composé l'épreuve. Cependant, la composition des épreuves est, ou devrait être, le fruit d'un compromis entre les différents pays participants. Ainsi, on dispose au final d'une « norme internationale ». Il est alors important de montrer comment la définition nationale s'écarte de cette norme pour expliquer les résultats.

Un point plus inquiétant concerne la motivation des individus. Il est possible que les individus répondent systématiquement moins bien par rapport à leur niveau réel, à cause de leur manque d'intérêt par rapport à l'évaluation. Dans le cas de IALS, l'examen des taux de non-réponses est particulièrement éclairant sur ce point : 45 % des ménages français de l'échantillon initial ont refusé de répondre à l'enquête ; 12 % n'ont répondu qu'au questionnaire portant sur leurs caractéristiques sociodémographiques, sans répondre aux exercices ; 6 % ont commencé à y répondre mais se sont arrêtés avant la fin. En définitive, 37 % seulement de l'échantillon initial est allé au bout du questionnaire d'évaluation, en omettant parfois certaines questions. Cela pose trois questions : les individus qui ont refusé de répondre sont-ils identiques aux répondants ? La non-réponse à une question est-elle le signe de l' » incompétence » du sujet ou de son désintérêt ? Est-on certain que ceux qui répondent l'ont fait avec assez de sérieux ? On peut penser que ces questions sont moins déterminantes dans le cadre scolaire, où les épreuves sont vite assimilées aux évaluations « à enjeu » habituelles. En revanche, quand, comme dans le cas de IALS, on cherche à mesurer les compétences d'individus ayant parfois quitté l'école depuis longtemps, on peut se demander s'ils accepteront de jouer le jeu et de consacrer des efforts de concentration souvent importants pour répondre aux questions.

5. Les inégalités de compétences

Il n'est sans doute jamais inutile de rappeler le caractère réducteur d'un jugement fondé sur le seul examen d'une moyenne. La plupart des études accordent une place tout aussi importante à la dispersion des résultats, qui donnent une idée des inégalités scolaires. On les caractérise souvent par les différences de taux d'accès à tel niveau de scolarisation entre enfants de cadres et enfants d'ouvrier, entre garçons et filles. On a vu ce que les évaluations de compétences permettaient d'apporter en plus par rapport à ces analyses. Il semble donc important de vérifier que la mesure des compétences se fait de façon suffisamment précise pour donner une idée des écarts individuels.

Les investigations que nous allons mener dans ce chapitre concernent la catégorie la plus « statistique » des indicateurs d'inégalités, ceux mesurant la dispersion brute (écart-type, écart interquartile). Nous essaierons de voir aussi dans quelle mesure on peut généraliser les résultats à des indicateurs plus « sociaux » (différences de scores entre enfants de cadres et enfants d'ouvriers par exemple)[14].

5.1. Inégalités des compétences et écart-type des performances

Nous avons déjà évoqué les difficultés à interpréter des écarts entre individus. En effet, si on utilise la simple proportion de bonnes réponses, le niveau de l'épreuve détermine la dispersion : faible si l'épreuve est très facile ou très difficile (taux moyen proche de 0 % ou de 100 %), élevée si l'épreuve est de difficulté moyenne pour la population considérée (taux moyen entre 25 et 75 %).

Cela suggère qu'il est difficile d'atteindre les inégalités de compétences et que la dispersion des scores que l'on va observer risque d'être très dépendante de l'épreuve que l'on a utilisée. En fait, la dispersion des performances va dépendre autant de la « qualité » de l'épreuve (ou, pour être plus précis, de son degré d'adaptation au public visé), que de la « réelle » dispersion des compétences. Comme contre-exemple, on peut imaginer une épreuve dont les questions seraient si mal posées que les élèves répondraient au hasard : la dispersion serait faible non parce que les compétences sont proches mais parce que l'épreuve les a mal mesurées.

---

[14] Cependant les indicateurs bruts de dispersions sont souvent utilisés dans le cadre des comparaisons internationales, en partie à cause des difficultés que l'on rencontre à comparer des écarts sociaux définis à partir de catégories rarement équivalentes d'un pays à l'autre.

L'objectif d'une évaluation statistique est de mettre en évidence des groupes en opposition bien tranchée, de distinguer les meilleurs des moins bons. Pour mieux comprendre cela, on peut songer à la situation du professeur dans sa classe : ce n'est pas parce que tous ses étudiants savent compter qu'un professeur d'université va leur mettre 10 sur 10. Inversement, un enseignant du primaire évitera d'interroger ses élèves sur la mécanique quantique. Le professeur adapte ses exigences afin de pouvoir distinguer les élèves selon leurs compétences. Il construira donc une épreuve de difficulté moyenne avec des questions faciles pour repérer les élèves en grande difficulté et des questions difficiles pour départager les meilleurs.

