HAUT CONSEIL DE L'EVALUATION DE L'ECOLE

La France et les évaluations internationales

Norberto BOTTANI
Directeur du SRED
Département de l'instruction publique
du Canton de Genève

Pierre VRIGNAUD
Maître de Conférences, HDR
Université Paris 10, Nanterre

N° 16
Janvier 2005

Rapport établi à la demande du
Haut Conseil de l'évaluation de l'école

Rapport établi à la demande
du Haut Conseil de l'évaluation de l'école

Directeur de la publication : Christian FORESTIER

Secrétariat général : 3-5, bd Pasteur 75015 – PARIS
Tel : 01 55 55 77 41
Mèl : hcee@education.gouv.fr

ISSN en cours
Conception et impression : DEP/Bureau de l'édition

TABLE DES MATIERES

SYNTHESE ET RECOMMANDATIONS

L'utilité des enquêtes internationales sur les acquis des élèves apparaît incontestable. En effet, dans une situation où la mise en place d'expérimentations est coûteuse et souvent impossible à concevoir, la comparaison avec d'autres pays ayant des fonctionnements différents sur le plan de l'organisation des cursus, les contenus des programmes, les pédagogies, etc. permet, toutes choses égales par ailleurs, de recueillir des informations et de conduire des recherches scientifiques sur l'effet de ces différentes options éducatives.

1. La France a participé aux enquêtes internationales depuis le début de ces enquêtes il y a 40 ans. Il faut distinguer deux époques dont le tournant se situe au début des années 1990 Dans un premier temps, jusqu'à la fin des années 1980, cette participation a été erratique dans le sens qu'elle n'était pas le produit d'un plan stratégique cohérent soit scientifique soit politique, A partir des années 1990, cette participation a été pilotée de manière plus cohérente.
2. Jusqu'à la fin des années 1980, l'implication dans les enquêtes internationales (celles de l'IEA) a été régulière sans être systématique. L'intensité de l'implication dépendait des organismes concernés. Dans certains cas, les données ont été l'objet d'analyses secondaires et de publications scientifiques, dans d'autres, la participation semble s'être limitée à l'élaboration du matériel français et à la collecte des données. Décideurs et Spécialistes de l'éducation ont peu utilisé les résultats de ces enquêtes.
3. Le système d'enseignement français est resté isolé par rapport au mouvement international dans le domaine des études et des recherches internationales comparées sur vaste échelle. La France a plutôt subi les enquêtes ou elle a suivi le mouvement sans vraiment concourir à le forger. Elle n'a pas joué un rôle moteur, n'a pas su influencer ni la conception ni le déroulement ni l'exploitation de ces enquêtes.
4. Il y a dix ans le système d'enseignement français a changé de stratégie par rapport à la participation aux enquêtes internationales sur les compétences et les acquis des élèves. Ce changement est lié à la prise de responsabilité de la DEP dans le pilotage de ces enquêtes en France. L'activité de ce service a permis une participation systématique aux différentes enquêtes. Depuis une décennie, la France a participé aux enquêtes internationales sur l'évaluation des acquis des élèves. On peut souligner la présence active de ses délégués en ce qui concerne l'élaboration des outils des enquêtes. Cette participation s'accompagne d'une position souvent critique voire sceptique vis-à-vis des orientations suivies. La France a participé à PISA1 et 2 de l'OCDE, et aux enquêtes TIMMS et PIRLS de l'IEA. Le cas de l'enquête IALS (enquête auprès des ménages sur les compétences en littéracie des adultes) est particulier comme nous l'avons souligné à plusieurs reprises. En ce qui concerne le pilotage international de ces enquêtes, la participation française aux travaux se limite à la présence de quelques experts. En effet, pour avoir une visibilité scientifique et politique dans le domaine des enquêtes internationales sur vaste échelle il est indispensable de disposer d'une masse critique importante de spécialistes dans le domaine des recherches psychométriques et de l'évaluation systémique, ainsi qu'une organisation administrative diversifiée et efficace pour programmer et piloter des interventions cohérentes au niveau de l'espace éducatif international. Or, comme nous l'avons souligné la recherche française dans les approches quantitatives de l'évaluation pédagogique et psychologique ainsi que dans les domaines psychométriques et édumétriques est réduite à quelques laboratoires et peu encouragée par les instances de pilotage et d'évaluation de la recherche scientifique française.

5. La France tente, par l'intermédiaire du RERPESE (Réseau Européen des Responsables des Politiques d'Evaluation des Systèmes Educatifs), d'impulser une présence européenne dans ce domaine en répondant aux appels d'offre internationaux pour le pilotage de ces enquêtes et en lançant des enquêtes comparatives sur les compétences des élèves entre pays européens intéressés.
6. De manière générale, les milieux de l'enseignement ignorent ces enquêtes ou les connaissent mal. La plupart du temps, leur connaissance des résultats se réduit à quelques idées sur le classement de la France et leurs interrogations portent sur la comparabilité des résultats. Il faut surtout déplorer l'absence d'intérêt et de connaissance de ces travaux et de leur utilité pour le pilotage du système éducatif d'une partie des instances décisionnelles du MEN, en particulier, de l'Inspection Générale et de la direction de l'enseignement scolaire.
7. La France a compris tardivement l'importance de ces enquêtes. En privilégiant les évaluations domestiques, elle en a sous-estimé l'importance aussi bien en tant que référentiel d'évaluation que comme instrument de formation d'une masse critique de chercheurs dans le domaine de la psychométrie.

A la lumière de ces observations, on peut affirmer que la participation aux études comparées internationales sur les compétences et les acquis des élèves et des adultes constituent des domaines de développement, de recherche et d'analyse incontournables. Une présence active, c'est-à-dire constructive et critique à la fois, dans les programmes internationaux d'évaluations des systèmes d'enseignement doit être programmée et soutenue d'une manière systématique. Ces programmes fournissent des termes de comparaisons qui sont tout d'abord un complément indispensable des opérations d'évaluation menées par la DEP sur le plan national ; ils se révèlent un levier efficace de l'action politique. La France a été jusqu'ici en partie protégée vis-à-vis des effets rétroactifs au niveau politique des résultats produits par ces enquêtes à la différence de ce qui s'est passé par exemple en Allemagne, au Royaume-Uni, aux Etats-Unis, en Suisse, en Espagne, probablement grâce à la politique d'évaluation de l'enseignement qu'elle a su développer et qui a déjà été examinée par le Hcéé. Cependant, plusieurs avis du Hcéé ont aussi mis en évidence les faiblesses et les incohérences de cette politique relativement unique par rapport à la plupart des pays occidentaux. En particulier, l'exploitation des résultats de ces études mérite, comme il a été plusieurs fois souligné, une plus grande attention.

Les enquêtes internationales d'évaluation des systèmes d'enseignement dans le but de comparer les prestations des ces systèmes ont acquis au cours de ces dernières quarante années une rigueur et une solidité méthodologique considérables. Il n'en reste pas moins que certains choix méthodologiques restent problématiques. En particulier, l'approche unidimensionnelle, reposant sur un score principal, peut être critiquée car elle réduit les comparaisons à un classement et, par ailleurs, ce caractère unidimensionnel des données ne semble pas complètement défendable. Il faut également souligner les questions liées à la signification des variables mesurées comme « les compétences pour vivre et travailler dans le monde moderne ». Ces variables semblent être davantage justifiées par leurs qualités psychométriques plutôt que par des construits théoriques en relation avec les champs scientifiques pertinents (psychologie, sociologie, économie, sciences de l'éducation).

Par ailleurs, ces enquêtes produisent une mine d'informations sur le système d'enseignement unique dont l'intérêt principal consiste dans la possibilité d'effectuer des comparaisons multiples sur des paramètres exerçant une influence vérifiée sur les

prestations des systèmes d'enseignement. Cependant, en disant ceci il ne faut pas supposer que des progrès et des modifications des plans d'enquêtes et des instruments mis au point au niveau international ne soient plus nécessaires. Au contraire, rien n'est acquis dans ce domaine, des limites importantes ont été signalées. La science des comparaisons internationales sur les acquis et compétences des élèves et des adultes est une science relativement jeune. Nous ne sommes encore, dans un certain sens, qu'au début d'une histoire scientifique et de ce fait des améliorations et des perfectionnements sont certainement à l'ordre du jour et il convient que la recherche française dans ce domaine soit présente dans ce processus.

L'éducation nationale en France doit prendre la décision de s'engager fermement dans ce secteur et de se doter des moyens pour le faire, moyens qui pour le moment manquent ou sont insuffisants. A ce propos, nous recommandons les points suivants :

1. Clarifier les responsabilités au sein de l'administration.
2. Mettre sur pied un dispositif étoffé spécialisé dans la programmation, le suivi et l'exploitation des enquêtes internationales sur l'évaluation des systèmes d'enseignement.
3. Promouvoir le développement des savoir faire dans le domaine de la psychométrie (rattraper les retards et former une masse critiques de scientifiques dans ce domaine).
4. Elaborer un plan cohérent de développement des analyses et des recherches comparées des résultats des enquêtes internationales de masse sur les acquis des élèves articulé avec la politique national d'évaluation de l'école.
5. Rendre compte et rendre visibles les résultats et la nature des études internationales auprès des enseignants.
6. Développer le partenariat européen :
    - Création d'un pôle européen de traitement des données en mesure de participer aux appels d'offre internationaux
    - Consolidation du financement à l'échelle européenne par l'inscription de ces études dans le programme cadre européen de promotion de la recherche.

CHAPITRE I - HISTORIQUE DES ENQUETES INTERNATIONALES SUR LES ACQUIS ET LES COMPETENCES DES ELEVES

1 - UNE HISTOIRE DE CINQUANTE ANS[1]

En 1952, l'UNESCO a créé à Hambourg l'Institut international de l'éducation qui existe encore aujourd'hui. C'est à partir de cette date que s'amorce l'évaluation des systèmes d'enseignement par le recours à de grandes enquêtes de masse sur les compétences et les acquis des élèves. Depuis lors, pendant une cinquantaine d'années, ont été progressivement mis au point les connaissances statistiques et méthodologiques nécessaires pour concevoir, planifier, organiser et développer des enquêtes sur des échantillons représentatifs de populations d'élèves, les méthodes pour construire les échantillons avec des marges d'erreurs réduites pour ne pas mettre en doute leur valeur représentative, les techniques pour permettre le déroulement d'enquêtes comparables sur le plan international, les modèles à appliquer dans l'analyse et le traitement des données ainsi que les formes de présentation des résultats. Un effort considérable a été accompli et a mobilisé une partie importante de la communauté scientifique travaillant dans le domaine de l'éducation.

Le début de cette histoire se situe au sein de l'Institut de l'éducation de Hambourg où, une fois par an, au cours des années cinquante, se tenaient des rencontres d'une durée d'une semaine qui réunissaient les figures les plus éminentes de l'époque de la recherche en éducation. Chaque année, ces rencontres étaient l'occasion de confronter les priorités dans le domaine de la recherche en éducation, d'échanger des avis et des opinions sur les méthodologies de travail et d'aborder des thèmes de recherche d'intérêt commun. Au départ donc, c'est au sein de la communauté de recherche que fut posé le problème de la comparaison des résultats des systèmes d'enseignement. On verra par la suite que ces intérêts rencontreront les préoccupations des décideurs politiques et seront repris à leur compte par les responsables des systèmes d'enseignement.

Entre 1952 et 1960, toute une série de questions de recherche ont été abordées lors des rencontres annuelles de Hambourg. C'est dans ce contexte qu'on s'est interrogé pour la première fois sur l'opportunité et les modalités permettant d'évaluer les systèmes d'enseignement en mesurant les acquis des élèves.

En 1954, le thème à l'ordre du jour fut celui des redoublements ; en 1955, logiquement, on s'intéressa aux tests pour évaluer les élèves ; en 1956, on fit un pas de plus vers la problématique de l'évaluation de masse et on traita des programmes d'évaluation ; en 1957, on aborda la question des méthodes et des outils d'évaluation pour mesurer dans quelle proportion on atteignait les objectifs des programmes d'enseignement ; enfin, en 1958 fut posé le problème de l'organisation d'évaluations comparées des acquis des élèves fréquentant des systèmes d'enseignement différents. En passant en revue la séquence des sujets traités lors de ces rencontres, on s'aperçoit que les responsables de la recherche en

[1] Plusieurs articles et publications ont été dédiés à l'histoire de l'IEA. Par exemple Husén T., Postlethwaite N. (1996) : A Brief History of the International Association for the Evaluation of Educational Achievement (IEA), in : Assessment in Education, vol. 3, No. 2, pp. 129-141.

éducation de l'époque ont progressivement mis sur rails le thème central des débats de la décennie suivante, l'évaluation comparée de masse. Lors de la rencontre de 1958, présidée par William Douglas Wall, directeur de la National Foundation of Educational Research (NFER) en Angleterre et dans le Pays de Galles, la décision fut prise de tester au moyen d'une enquête pilote la faisabilité d'une évaluation comparée entre les systèmes d'enseignement de différents pays. Cette décision a représenté une rupture conceptuelle par rapport à la manière avec laquelle, jusqu'à ce moment-là, on appréciait l'efficacité d'un système d'enseignement. Les données qu'on prenait en compte étaient des données simples, comme par exemple le pourcentage des taux de participation aux différents niveaux d'instruction, les taux de scolarisation au-delà de l'enseignement obligatoire, la prolongation de la durée de la scolarité obligatoire. Les responsables de la recherche en éducation ont estimé qu'on ne pouvait plus se borner à ces données pour évaluer les prestations d'un système et qu'il fallait prendre en compte les résultats scolaires. Pour ce faire, il était cependant nécessaire d'inventer des instruments appropriés, bien plus élaborés que ceux utilisés par les économistes.

Benjamin Bloom de l'Université de Chicago, un des participants aux meetings de 1958 et 1959, pendant lesquels fut discutée et décidée la réalisation de l'étude pilote, se chargea de rédiger la proposition décrivant le type d'étude qu'il aurait été nécessaire d'effectuer pour examiner s'il était possible de réaliser une enquête internationale dans ce domaine. Parmi les autres figures influentes qui eurent un rôle déterminant dans la clarification et l'identification de ce projet, on trouve Arnold Anderson, directeur du Centre d'éducation comparée de l'Université de Chicago, Bob Thorndike du Teachers College de la Columbia University de New York, Douglas Pidgeon de la NFER, David Walker qui était à l'époque directeur du Conseil écossais pour la recherche en éducation (SCRE), Fernand Hotyat pour la Belgique, Gaston Mialaret pour la France, Walter Schukze pour l'Allemagne, Torsten Husén pour la Suède, Wellesley Foshay et Harry Passow du Teachers College également. Cette liste comporte grosso modo toutes les figures les plus importantes de la recherche en éducation de l'époque. Malgré le fait que les anglophones étaient majoritairement représentés, les francophones n'étaient pas absents comme le montre la présence de Mialeret (voir annexe) et de Hotyat.

En 1959, le projet initial de Bloom fut adopté après révision et réécriture par le comité directeur de l'institut de l'éducation de l'UNESCO de Hambourg. Les objectifs de l'étude pilote y étaient définis de la manière suivante :

(1) mieux comprendre le fonctionnement intellectuel, en se servant du test à choix multiple construit pour permettre de dégager des tendances communes dans les réponses d'élèves de plusieurs pays ;
(2) découvrir les possibilités et les difficultés de réalisation d'une étude internationale de masse sur les acquis des élèves.

1.1 - UN CONTEXTE INTERNATIONAL FAVORABLE

Pour mieux comprendre comment s'est développée l'exigence d'effectuer des analyses comparées de systèmes différents, il est opportun de rappeler le contexte politique de l'époque. Le 14 octobre 1957, l'Union soviétique avait réussi à mettre en orbite le premier satellite artificiel de la Terre, le Spoutnik I. Cette opération spectaculaire marqua l'opinion publique occidentale. Le choc fut surtout énorme aux Etats-Unis où la population et les

autorités eurent l'impression que le pays n'était plus à la pointe du développement technologique et scientifique et qu'il pouvait perdre la guerre froide avec l'Union soviétique. La réaction fut donc immédiate et violente. Elle déclencha entre autres un vaste programme de recherches scientifiques et plusieurs initiatives d'amélioration de l'enseignement scientifique au niveau secondaire, accusé d'être peu exigeant et mauvais. Ce facteur fut certainement déterminant dans l'intérêt des Etats-Unis pour la mise en œuvre d'un programme international d'évaluation des systèmes d'enseignement permettant de comparer sur une base objective les niveaux d'instruction des élèves.

1.2 - LA PREMIERE ETUDE PILOTE

La direction de l'étude pilote fut attribuée à Arthur Wellesley Foshay du Teachers College de New York. La préparation et la réalisation de l'étude s'étalèrent sur deux ans (1959-1961). Douze pays y participèrent, avec un échantillon non représentatif de 1000 élèves ayant 13 ans révolus au début de l'année scolaire, indépendamment de la classe fréquentée. Au total donc, 12'000 élèves furent concernés par ce test, ce qui à l'époque était une population considérable, compte tenu des moyens techniques existant pour le traitement des données[2]. Deux problèmes s'imposèrent d'emblée : celui des critères de définition de la population concernée (l'âge ou le degré d'école) et celui de la construction d'un échantillon représentatif. Lors de ce premier exercice, les domaines testés furent la compréhension de la lecture, les mathématiques, les sciences et la géographie. Par ailleurs, un test non verbal, mis au point en Angleterre, fut également soumis à tous les élèves. Aucune limitation de temps ne fut imposée pour la passation du test : les élèves pouvaient prendre tout le temps qu'il leur fallait pour répondre aux questions. Par ailleurs, on avait prévu également un mini-questionnaire pour les étudiants et les directeurs d'établissements. Les données recueillies en 1961 furent traitées et analysées par Bob Thorndike et les résultats publiés en 1962.

Un des buts de l'étude pilote était de tester la possibilité de construire des instruments utilisables d'une manière uniforme dans différents systèmes d'enseignement (comme pouvaient l'être par exemple les systèmes suédois et japonais) tout en parvenant à recueillir des données comparables entre elles et susceptibles d'être traitées de manière homogène. Beaucoup d'éléments, qui par la suite se retrouveront régulièrement dans toutes les enquêtes internationales, ont été testés et mis au point dans cette première étude pilote.

On retrouve dans cette étude pilote, sous des formes embryonnaires, presque toutes les composantes des futures enquêtes. Un des éléments qui a par contre disparu est le test non-verbal. Il faudra attendre presque quarante ans pour voir inclure dans le jeu d'épreuves de l'enquête PISA effectuée dans certains pays, comme l'Allemagne ou la Suisse, des tests d'aptitude cognitive qui ne mobilisent pas des compétences verbales ou linguistiques et qui sont donc moins susceptibles d'être influencés par les connaissances scolaires ou les différences socio-culturelles.

Les résultats de l'étude pilote ont été tellement concluants que le groupe des promoteurs décida immédiatement de poursuivre l'expérience et de réaliser une étude à grande échelle, permettant de tester ce qui avait été appris à l'école à un moment déterminé. En tirant les leçons de l'étude pilote, on s'accorda tout d'abord à prêter une grande attention à la

[2] Les techniques d'évaluation comparée ont été développées au cours des années trente aux Etats-Unis, comme par exemple lors de l'enquête « Eight-year Study » (Huit années d'études) qui s'est déroulée entre 1936 et 1941 sous le pilotage de Ralph Tyler (voir « International Encyclopedia of Educational Evaluation », Pergamon Press, 1990, 170-171).

construction de l'échantillon de population dont la qualité apparut déterminante pour la crédibilité des comparaisons[3]. Par ailleurs, le choix d'une seule discipline s'imposa comme une évidence, si l'on voulait faire passer des tests plus longs pour parvenir à une meilleure connaissance des acquis des élèves. Ces décisions ont été prises par les directeurs et les responsables des instituts de recherche en éducation et non par des bureaucraties gouvernementales ou des décideurs ayant un mandat politique.

Benjamin Bloom, de l'Université de Chicago et Arnold Anderson, directeur du Centre d'éducation comparée de la même université, furent chargés de soumettre une demande de subvention à l'Office de l'éducation des Etats-Unis (US Office of Education) afin de trouver un financement susceptible de couvrir les coûts de la partie internationale de l'étude ainsi que ceux de la participation des Etats-Unis. Cette demande ayant été acceptée, il fut possible d'engager un coordinateur à plein temps, l'Anglais Neville Postlethwaite[4]. Douze pays participèrent à cette première étude sur les connaissances en mathématiques, qui fut par la suite appelée la « première étude internationale sur les mathématiques » (FIMS/First International Mathematic Study).

2 - LA NAISSANCE DE L'IEA ET LA MISE EN PLACE D'UN MONOPOLE DE LA COMMUNAUTE DE RECHERCHE SUR L'EVALUATION

A la conclusion de l'étude pilote, il était devenu clair que des opérations semblables ne pouvaient être menées que par une organisation spécialisée dans la conduite de projets de grande envergure, rassemblant des expertises multiples pour pouvoir coordonner la participation de plusieurs instituts, planifier les travaux, assurer le financement et la gestion, coordonner le calendrier des initiatives, superviser la comparabilité des démarches, vérifier l'uniformité des outils, du traitement des données et des calculs, et pour finir, assurer une analyse cohérente des résultats. C'est ainsi qu'en 1961 fut constituée l'IEA (International Association for the Evaluation of Educational Achievement) dont le siège fut établi auprès de l'Institut International de l'éducation de Hambourg. Le premier président de l'IEA fut William Douglas Wall, qui était aussi le président de la NFER (National Foundation for Educational Research) anglaise. Très rapidement, en 1962, une première crise éclata entre Wall et le directeur de l'Institut International de l'éducation de Hambourg, l'Israélien Saul B. Robinson. Après seulement une année d'activité, Wall donna sa démission et fut remplacé par le Suédois Torsten Husén, qui dirigea par la suite l'IEA pendant plus de quinze ans, jusqu'à fin 1978. Dans le domaine de l'évaluation comparée, l'IEA a occupé une place prédominante. C'est grâce à elle qu'on est arrivé à produire des comparaisons crédibles entre différents systèmes d'enseignement, défi qui n'était pas gagné d'avance.

Il faut reconnaître à Husén un rôle déterminant dans le développement de l'IEA[5]. Si pendant presque un quart de siècle, l'IEA a réussi à jouer un rôle primordial sur la scène internationale en ce qui concerne la conduite d'études d'évaluation comparée, l'avancement de la recherche dans le domaine des évaluations de masse et l'établissement d'un équilibre harmonieux entre recherche scientifique et politique de l'éducation, cela doit être attribué au

[3] Pour parvenir à obtenir des échantillons probabilistes valables, l'Anglais Gilbert Peaker mit au point un manuel d'échantillonnage, devenu depuis un texte de référence pour toutes les études successives de l'IEA.
[4] Président de l'IEA entre 1978 et 1986 et coordinateur de la recherche IEA sur la compréhension de la lecture en 1990-1991.
[5] Voir par exemple le livre-interview de Arild Tjeldvall (2000) : Torsten Husén. Conversations in Comparative Education. Phi Delta Kappa. Educational Foundation, Bloomington.

flair et aux qualités de Husén qui a su conduire de main de maître une organisation dont le succès n'était pas assuré d'emblée. En naviguant entre plusieurs obstacles, Husén a réussi à préserver d'un côté l'indépendance du travail scientifique et de l'autre à satisfaire les attentes des responsables politiques des systèmes d'enseignement. Husén a été toujours conscient du fait que les enquêtes de l'IEA ne pouvaient pas se justifier uniquement sur la base d'un intérêt scientifique, mais qu'elles devaient fournir aux décideurs des informations utiles pour orienter les politiques de l'éducation. L'IEA, en effet, n'opérait pas sur un terrain vierge et ne pouvait pas échapper aux clins d'œil ou aux influences des milieux politiques. Tout en essayant de préserver sa neutralité, les études de l'IEA, par le fait même qu'elles étaient centrées sur les résultats de l'enseignement, concernaient la volonté de vérité des dispositifs de pouvoir et pouvaient faire évoluer les systèmes d'enseignement. Dans certains cas, les études de l'IEA ont indéniablement influencé la politique, quoiqu'en général il ne soit pas aisé de démontrer la présence d'une corrélation directe entre ces études et le changement en éducation.

A côté de Torsten Husén, une deuxième personnalité a joué un rôle important dans cette phase de développement et d'expansion de l'IEA : Neville Postlethwaite, recruté par Husén en 1962 grâce à la contribution extraordinaire de l'Office de l'Education des Etats-Unis (USOE), comme coordinateur à plein temps de la première étude. A l'instar de Husén, Postlethwaite marqua de son empreinte l'IEA. Tout naturellement, pourrait-on dire, il a été désigné en 1978 pour remplacer Husén à la tête de l'Association. Le tandem Husén-Postelthwaite a constitué le noyau autour duquel s'est structurée l'IEA et se sont formés des dizaines de spécialistes qui ont contribué à la réalisation des enquêtes de l'IEA, pratiquement jusqu'à la fin du XXe siècle. Ce n'est qu'en 1986 que Postlethwaite quitta à son tour la présidence de l'Association et fut remplacé par le Hollandais Tjeerd Plomb, qui appartenait à une génération différente de celle des pionniers de l'IEA.

Le mérite scientifique majeur de l'IEA a été celui de mettre au point des instruments rendant possible la plus grande comparabilité des données. Des efforts et une attention considérables ont été prêtés à la traduction des tests et des questionnaires, à la constitution d'échantillons non seulement représentatifs de la population scolaire, mais également et surtout comparables entre eux, et à la mise au point des méthodologies d'analyse des items sensibles aux variations observées. De ce fait, l'IEA a réussi à constituer des bases de données d'une grande richesse d'informations sur l'école et à stimuler le travail de recherche dans plusieurs pays, en exploitant les opportunités représentées par la possibilité de comparer des données non seulement en terme de moyennes brutes, mais aussi et surtout en termes de dispersions et de variation entre groupes d'élèves d'un même pays. Beaucoup de critiques ont été adressées aux études de l'IEA, en leur imputant notamment des objectifs qu'elles n'avaient pas, comme par exemple ceux d'établir des explications causales des différences entre systèmes d'enseignement. Or, très vraisemblablement, encore maintenant, il est prétentieux de se fixer un objectif de ce type, car d'autres études et d'autres données seraient nécessaires pour fournir des explications causales. Elles ne suffisent pas pour donner des explications, mais elles sont irremplaçables pour ouvrir des pistes de réflexions et stimuler des échanges susceptibles de mettre en évidence des failles et des points critiques dans les systèmes d'enseignement. Ces remarques ne doivent cependant pas amener à tirer une conclusion erronée qui considérerait les études de l'IEA comme insignifiantes. L'effort réalisé pour développer des tests sur les acquis des élèves et des questionnaires sur leur environnement familial et leurs stratégies d'apprentissage, des instruments pour appréhender la proportion des curriculums réellement enseignés ainsi que le temps réel pendant lequel les élèves en classe se trouvaient dans des conditions

d'apprentissage et d'enseignement, a été considérable et le matériel rassemblé a une valeur inestimable. Ces études ont sans doute stimulé le progrès scientifique et rendu possible la réalisation au niveau mondial de recherches sur les systèmes d'enseignement, qui ont modifié la représentation et la compréhension de leurs fonctions et modalités de fonctionnement. Le rôle et la place des études de l'IEA en ce qui concerne l'évaluation de l'enseignement est de ce fait incontournable. Or, certains pays ont mieux profité de cette aventure que d'autres. Aussi bien au niveau des autorités politiques qu'à celui des instituts de recherche, dans certains pays on a compris qu'il fallait prendre le train en marche et coopérer sans aucune réserve au développement de ces opérations. Tel n'a pas toujours été le choix adopté en France, aussi bien dans le secteur de la recherche en éducation que dans celui de la politique de l'enseignement.

3 - LA FIN DU MONOPOLE DE L'IEA

L'IEA a été pendant une trentaine d'année la seule organisation scientifique mondiale spécialisée dans le domaine de la réalisation d'enquêtes internationales de masse sur les acquis des élèves. Ce monopole a été brisé en 1988. La responsabilité de cet acte incombe à l'administration fédérale américaine qui, ayant été pendant des décennies le principal pourvoyeur de fonds de l'IEA, décida soudainement de mandater à l'Educational Testing Service de Princeton (ETS) la réalisation d'une enquête internationale sur les acquis des élèves concurrentielle à celle de l'IEA. Cette enquête, désignée comme IAEP1 (International Assessment of Educational Progress), a porté sur deux disciplines, les mathématiques et les sciences[6]. Dupliquée en 1991, elle a représenté une tentative d'innovation des études internationales à plusieurs égards originale, mais aussi contestable. La France[7] a participé ä la deuxième en 1991. Ces deux enquêtes ont été critiquées sur le plan scientifique et ont donné lieu à un large débat qui a abouti à l'abandon de l'entreprise de la part de l'administration américaine, laquelle renonça à en assurer le financement. La contestation de ces deux enquêtes portait en particulier sur la transposition au niveau international des items des tests du NAEP (National Assessment of Educational Progress)[8], c'est-à-dire de l'évaluation américaine des acquis des élèves réalisée par l'ETS. Avec l'IAEP on mesurait les compétences et les acquis des élèves à l'aune de critères américains, ce qui permettait d'éviter de longues discussions sur l'adéquation des tests par rapport aux curriculums nationaux. Ces enquêtes ont cependant constitué une avancée méthodologique considérable car elles ont utilisé, pour la première fois au niveau international, la méthode de la théorie des réponses aux items (Items Response Theory) qui permettait d'organiser la passation des tests en généralisant la valeur prédictive des résultats sans obliger tous les élèves à passer la totalité de l'épreuve[9].

[6] La première enquête IAEP en 1988 a concerné cinq pays (Corée, Espagne, Irlande, Royaume-Uni, Etats-Unis).

[7] Les responsables de l'organisation de la participation de la France à l'enquête IAEP2 ont été : Pierre Jouvanceau (Ministère Education Nationale) et Martine Le Guen (Ministère Education Nationale-DEP).

[8] Le NAEP est le programme d'évaluation mis sur pied aux Etats-Unis en 1964 pour observer et mesurer les progrès des élèves américains de quatre groupes d'âge. La première enquête sur un échantillon national a été effectuée en 1969. Le budget initial du NAEP était d'environ 2 millions de dollars américains. Après 2001, avec la réforme scolaire proposée par le président Bush, le NAEP est devenu une pièce centrale du programme d'amélioration de l'enseignement. Son budget a atteint et dépassé les 100 millions de dollars. Sans aucun doute, le NAEP a été un laboratoire formidable de rénovation et d'expérimentation de nouvelles approches en matière d'évaluation. Voir : Jones, Lyle V. et Olkin, I. (2004) : The Nation's Report Card. Evolution and Perspectives. Phi Delta Kappa Educational Foundation, Bloomington.

[9] La première étude de l'IEA à utiliser l'IRT a été celle sur la compréhension de la lecture (Reading literacy) de 1991, pilotée par Neville Postlethwaite.

L'opération IAEP a considérablement affaibli l'IEA. Celle-ci a réagi et relevé le défi méthodologique et organisationnel de l'ETS en réalisant, au cours des années quatre-vingt-dix, trois grandes enquêtes : l'enquête sur la compréhension de la lecture en 1990-1991, la troisième enquête internationale sur les mathématiques et les sciences en 1994-1995 (TIMSS), qui a été à ce moment-là la plus grande enquête conduite sur le plan international, avec la participation d'une quarantaine de systèmes d'enseignement différents, et enfin l'enquête sur l'éducation civique en 1999. Avec ces trois enquêtes, l'IEA a démontré qu'elle avait les ressources et l'énergie pour relever le défi organisationnel de l'ETS, mener des enquêtes innovantes et de grande envergure et traiter les données dans des délais relativement courts, mais cette réaction a été probablement tardive.

En effet, la sortie de scène de l'ETS n'a pas été une victoire pour l'IEA, car la place laissée libre par l'ETS fut occupée par un autre concurrent, l'OCDE, autrement plus armée et dotée de plus de moyens de conviction et de ressources que l'ETS. Dès 1993, l'OCDE a commencé à réfléchir sur l'organisation d'une enquête internationale sur les acquis des élèves, pour obtenir une source régulière de données sur les résultats de l'enseignement capable d'alimenter l'ensemble d'indicateurs internationaux que l'OCDE avait commencé à produire en 1992. Les tentatives de mettre en œuvre une collaboration avec l'IEA, absorbée à cette époque par la réalisation de TIMSS, ont échoué pour des raisons multiples qu'il faudra encore clarifier. En 1997, les ministères de l'éducation des pays de l'OCDE décidèrent donc de lancer un cycle d'enquêtes autonome d'un nouveau genre sur les acquis des élèves, indépendant de l'IEA. La préparation de l'opération débuta en 1998. Celle-ci fut menée au pas de charge puisqu'au printemps 2000, une trentaine de pays participèrent à la première étude du Programme International sur les Acquis des élèves (PISA).

L'OCDE a ainsi obtenu ce que l'IEA n'avait jamais réussi à réaliser pendant quarante ans, c'est-à-dire non seulement à attirer l'attention des décideurs sur ces enquêtes mais à réorienter les politiques de l'enseignement dans plusieurs pays, PISA étant devenu un référentiel pour justifier toutes sortes de décisions ou de réformes (voir OECD : What Makes School Systems Perform ? Seeing School Systems Through the Prism of PISA. Paris, 2004). On peut presque affirmer que les enquêtes du PISA ont sensiblement modifié le paysage des politiques de l'enseignement au niveau mondial.

4 - LES NOUVEAUTES DU PISA

Il serait erroné d'affirmer qu'il n'y a pas de continuité entre PISA et les enquêtes antérieures de l'IEA. PISA, tout en reprenant plusieurs acquis méthodologiques mis au point par l'IEA, présente des éléments innovateurs qui rompent avec les enquêtes précédentes. Tout d'abord, l'OCDE a réussi à imposer le principe d'une enquête régulière sur une base périodique avec une récurrence régulière. Les raisons du choix de la périodicité triennale ne sont pas claires, mais on peut supposer que le but de l'OCDE était de démontrer non seulement qu'il était possible d'organiser des enquêtes à grande échelle à des intervalles rapprochés comme l'avait jadis tenté de faire l'ETS, et qu'elle avait aussi la capacité de le faire. Une périodicité de trois ans permet de tenir les pays sous pression dans un certain sens et de forcer une communauté scientifique plutôt récalcitrante à tenir un rythme soutenu

de production et d'analyse de données[10]. Ce faisant, l'OCDE savait pertinemment qu'elle aurait l'appui des décideurs politiques qui, eux, au contraire, tenaient à disposer de données récentes permettant d'apprécier l'état du système d'enseignement presque instantanément[11]. On peut émettre l'hypothèse qu'un tempo aussi élevé était le prix à payer pour obtenir le financement des gouvernements.