On met ainsi en évidence une apparente contradiction d'objectifs : si l'un des buts principaux d'une bonne politique éducative est de maintenir des écarts peu importants de compétences entre individus, celui des concepteurs d'évaluation est de chercher la meilleure épreuve possible, celle qui distingue le mieux les meilleurs des moins bons, les pays performants des moins performants. Ils visent donc la plus grande dispersion possible, afin de pouvoir plus facilement expliquer les écarts. En ce sens le fait de trouver un écart-type élevé peut être le fait d'une épreuve bien adaptée autant que d'une population variée.

Ce résultat doit rendre extrêmement prudent quand on compare deux populations n'ayant pas passé la même épreuve de mathématiques par exemple (ne parlons pas des comparaisons entre disciplines dont le sens est encore plus problématique). Le fait de trouver dans un cas un écart-type des performances plus élevé signifie-t-il que la population est plus dispersée ou que l'épreuve proposée est mieux adaptée ? Mais le problème peut aussi se poser quand on utilise une seule épreuve sur des publics différents : les écarts de dispersions ne relèvent-ils pas alors davantage du degré d'adaptation de chaque épreuve au public plutôt que de réels écarts de compétences ? C'est ce dernier cas que nous allons plus particulièrement étudier dans le cas des comparaisons internationales : l'usage d'un même protocole partout permet-il d'avoir une bonne image de la dispersion des compétences au sein de chaque pays ?

5.2. Dispersion des résultats et qualité de l'épreuve

On a vu que l'une des façons d'apprécier la qualité d'un test était l'examen de sa fidélité (voir chapitre 2.3). Rappelons que la fidélité mesure le degré de corrélation entre items, afin de vérifier qu'ils dépendent tous partiellement d'un même facteur. Il peut être intéressant de confronter cette notion, et les indicateurs qui en découlent, avec la mesure de la dispersion.

5.2.1. Des notions interdépendantes

Tout d'abord, on peut noter une dépendance mathématique entre la plupart des indicateurs de fidélité et les indicateurs de dispersion. Pour prendre le plus courant, le α de Cronbach (voir chapitre 3.2.), il apparaît qu'à structure d'épreuve donnée en terme de taux de réussite, plus l'écart-type est élevé, plus le α le sera. Et inversement. Ceci concerne l'écart-type du taux de réussite et ne se généralise pas forcément aux scores issus des MRI. Cependant, sur le plan empirique, il n'est pas difficile de mettre en évidence une telle relation : dans le cas de TIMSS, quand on compare l'écart-type du pays avec la mesure de la fidélité de l'épreuve dans ce pays, la corrélation linéaire est extrêmement élevée (elle est par exemple de 0,95 en sciences pour la septième année de scolarité sur 39 pays participants). Les corrélations sont du même ordre dans l'enquête IALS.

Une telle ampleur dans les corrélations est assez préoccupante. En effet, on peut se demander si l'un de ces indicateurs n'est pas « mensonger » et simple décalque de l'autre. Il est possible par quelques expériences de montrer que les torts sont partagés. Par exemple, quand on travaille sur des sous-groupes a priori peu dispersés, on note des α beaucoup plus faibles que pour l'ensemble de la population. Il apparaît ainsi que les α calculés pour des groupes définis selon le niveau de diplôme des parents sont tous inférieurs à ce que l'on note sur l'ensemble de la population. L'épreuve serait donc adaptée à toute la population sans être mieux adaptée à aucun sous-groupe qui la compose. En d'autres termes, discriminer entre eux des polytechniciens ou des illettrés est plus difficile que faire le même travail sur l'ensemble de la population. De plus, quand on propose une épreuve à une population a priori très dispersée (la population totale par exemple, avec ses niveaux d'instruction très variables) on risque de voir son épreuve créditée de coefficient de fidélité important, quelle que soit sa réelle qualité. La plupart des participants des JMS savent résoudre l'équation $3x+2=5$ et répondre à la question : « A quel temps est la phrase : « j'aurais mangé » ? ». Ces deux questions risquent de poser problème à beaucoup de personnes ayant quitté l'école assez vite, il y a longtemps. Par conséquent, la corrélation entre les réussites à ces deux questions sera sans doute significative. De là à dire qu'elles mesurent la même chose...

5.2.2. Des dispersions variables

La dispersion de la population a donc une incidence directe sur l'indice de fidélité : plus les écarts sont élevés, meilleurs seront les α. Inversement, l'image que l'on aura des inégalités de compétences au sein d'un pays risque de dépendre de la qualité de l'épreuve. Ainsi, en travaillant sur les « meilleurs » items (c'est-à-dire les plus discriminants, les mieux corrélés avec le score global) on aboutit à des indicateurs de dispersions beaucoup plus élevés. Il est aussi possible de construire des épreuves discriminantes pour certains pays mais pas pour d'autres et de montrer la sensibilité des classements obtenus. Prenons un exemple. On va comparer les résultats de l'Australie et de la Grèce en mathématiques, en fin d'étude primaire. Quand on

travaille avec les scores MRI calculés au niveau international, on trouve respectivement des scores moyens de 530 et 453 et des écarts-type de 98 et 95 : si les élèves grecs ont de moins bons résultats ils ne paraissent ni plus ni moins dispersés que les élèves australiens. Nous allons maintenant raisonner sur deux sous-épreuves du cahier 3 construites de façon à maximiser l'écart. L'une contient les items qui sont très discriminants en Grèce sans l'être en Australie ; l'autre contient les items qui sont discriminants en Australie sans l'être en Grèce. Nous présentons ci-dessous (voir tableau 5) les scores moyens sous forme de pourcentages puis mis sous une forme logistique (on applique donc au score individuel la transformation log(p/(1-p)) ce qui donne des résultats très proches, on l'a vu, des scores issus d'un modèle de réponse à l'item).

Tableau 5 : Dispersions en Grèce et en Australie selon l'épreuve

	Australie		Grèce	
	Moyenne	Ecart-type	Moyenne	Ecart-type
Score total (en %)	61,6 %	19,9 %	43,7 %	19,8 %
Score total (logit)	0,191	0,334	-0,101	0,319
Score « à la grecque » (en %)	69,4 %	19,5 %	52,7 %	22,8 %
Score « à la grecque » (logit)	0,327	0,345	0,048	0,379
Score « à l'australienne » (en %)	49,0 %	24,6 %	29,0 %	18,7 %
Score « à l'australienne » (logit)	-0,020	0,423	-0,354	0,336

L'analyse du score global (sous forme % ou logistique) confirme ce que l'on trouve avec le score MRI : la moyenne de la Grèce est inférieure et les écarts-type sont proches. Quand on travaille sur les deux sous-épreuves, les résultats diffèrent, pas tellement en terme de moyenne (l'écart entre les pays reste à peu près du même ordre) mais plutôt en terme de dispersion. Sur l'épreuve construite à partir d'items particulièrement discriminants en Grèce, ce pays apparaît légèrement plus dispersé, tandis que sur l'autre épreuve, les écarts sont plus importants en Australie.

Ces résultats peuvent être sans peine généralisés à d'autres couples de pays, à d'autres cahiers, à d'autres disciplines. Il convient cependant d'en saisir la portée. Comme dans l'étude des moyennes, on montre l'ampleur du problème sans l'expliquer et en se plaçant dans le pire des cas, sans dire si cette hypothèse est réaliste. En d'autres termes, nous sommes capables d'extraire des sous-épreuves qui ne fonctionnent pas bien à partir de l'existant. Cela ne prouve pas pour autant que les constructeurs de tests n'ont pas réussi à établir une épreuve globale qui soit valable pour tous, même si elle contient des sous-parties qui posent problème.

En effet, tout ce que nous développons ici n'a d'intérêt que si l'on peut montrer qu'il existe un risque qu'une épreuve donnée s'adapte plus ou moins bien selon les pays. Quels sont les facteurs qui peuvent influer sur la fidélité nationale d'une épreuve ? L'examen de cette question est encore peu avancé mais il apparaît vite un facteur

assez accessible et très déterminant sur le degré d'adaptation d'une épreuve : sa difficulté.

5.3. Moyenne et écart-type

5.3.1. Relation entre niveau moyen et dispersion

Les modèles psychométriques du type MRI ont beau être très élaborés, il est sans doute impossible de recruter les jeunes polytechniciens à l'aide d'une épreuve composée d'une centaine d'additions. De même si l'on propose une épreuve par QCM sur les équations du second degré à des élèves de CE2 et si le score varie entre 0 et 20 % de réussite, il est tout de même un peu abusif d'en conclure que les compétences en mathématiques à ce niveau sont faibles et peu dispersées. Exemples caricaturaux ? C'est à voir...