Le principe de la répétition d'une enquête dans le même champ disciplinaire n'est pas en soi une nouveauté. L'IEA avait aussi mis en œuvre un programme d'enquêtes successives de mathématiques et dans le domaine de la lecture, car l'intérêt de pouvoir procéder à des comparaisons diachroniques des résultats obtenus pour un même groupe d'âge à l'intérieur d'un même système d'enseignement (mais à des dates différentes) est évident pour tout le monde. Cependant, avec PISA, c'est la première fois qu'on lance un programme fixant d'emblée la périodicité des tests. Les bénéfices de la participation à PISA pourront donc être engrangés seulement si un système d'enseignement s'engage tout au long du cycle d'enquête. Selon le plan d'enquête, ce ne sera qu'après neuf ans que seront testées à nouveau avec une même profondeur analytique les compétences dans la compréhension de la lecture, domaine principal testé au printemps 2000. De ce fait, l'OCDE devrait – le conditionnel est obligatoire – fidéliser les pays sur une longue période. L'IEA, avec l'enquête sur la lecture « Reading Literacy », qui a eu lieu en 1991 et qui a été répétée en 2001, semble avoir aussi adopté un rythme décennal, qui permet d'analyser les résultats des réformes sur le long terme.

La troisième nouveauté du programme PISA est le choix de trois domaines testés simultanément, quoique dans un ordre d'importance différent. Les trois domaines pris en compte par l'OCDE sont la compréhension de la lecture, la culture mathématique et la culture scientifique. A chaque fois, les trois domaines font l'objet d'un test, mais la place réservée dans le test à chacun de ces domaines est différente. Dans le programme PISA, on parle de domaines majeurs et de domaines mineurs. Lors de chaque enquête, l'une des matières sera évaluée en détail, son étude représentant près des deux tiers de la durée totale des tests[12]. Le temps total qui sera consacré aux tests pour chaque élève est de deux heures, mais la collecte d'information se fondera sur une batterie d'items correspondant à près de sept heures de tests. Par ailleurs, comme c'était déjà le cas dans les enquêtes IEA, les élèves passeront également 20 minutes à répondre à un questionnaire contextuel sur leur motivation, leur environnement familial et culturel, leurs fréquentations, leurs intérêts.

Une autre originalité du programme consiste dans la composition de la population testée. L'OCDE a décidé de tester les élèves de 15 ans indépendamment de la classe qu'ils fréquentent. Cela signifie que dans l'échantillon représentatif de la population d'élèves de 15 ans fréquentant un système d'enseignement se retrouveront à côté des élèves de 2e et de 3e, fréquentant des degrés différents, aussi bien des élèves de 5e ou de 4e ou dans certains cas des élèves de première[13]. Etant donné le recours important au redoublement pratiqué au

[10] On peut s'interroger sur la pertinence de ce rythme, non seulement à cause des coûts qu'il impose mais aussi de l'intérêt de comparaisons à aussi faibles distances, à moins que PISA ne remplace, dans les systèmes d'enseignement qui ne l'ont pas, un dispositif propre d'évaluation.

[11] Il est intéressant de noter que dans certains pays, au moment de la présentation des résultats du PISA 2000 et PISA 2003, on est allé vérifier quelles réformes de l'enseignement les élèves testés avaient vécu. Par exemple, au moment de la présentation des résultats du PISA 2003, le président de la Conférence des chefs de département de l'éducation des cantons helvétiques, M. Stöckling, a déclaré que le positionnement relativement bon des élèves suisses était la preuve de l'efficacité des réformes de l'enseignement entreprises dans les années 90.

[12] L'organisation des tests avec des carnets tournants est expliquée dans la deuxième partie.

[13] La population modale de 15 ans au moment du test peut comprendre aussi bien des élèves de seconde que de 3e selon le mois de naissance. Lors de l'enquête PISA 2003, 49,6% des élèves de l'échantillon français étaient en seconde et 34,5% étaient dans une classe de 3e.

sein du système d'enseignement français, cela signifie que dans l'échantillon français on trouvera une proportion plus ou moins importante d'élèves de 15 ans qui ne sont pas en troisième, la classe théorique où se trouve la majorité des élèves de 15 ans ayant eu une scolarité régulière, mais plutôt en quatrième. Ces élèves passeront ainsi le test sans avoir parcouru toute la scolarité obligatoire, ce qui représente un désavantage pour l'échantillon français comparé à l'échantillon des élèves de 15 ans représentatifs d'autres systèmes d'enseignement dans lesquels on ne pratique pas, ou presque pas, le redoublement. L'option de préférer le groupe d'âge comme critère de composition de l'échantillon testé, au lieu du degré, présente une autre particularité qui est en contradiction, soit dit en passant, avec un des objectifs affichés par PISA. En effet, le choix du groupe d'âge de 15 ans est justifié par le fait d'évaluer les compétences en fin de scolarité obligatoire, en estimant que la plupart des jeunes gens des pays de l'OCDE complètent leur scolarité obligatoire à l'âge de 15 ans. Or, cette supposition n'est pas totalement correcte. En effet, si on voulait vérifier le niveau des compétences atteint à l'intérieur d'un système d'enseignement déterminé par l'ensemble des élèves achevant leur scolarité obligatoire, il aurait fallu choisir non pas le groupe d'âge pour constituer la population testée, mais le dernier degré de l'enseignement obligatoire, c'est-à-dire la troisième en France. Cette option aurait impliqué la constitution d'un échantillon d'élèves composé d'individus ayant un âge différent car, comme on peut le constater dans le système d'enseignement français, en troisième se retrouvent des élèves qui ont en majorité 15 ans, mais aussi des élèves de 16, voire 17 ans, et quelques élèves plus jeunes de 14 ou peut-être même 13 ans. Cette option, statistiquement plus complexe à gérer, n'a pas été choisie par l'OCDE, mais a été adoptée par quelques pays comme par exemple la Suisse, qui a participé avec un double échantillon : l'échantillon officiel des élèves de 15 ans imposé par l'OCDE et un deuxième échantillon optionnel composé d'élèves de la dernière année de l'école obligatoire, indépendamment de leur âge.

Une spécificité ultérieure de PISA est représentée par le fait que le test ne vise pas à évaluer l'acquisition des connaissances fixées dans les programmes scolaires, mais les compétences ou aptitudes jugées indispensables pour mener une existence autonome et indépendante dans des sociétés démocratiques avec une économie de marché, comme le sont les sociétés des pays membres de l'OCDE. Ceci signifie que PISA se démarque nettement des enquêtes de l'IEA, lesquelles avaient toujours été fabriquées pour tester le degré de maîtrise des programmes d'enseignement à un âge déterminé ou dans un degré spécifique, ce qui obligeait les responsables des enquêtes de l'IEA à identifier un soi disant « méta-programme » représentant un dénominateur commun des programmes d'enseignement officiels des différents systèmes d'enseignement participant aux enquêtes. Pour cette raison, l'IEA a dû prévoir des analyses comparées des programmes d'enseignement qui étaient indispensables, soit pour fabriquer des tests susceptibles d'être acceptés par les responsables des différents systèmes d'enseignement, soit pour interpréter les résultats. La recherche d'un dénominateur commun entre les programmes de plusieurs dizaines de systèmes d'enseignement est une opération compliquée, comme l'a montré la mise au point du test de l'enquête TIMSS en 1994 à laquelle ont participé 45 systèmes d'enseignement. Lors de l'interprétation des résultats, les scores insatisfaisants obtenus par les élèves d'un système d'enseignement dans le calcul algébrique ou la géométrie spatiale pouvaient être expliqués par le fait que ces chapitres n'avaient pas été traités à l'école ou n'étaient même pas inclus dans le programme d'enseignement prévu pour des élèves de 13 ans. De ce fait, l'IEA a régulièrement été confrontée au problème de la définition du champ à évaluer afin de placer les élèves participant au test, mais suivant des programmes d'enseignement différents, dans des conditions comparables. Pour résoudre cette question, on pouvait prendre deux options entre lesquelles l'IEA n'a jamais clairement tranché : d'un

côté, se concentrer uniquement sur les contenus communs des différents programmes d'enseignement et de l'autre, au contraire, mélanger tous les programmes et construire des épreuves avec des éléments extraits de cet ensemble. Ces deux options ont chacune leurs inconvénients. La première restreint considérablement l'ampleur de l'éventail des thèmes qui auraient pu être abordés dans le test, tandis que la deuxième présente des risques de décalage entre test et enseignements scolaires. Par ailleurs, il ne faudrait pas non plus sous-estimer les difficultés considérables pour bâtir un test axé sur une option ou sur une autre. En effet, il n'est pas du tout évident de parvenir à analyser une trentaine, voire une quarantaine de programmes d'enseignement de mathématiques de quatrième pour en dégager les thèmes communs. En général, le champ couvert par les évaluations internationales est négocié à chaque fois entre systèmes d'enseignement avec l'espoir de parvenir à un compromis acceptable pour tous les participants. Lors de la première enquête internationale sur les mathématiques (FIMS) réalisée en 1964, l'IEA avait entrepris des analyses approfondies des programmes de mathématiques en mettant au point une grille permettant de mettre en relation les différents arguments du programme avec les comportements cognitifs relatifs. Après avoir complété pour chaque programme d'enseignement ces grilles, conçues pour saisir en détail les objectifs d'apprentissage, on a mis au point une grille internationale comprenant les éléments communs pour les douze systèmes d'enseignement participant à l'enquête. Le produit final, avant d'être utilisé pour la construction des épreuves, a été validé par des comités nationaux de révision qui avaient la responsabilité de vérifier l'adéquation entre les objectifs d'apprentissage de la grille internationale et les programmes d'enseignement nationaux.

L'analyse des programmes d'enseignement a reçu une attention particulière dans le cas de l'enquête TIMSS qui s'est déroulée en 1994-1995. A cette occasion, un groupe spécifique de travail dirigé par William H. Schmidt de l'Université de Michigan a été constitué pour analyser méthodiquement les programmes d'enseignement officiels et les manuels de mathématiques en usage dans les systèmes d'enseignement participant à l'étude. Le projet SMSO (Survey of Mathematics and Science Opportunities) a permis de constituer une banque de données sur les programmes de mathématiques et les manuels scolaires utilisés en classe lors de l'enseignement des mathématiques[14].

Cette analyse des programmes n'a concerné qu'un très petit nombre de systèmes d'enseignement et, surtout, elle s'est déroulée parallèlement à l'enquête TIMSS et ne l'a pas précédée comme l'aurait voulu la logique. Cependant, le matériel rassemblé par le projet SMSO a permis d'effectuer des interprétations poussées des scores de TIMSS. Il n'empêche qu'au sein de l'IEA, le problème de l'adéquation entre programmes d'enseignement et épreuves internationales a constitué une pierre d'achoppement permanente.

Pour ces raisons, les entreprises concurrentes de l'IEA dans le champ des évaluations internationales des compétences ont suivi un autre chemin. Les deux enquêtes de l'ETS connues sous les acronymes IAEP1 et 2 (International Assessment of Educational Progress), réalisées en 1988 et 1991, ont été conçues pour évaluer les connaissances dans deux disciplines, les mathématiques et les sciences. Or, pour neutraliser le lien entre épreuves et programmes nationaux d'enseignement, l'ETS a adopté et partiellement adapté les items du test du programme d'évaluation fédéral américain NAEP (National Assessment of Educational Progress). Ce faisant, le problème aigu de la concordance entre test

[14] Voir Schmidt H.W. et autres: Characterising Pedagogical Flow : An Investigation of Mathematics and Science Teaching in Six Countries. Dordrecht, Kluwer Academic Publishers, 1996.

international et programmes d'enseignement était résolu en ayant comme unique référentiel le programme d'enseignement en vigueur dans un seul des systèmes d'enseignement. La prise en compte des spécificités des programmes d'enseignement en vigueur dans différents pays n'est entrée en ligne de compte qu'au moment de l'interprétation des résultats, en pondérant les scores en fonction des programmes.

L'OCDE, lorsqu'elle a pris la décision de réaliser des évaluations à grande échelle au niveau international à un rythme très soutenu, a été aussi confrontée au problème posé par les différences existant entre les programmes d'enseignement. Pour procéder rapidement dans la réalisation de l'enquête et sa mise en œuvre, il fallait donc au préalable régler cette question. Etant donné que la piste suivie par l'ETS s'était révélée plus que problématique et avait soulevé un tollé sur le plan international, l'autre solution était celle de s'affranchir totalement, lors de la construction des tests, de tous les programmes d'enseignement nationaux. L'OCDE a suivi cette voie consistant à ne plus mesurer ce que les élèves apprennent à l'école, mais à tester le niveau de compétences qu'on considère indispensable à 15 ans pour pouvoir vivre dans des sociétés démocratiques et à économie de marché. De ce fait, les tests PISA ont été conçus comme des tests « curriculum free », c'est-à-dire des tests neutres par rapport aux programmes d'enseignement. Le test PISA vise à évaluer la présence ou l'absence d'aptitudes jugées « essentielles » pour une vie d'adulte[15]. Selon les propres termes de l'OCDE, « il s'agit là de l'apport le plus important et le plus ambitieux du programme OCDE/PISA »[16].

Ce choix n'a pas été uniquement dicté par une exigence pragmatique (ne pas perdre du temps dans la construction des tests avec la recherche des concordances entre les programmes d'enseignement compte tenu des délais serrés de réalisation), mais est aussi l'apogée d'un processus enclenché dans l'ensemble des systèmes d'enseignement du monde occidental il y a longtemps, caractérisé par l'abandon progressif d'une organisation des programmes d'enseignement autour de champs disciplinaires rigoureusement distincts pour adopter une organisation des programmes d'enseignement articulée autour de compétences et de savoirs génériques[17]. Ce n'est donc pas un hasard si les outils élaborés pour réaliser PISA ont été construits en fonction du concept de littératie, c'est-à-dire de compétences génériques remplaçant la série d'informations et techniques structurant le cadre de chaque matière scolaire. Le programme PISA évalue la compréhension de l'écrit, la culture mathématique et la culture scientifique d'un étudiant de 15 ans par rapport au niveau de compétence en lecture, en mathématiques et en sciences identifiées comme nécessaires pour pouvoir comprendre le monde dans lequel les jeunes se développent et grandissent. Le programme OCDE/PISA ne mesure donc pas le degré de maîtrise des connaissances

[15] OCDE : Mesurer les connaissances et compétences des élèves. Un nouveau cadre d'évaluation. Paris. 1999. Dans les documents de l'OCDE on mentionne un nouveau référentiel : les compétences nécessaires pour réussir sa vie (« for a successful life »), qui est plus vague que les connaissances ou compétences fixées dans les programmes d'enseignement.

[16] Par la suite, l'OCDE a sponsorisé un projet financé par les Etats-Unis et la Confédération helvétique dont le but était de définir des composantes permettant de créer à l'intérieur des comptes nationaux un panier de biens éducatifs qui aurait dû permettre de mettre au point un index PPA (parité pouvoir d'achat) pour l'éducation, index dont le besoin s'est fait sentir en produisant les indicateurs financiers de « Regards sur l'éducation » qui sont comparables seulement en pondérant les valeurs nationales en fonction du coût de la vie existant dans les différents pays. Or, l'index PPA produit et utilisé par l'OCDE est établi sur la base du calcul du coût d'un panier de biens essentiels dans lequel il n'y a pas de biens relatifs aux dépenses d'éducation. C'est pour identifier ces biens éducatifs que le projet DESECO (Définition et sélection des compétences clés) a été lancé. Par la suite, le projet est devenu un enjeu pour l'élaboration des tests des enquêtes sur la littératie des adultes (projet ALL) et des élèves de 15 ans (projet PISA). Voir Salganik L., Rychen D. (ed.) (2001): « Defining and Selecting Key Competencies », Hogrefe & Huber Publ., Seattle; Rychen D. Hersh Salganik L.(ed.)(2003): « Key Competencies for a Successful Life and a Well-Functioning Society », Hogrefe & Huber Publ., Seattle. Voir annexe.
Dans le site du projet on peut trouver tous les documents in extenso: http://www.portal-stat.admin.ch/deseco/deseco_doc_strategique.pdf

[17] Voir Bernstein B., 1971, 1975, 1977, 1990. Class, Codes and Control. Vol. I-IV, Routledge & Kegan Paul, London.

scolaires prévues dans les programmes d'enseignement et qu'on suppose enseignées à l'école, mais évalue l'acquisition de notions et aptitudes à caractère général qui exploitent les connaissances éventuellement apprises à l'école.

Le programme PISA s'affranchit des limites imposées par la nécessité de trouver un dénominateur commun constitué par le contenu spécifique de l'enseignement dispensé dans les écoles des pays participants. Cette approche de l'évaluation en termes de maîtrise de grands concepts est justifiée par l'OCDE par l'exigence de valoriser le capital humain défini comme « les connaissances, qualifications, compétences et autres qualités possédées par un individu et intéressant le bien-être personnel, social et économique ». L'ambition de l'OCDE est de parvenir à mesurer le stock de capital humain d'un pays autrement que par de variables telles que le niveau d'étude atteint. Le recours à cette justification pour légitimer la disjonction entre tests et programmes d'enseignement a parfaitement fonctionné, car la totalité des pays membres de l'OCDE a accepté de participer au programme PISA et de le financer sur la base de ces arguments. De ce point de vue, PISA marque un virage par rapport à toutes les enquêtes antérieures à grande échelle. Il consacre le fait que l'évaluation des connaissances et des savoirs scolaires n'est plus en soi suffisante et nécessaire pour évaluer les prestations d'un système d'enseignement. Le programme PISA transmet un tout autre message, c'est-à-dire que l'efficacité des systèmes d'enseignement ne se mesure pas par rapport à l'enseignement et à l'acquisition de connaissances, mais plutôt par des évaluations qui tiennent compte d'autres qualités et d'autres compétences dont la justification est formulée en référence aux théories du développement du capital humain. Il est trop tôt pour savoir si ce virage marquera un point de non-retour dans les pratiques des évaluations à grande échelle. Il faudra attendre encore quelques années pour parvenir à apprécier correctement quelles seront les retombées sur l'organisation et le fonctionnement des systèmes d'enseignement d'un dispositif d'évaluation qui prend ses distances avec l'appareil scolaire.

Une autre spécificité du programme OCDE/PISA consiste dans son origine. Le programme d'enquête de l'IEA, lu à posteriori, semble construit logiquement, mais sa cohérence est plus fictive que réelle. Le programme de l'IEA est en effet le résultat de choix plutôt aléatoires ; il n'est pas le produit d'un projet cohérent prévoyant une succession d'enquêtes complémentaires ou ancrées l'une dans l'autre pour permettre d'effectuer des comparaisons dans le temps. Par ailleurs, la décision de planifier une enquête et de la réaliser à toujours été, au sein de l'IEA, fortement conditionnée par de pressants impératifs financiers. L'organisation, ne disposant de sources de financement assurées, a été obligée de tout temps d'accepter de réaliser les enquêtes pour lesquelles le financement était garanti. De ce fait, les financeurs ont passablement conditionné le programme de l'IEA. Le lien entre les enjeux du financement et les impératifs de la recherche dictant de programmer des enquêtes différenciées a constitué un enjeu récurrent dans les démarches de l'IEA. Selon les sources de financement, les projets devaient être modifiés pour tenir compte des exigences ou des objectifs des financeurs. Par ailleurs, faute de financement assuré et suffisant, l'IEA a été aussi obligée de refuser de mettre en chantier des enquêtes parfaitement justifiées d'un point de vue scientifique. Deux cas méritent d'être signalés à ce propos : l'échec des propositions d'évaluation des compétences linguistiques dans une deuxième langue, déterminé par le désintérêt des financeurs les plus puissants, comme par exemple le gouvernement des Etats-Unis, qui n'avaient aucun intérêt à réaliser une étude comparée de ce type ; le projet d'évaluation des compétences en écriture, réalisé une seule fois en 1984-85 et qui n'a plus été répété faute de financement. Au contraire, l'évaluation des compétences ou des connaissances en mathématique est récurrente, non seulement à cause de la relative facilité

de la construction des épreuves de mathématique, mais aussi parce qu'on a toujours trouvé les fonds nécessaires pour réaliser des enquêtes sur les niveaux de compétence en mathématique, en particulier aux Etats-Unis où l'objectif de s'imposer comme le premier système d'enseignement au monde dans le domaine des mathématiques a été plusieurs fois souligné et affirmé par les Présidents successifs de ce pays. La répétition de l'enquête TIMSS en 1999 et en 2003 a été pratiquement imposée par les autorités américaines qui n'ont pas lésiné sur les moyens pour aider toute une série de pays à participer à ces répétitions programmées pour vérifier l'évolution du niveau de compétences des étudiants américains en mathématique et en sciences par rapport aux résultats atteints par les étudiants d'autres pays. L'IEA a dans ce cas reçu les ressources financières nécessaires pour planifier, organiser et conduire au niveau international ces évaluations, malgré le fait qu'à peu près dans les mêmes années l'OCDE avait commencé le programme PISA dont l'un des domaines d'évaluation était la culture mathématique des élèves à 15 ans. Bien sûr, le champ et la population des élèves de l'enquête de l'IEA ne sont pas les mêmes que ceux de l'enquête PISA, mais néanmoins on ne peut pas éviter de s'interroger sur la raison d'être d'une mobilisation de la communauté éducative internationale autour des connaissances en mathématiques à travers la réalisation d'évaluations comparées apparemment concurrentes.

On pourrait par ailleurs s'interroger sur les raisons profondes justifiant la récurrence des évaluations sur les compétences en lecture. Ce domaine a été testé trois fois par l'IEA, la première en 1971, la deuxième en 1991 et la troisième en 2001. Une quatrième évaluation est prévue en 2005. Le programme PISA a démarré en 2000 son cycle d'évaluation avec une enquête également centrée sur les compétences en lecture. Nous devons certainement supposer que ces coïncidences et ces périodicités ne sont pas le fruit du hasard ni non plus de facilités de nature financière. Cependant, dans le cadre de ce rapport, nous ne pouvons pas approfondir les hypothèses explicatives d'un semblable florilège d'enquêtes sur la lecture. L'aspect qui nous intéresse ici est de nature politique et nous amène à chercher une liaison entre configuration du programme d'évaluation et les facteurs le rendant possible, comme par exemple les conditions de financement ou la mise en œuvre de dispositifs de contrôle ou de techniques du gouvernement des écoles.

La recherche de financeurs garantissant la possibilité de mener des évaluations internationales à grande échelle a amené l'IEA à ouvrir progressivement la porte aux gouvernements, c'est-à-dire aux responsables des systèmes d'enseignement publics qui seuls pouvaient garantir des contributions substantielles pour permettre le démarrage d'évaluations comparées à grande échelle. Cette évolution a transformée l'IEA : au lieu de rester une organisation autonome de recherche, indépendante et déterminant elle-même les objectifs des évaluations, les outils d'évaluation, les rythmes des enquêtes ainsi que les modalités de divulgation des résultats, elle est devenue une organisation hybride de plus en plus infiltrée par les représentants des gouvernements sachant utiliser la politique de l'éducation comme un champ diplomatique d'influence sur le plan international ou ayant les moyens suffisants pour conditionner la réalisation du programme d'évaluation. A la fin des années 80, il était devenu évident qu'il n'était pas possible de réaliser une évaluation internationale sans négocier sa configuration avec les autorités des systèmes d'enseignement les plus puissants du monde, afin d'obtenir les ressources nécessaires pour la réaliser. Or, une des conditions principales posées par les responsables politiques était le respect de délais serrés entre le moment de la passation des tests et celui de la présentation des résultats. A cet égard, l'enquête réalisée par l'IEA en 1991 sur les compétences en lecture a représenté un véritable virage, car l'enquête fut financée par l'administration américaine et par d'autres gouvernements à la condition que les résultats aient été

disponibles deux ans après la passation des épreuves. C'est d'ailleurs la même exigence qui a incité l'administration fédérale américaine à financer l'opération IAEP de l'ETS visant justement à vérifier s'il était possible de tenir un rythme soutenu d'évaluation comparée à grande échelle au niveau international. Nous pouvons estimer que la question du financement et des délais de réalisation des enquêtes ont été l'épée de Damoclès qui a conditionné les réactions et les choix de la communauté scientifique internationale engagée dans la réalisation des évaluations à grande échelle des systèmes d'enseignement. La planification et la réalisation de l'enquête TIMSS en 1994-95 ont été possibles grâce à l'ingérence prévisible des responsables politiques de plusieurs pays et notamment des Etats-Unis qui ont posé des conditions strictes de réalisation en échange de sommes importantes rendant possible la réalisation du programme d'enquête.

Cette ambiguïté a été levée avec l'entrée en action, plus ou moins à la même période, de l'OCDE dans le domaine de l'évaluation des compétences et connaissances des élèves. Lorsque, en automne 1997, les gouvernements de l'OCDE ont décidé de lancer PISA et de répartir entre eux les coûts de son financement, une nouvelle époque dans l'histoire des évaluations internationales à grande échelle s'ouvre. Pour la première fois, en effet, les gouvernements assument directement la responsabilité de programmer des évaluations à grande échelle, pour obtenir des informations estimées nécessaires pour réguler les systèmes d'enseignement et non plus seulement pour des objectifs de connaissance proposés par la communauté scientifique. Sous-jacent à cette décision, on trouve l'enjeu des relations entre recherche en éducation et politique de l'éducation et d'une manière plus spécifique celui de l'autonomie de la recherche scientifique par rapport aux visées des décideurs et responsables des systèmes d'enseignement. En finançant PISA, les autorités scolaires responsables des systèmes d'enseignement espéraient non seulement obtenir rapidement des données les orientant sur les effets des politiques éducatives, mais surtout contrôler l'efficacité des systèmes, obtenir des informations sur les modalités de régulation de l'enseignement et réguler les systèmes avec ce programme. On peut dire, après la réalisation du premier cycle d'enquête PISA en 2000, que ce pari a été amplement gagné, si on se borne simplement à relever les retombées de l'enquête PISA sur les politiques nationales d'éducation de plusieurs pays[18].

Enfin, une dernière particularité du PISA est constituée par la possibilité laissée aux responsables des systèmes d'enseignement d'ajouter au programme de base des tests ou des questionnaires optionnels. Lors de la première volée PISA en 2000, deux questionnaires optionnels ont été proposés aux pays, l'un sur les compétences transdisciplinaires et l'autre sur les compétences informatiques. Le premier de ces deux questionnaires a été préparé au sein du projet INES de l'OCDE-CERI par un réseau de scientifiques qui s'interrogeaient sur les produits de l'éducation scolaire. Après environ sept années de travail, ce groupe est parvenu à faufiler un questionnaire dans le paquet d'instruments proposés par le programme PISA. Une démarche analogue a débouché sur l'élaboration du questionnaire sur les compétences en informatique. Vingt-six pays sur 32 ont utilisé les questionnaires sur les compétences transdisciplinaires et 20 sur 32 celui sur l'utilisation des ordinateurs, la disponibilité d'ordinateurs dans les écoles et à la maison, l'appréciation que les étudiants donnaient de leurs compétences en informatique. A ce propos, il serait aussi intéressant de relever les modules qui n'ont pas abouti, c'est-à-dire les questionnaires qui ont été refusés,

[18] Voir par exemple en France les rapports rédigés pour le Haut Conseil de l'évaluation de l'école, comme le rapport « L'appréciation des compétences des élèves et des jeunes en lecture et en écriture et l'évolution de ces compétences dans le temps » de Marie-Thérèse Ceard, Martine Remond, Michelle Varier (décembre 2003), ou le rapport de Michel Salines et Pierre Vrignaud « Apprécier et certifier les acquis des élèves en fin de collège : diplôme et évaluations-bilans » (juin 2001).

comme par exemple celui sur les compétences dans une deuxième langue. Cependant, il est ardu d'obtenir des informations sur les modules avortés et de ce fait de formuler une appréciation sur le programme complémentaire de PISA dont un des buts est de parcourir des champs de compétences peu explorés. Questionnaires et modules complémentaires, comme par exemple le questionnaire sur les stratégies d'apprentissages des élèves, ont été beaucoup exploités dans l'interprétation des résultats du test principal. Cette remarque nous permet de mettre en évidence une caractéristique du PISA qui le différencie passablement des enquêtes similaires menées par l'IEA : le programme OCDE/PISA est une démarche omnivore qui tend à englober toutes sortes de thèmes que l'IEA, au contraire, avait tendance à traiter séparément.

5 - LE CONTEXTE HISTORIQUE ET SOCIAL DES ENQUETES INTERNATIONALES SUR LES ACQUIS ET LES COMPETENCES DES ELEVES

La mise en œuvre, le développement et le succès grandissant des évaluations à grande échelle des acquis et des compétences des élèves au cours de ces dernières quarante années ne s'expliquent pas si l'on ne tient pas compte du contexte dans lequel ont été pensées, façonnées et mises en œuvre les politiques de l'éducation des pays de la zone de l'OCDE. On peut sans doute soutenir que la réalisation d'opérations d'envergure comme celle menée par l'IEA, l'ETS et l'OCDE dans des dizaines de systèmes d'enseignement pour en évaluer les prestations n'a été rendue possible que par la présence de conditions favorables qui ont labouré le terrain sur lequel ces évaluations ont passé. Trois ingrédients nous paraissent avoir joué à cet égard un rôle déterminant :

- l'expansion des systèmes éducatifs dans la zone de l'OCDE ;
- l'uniformisation des modèles éducatifs qui a réduit les divergences entre systèmes d'enseignement en les alignant les uns sur les autres ;
- la dichotomie grandissante entre objectifs scientifiques d'une part et objectifs politiques d'autre part dans le domaine de l'éducation.

5.1 - L'EXPANSION DES SYSTEMES EDUCATIFS DANS LA ZONE DE L'OCDE

Au moment du lancement des premières enquêtes de l'IEA et de la constitution de l'IEA, les pays européens venaient juste de sortir de la Deuxième Guerre mondiale et commençaient une phase de croissance économique après la dure période de reconstruction industrielle et politique des pays ravagés par la guerre. A la fin des années 50, les différences entre les systèmes d'enseignement des pays de l'OCDE étaient considérables et l'étendue de ces systèmes était encore relativement réduite. D'un côté, il fallait reconstituer ces systèmes et de l'autre engager une politique de l'éducation expansionniste dont le *leitmotiv* était le développement du capital humain nécessaire pour soutenir l'expansion économique et sociale. Dans plusieurs pays européens par ailleurs existait encore un analphabétisme de base diffus et dans tous les systèmes, la démocratisation de l'enseignement restait un objectif à réaliser. Quarante ans plus tard, à l'orée du 21ème siècle, le contexte social, économique et culturel de la zone de l'OCDE a totalement changé. On peut ainsi dire que le développement des programmes d'évaluation à grande échelle est associé aux changements subis par les systèmes d'enseignement dans une phase d'expansion sans

précédent et n'est que le reflet des exigences ressenties par les responsables des systèmes d'enseignement pour piloter la croissance de ces systèmes.

Celle-ci a été caractérisée essentiellement par quatre phénomènes[19] :

- l'augmentation des taux de scolarisation dans l'enseignement secondaire de deuxième cycle et dans toutes les formes d'enseignement tertiaire ;
- la prolongation de l'espérance de scolarisation ;
- l'augmentation du nombre de diplômés achevant l'enseignement secondaire de deuxième cycle et des diplômés de l'enseignement tertiaire ;
- la différenciation de l'offre éducative à l'intérieur de l'enseignement secondaire de deuxième cycle et de l'enseignement tertiaire.

Ces quatre phénomènes ont déterminé des bouleversements dans le dispositif de financement de l'éducation, mais à cet égard on peut aussi faire remarquer que l'augmentation des coûts de l'éducation n'a pas été aussi rapide que celle des paramètres que nous venons de signaler. En effet, si on prend en compte comme indicateur la part de dépenses de l'éducation par rapport au produit intérieur brut (PIB), on constate que les changements dans la distribution de la richesse nationale au bénéfice des politiques de l'éducation n'ont pas été aussi considérables que l'on aurait pu s'y attendre par rapport à l'ampleur des autres phénomènes déclenchés par l'expansion des systèmes d'enseignement[20]. En ce qui concerne cette expansion, il est intéressant de souligner le décalage existant entre la baisse des naissances dans les pays de l'OCDE d'un côté et l'augmentation des taux de scolarisation imputable à la prolongation de la scolarité de l'autre. Par ailleurs, il convient aussi d'attirer l'attention sur un autre phénomène contradictoire, à savoir l'augmentation du nombre du personnel enseignant malgré la diminution des effectifs des élèves. Ces deux phénomènes sont en soi susceptibles de justifier au niveau politique la pression en faveur d'une évaluation des résultats des systèmes d'enseignement, car on peut supposer que même si on n'a pas braqué les projecteurs sur ces deux situations, leurs effets indirects engendrent des questionnements à propos des investissements en éducation.

Il est aussi intéressant de relever que les changements significatifs des systèmes d'enseignement affectent essentiellement l'enseignement secondaire de deuxième cycle ou le secteur tertiaire, tandis que la plupart des évaluations à grande échelle réalisées entre-temps ne concernent que l'enseignement obligatoire.

Un des premiers indices de l'expansion de l'enseignement est constitué par la prolongation de la durée de la scolarité obligatoire. La diversification de l'offre éducative au niveau de l'enseignement secondaire de deuxième cycle n'est dans un certain sens qu'un corollaire de ce phénomène. Dans les pays de l'OCDE, l'âge de fin d'obligation scolaire était au début du 21ème siècle en moyenne établi à 16 ans[21]. Dans certains pays de l'OCDE la fin de l'obligation scolaire est à 18 ans (Allemagne, Belgique, Pays-Bas), dans d'autres à 14 ans (Corée, Portugal, Turquie). Aucun pays de la zone de l'OCDE n'a une durée d'obligation

[19] Voir Bottani N., Pegoraro R. (2004): La situation de l'enseignement secondaire dans les pays de l'OCDE. Document préparé à l'occasion du séminaire « L'enseignement secondaire à l'échelle mondiale : bilans et perspectives » organisé dans le cadre du BIE, Genève 5-7 septembre 2004 (sous presse).