Les deux enquêtes que nous avons prises comme exemples conduisent parfois à des situations assez similaires. Dans TIMSS, on a cherché à mesurer la progression des élèves entre le milieu de l'école primaire et le milieu du secondaire, en les plaçant donc sur une même échelle. On avait même envisagé de prolonger la comparaison jusqu'à la fin du secondaire. Quant à IALS, le but de cette enquête était clairement d'évaluer les compétences en lecture de l'ensemble de la population, ce qui inclue en France les polytechniciens, les énarques et les personnes n'ayant pas eu le certificat d'étude. De plus, ces diversités des publics au sein de chaque pays s'additionnent quand on cherche à construire une épreuve valable pour tous les pays. La question se pose véritablement de savoir s'il est possible de construire une épreuve qui convienne à tous. Ne risque-t-on pas, en proposant une épreuve facile, de trouver une faible dispersion dans les pays performants, tandis que les pays les moins bons apparaîtraient comme particulièrement inégalitaires[15] ?

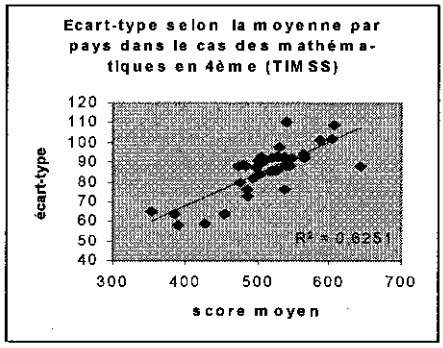
Prenons deux exemples. Le premier concerne les résultats en mathématiques des élèves en huitième année d'étude (niveau quatrième) dans une trentaine de pays

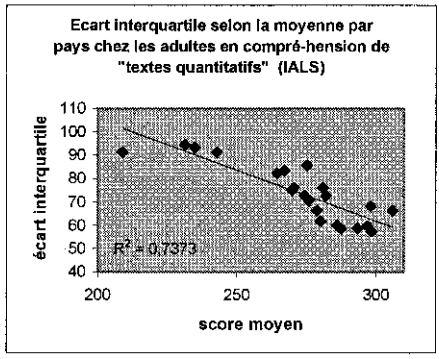
---

[15] On peut exprimer ce problème au niveau des items (ce qui selon les psychométriciens donne des résultats équivalents aux considérations sur les individus). Y a-t-il plus d'écarts en terme de difficulté entre la résolution d'une addition et d'une règle de 3, d'une part, ou entre une équation du second degré et la résolution d'un algorithme de maximisation de la vraisemblance d'autre part Il est probable qu'un traitement statistique nous conduirait à affirmer que le premier écart est plus grand (beaucoup d'individus réussissent l'addition sans réussir la règle de 3 alors que dans une population quelconque les deux autres sembleront également difficiles). Une analyse pédagogique ne donnerait pas forcément le même résultat. Les deux perspectives apparaissent d'ailleurs également légitimes mais se distinguent suivant le public et le niveau de compétence visé.

(enquête TIMSS). Le graphique 8 présente la relation entre l'écart-type du score et la moyenne par pays. Une corrélation fortement positive apparaît (r^2=63 %) : meilleure est la moyenne, plus grandes sont les inégalités. Si nous étudions maintenant les résultats de la population adulte en compréhension de textes quantitatifs (données issues de IALS, graphique 9), on note au contraire une corrélation fortement négative entre l'écart interquartile et la moyenne (r^2=74 %) : meilleure est la moyenne, moins grandes sont les inégalités ! Voilà deux résultats fort différents ! Bien sûr les notions mesurées ne sont pas identiques. De plus, la population n'est pas la même. Cependant, une telle contradiction dans les relations est troublante. La question est de savoir si ce résultat provient de la nature des épreuves (difficile dans le premier cas, facile dans le second) ou s'il touche réellement au lien entre inégalités et niveaux moyens (problème dont on voit sans peine l'importance car on touche à la question du lien entre démocratisation et évolution du niveau moyen).

Graphiques 8 et 9 : relation entre niveau moyen et dispersion pour TIMSS et IALS





5.3.2. Approche empirique

Pour mieux comprendre le lien entre la difficulté de l'épreuve et sa fidélité, nous allons extraire d'une épreuve donnée (le cahier 3 de mathématiques de TIMSS sur la population 1) différentes sous-épreuves de difficulté variée. Notons que cette méthode ne souffre plus des défauts précédemment exprimés : on ne se place plus, a posteriori, dans le pire des cas mais on teste une hypothèse sur un élément (la difficulté) indépendant du facteur étudié (la dispersion).