[20] Voir par exemple l'indicateur B2 de l'ensemble d'indicateurs 2004 de l'enseignement produit par l'OCDE qui porte sur les dépenses au titre des établissements scolaires en pourcentage du PIB. Selon l'OCDE, dans 17 pays sur 18 pour lesquels des données sont disponibles, les dépenses publiques et privées au titre des établissements scolaires ont augmenté de plus de 5% entre 1995 et 2001, mais la croissance des dépenses d'éducation n'a pas suivi celle de la richesse nationale, contrairement à ce qui avait pu être observé au début des années 90.

[21] Ces données sont fournies dans la série *Regards sur l'éducation. Les indicateurs de l'OCDE*, Paris.

scolaire inférieure à l'âge de 14 ans, mais aucun non plus ne dépasse l'âge de 18 ans. On peut donc se demander si l'âge de 18 ans est une limite vers laquelle tendrait l'ensemble des systèmes d'enseignement des pays de l'OCDE. Il est difficile de répondre à cette question ; cependant, si nous prenons en compte l'âge de fin d'obligation scolaire pendant laquelle plus de 90% de la population est scolarisée, on peut alors supposer que tôt ou tard cet âge pourrait devenir l'âge standard de la fin de la scolarisation obligatoire. Or, les données statistiques disponibles indiquent que déjà maintenant il est difficile de scolariser plus de 90% de la population au-delà des 18 ans, tandis que dans les systèmes d'enseignement dans lesquels la fin de l'obligation scolaire est inférieure à 18 ans (par exemple à 16 ou 15 ans), la scolarisation dépasse la limite de 90% pendant les années qui sont situées au-delà de l'âge légal de fin d'obligation scolaire. Ceci nous amène à penser que pour le moment, dans la pratique plus que sur le plan juridique, dans les pays de l'OCDE, la scolarisation est presque universelle jusqu'à l'âge de 18 ans. Ces considérations devraient avoir une répercussion sur la structure des programmes des enquêtes d'évaluation à grande échelle qui visent à appréhender le niveau de connaissances et de compétences à la fin de la scolarisation formelle initiale. Ce n'est pas un hasard si l'OCDE a adopté l'âge de 15 ans pour son programme d'évaluation, tandis qu'il reste surprenant de constater que l'IEA continue à privilégier comme cible de ses évaluations la population des élèves de 13 ans[22]. On peut à cet égard peut-être justifier ce choix avec un argument d'opportunité : la focalisation sur la population de 13 ans permettait d'effectuer des comparaisons avec les enquêtes précédentes et de mesurer ainsi les effets des réformes de l'enseignement entreprises entre-temps ou le progrès réalisé entre deux évaluations successives, ce qui exige la prise en compte d'une même population. Cependant force est de noter que si les responsables scolaires, comme l'affirme l'OCDE, souhaitent disposer d'informations sur les résultats à la fin de la scolarisation, il serait plus qu'opportun d'envisager la mise en œuvre d'enquêtes à un âge plus élevé que l'âge pris en compte jusqu'ici, par exemple des évaluations sur les niveaux de compétences des étudiants ou des jeunes à 18 ans. Or, l'organisation d'évaluations de ce type, lorsque la population cible se trouve dispersée dans plusieurs établissements ou lorsqu'une partie de cette population ne fréquente même plus l'école, est manifestement plus difficile à réaliser et donc plus coûteuse, mais c'est dans cette direction qu'il faudra probablement s'orienter à l'avenir.

En résumé, en 2000, dans 25 pays de l'OCDE sur 27, la scolarisation institutionnelle durait en moyenne entre 15 et 20 ans et l'espérance de scolarisation avait augmenté dans 18 pays de l'OCDE sur 20 au cours de la période 1995-2000. Dans certains pays de l'OCDE (Australie, Corée, Finlande, Grèce, Hongrie, Pologne, République Tchèque, Royaume-Uni), en cinq ans l'espérance de scolarisation moyenne s'était accrue de plus d'un an. Ces indications permettent de dire que la poussée expansionniste reste toujours forte dans le domaine de l'éducation. C'est certainement pour cette raison que les autorités politiques de tous ces pays ont manifesté une attitude favorable vis-à-vis d'un programme d'évaluation systématique et rapide des résultats des systèmes d'enseignement, car il est vraisemblablement difficile au niveau politique de soutenir les requêtes d'un financement et d'un développement ultérieur des systèmes d'enseignement sans fournir la preuve que ces requêtes sont justifiées par l'amélioration des résultats et des performances des systèmes.

[22] Des raisons géopolitiques concourent probablement au choix de l'IEA. En effet, à l'IEA adhèrent plusieurs membres opérant dans les systèmes d'enseignement de pays en voie de développement, où l'obligation scolaire n'atteint pas 15 ans.

5.2 - L'UNIFORMISATION DES SYSTEMES D'ENSEIGNEMENT

Un autre grand processus qui s'est mis en place dans la deuxième moitié du 20ème siècle, après la fin de la Deuxième Guerre mondiale, a été celui de l'homogénéisation progressive des systèmes d'enseignement. Tous ces systèmes ont évolué dans une même direction à des vitesses différentes bien sûr, en adoptant néanmoins les traits d'un modèle dominant d'organisation de l'enseignement et des finalités de l'éducation. Les différences entre systèmes d'enseignement s'estompent tandis que les similitudes se renforcent. Les particularités restent prononcées en ce qui concerne les structures administratives tandis que les affinités s'imposent en ce qui concerne la pédagogie, l'évolution des programmes d'enseignement, les modalités de fonctionnement des établissements scolaires. On peut donc estimer que ce processus d'unification a offert un terrain favorable à l'implantation des évaluations à grande échelle sur le plan international.

L'unification des programmes d'enseignement est en soi un objet de recherche qui a été particulièrement travaillé par des auteurs comme Meyer, Ramirez, Benavot, Boli, Muller, Ringer, etc.[23]

Un projet en cours aux Etats-Unis sous les auspices de l'American Academy of Arts and Science a l'ambition de mettre en évidence les similitudes dans l'expansion rapide de l'enseignement au niveau mondial (projet UBASE - Universal Basic and Secondary Education)[24]. De même, Aaron Benavot, sociologue américain, qui analyse dans une perspective comparativiste les systèmes d'enseignement, est en train d'étudier comment ont évolué les programmes d'enseignement de l'école obligatoire entre 1980 et 2000 sur le plan mondial, en exploitant les archives des données du Bureau international de l'Education[25]. Ce même sujet a été exploré sur le plan européen, particulièrement par Antonio Novoa qui depuis plusieurs années analyse la restructuration de l'espace éducatif européen et met en évidence l'émergence de phénomènes transversaux qui sont communs à l'ensemble des systèmes d'enseignement du continent[26]. Ce n'est pas le lieu ici d'analyser les facteurs qui sont en train de façonner l'émergence de systèmes d'enseignement similaires ni d'esquisser les processus de transferts à l'œuvre à l'intérieur de la zone de l'OCDE, processus certainement favorisés par l'action plus ou moins coordonnée des organisations internationales telles que l'OCDE et l'Union Européenne, dont l'influence considérable sur les politiques de l'enseignement dans les différents pays est considérable.

Le rapprochement entre systèmes d'enseignement au cours de ces dernières décennies peut être lu comme une manifestation du processus de la mise en œuvre de politiques inspirées par la « new economy » des réformes de la gestion publique, des pressions en faveur de la réduction du rôle de l'Etat[27]. Si nous considérons le fait que les systèmes

[23] Ramirez Francisco O., Meyer John W., 2002: *National curricula: World models and national historical legacies*. Manuscrit non publié, Stanford, Département de sociologie de l'Université de Standford.

[24] Voir Bloom E. D., Cohen E.J. (2002): Education for ALL: An Unfinished Revolution. In: Daedalus, Summer 2002, pp. 84-94; American Academy of Arts and Sciences (2003): Project on Universal Basic and Secondary Education. The UBASE project: A Progress Report; Salganik H. L., Provasnik J. S. (2004): Defining Quality Education for Universal Basic and Secondary Education (UBASE). Document interne non publié.

[25] Benavot Aaron, 2004: *A Global Study of Intended Instructional Time and Official School Curricula, 1980-2000*. Manuscrit non publié, BIE-Genève.

[26] Antonio Novoa et Martin Lawn, 2002 : *Fabricating Europe*. Kluwer Academic Publishers; Antonio Novoa, 1998: *Histoire et comparaison (essai sur l'éducation)*. Lisbonne, Educa, 1998;

[27] La bibliographie à ce sujet est inépuisable. Voir par exemple les publications de Stephen Ball : « Big Policies/Small World : An Introduction to International Perspectives in Education Policy », in Comparative Education, 34(2)1998, pp. 130-199 ; « Education Reform : A Critical and Post-Structural Approach », Buckingham, Open University Press 1998; (avec A. Van

d'enseignement ont été un des outils les plus importants pour constituer, forger et renforcer l'Etat-Nation, on peut mieux interpréter les processus qui actuellement sont à l'œuvre pour atténuer les différences nationales dans l'enseignement et favoriser l'émergence d'un espace éducatif mondialisé. La crise de l'Etat-Nation se répercute au niveau éducatif en délégitimant toute une série d'institutions, de pratiques, de programmes conçus dans le but de produire l'identité nationale et de créer une communauté de citoyens autour d'un noyau de traditions et de valeurs communes. Le développement de la littérature nationale et son enseignement ont été de ce point de vue un des éléments constitutifs du processus éducatif. Cette fonction d'éducation de la société par la société attribuée à l'école est désormais entrée en crise comme le démontre par ailleurs la mise en parenthèse des programmes d'enseignement nationaux dans les enquêtes internationales. L'existence d'une sorte de discours mondialisé de l'éducation et donc d'une idéologie unificatrice de l'espace éducatif et des réformes des divers pays a été particulièrement étudiée par Jürgen Schriewer qui invite à avancer avec prudence dans ce domaine[28]. Le rapprochement entre systèmes d'enseignement, l'adoption de programmes d'enseignement de plus en plus similaires, la fin de la mission éducative de la société que l'Etat-Nation attribuait à l'enseignement, ont rendu possible la constitution d'un terrain commun favorable à l'implantation des évaluations internationales à grande échelle et ont été favorisés par ces évaluations. L'existence d'un discours international sur l'éducation a favorisé la constitution d'une communauté scientifique internationale qui a réussi à trouver un terrain d'entente pour réaliser des évaluations à grande échelle au niveau international. En même temps, cette évolution pousse à valoriser les spécificités éducatives nationales qui, telles des reliques, sont scrutées et évaluées pour en apprécier les influences marginales qu'elles auraient sur la productivité de l'éducation. Le dispositif d'évaluation français se place au cœur de cette ambivalence : bien installé depuis des décennies à l'intérieur de l'Education nationale, il s'articule relativement mal avec les comparaisons internationales opérées par les évaluations de masse qui représentent un autre référentiel déstabilisateur des pratiques et certitudes nationales.

5.3 - LA DICHOTOMIE ENTRE OBJECTIFS SCIENTIFIQUES ET OBJECTIFS POLITIQUES

Comme nous venons de le voir en abordant les nouveautés caractérisant le programme PISA, les tensions entre priorités politiques et préoccupations ou intérêts scientifiques dans le domaine de l'évaluation à grande échelle sont considérables. Le mérite, si ainsi on peut s'exprimer, du programme PISA a été celui de trancher dans le vif en donnant la préférence aux attentes et exigences des milieux politiques. La subordination de la recherche dans le domaine de l'éducation aux programmes politiques et aux besoins de connaissances des responsables des systèmes d'enseignement a pour le moment contribué à imposer le projet PISA sur la scène internationale. Lorsque les évaluations internationales à grande échelle étaient déterminées et programmées par l'IEA, les responsables politiques des systèmes d'enseignement n'y prêtaient pas, a vrai dire, beaucoup d'attention. Ces études, malgré leur qualité, n'alimentaient pas les débats politiques comme nous avons pu l'observer avec les enquêtes PISA. On pourrait même dire que probablement, à cause même de leur rigueur et de leurs préoccupations méthodologiques, ces évaluations ne suscitaient de l'intérêt qu'auprès des spécialistes du domaine et du champ de l'évaluation comparée sur le plan

Zanten) « Logiques de marché et ethniques contextualisées dans les systèmes français et britanniques » in *Education et société*, 1, 1998, pp. 47-71.

[28] Jürgen Schriewer, 2004 : L'internationalisation des discours sur l'éducation : adoption d'une « idéologie mondiale » ou persistance du style de « réflexion systémique » spécifiquement nationale ? dans *Revue française de pédagogie*, No 146, Janvier-Février-Mars 2004, 7-26.

international qui était en train de se constituer et de se structurer comme un champ scientifique propre.

Les divergences séparant les intérêts des milieux politiques et ceux du champ scientifique peuvent en partie être reconduites à des conceptions différentes de la nature et des finalités des évaluations. A cet égard, on peut distinguer grosso modo quatre types d'évaluation :

– l'évaluation pour l'apprentissage ;
– l'évaluation pour rendre compte (redevabilité) ;
– l'évaluation pour la classification ;
– l'évaluation pour la certification.

Les milieux politiques ont toujours vivement réagi face aux évaluations internationales lorsque celles-ci aboutissaient à un classement des systèmes d'enseignement. Cette information, qui n'est que la manifestation d'une certaine manière de traiter l'information, a été immédiatement exploitée par les responsables politiques ainsi que par les médias. Or, on ne peut pas affirmer que les évaluations internationales à grande échelle aient été conçues et menées dans le but d'établir un classement entre systèmes d'enseignement. Même dans le cas du programme PISA qui a été amplement exploité sur ce plan aussi bien par les milieux politiques que par les médias, on ne trouve dans les documents préparatoires du programme de justifications basées sur l'argument de la classification. Cependant, nous devons admettre que les concepteurs de ces évaluations et les institutions qui les ont réalisées n'ont pas été suffisamment prudents avec le traitement des données. Le recours systématique à l'analyse univariée lors de la présentation des résultats a contribué à valoriser le classement et à affaiblir l'analyse et l'interprétation. Il est regrettable qu'on n'ait pas essayé de mettre au point des formes de présentation des résultats différents du classement[29]. Dès les tous premiers volumes des indicateurs de l'enseignement publiés par l'OCDE (voir par exemple « Regard sur l'éducation » 1992, 1993 ou 1995), les scores des pays dans les évaluations internationales ont été présentés sous la forme d'une grille utilisant la méthode de Bonferroni qui consiste à ajuster la signification statistique des données. Par ailleurs, un effort systématique a été opéré pour fournir l'erreur type et l'écart type dans les tableaux des résultats. L'exigence de respecter les priorités des systèmes d'enseignement et d'éviter leur classement en fonction des résultats obtenus par les élèves dans les tests a amené à présenter des graphiques sans l'indication des noms des pays (voir par exemple le graphique A9.1, page 140, dans « Regards sur l'éducation 2004 »). Cependant, des progrès restent à réaliser dans le traitement et la présentation des données pour éviter d'induire en erreur les interprètes de ces résultats et surtout pour éviter de fournir aux responsables politiques des arguments servant de prétexte pour justifier ou valider des choix politiques ou des programmes de réformes de l'éducation. Si on analyse la manière avec laquelle, dans certains pays, on s'est servi des résultats PISA, on ne peut pas ne pas être alerté face à une tendance qui privilégie les classements à tout autre travail d'interprétation des données.

L'exploitation des évaluations internationales des acquis et des compétences des élèves sous la forme de classements n'est que la réponse du berger à la bergère, car les responsables politiques des systèmes d'enseignement ont accepté de financer ces opérations de grande envergure non pas seulement pour comprendre les origines des

[29] Le réseau A du projet INES du CERI/OCDE a travaillé sur cette question et fait des propositions alternatives. Voir : Moskowitz J., Garet M., Herman R., Stephens M., 1997: *Implementing the Data Strategy: a plan for the analysis and presentation of outcomes indicators*. Pelavia Research Institute, Washington. Doc. non publié.

différences constatées entre systèmes d'enseignement ou pour améliorer la connaissance aussi bien du fonctionnement des systèmes que des modalités d'apprentissage, mais pour connaître la productivité des systèmes dont ils ont la charge. Ce n'est pas un hasard si les résultats des évaluations internationales ont été d'emblée insérés dans les chapitres des indicateurs internationaux de l'enseignement réservé aux résultats. Le financement massif par les gouvernements de ces programmes d'évaluation de grandes dimensions a été alloué en contrepartie de l'engagement des milieux de la recherche scientifique sur l'évaluation et la comparaison des systèmes d'enseignement de fournir des informations sûres, exploitables et comparables des produits et des résultats des systèmes d'enseignement[30]. Le but prioritaire de PISA est d'acquérir des informations pour rendre compte de l'utilisation des investissements et des moyens mis à disposition des systèmes d'enseignement, ce qui correspond à une demande explicite des gouvernements.

Par ailleurs, depuis les toutes premières enquêtes de l'IEA jusqu'aux dernières de l'OCDE dans le cadre du programme PISA, le souci constant des chercheurs a été celui d'exploiter ces évaluations pour essayer de comprendre comment on parvenait aux résultats collectés et comment ces résultats étaient obtenus. Ces opérations d'évaluation ont obtenu l'adhésion des scientifiques car elles étaient appréhendées comme une occasion unique pour améliorer la connaissance du fonctionnement des systèmes d'enseignement, de l'organisation des établissements, des pratiques didactiques, et des stratégies d'apprentissage en classe. Ces enquêtes ont effectivement contribué à améliorer la base de connaissances sur les mécanismes d'éducation scolaire et à attirer l'attention sur des situations problématiques du point de vue des politiques de l'éducation, comme par exemple les différences de performance entre garçons et filles, les variations des attentes professionnelles, l'engagement à l'égard de l'école (voir par exemple les indicateurs A8 et A9 dans « Regards sur l'éducation », OCDE, 2004).