A partir des 39 items du cahier que nous avons retenu, nous avons construit 3 sous épreuves différentes de 10 items chacune (bien sûr le fait de travailler sur relativement peu d'items est un peu embarrassant mais l'objectif est ici de dégager des tendances plutôt que d'à aboutir à un chiffrage précis du phénomène) en considérant trois niveaux de difficulté (difficile/moyen/facile). Le tableau 6 donne les caractéristiques des scores sur l'ensemble de la population.

Tableau 6 : Dispersion selon la difficulté de l'épreuve

	Score en %		Score en logit	
	Moyenne	Ecart-type	Moyenne	Ecart-type
Ed=Epreuve difficile	31,6 %	24,4 %	-0,975	1,334
Em=Epreuve moyenne	52,6 %	26,8 %	0,138	1,407
Ef=Epreuve facile	78,6 %	20,8 %	1,563	1,210

Ce tableau présente les scores sous forme de pourcentages mais dans la suite nous n'utiliserons plus que les scores sous forme logistique. On retrouve les mêmes résultats concernant les scores en pourcentage sauf qu'il s'ajoute une dépendance due aux bornes 0 et 1 qui rend les relations (généralement quadratiques) entre moyennes et écarts-type encore plus fortes. L'examen de ce tableau permet de voir que nos épreuves sont bien distinctes en terme de difficulté. La deuxième étape consiste à calculer les moyennes et écarts-type par pays et à étudier les corrélations entre les valeurs obtenues. Il est tout d'abord intéressant de constater que lorsque l'on confronte les classements de moyennes obtenus selon les épreuves, les corrélations sont assez élevées (rappelons que l'on travaille maintenant sur une population de 27 pays) : la moins bonne est de 0,88 (entre Ed et Ef). En particulier, les corrélations avec le score moyen MRI calculé et publié par l'IEA sont toujours supérieures à 0,95. Cela semble indiquer que pour effectuer un classement des pays selon leur niveau moyen, la difficulté de l'épreuve importe peu, les classements sont assez stables. En revanche les résultats sont beaucoup moins bons quand on compare les écarts-type obtenus sur les différentes épreuves. Par exemple, il n'y a pas de lien entre les écarts-type calculés par pays sur l'épreuve Ed et ceux calculés sur l' épreuve Ef. Les pays qui paraissent très dispersés pour la première épreuve ne le sont pas forcément pour l'autre. La mesure de la dispersion au sein de chaque pays semble donc très sensible à l'épreuve utilisée et les classements obtenus peu fiables.

Notre hypothèse est que la difficulté des épreuves influe sur le public auquel l'épreuve s'adapte le mieux et confère donc la plus grande dispersion de résultats. Nous allons tester cette hypothèse en cherchant pour chaque épreuve la relation entre l'écart-type et la moyenne. Le tableau 7 présente le R^2 issu d'une régression linéaire simple de l'écart-type sur la moyenne puis le R^2 obtenu en ajoutant le carré de la moyenne (cet ajout permet de tenir compte du fait que la dépendance peut avoir la forme d'une cloche : pour une épreuve de difficulté moyenne en particulier, l'écart-type sera faible pour les pays les moins bons, élevé pour les pays moyens et faible pour les pays les meilleurs). Pour la relation linéaire, nous indiquons par une flèche, si la corrélation est positive ou négative.

Tableau 7 : Relation moyenne-écart-type selon la difficulté de l'épreuve

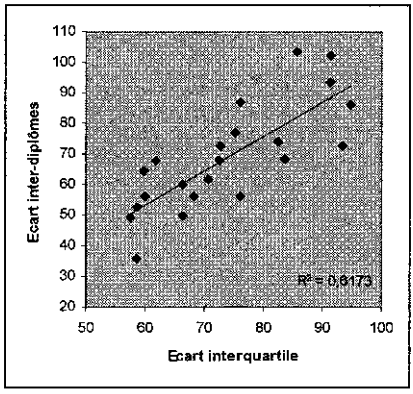
	relation linéaire	relation quadratique
Ed=Epreuve difficile	55,9 % ↗	61,7 %
Em=Epreuve moyenne	15,7 % ↗	67,8 %
Ef=Epreuve facile	66,7 % ↘	78,4 %

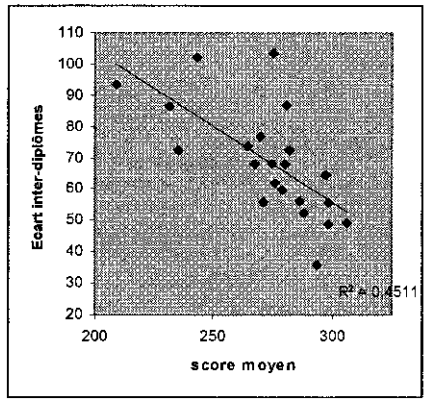
On observe des relations assez fortes entre moyennes et écarts-type : la dépendance est assez nettement linéaire et positive pour l'épreuve difficile (meilleur est un pays plus ses résultats sont dispersés). La dépendance est négative pour l'épreuve facile. La relation linéaire est non significative pour l'épreuve moyenne mais l'ajout d'un terme quadratique fait augmenter fortement la valeur de la prédiction, ce qui rend compte de la forme en cloche de la dépendance.

5.4. Ecarts sociaux

Les problèmes que nous venons d'exposer concernant la mesure de la dispersion brute ont évidemment une influence sur la mesure des inégalités sociales. En effet, on aurait tendance à voir une causalité inverse (s'il y a de grands écarts entre groupes sociaux, il risque d'y avoir de forts écarts absolus) mais comme nous venons de voir que les écarts entre individus sont appréhendés de façon toujours un peu déformée, il est probable que cette déformation affecte la mesure des inégalités sociales. Ainsi, on peut craindre que les conclusions quant à savoir quels sont les pays où il y a le plus d'écarts suivant l'origine sociale, soient très sensibles aux caractéristiques de l'épreuve choisie. Reprenons l'exemple de IALS. Le premier des deux graphiques ci-dessous permet de voir que l'écart « inter-diplômes » (i.e. la différence de scores entre les personnes ayant achevé des études supérieures et celles n'ayant pas fini des études secondaires) est très lié à la mesure brute de dispersion que constitue l'écart interquartile. Le graphique 11, quant à lui, montre le lien entre le premier écart et le score moyen : la forme de la dépendance rappelle celle que l'on trouvait pour l'écart interquartile (cf graphique 9). On peut se demander si cette dépendance a un sens, si les pays qui ont le meilleur niveau moyen sont aussi ceux qui maintiennent des écarts faibles entre les groupes sociaux ou si cette corrélation n'est qu'un artefact statistique dû au fait que l'épreuve était assez facile. En effet, si l'on effectue le même genre d'analyse sur TIMSS, en mathématiques, en utilisant l'origine sociale des parents des élèves interrogés, on trouve une dépendance positive : les pays les meilleurs sont ceux où il y a le plus d'écart, résultat qui ressemble à ce que l'on trouve pour l'écart-type et qu'il nous semble possible d'expliquer par le fait que les exercices étaient plutôt difficiles.

Graphiques 10 et 11 : écarts sociaux et dispersions, écarts sociaux et niveaux moyens





Les remarques développées dans ce chapitre amènent donc à s'interroger sur la validité de tous les indicateurs cherchant à mesurer les écarts entre individus. Est-il possible de conclure que tel pays est plus inégalitaire que tel autre, sachant qu'un ensemble différent d'exercices aurait pu conduire à des constats différentes. Parmi les facteurs qui peuvent perturber les classements, on a vu que la difficulté de l'épreuve jouait un rôle important ce qui est un problème en soi. En effet, le lien entre le niveau moyen atteint par l'ensemble de la population et les inégalités entre individus est un point crucial du débat sur l'école. Pour dire les choses en termes plus politiques, est-il possible de démocratiser le système éducatif, en élevant le niveau ? On a vu que selon qu'on travaillait sur TIMSS ou sur IALS, on aboutissait sur ce point à des conclusions divergentes.

Enfin, nous n'avons pas évoqué la question des indicateurs relatifs qu'on utilise parfois. Dans l'exemple précédent, on peut en effet se demander si le rapport entre l'écart inter-diplôme et l'écart interquartile donne une image assez bonne de l'influence du niveau d'étude. Loin d'être une évidence, cette question est en cours d'exploration.

6. Quelques pistes de recherche

Il n'a pas été possible de présenter dans le cadre de ce colloque l'ensemble de nos recherches, ni de mener à bien tout ce que nous avons entrepris. Nous allons donc présenter succinctement les pistes qu'il nous semble intéressant d'explorer.