Plusieurs dimensions des évaluations internationales à grande échelle répondent à l'exigence d'évaluations pour l'apprentissage. Une évaluation pour l'apprentissage est une évaluation pour laquelle la première priorité prise en considération lors de sa conception et de la définition de son architecture ainsi que de sa réalisation est d'être au service de la promotion de l'apprentissage des élèves. Une évaluation peut aider à apprendre si elle fournit des informations que les enseignants et les étudiants peuvent utiliser pour juger de leurs prestations et pour modifier leur manière d'enseigner ou d'apprendre. Il n'est pas facile de se rapprocher de cet idéal surtout lorsqu'on organise des évaluations à grande échelle sur le plan international, car les bénéficiaires principaux de ces évaluations sont les gouvernements ou les ministères de l'éducation et non les écoles ou les enseignants pour qui ces évaluations se présentent souvent comme des corvées sans aucun intérêt immédiat ou sans aucune conséquence pour la vie quotidienne à l'intérieur des établissements. On peut ainsi se demander s'il n'y a pas deux évaluations antinomiques, selon les analyses de Ernest House : une évaluation hostile aux écoles et une évaluation qui leur est amicale. Selon House, il ne faut pas cultiver trop d'illusions : « les tests standardisés de résultats sont hautement sensibles aux influences politiques. Les différentes utilisations des données fournies par des tests sont en relation avec les demandes des gouvernements qui doivent justifier leurs actions »[31]. La mise en œuvre des programmes d'évaluation doit donc être interprétée et lue dans la perspective du contexte politique duquel émanent ces programmes. Selon House, aux Etats-Unis, les résultats de tests dans les années cinquante

[30] Voir par exemple l'appel d'offre international du programme PISA.
[31] House E. Aspects politiques des évaluations à grande échelle aux Etats-Unis. Dans : Politiques d'éducation et de formation, Analyses et comparaisons internationales n°11, 2004/2, pp. 94-101.

étaient utilisés pour informer les individus. L'économie de l'époque était florissante et le gouvernement bénéficiait d'une haute crédibilité. Le but des évaluations à l'époque était donc relativement simple : elles servaient à renseigner des individus sur leurs compétences, leurs capacités et à les aider à prendre conscience de leur potentialité de développement. Lorsque la configuration économique change et devient médiocre, la crédibilité du gouvernement, toujours selon House, s'affaiblit. Dans ces circonstances, les gouvernements éprouvent le besoin de redorer leur blason et de récupérer leur légitimité. Les programmes d'évaluation sont mis en œuvre pour atteindre cet objectif. Enfin, lorsque l'économie est en crise et que le gouvernement perd une bonne partie de sa crédibilité, on lance des programmes d'évaluation dans un double but : « faire peser le poids de l'échec sur ce qui était au bas du système éducatif et les en rendre responsables » ; « contrôler l'information ». Dans cette configuration, les tests deviennent des moyens de pression et de coercition. Les programmes d'évaluation ne sont alors plus des opérations pour l'apprentissage et au service des écoles, mais des opérations bureaucratiques pourrait-on dire, qui mettent la pression sur les écoles et les enseignants, les rendant responsables des échecs et les obligeant à rendre compte de leur action comme s'il s'agissait de les forcer à une confession publique. D'ailleurs, la publication des résultats bruts des tests par école en Angleterre est comparable à une opération de ce genre et a tout au moins un sens proche d'une confession imposée aux écoles, forcées à étaler leurs défauts ou leurs vertus.

Une évaluation amicale selon House est une évaluation qui devrait être « démocratique, fondée sur le dialogue et la discussion. Elle devrait être démocratique, c'est-à-dire prendre en considération les intérêts et les points de vue de toutes les parties prenantes. Elle devrait être fondée sur le dialogue, car c'est par le dialogue avec les parties prenantes qu'elles peuvent représenter de manière authentique leurs intérêts. Elle devrait être fondée sur les discussions pour permettre aux parties prenantes de s'assurer de leurs intérêts véritables et de parvenir à de sages décisions »[32].

6 - PUISSANCE ET ATTRAIT DE LA COMPARAISON

Les enquêtes internationales à grande échelle ont marqué un renouveau des études comparées en éducation et ont mis en évidence l'utilité qu'on pouvait tirer de ces études lorsqu'elles étaient organisées selon des critères scientifiques et non plus seulement sur la base de visites et d'impressions subjectives. Ce qui est en jeu ici est la méthode de la comparabilité et plus précisément la disposition à admettre qu'on apprend quelque chose des analyses comparées, qu'il est possible d'améliorer la connaissance de son propre système d'enseignement en le comparant avec d'autres, et enfin qu'on peut utiliser les connaissances dérivant des études comparées comme un levier pour faire évoluer les systèmes d'enseignement, pour lancer des réformes, pour mettre en œuvre et soutenir un processus politique de changement de l'enseignement. La comparaison est à la fois un instrument de connaissance et un levier du changement, car elle sert pour légitimer des réformes de l'enseignement, en apportant par exemple des preuves validées scientifiquement de l'utilité d'une démarche ou de la nécessité d'une réforme, contrecarrant les résistances d'ordre idéologiques.

[32] House E., 1998 : Les mécanismes institutionnels de l'évaluation. In : Perspectives. Revue trimestrielle d'éducation comparée, Bureau international d'éducation, Volume 28, n° 105, pp. 123-131.

Sans la foi dans l'utilité et dans la faisabilité des comparaisons entre systèmes d'enseignement, il n'est pas envisageable de programmer ou de participer à des évaluations à grande échelle et il n'est pas possible de tirer profit de ces opérations scientifiques. De ce fait, la compréhension de la nature des enquêtes internationales à grande échelle implique une discussion de la portée épistémologique des analyses comparées tout court et non seulement en éducation.

Parmi les pères fondateurs de la comparaison entre systèmes d'enseignement, nous devons rappeler ici la place de premier plan occupée par deux scientifiques français : Condorcet d'une part, à qui l'on doit la mise au point de la méthode des statistiques sociales, et Marc-Antoine Jullien, qui a proposé le premier prototype de questionnaire (comprenant 266 questions) pour recueillir des informations comparables sur les systèmes d'enseignement de son époque, celle de l'Empire et du début du 19^{e} siècle. C'est à Marc-Antoine Jullien qu'on attribue la première tentative de découpage d'un champ empirique d'observation en éducation et d'élaboration de techniques d'investigation et d'utilisation de modèles d'analyse formels[33]. Au tout début de la science comparée en éducation apparaît déjà une pratique, voire une interrogation, largement suivie par la suite par les responsables des systèmes d'enseignement, consistant à rechercher les meilleures pratiques, c'est-à-dire les solutions associées aux meilleurs résultats, pour pouvoir les copier ou les transposer. Il est à cet effet amusant de constater qu'environ 200 ans après les premiers pèlerinages pédagogiques pour visiter des systèmes d'enseignement performants proposés par Marc-Antoine Jullien, cette même pratique a été appliquée à grande échelle après la publication des résultats de la première enquête PISA en 2001. Dans les années suivantes, un nombre élevé de commissions et d'experts se sont rendus par exemple en Finlande ou au Canada, deux pays dans lesquels les élèves ont particulièrement bien réussi dans les tests du PISA, pour trouver les raisons d'un tel succès. Une manifestation de même nature a été le séminaire organisé à Rome en coïncidence avec la rentrée scolaire 2004-2005 sur la réussite de la Finlande dans l'enquête PISA 2001, sous le patronage du Ministère de l'Education nationale par une fondation privée (TRELLE).

Il n'y a aucun doute sur le fait que la formulation d'une science comparée des systèmes d'enseignement a toujours été fortement marquée par l'attrait des sciences expérimentales et l'approche positiviste, comme l'a montré de manière éloquente Jürgen Schriewer[34]. A cet égard, il est approprié de mentionner que le premier directeur du Centre d'éducation comparée de l'Université de Chicago, C.A. Anderson, qui a joué un rôle majeur dans la constitution de l'IEA et dans la réalisation des premières enquêtes à grande échelle de l'IEA, estimait que le but ultime de l'éducation comparée est la connaissance systématique de la causalité en éducation. Les comparatistes comme Anderson, auquel on peut associer d'autres pionniers de l'IEA, estimaient qu'il était possible et même nécessaire de parvenir à des lois de portée générale en partant des résultats des études comparées. Pour Anderson, l'éducation comparée devait utiliser le modèle des sciences naturelles marqué par la mise à l'épreuve des hypothèses et l'analyse de la co-variance[35]. Cet esprit a imprégné les travaux de l'IEA qui, dès le début, ont visé à réaliser des comparaisons scientifiques des pratiques d'enseignement dans les écoles de différents systèmes d'enseignement.

[33] Jullien Marc-Antoine, 1817 : Esquisse d'un ouvrage sur l'éducation comparée.
[34] Schriewer J. (2000) : *Discourse formation in comparative education*, Berne : P. Lang.
[35] L'analyse de la co-variance est une méthode qui permet d'éliminer d'une manière statistique les effets de variables qu'on ne veut pas examiner dans une étude. Ces variables sont appelées co-variées ou variables de contrôle. Avec une méthode statistique appropriée on peut éliminer les co-variées de la liste par les variables dépendantes qui concourent à expliquer la variance des résultats.

Avec les évaluations à grande échelle proposées par l'IEA, on a essayé pour la première fois de mesurer les différences internationales dans les acquis des élèves en utilisant des tests permettant de rendre compte des variations des résultats. On estimait que ce faisant, il aurait été possible d'acquérir une meilleure connaissance des relations existant entre input éducatif, processus scolaire et résultats, jusqu'à identifier les variables les plus malléables sur lesquelles pouvait agir, à travers des réformes appropriées, la politique de l'éducation. Les enquêtes de l'IEA ont ainsi produit un modèle de travail qui est à la base du renouveau des études comparées en éducation. L'aspiration plus ou moins larvée de ces évaluations a été la production d'explications causales des effets observés grâce à l'analyse des corrélations s'établissant entre résultats et variables contextuelles. Cette aspiration est toujours présente dans les évaluations à grande échelle et on peut en déceler la présence dans les interprétations produites par l'OCDE des résultats PISA. Cependant, il convient de relativiser considérablement ces attentes, car, comme le dit Paulston dans son article pour l'encyclopédie internationale de l'éducation, les évaluations à grande échelle, plus que de mettre en évidence des corrélations et des liens causaux, ont surtout contribué à révéler la puissance des facteurs non intentionnels à l'œuvre dans l'enseignement, les dangers représentés par une abondance excessive de données et la pauvreté de la modélisation conceptuelle, ou les problèmes inhérents aux comparaisons elles-mêmes, étant donné que chaque école et chaque système d'enseignement est incrusté dans un réseau complexe et unique de relations culturelles, économiques et politiques[36].

Sans entrer dans l'examen détaillé de l'univers particulièrement riche et foisonnant de la recherche scientifique dans le champ de l'éducation comparée contemporaine, on peut affirmer que, malheureusement, la France, qui avait été à l'avant-garde dans le développement des études comparées en éducation, a perdu depuis longtemps cette position. Par exemple, Paulston, dans sa grille des paradigmes et des théories dans l'éducation comparée au niveau international, cite uniquement trois auteurs français (Althusser, Bourdieu et Passeron), qui sont certainement des figures de premier plan de la scène intellectuelle française, mais qui ne peuvent pas être considérés comme des représentants « stricto sensu » du monde de la pédagogie et de l'éducation. Il y a donc eu un hiatus important entre l'évolution de la recherche en éducation en France d'un côté, et celle des sciences comparées dans le domaine de l'éducation au niveau international. On peut estimer qu'une des causes de la position particulière occupée par la France dans le champ des évaluations internationales à grande échelle trouve sa raison d'être dans la crise de la recherche française dans le domaine de l'éducation comparée qui, à son tour, aurait déterminé la perte de crédibilité et de confiance en la validité épistémique des études comparées en éducation. La centration de la France sur elle-même, la centralisation du système d'enseignement français, sa robustesse, sont des facteurs qui ont accentué la méfiance ou la relative indifférence autour des enquêtes internationales à grande échelle ; de même, le nombre réduit de scientifiques français de renommée internationale dans le domaine des études comparées quantitatives est un autre facteur explicatif du déclin français au moment même du renouveau de ce champ de recherche[37]. Enfin, il ne faut pas non plus sous-estimer le fait que l'évaluation du système d'enseignement en termes de compétences est acquise depuis longtemps en France, ce qui a probablement induit un certain aveuglement à l'égard de la valeur ajoutée représentée par les comparaisons internationales.

[36] Paulston R.G. (1988) : Comparative and International Education : Paradigms, Theories, and Debates. In: Education: The Complete Encyclopedia, version 1.1/Torsten Husén…(et al.). New York: Elsevier Science, cop. 1998.

[37] Cacouault M., Orivel F. (ed). L'évaluation des formations : points de vue comparatistes. Actes du 15e Congrès de l'Association européenne d'éducation comparée, Dijon, 27 juin/2 juillet 1992, Volume 1, IREDU-CNRS-Université de Bourgogne, France. Janvier 1993.

Le développement des évaluations internationales à grande échelle au cours des quatre dernières décennies s'est fait en coïncidence avec un accroissement de l'intérêt des gouvernements, et plus particulièrement des autorités politiques ou administratives en charge des systèmes d'enseignement pour les résultats des comparaisons internationales. Il est difficile de dire si l'intérêt des autorités pour les comparaisons effectuées grâce à des tests sur les acquis et les compétences des élèves soumis à des échantillons comparables d'élèves est l'effet ou la cause du développement des études internationales comparées à grande échelle. La chronologie suggérerait que la montée de l'intérêt des gouvernements a fait suite à la mise en œuvre des enquêtes internationales, mais ces enquêtes n'auraient pu être effectuées sans qu'un certain nombre de gouvernements les aient appuyées et rendues possibles grâce à de substantielles contributions financières. Thomas Kallaghan du St. Patrick's College de Dublin, un des spécialistes internationaux dans le domaine de l'évaluation à grande échelle et de l'application des tests, estime que plusieurs raisons interviennent dans l'explication de l'intérêt des gouvernements et des autorités pour ces études :

- la fin de l'expansion quantitative de la scolarisation, qui a permis l'émergence de préoccupations qualitatives dictées par la nature des résultats atteints avec les investissements réalisés dans le domaine de l'éducation ;
- l'augmentation des coûts unitaires en éducation, d'où découlent les questions sur le rapport coût-bénéfice ou coût-efficacité dans l'enseignement ;
- le phénomène de la mondialisation et de la globalisation qui ont relancé la compétition entre pays ou entre économies différentes en accentuant la place du capital humain et donc du rôle de l'éducation dans l'offre du capital humain ;
- la recherche d'une légitimation que les résultats de ces comparaisons fournissent aux décisions des autorités ou aux programmes de réforme lancés dans le domaine de l'éducation. Ces évaluations peuvent aussi bien justifier des réformes mises en œuvre dans le passé ou légitimer la nécessité d'entreprendre des réformes en fournissant des données sur les acquis des élèves d'un système d'enseignement par rapport aux résultats obtenus par d'autres systèmes d'enseignement.

Toutes ces raisons réunies contribuent à déterminer l'intérêt pour les études comparées sur les résultats des systèmes d'enseignement[38].

7 - LA PARTICIPATION DE LA FRANCE AUX EVALUATIONS INTERNATIONALES SUR LES ACQUIS DES ELEVES

7.1 - PARTICIPATION FRANÇAISE AUX ENQUETES DE L'IEA

La présence de la France dans les enquêtes internationales sur les acquis des élèves au cours de ces dernières quarante années, entre 1964 et 2003, a été constante quoique irrégulière. La France est présente dans ces opérations dès le début, mais elle n'a pas participé à toutes les enquêtes.

[38] Thomas Kallaghan, 1996 : IEA Studies and Educational Policy. In: Assessment in Education, Vol.3, n°2, pp.143-160.

Tableau 1 : Principales enquêtes internationales sur les acquis des élèves (1964-2003)

Sponsor	Titre de l'enquête	Nombre de systèmes d'enseignement concernés	Années	Présence de la France	Responsabilité du pilotage en France
UNESCO	Etude pilote en douze pays	12	1959-1962	Oui	
IEA	Première enquête internationale sur les mathématiques (FIMS)	12	1964	Oui	Gaston Mialaret, Univ. de Caen
IEA	Enquêtes sur six matières (Six Subjects Study): ▪ Sciences ▪ Compréhension de la lecture ▪ Littérature ▪ Français comme langue étrangère ▪ Anglais comme langue étrangère ▪ Education civique	 19 15 10 8 10 10	1970-73	 Oui Non Non Non Non Non	Françoise Bacher Denis Bonora INETOP, Paris
IEA	Deuxième enquête internationale sur les mathématiques (SISS)	10	1982	Oui	Daniel Robin et Emilie Barrier CIEP, Sèvre
IEA	Deuxième enquête internationale sur les sciences (SISS)	19	1983-84	Non	
IEA	Enquête sur la composition écrite	14	1980-88	Non	
ETS	Première évaluation internationale des progrès en éducation (IAEP-I, mathématiques et sciences)	6 pays (12 systèmes d'enseignement)	1988		
ETS	Deuxième évaluation internationale des progrès en éducation (IAEP-II, mathématiques et sciences)	20	1991	Oui	Martine le Guen Thierry Soupault Min. de l'Educ. Nat. DEP.
IEA	Compréhension de la lecture (Reading Literacy)	32	1990-91	Oui	
IEA	Première enquête internationale sur les nouvelles technologies à l'école (Computers in Education)	1er groupe : 21 2ème groupe : 12	1988-89 1991-92	Oui	
OCDE	Enquête internationale sur la littératie des adultes (IALS) I	22 en trois phases successives	1994 - 2000	Oui	J.P. Jeantheau

IEA	L'éducation pré-scolaire (PPP : Pre Primary Education) Première phase : Deuxième phase : Troisième phase (enquête longitudinale) :	 11 15 15	 1989-91 1991-93 1994-96	Non	
IEA	Troisième enquête internationale sur les mathématiques et les sciences (TIMSS)	45 49 (8ème année) et 27 en CM1)	1994-95 2003	Oui	Anne Servant Min. de l'Educ. Nat. DEP.
IEA	Troisième enquête internationale sur les mathématiques et les sciences – Réplique (TIMSS-R)	40	1997-98	Non	
IEA	Deuxième enquête internationale sur les nouvelles technologies à l'école (Computers in Education) (IEA-SITES)	26	1998-99	Oui	
IEA	Education civique (CIVED)	28	1999	Non	
OCDE	Programme pour l'évaluation international des acquis des élèves (PISA - Program for International Student Assessment)	32	2000 (centre de gravité : lecture)	Oui	J.P. Jeantheau
IEA	Progrès dans la compréhension de la lecture à 9 ans (CM1 (enquête PIRLS)	35	2001	Oui	
OCDE	Enquête internationale sur la littératie et les compétences de base des adultes (ALL)	1er groupe : 5	2003	Non	
OCDE	Programme pour l'évaluation international des acquis des élèves (PISA - Program for International Student Assessment)		2003 (centre de gravité : mathématiques)	Oui	Anne-Laure Monnier, Min. de l'Educ.Nat., DEP.
IEA	Troisième enquête internationale sur les mathématiques et les sciences – Réplique (TIMSS-R)	49 pour la 8ème année (13 ans) 27 pour la 4ème année (10 ans)		Non	

Note : Les cellules vides dans la colonne des responsables français sont imputables à des données manquantes non communiquées aux auteurs.

Sur 27 évaluations répertoriées (voir tableau 2), on constate que la France est présente dès la première étude pilote, qui s'est déroulée entre 1959 et 1962. Cependant, la France n'a pas participé à toutes les évaluations conduites au cours de ces quatre décennies. Le deuxième constat nous amène à observer que le total des participations françaises pour l'ensemble des évaluations prises en considération s'élève à treize présences, ce qui équivaut, grosso modo, à une présence sur deux. Donc, le choix de participer à des évaluations internationales a été plutôt sélectif, car l'Education nationale a écarté la possibilité d'être présente dans la moitié des opérations internationales d'évaluation des élèves. Dès lors, une première question se pose : qui a pris ces décisions, et en deuxième lieu selon quels critères les choix ont-ils été opérés ? On peut supposer que la participation française répond à une stratégie délibérée ou explicite, compte tenu du fait que les présences de la France s'étalent tout le long de la période observée.

L'analyse des treize participations françaises met en évidence le fait que la France a systématiquement choisi d'être engagée dans les évaluations à grande échelle concernant ce qu'on peut appeler les disciplines fondamentales, c'est-à-dire les évaluations des connaissances et compétences en mathématiques, en sciences et en lecture. De ce point de vue, la France a manqué uniquement la première enquête sur la compréhension de la lecture en 1970-71. Les seules exceptions concernent la participation aux deux enquêtes internationales sur les nouvelles technologies à l'école (Computers in Education) dont la première a eu lieu en 1988-89 et la deuxième dix ans après, en 1998-99. On peut dire que le choix de la France a été scellé par l'orthodoxie pédagogique et que les chercheurs français ou les autorités françaises, point encore à clarifier, ne se sont pas engagées dans les enquêtes plus expérimentales comme par exemple celle sur l'éducation préscolaire, la composition écrite, l'enseignement en classe ou même l'éducation civique. Le système d'enseignement français a délaissé les enquêtes internationales mobilisant surtout la recherche en éducation et l'originalité méthodologique, comme le montre le tableau 2 sur les présences françaises dans les enquêtes internationales de ces dix dernières années. La seule exception à cet égard est la participation à l'enquête internationale sur la littératie des adultes (enquête IALS).

Tableau 2 : Participation de la France aux dernières grandes enquêtes internationales

Enquêtes internationales sur l'évaluation des acquis des élèves				
	TIMSS 1994	**PISA 2000**	**PIRLS 2001**	**PISA 2003**
Patronnage	IEA	OCDE	IEA	OCDE
Coordination internationale	ISC	ACER	ISC	ACER
Coordination en France	DEP	DEP	DEP	DEP
Périodicité	4 ans	3 ans	5 ans	3 ans
Années de collecte de données	1995, 1999, 2003	2000	2001, 2006	2003
Domaines évalués	Mathématiques et sciences	Lecture, mathématiques et science	Compréhension de la lecture	Maths, sciences et lecture

Questionnaires	Etudiants, enseignants et école			Etudiants et école	Etudiants, enseignants et école	Etudiants et école
Echantillon des écoles	Probabilité proportionnelle à la taille du système			Probabilité proportionnelle à la taille du système	Probabilité proportionnelle à la taille du système	Probabilité proportionnelle à la taille du système
Echantillon des élèves	2 classes du degré par école			35 élèves de 15 ans par école	1 classe par école	35 élèves de 15 ans par école
Participation de la France	1995	1999	2003	oui	2001	oui
Degrés ou population	8,12			15 ans	4	15 ans
Notes : IEA : International Association for the Evaluation of Educational Achievement ISC : International Study Center, Boston College OCDE : Organisation de coopération et de développement économiques ACER : Australian Council for Educational Research						

7.2 - L'ENQUETE IALS (INTERNATIONAL ADULT LITERACY SURVEY)

Une seule exception, comme on vient de le dire, dans ce panorama : la participation de la France à l'enquête internationale sur la littératie des adultes en 1994 qui était une nouveauté absolue au niveau mondial, car il s'agissait de tester une population d'adultes à domicile et non plus d'élèves en classe. Or, cette exception a été payée à un prix élevé. Comme tout le monde le sait, la France, après avoir participé à la préparation de l'enquête, et après avoir réalisé l'enquête auprès d'un échantillon d'adultes représentatif de la population civile âgée de 16 à 65 ans, a contesté la validité des résultats la concernant et a exigé « in extremis », lorsque le rapport conclusif était prêt à aller sous presse, en automne 1995, son épuration afin d'éliminer toute référence à la France. Cette décision, prise par le Ministre de l'éducation François Bayrou a soulevé beaucoup de vagues. Pour la justifier, la France s'est engagée par la suite dans un long processus de vérification de la validité de l'enquête, en mobilisant la communauté scientifique nationale et internationale ainsi que l'Union européenne pour examiner sa pertinence méthodologique. Cette démarche, par ailleurs inattaquable du point de vue scientifique, marque un tournant dans l'histoire de la présence française dans les enquêtes internationales à grande échelle, car elle a forcé la France à prendre la tête d'une cordée et à jouer un rôle actif dans ce champ. On peut donc distinguer une période antérieure à l'enquête IALS et une période postérieure à cette enquête, un « avant 1994 » et un « après 1994 ». Pendant trente années, entre 1964 et 1994, la France, comme on vient de le voir, avait plutôt adopté un profil bas sur la scène des évaluations internationales à grande échelle. Tout en participant à un certain nombre d'enquêtes, la France, c'est-à-dire d'un côté les autorités responsables du système d'enseignement et de l'autre la communauté scientifique travaillant dans la recherche en éducation, n'a pas tenu de rôle de premier plan[39].

[39] Jusqu'au tournant des années 90, la participation de la France aux enquêtes internationales sur les acquis des élèves a été assumée par trois centres : l'Université de Caen (avec Gaston Mialreet et François Beaufils), l'INETOP (avec Maurice Reuchlin, Françoise Bacher et Denis Bonora) et le CIES de Sèvre (avec Daniel Robin et Emilie Barrier).

Or, avec la contestation, largement exploitée par les médias, de l'enquête IALS et surtout avec le geste spectaculaire de retrait du rapport international, tout à coup, en 1995, la France sort de l'ombre et s'installe sur le devant de la scène. Or, sa posture est délicate, les scientifiques français n'ayant pas spécialement contribué au développement des méthodes d'enquêtes à grande échelle pour comparer les prestations de plusieurs systèmes d'enseignement, en particulier les acquis des élèves ou les compétences des adultes. La recherche française en éducation au sens large n'avait pas d'emblée la crédibilité ou l'autorité scientifique pour interpeller la pertinence et la validité de l'enquête IALS ; d'autre part, le retrait des données français, décidé par le ministre, obligeait la France à justifier ses réserves en mobilisant la communauté scientifique internationale pour fournir les preuves susceptibles d'étayer ses critiques. La dénonciation de la validité des résultats de cette enquête et de ses biais méthodologiques laissait entrevoir une critique de fond de la conception psychométrique de l'enquête, de son modèle organisationnel et des procédures adoptées pour l'effectuer.

L'entrée en scène fracassante de la France dans le domaine de la recherche et du développement des évaluations à grande échelle s'opérait autour d'une enquête tout à fait spéciale, entièrement différente par sa conception et sa méthode de déroulement des enquêtes menées auparavant aussi bien par l'IEA que par l'ETS. En effet, l'enquête IALS ne concernait pas une population d'élèves testés dans les classes, mais une population d'adultes testés chez eux à la maison, ce qui modifiait non seulement le cadre de l'enquête, mais sa conception, sa structure, son déroulement et donc aussi l'ensemble des opérations connectées au traitement des données[40]. L'enquête IALS avait un caractère expérimental, car elle abordait une population différente de celle scolaire ; elle testait non pas des connaissances scolaires, mais une compétence générique pour définir laquelle il a fallu forger un nouveau terme en français : la littératie, c'est-à-dire la capacité à comprendre le contenu de différents types de textes écrits. La littératie dans l'enquête IALS est définie comme la capacité à « utiliser des imprimés et des écrits nécessaires pour fonctionner dans la société, atteindre ses objectifs, parfaire ses connaissance et accroître son potentiel » (Rapport IALS 1995 - 16). Cette définition tente d'englober un ensemble de compétences mobilisées pour traiter l'information que des adultes peuvent être appelés à utiliser pour accomplir un grand nombre de tâches différentes au travail, à la maison et dans leur collectivité. Trois catégories de textes, qu'on retrouvera par la suite dans l'enquête PISA, ont été constituées, chacune faisant appel à l'ensemble commun de compétences pertinentes liées à diverses tâches :

[40] L'enquête internationale sur l'alphabétisation des adultes ou plus précisément sur les capacités de lecture et d'écriture des adultes s'est déroulée en trois vagues : une première en 1994 composée de 9 pays (Allemagne, Canada (population anglophone et francophone), Etats-Unis, France, Irlande, Pays-Bas, Pologne, Suède et Suisse (régions germanophone et francophone) ; une deuxième en 1996 composée de l'Australie, la communauté flamande de Belgique, la Grande-Bretagne, l'Irlande du Nord et la Nouvelle-Zélande ; une troisième en 1998 composée de 9 autres pays ou régions (Chili, Danemark, Finlande, Hongrie, Italie, Norvège, République tchèque, Slovénie et région italophone de la Suisse). Chaque vague d'enquête a été suivie par la publication d'un rapport par l'OCDE :
- en décembre 1995 : Littératie, économie et société : résultats de la première enquête internationale sur l'alphabétisation des adultes ;
- en novembre 1997 : Littératie et société du savoir : nouveaux résultats de l'enquête internationale sur les capacité de lecture et d'écriture des adultes ;
- en 2000 : le rapport final : La littératie à l'ère de l'information.
Seule la France a refusé la publication de ces résultats. L'Italie a aussi en partie contesté la validité des résultats la concernant et a refusé l'inclusion des données italiennes dans le rapport final, mais elle a publié ses données dans un rapport en italien : La competenza alfabetica in Italia : una ricerca sulla cultura della popolazione. Editore Franco Angeli, Milano.

- textes suivis tels que des éditoriaux, des nouvelles, des brochures et des modes d'emploi que les adultes doivent lire et comprendre et dans lesquels les lecteurs sont invités à identifier des informations ;
- textes schématiques tels que les demandes d'emploi, les fiches de paie, les horaires de transport, les cartes routières, les tableaux et les graphiques, c'est-à-dire un ensemble de textes d'usage courant dans la vie quotidienne, dans lesquels les lecteurs doivent repérer des informations et que les adultes doivent pouvoir utiliser pour réaliser des tâches qui se présentent régulièrement dans la vie de tous les jours ;
- textes au contenu quantitatif tels que le solde d'un compte-chèques, le calcul d'un pourboire, un bon de commande, le calcul de l'intérêt d'un emprunt à partir d'une annonce publicitaire, c'est-à-dire un ensemble d'imprimés qui contiennent des informations numériques qui font appel à une compétence arithmétique ou mathématique, à partir desquelles le lecteur doit parvenir à extraire des informations ou des renseignements qu'on applique pour organiser par exemple l'existence, les déplacements, les décisions sur le choix de l'emploi, la demande d'un prêt ou la souscription d'une assurance.

Comme on a déjà eu l'occasion de le dire, non seulement la population ciblée avec cette enquête était différente de la population des enquêtes de l'IEA, mais le déroulement de l'enquête se démarquait entièrement des évaluations précédentes à grande échelle. Pour tester directement les capacités de lecture et d'écriture des adultes d'une manière semblable à celle que l'on adopte dans les écoles, il est nécessaire de se rendre au domicile des gens. L'enquête IALS sortait donc des sentiers battus du fait qu'elle combinait les techniques d'une enquête-ménage à celles d'un test scolaire. Cependant, contrairement à la plupart des tests normés, on a évité les questions à choix multiples car les adultes préfèrent des questions à réponse libre. Le test était mené par des intervieweurs expérimentés qui procédaient en trois temps :

- tout d'abord en posant aux répondants une série de questions afin d'obtenir des renseignements contextuels (données démographiques, antécédents professionnels, etc.) ;
- ensuite, l'intervieweur remettait aux répondants un livret renfermant six tâches de lecture simples. Si le répondant se trouvait dans l'incapacité d'effectuer au moins deux tâches, l'interview n'allait pas plus loin ;
- enfin, si le répondant effectuait correctement deux exercices ou plus, il recevait un livret distinct renfermant un nombre beaucoup plus important de tâches.

L'évaluation n'était pas minutée et les répondants étaient invités à essayer chaque exercice pour lui accorder le maximum de chances de prouver ses compétences.

7.3 - PROBLEMES METHODOLOGIQUES POSES PAR L'ENQUETE IALS

En ce qui concerne l'enquête IALS, la réaction française a permis d'attirer l'attention sur deux dimensions problématiques communes à toutes ces enquêtes : le problème de la traduction des items de test, mais aussi de la formulation des questions dans les questionnaires et le problème des biais de nature culturelle. C'est sur ces deux aspects que s'est focalisée l'action de la DEP au cours des années qui ont suivi la publication des premiers résultats de l'enquête IALS. Les questions techniques relatives à la neutralisation des biais culturels et à la réalisation de traductions comparables, mettant les textes en

différentes langues sur le même plan de difficulté, seront abordées plus tard dans ce rapport. Cependant, nous pouvons remarquer que ces enjeux n'étaient pas nouveaux et étaient bien connus de la communauté scientifique engagée dans la mise au point des évaluations à grande échelle au sein de l'IEA. Il est vrai que les solutions adoptées pour parer à des erreurs de traduction ou à des biais de nature culturelle n'ont pas toujours été à la hauteur, les faits ne suivant pas les intentions, parfois par manque de moyens ou parfois aussi à cause des compromis passés dans les négociations entre responsables éducatifs et spécialistes des enquêtes. En ce qui concerne l'enquête IALS, le problème a été abordé au cours de la préparation de l'enquête. Pour obtenir une estimation valide du niveau de capacité de chaque personne, les promoteurs de l'enquête avaient convenu qu'il aurait été nécessaire d'élaborer un vaste éventail de tâches reflétant la diversité culturelle de la population concernée par l'enquête. Dans l'ensemble, quelque 175 tâches de lecture et d'écriture ont été élaborées pour les tests sur le terrain. De ce nombre, 114 qui se sont avérés valides pour toutes les cultures ont été retenus aux fins de l'évaluation principale. Environ la moitié de ces tâches étaient fondées sur des documents provenant de l'extérieur de l'Amérique du Nord. A la suite de la publication des résultats en automne 1995, la DEP a estimé que ces précautions n'avaient pas été suffisantes et que la composition du test était susceptible de fausser les réponses des interviewés. Le lancement de toute une série d'initiatives visant à apprécier les risques de biais présents dans la méthodologie appliquée jusqu'à 1995 dans les enquêtes scolaires et surtout les conséquences de la transposition aux adultes de cette méthodologie testée auparavant avec les élèves en classe a engendré un important débat scientifique, produit beaucoup de publications sous forme de rapports et d'articles, mais n'est pas parvenu à fournir des preuves irréfutables d'invalidité de ces enquêtes ou à mesurer précisément le degré de distorsion produit par les biais culturels et linguistiques sur les résultats.