- ***Confronter les résultats d'enquêtes entre eux.*** Bien sûr il n'existe pas au niveau international d'enquêtes concurrentes qui chercheraient à mesurer au même moment, dans la même matière, au même niveau de la scolarité les compétences des individus. Cependant, des enquêtes plus ou moins comparables sont effectuées à des époques différentes ou à des niveaux différents. Il peut être intéressant de croiser le classement des différents pays pour voir s'ils sont stables (relativement stables, évidemment car, si l'enquête a un intérêt, c'est justement parce qu'on s'attend à ce qu'ils bougent). Dans le cas de TIMSS, par exemple, il est important de noter que deux niveaux successifs étaient ciblés par les enquêtes : CE2-CM1 pour la population 1 ; cinquième-quatrième pour la population 2. Le fait de trouver des classements plus stables quand on compare deux niveaux successifs peut s'interpréter à l'aide de cette proximité mais le fait que ces niveaux sont évalués par le même protocole doit jouer un rôle. Il est à noter que la comparaison des inégalités bruts (écarts-type des scores) montrent une grande instabilité des classements quand on compare deux niveaux « éloignés », alors qu'ils sont assez proches pour les comparaisons CE2-CM1 et cinquième-quatrième.
- ***Décomposer une épreuve.*** Une méthode pour étudier de façon empirique la sensibilité des indicateurs peut être de tirer au hasard dans un ensemble d'items, une sous-épreuve dont on va confronter les résultats avec ce que donne la totalité. En procédant ainsi un assez grand nombre de fois, en imposant de plus quelques contraintes pour obtenir des épreuves de caractéristiques assez variées (en particulier en terme de difficulté) on pourra se faire une idée de la stabilité des classements. Les premiers essais que nous avons effectués montrent une bonne stabilité des classements de pays par niveaux moyens (on est à chaque fois très proche du classement obtenu avec l'ensemble des items) alors que les classements de dispersion sont beaucoup moins constants (il n'est pas rare de construire ainsi des épreuves qui donnent un classement corrélé autour de .50 avec le classement sur l'ensemble des items). De même, il apparaît que conformément à ce que l'on a dit, le lien entre moyenne et écart-type dépend fortement du niveau de difficulté de l'épreuve.
- ***Les cahiers tournants.*** Dans le même ordre d'idée, il peut être intéressant d'utiliser ces sous-épreuves déjà définies que sont les cahiers tournants. En comparant les classements de pays obtenus sur chaque cahier indépendamment, on a une autre image, un peu moins négative, de leur robustesse. Les classements de moyennes sont corrélés à plus de .95 entre eux. Encore une fois, les corrélations sont moins bonnes pour les indicateurs de dispersion.

Cependant, les premières analyses montrent que l'usage des modèles MRI permet d'augmenter un peu la stabilité des classements d'un cahier à l'autre.

- ***Approfondir la connaissance des modèles MRI.*** Il semble indispensable de chercher à mieux comprendre le fonctionnement de ces modèles, pour déterminer ce qu'ils peuvent apporter. La technique des valeurs plausibles permettra peut-être d'améliorer l'estimation des écarts individuels. De plus, la confrontation des modèles à 1, 2 et 3 paramètres s'avérera instructive. En début d'année, les responsables de TIMSS ont ainsi fourni de nouvelles estimations des scores de compétences calculées à partir d'un modèle à 3 paramètres au lieu de 1 jusqu'à présent. Ce modèle donne des résultats sensiblement différents. En particulier, il n'y a plus de corrélation positive entre la moyenne et l'écart-type en mathématiques, pour la population d'élèves de collège. Au contraire, la corrélation est significativement négative, comme pour les autres scores ! Une telle inversion dans un résultat fondamental doit amener à s'interroger sur l'effet de la modélisation sur un même ensemble de données.

Conclusion

Le tableau dressé ici peut paraître globalement négatif, du fait que notre démarche critique s'est attachée à montrer les points sur lesquels des progrès restent à faire. Il convient de nuancer notre propos. Tout d'abord, la plupart des remarques concernent en premier lieu le cas des comparaisons internationales et ne s'appliquent que partiellement aux évaluations faites dans un seul pays. D'autre part, tout n'est pas à rejeter dans les comparaisons internationales et sur bien des points, elles apportent un éclairage intéressant sur le fonctionnement des systèmes éducatifs. Cependant, nous avons voulu montrer qu'il fallait les utiliser avec d'extrêmes précautions.

De plus, il va nous être impossible de rester dans le domaine de la critique. En effet, l'INSEE, en collaboration avec la DPD, l'INED et l'INETOP, s'est vue charger de mener une enquête méthodologique sur la question la plus délicate : l'évaluation des compétences des adultes. Quand l'enquête IALS a commencé, l'expérience française sur le sujet était à peu près nulle. Il existe certes de nombreuses évaluations menées auprès d'adultes, dans le cadre de la formation professionnelle par exemple, mais aucune ne visait à aller « chez les gens » pour évaluer leurs compétences. Le pari de faire de cette première expérience dans le cadre d'une enquête internationale était risqué. Effectivement, devant les difficultés rencontrées concernant la passation des épreuves et le traitement des réponses, il a semblé préférable d'ajourner la participation française. En définitive, l'INSEE a certainement un rôle important à jouer pour faire progresser la connaissance française sur la question, en menant une enquête méthodologique à la fois sur le terrain et dans les modèles.

Eléments de bibliographie

Publications de la DPD

- La Direction de la Programmation et du Développement donne régulièrement dans ses publications annuelles des indicateurs relevant de la mesure de compétence (voir *Etat de l'Ecole, Géographie de l'Ecole, Repères et Références*).

- On trouvera dans les autres publications (les *Notes d'information,* la revue *Education et Formations*, les *dossiers d'Education et Formations*) des informations plus détaillées sur des opérations d'évaluation de compétences ayant été mené au CP, au CE2, au CM2, en sixième, en 4ème et 5ème, en 3ème générale et technologique, en Terminale, ainsi que des études spécifiques sur les jeunes de 17 ans passant les tests de lecture de la journée d'Appel Préparation Défense, le recrutement de l'élite scolaire (Grandes Ecoles), une comparaison fondée sur les résultats d'un échantillon de 1920 au certificat d'étude, etc.

Ouvrages généraux sur le système éducatif.

- C. THELOT, « *L'évaluation du système éducatif* », Paris, Nathan, 1993.

- M. DURU-BELLAT, A. VAN ZANTEN, « *Sociologie de l'école* », Paris, Armand Colin, 1999.

Articles récents sur les inégalités devant l'école

- P. MERLE, « *Concept de démocratisation de l'institution scolaire* », Population n°1-2000, janvier-février.

- D. GOUX, E. MAURIN, « *La persistance du lien entre pauvreté et échec scolaire* », France Portrait Social 2000-20001.

- C. THELOT, L.-A. VALLET, « *La réduction des inégalités sociales devant l'école depuis le début du siècle* », Economie et Statistique n°334, 2000-4.

Psychométrie

- S. J. GOULD, « *La Malmesure de l'homme* », Odile Jacob, 1997.

- M. HUTEAU, J. LAUTREY, « *Evaluer l'intelligence* », PUF, 1999.

- M. REUCHLIN, « *La psychologie différentielle* », PUF, 1997.

- P. DICKES & al., « *La psychométrie* », PUF, 1994.

- F.M LORD & M. R. NOVICK , « *Statistical theories of mental test scores* » Addison-Wesley, Reading, 1968.

Publications internationales

- A. E. BEATON & al., « *Science achievement in the middle school years : IEA's third international mathematics and science study (TIMSS)*, » MA : Boston College, Chestnut Hill, 1996.

- A. E. BEATON & al., « *Mathematics achievement in the middle school years : IEA's third international mathematics and science study (TIMSS)*, » MA : Boston College, Chestnut Hill, 1996.

- OCDE, « *Regards sur l'éducation : Les indicateurs de l'OCDE* », OCDE, Paris, (1995 à 1998).

- OCDE & Statistique Canada, « *Littératie, Economie et Société : résultats de l'enquête internationale sur les capacités de lecture et d'écriture des adultes* » OCDE, Paris, 1995.

- OCDE & Développement des ressources humaines Canada, « *Littératie et société du savoir : nouveaux résultats de l'enquête internationale sur les capacités de lecture et d'écriture des adultes* », OCDE, Paris, 1997.

- NCES, « *Technical Report on the First International Adult Literacy Survey* », US Department of Education, Washington DC, 1998.

Pour une vision critique des comparaisons internationales

- H. GOLDSTEIN, « *Résultats scolaires : interprétation des comparaisons internationales* », Etudes et documents d'éducation, 63, Ed by UNESCO, Paris, 1995.

- F. GUERIN-PACE., A. BLUM, « *L'illusion comparative : Les logiques d'élaboration et d'utilisation d'une enquête internationale sur l'illettrisme* », *Population*, 54 (2), 271-302, INED, Paris, 1999.

- F. GUERIN-PACE., A. BLUM, « *Des lettres et des chiffres* », Fayard, 2000.