L'autre problème sur lequel s'est focalisée la critique française a été celui de la traduction en plusieurs langues des tests et des questionnaires tout en préservant leur comparabilité. Cet aspect sera aussi abordé par la suite. Il n'était pas nouveau non plus car il avait préoccupé les promoteurs des enquêtes de l'IEA dès le début. Cependant, l'enquête IALS a offert l'occasion pour inciter la communauté scientifique à analyser de plus près les enjeux de nature linguistique et culturelle inclus dans les problèmes de traduction, ainsi qu'à s'interroger sur la validité des méthodologies alternatives évitant les problèmes posés par la traduction en plusieurs langues tout en préservant la comparabilité des tests[41]. En ce qui concerne par exemple le cas particulier de IALS, il s'est avéré que la traduction en français des items du questionnaire pour les adultes effectuée en France était différente de la traduction en français des mêmes items effectuée en Suisse romande. L'analyse des réponses aux différents items a montré que la traduction de la Suisse romande était plus précise et moins ambiguë. De ce fait, les réponses des deux populations testées en français devaient nécessairement présenter des variations statistiques significatives. Par ailleurs, une vérification effectuée en France sur une population adulte réduite, à laquelle on a soumis le questionnaire en français utilisé en Suisse romande, a permis de constater la meilleure qualité de la traduction effectuée en Suisse romande par rapport à celle réalisée en France. Ce constat est déjà en soi suffisant pour confirmer les incidences de la traduction dans le but de parvenir à un test commun rédigé en différentes langues permettant néanmoins la comparaison des résultats.

[41] Voir par exemple Bonnet, G. et al., 2004 : Culturally balanced assessment of reading (c-bar). Ministère Education Nationale, DEP, Paris ; Baye, A., 2004 : la gestion des spécificités linguistiques et culturelles dans les évaluations internationales de la lecture. In : Politiques d'éducation et de formation, n° 11, pp. 55-70.

Enfin, une troisième composante a une incidence sur la comparabilité des résultats : l'échantillonnage de la population à laquelle on fait passer le test. Les questions techniques d'échantillonnage sont aussi bien connues et relativement bien maîtrisées, mais il est évident que lorsqu'on sort du contexte scolaire dans lequel les élèves sont bien identifiés pour passer à un contexte externe à l'école, la construction d'un échantillon probabiliste représentatif de l'ensemble de la population adulte devient un problème autrement plus complexe. Dans le cas de l'enquête IALS on a moins parlé de cet enjeu, mais on peut supposer que dans le cas de la France, la constitution de l'échantillon doit aussi avoir eu un effet sur les résultats. L'effet combiné des biais culturels, des problèmes de traduction, et de la construction de l'échantillon a probablement multiplié les déformations des résultats finaux. Il ne s'agit ici que de suppositions qui restent à être démontrées. Nous ne disposons pas encore de mesures sur l'incidence de l'un ou de l'autre de ces facteurs sur les résultats de l'enquête. Nous savons que des erreurs sont possibles, que l'importance de ces erreurs dans certains cas, comme par exemple dans celui de l'échantillonnage, peut être estimée ou calculée, mais nous ne maîtrisons pas encore suffisamment l'incidence de l'ensemble de ces paramètres sur les résultats d'une évaluation comparée à grande échelle.

Enfin, un dernier aspect mérite d'être relevé : les conditions dans lesquelles on effectue les interrogations des adultes. S'agissant d'adultes testés chez eux, il était indispensable de s'appuyer sur des interviewers professionnels formés à ce type d'exercice et de les encadrer en suivant de près la progression de la collecte des informations. Le déroulement de l'enquête était en effet en soi fort délicat. Or, la DEP n'avait pas l'expérience de ce genre d'exercice, car sa compétence couvrait les évaluations des élèves en classe, domaine dans lequel l'expertise de la DEP était indéniable, mais ne s'étendait pas au monde des adultes. Il se peut donc que des erreurs aient été commises lors de la réalisation de l'enquête en France ou que des difficultés soient apparues sans qu'on parvienne à les maîtriser convenablement, en causant de ce fait des distorsions importantes des résultats.

7.4 - LE RESEAU EUROPEEN DES RESPONSABLES DES POLITIQUES D'EVALUATION DES SYSTEMES EDUCATIFS (RERPESE)

L'activisme français dans le champ des évaluations internationales à grande échelle dans le domaine éducatif aboutit, sur le plan institutionnel, à la constitution du réseau européen des responsables des politiques d'évaluation des systèmes éducatifs, décidée lors d'une réunion des hauts fonctionnaires de l'éducation de l'Union européenne en 1995, au moment de l'éclatement de la crise IALS, lorsque l'Union européenne était présidée par la France. Le dispositif proposé par cette initiative française – un réseau intergouvernemental dont les membres sont désignés par les gouvernements par l'intermédiaire de leur représentant au comité de l'éducation de l'Union européenne – se démarque nettement du seul dispositif qui, jusqu'à ce moment, gérait les enquêtes internationales en comparant les résultats des systèmes d'enseignement, à savoir celui de l'IEA qui était structuré en fonction de l'organisation de la communauté scientifique dont le seul principe était de neutraliser autant que possible l'ingérence des gouvernements dans la programmation des enquêtes. Pendant qu'au sein de l'OCDE on envisageait sérieusement la réalisation d'un programme d'évaluation sur les acquis des élèves, la crise éclatée avec l'enquête IALS a probablement montré la nécessité d'une collaboration intracommunautaire dans le domaine de l'évaluation des systèmes éducatifs – étant rappelé que l'éducation n'est pas une prérogative communautaire. Seule une initiative des gouvernements pourrait faire avancer les dossiers en la matière. Ainsi, au même moment, se prépare au sein de deux institutions

intergouvernementales un tournant décisif qui va modifier la nature des enquêtes comparées sur vaste échelle des acquis des élèves.

Le Réseau européen des responsables des politiques d'évaluation des systèmes éducatifs (que nous désignerons désormais par l'acronyme RERPESE) a été créé lors d'une réunion des hauts fonctionnaires de l'éducation pendant la Présidence française de l'Union européenne de 1995. Ce réseau intergouvernemental regroupe des membres, désignés par l'intermédiaire des représentants des pays au Comité de l'éducation, qui exercent des responsabilités dans le domaine de l'évaluation et du pilotage de leur système éducatif. À ce jour, le Réseau, qui se réunit deux fois l'an dans une capitale européenne, regroupe des représentants des États membres (dont les deux Communautés belges et l'Écosse) ainsi que l'Islande et la Norvège. La Commission européenne et la Suisse y sont représentées et les dix nouveaux États membres ont été invités à y désigner leurs représentants. La France, par le truchement de la Direction de l'évaluation et de la prospective du ministère de la Jeunesse, de l'Éducation nationale et de la Recherche, assure son animation et sa présidence depuis sa création, ainsi que la rédaction en chef, la publication et la diffusion d'ÉVALUATION, lettre d'information du Réseau qui paraît deux fois chaque année. Les objectifs du Réseau sont, d'une part, de permettre des échanges d'informations sur l'évaluation et le pilotage du système éducatif dans chacun des pays et au plan communautaire ; et d'autre part, de définir, d'initier et de conduire des actions de coopération européenne dans ce même domaine. L'étude sur les compétences des élèves européens en anglais est l'un des nombreux projets qu'il a menés à ce jour[42].

Comment interpréter ce changement de stratégie adopté par la France en 1995 ? Le choc produit par les résultats de IALS avait probablement fait comprendre à la DEP qu'il lui était impossible d'influencer efficacement les enquêtes internationales qui étaient des opérations très complexes et d'une grande envergure. La DEP, après avoir participé activement à la conception de IALS et à toutes les phases de réalisation de l'enquête, découvre à la fin des résultats surprenants, inattendus. Quelque chose, manifestement, lui avait échappé. Or, l'évaluation des systèmes d'enseignement était en train de devenir un enjeu international, en partie à cause du battage médiatique entourant la publication des résultats des enquêtes et des indicateurs de l'enseignement produits par l'OCDE. Il était donc urgent pour la DEP et pour le Ministère de l'éducation nationale de se positionner face à ce sujet pour éviter des situations politiquement gênantes.

Le RERPESE a organisé en 1996 une étude comparative des compétences en anglais des élèves de 15 ans dans trois pays, la France, la Suède et l'Espagne sur la base d'un protocole d'évaluation commun (voir Levasseur & Shu, 1997 ; ainsi que Bonnet, 1998). On a signalé que le domaine des langues secondes, en particulier de l'anglais, n'avait pas intéressé les organismes promoteurs des enquêtes internationales (en ce qui concerne l'anglais, on comprend le manque d'intérêt des USA et de la Grande-Bretagne). Cette enquête palliait donc un manque dans l'éventail des compétences évaluées. Une réplication de cette enquête a eu lieu en 2002/2003 toujours sous l'égide du Réseau européen des responsables des politiques d'évaluation des systèmes éducatifs. Huit pays s'y sont associés : les trois participants à la première étude, auxquels sont venus s'ajouter la Finlande, la Norvège, les Pays-Bas, le Danemark et l'Allemagne (pour des raisons de

[42] Le site http://cisad.adc.education.fr/reva/ donne toutes les informations utiles sur ses activités et permet d'accéder à ses publications.

comparabilité, les résultats allemands n'ont pas été pris en compte dans le document final). On trouvera la description et les résultats de cette recherche dans Bonnet et al., 2004a.

Comparaison des performances en littératie sans épreuves communes

A la suite des rencontres pour répondre à l'appel d'offre PISA et avec l'appui du RERPESE, des partenaires de quatre pays (Angleterre, Finlande, France, Italie ; on trouvera en annexe la liste des institutions représentées) ont décidé de lancer une enquête visant à comparer les performances en lecture d'élèves de 15 ans dans leurs pays respectifs sans recourir à des épreuves traduites (Bonnet et al., 2001 ; Vrignaud et Rémond, 2002). Cette première recherche porte sur la faisabilité de la comparaison des compétences en littéracie d'élèves en fin de scolarité obligatoire dans ces quatre pays en utilisant des épreuves originales dans la langue nationale de chacun des pays sans recourir à une épreuve commune traduite.
Les travaux ont porté sur :

- la présentation des épreuves existantes dans les différents pays, leur conception et la fiabilité de les mettre en regard ;
- la construction d'une grille de compétences commune permettant d'identification des compétences évaluées par les exercices et les items des différentes épreuves nationales (on s'est ici largement appuyé sur la grille de compétences élaborée dans le cadre du consortium) ;
- les méthodologies statistiques permettant d'effectuer des comparaisons des performances sans disposer d'épreuve commune pour assurer un passage entre les différentes épreuves nationales.

Le projet C-BAR

Les résultats de cette étude ont servi de base à une seconde recherche sur l'élaboration de nouvelles approches comparatives pour les évaluations internationales des compétences en littéracie le projet C-BAR (*Culturally Balanced Assessment of Reading*). Huit pays européens (Belgique : Communauté flamande), Angleterre, Finlande, France, Italie, Pays-Bas, Norvège et Suède) ont participé à ce projet, piloté par le RERPESE. L'objectif était de tirer les enseignements de l'étude de faisabilité de l'évaluation de la littéracie à partir d'épreuves nationales. Il ne s'agissait pas comme dans l'étude précédente de réaliser une enquête mais de conduire une réflexion sur les principaux éléments nécessaires pour mettre en place une évaluation comparative : les compétences, les épreuves, les échantillons et bien sûr les méthodes pour assurer l'équivalence. Pour chacun de ces points, les apports des participants ont permis d'élaborer des procédures qui pourront être proposées soit lors d'appels d'offre soit pour rechercher des financements permettant de conduire un tel projet au niveau européen ou international (Bonnet et al., 2004b).

Autres projets du RERPESE

Parmi les autres projets conduits par le RERPESE et pilotés par la France, on peut citer l'évaluation de la lecture à l'école primaire en Angleterre, France et au Grand-duché de Luxembourg (Andrieux et al., 2002). L'intérêt majeur de cette enquête était de mettre en relation les variables de performance en lecture avec les méthodes pédagogiques employées observées in situ. D'autres projets ont été réalisés par le RERPESE : une enquête sur l'identification d'indicateurs pour réussir l'enseignement des langues dans les

lycées pilotée par l'Irlande est en cours, une enquête sur les compétences en sciences physiques, également pilotée par l'Irlande a été effectuée en 1998. L'Ecosse a piloté une enquête sur les parents et l'école dans les pays européens. Enfin, signalons une enquête pilotée par la Grèce, en 1997, sur le thème « auto-évaluation des établissements scolaires et régionalisation ». Enfin, un projet pour l'évaluation des compétences des adultes est en préparation : *Project for international assessment of adult competencies.*

CHAPITRE II - PLANIFICATION, CONSTRUCTION ET MISE EN ŒUVRE DES ENQUETES INTERNATIONALES

La réalisation d'une enquête internationale à grande échelle comportant une évaluation des connaissances et des compétences des élèves exige une planification détaillée pour articuler entre eux des domaines complémentaires tels que :

- le pilotage et la gestion de l'exercice ;
- le financement initial, le plan comptable et le bilan final ;
- la construction des instruments de l'enquête ;
- le travail des centres de calcul et traitement des données ;
- l'analyse des résultats et leur divulgation.

Chacune des ces dimensions a un double volet : un volet international pour les problèmes transversaux communs dont le rôle est crucial pour assurer la cohérence de l'ensemble et valider les comparaisons ; un volet national qui concerne chaque système d'enseignement participant à l'enquête ou chaque pays qui décide de soumettre son système d'enseignement à l'évaluation internationale.

Tableau 3 : Composante technique d'un projet d'évaluation comparée à grande échelle au niveau international : schéma synthétique

Niveau	**Composantes structurelles**			
	Organisation **Pilotage** **Gestion**	**Financement et comptabilité**	**Opérationnalisation : construction des instruments**	**Centre de calculs**
Niveau international	Coordination internationale	Gestion du financement de la partie internationale (ticket d'entrée par système d'enseignement ; répartition des frais etc.)	Construction des instruments : - cadre théorique - typologie des problèmes - recueil d'items - sélection des items Validation des instruments Traductions Echantillonnage	Traitement des données et production des résultats
Niveau national	Coordination nationale	Financement des frais nationaux	Proposition d'items Modules nationaux Administration des tests Approfondissements analytiques	Vérification, nettoyage, extraction des données

L'articulation entre les niveaux international et national (Tableau 3) est un point critique non seulement pour les besoins de l'enquête, mais aussi et surtout pour sa configuration en fonction des attentes des pays. Pour chacune des composantes structurelles (le pilotage

d'ensemble, le financement, l'opérationnalisation, les calculs), un pays qui a l'ambition d'infléchir le cours des travaux au niveau des priorités, du choix des instruments, des critères de gestion, des échéances à tenir, etc. doit être en mesure d'être présent au niveau international avec des représentants capables d'intervenir dans le processus décisionnel selon une stratégie non pas individuelle mais collective, c'est-à-dire décidée par une instance nationale en mesure de fournir des indications cohérentes aux représentants du système d'enseignement qui opèrent dans les organisations et groupes internationaux. Un système d'enseignement qui n'a pas cette capacité ne peut que se borner à exécuter des décisions qui lui sont imposées et à subir un programme qu'il paie avec ses contributions mais sur lequel il n'a aucun mot à dire. On pourrait parler dans ce cas d'un gâchis ou peut-être d'un bénéfice minimal qui consiste dans l'acquisition du « know-how » requis pour mener des évaluations dans le système d'enseignement.

Ce processus comporte donc la mise en œuvre simultanée ou successive de toute une série d'instances nationales et internationales ayant les compétences pour intervenir dans chacune des séquences d'une enquête (voir Tableau 4) dont les enchaînements et le suivi sont une condition essentielle de succès de l'enquête, c'est-à-dire d'un déroulement sans entrave dans le respect de critères méthodologiques qui permettent d'assurer la comparabilité des données.

Tableau 4 : Séquences de réalisation d'une enquête internationale

1. Décisions préliminaires :
    - Définition du cadre de l'enquête
    - Questions à poser
    - Population concernée
    - Objectifs à atteindre
    - Calendrier à tenir
2. Préparation de l'appel d'offre international
3. Choix du/des maître(s) d'ouvrage
4. Planification de l'enquête : construction des instruments
5. Administration des tests
6. Traitement des données
7. Vérifications nationales
8. Analyses et diffusion des résultats

Chacune de ces huit séquences sollicite des organismes spécifiques et des instances décisionnelles différentes aussi bien au niveau national qu'international. Les séquences ont donc une double face et un système d'enseignement qui souhaite participer activement à l'enquête et exercer une influence sur son déroulement doit être en mesure de gérer en synchronie ces interfaces. Faute de cette capacité, un système d'enseignement ou un pays ne pourra pas être en mesure d'agir sur le projet et d'en tirer un bénéfice majeur. Aussi bien les enquêtes de l'IEA que le programme PISA ou le programme ALL (Adult Literacy and Life Skills Survey) comportent ces séquences.

Nous allons maintenant examiner plus en détail le cas de l'IEA et celui de PISA pour mettre en évidence le comportement de la France dans ces deux programmes.

1 - ORGANISATION ET MODES DE FONCTIONNEMENT DE L'IEA DANS L'OPTIQUE DE LA PRESENCE FRANÇAISE

1.1 - L'ASSEMBLEE GENERALE DE L'IEA

L'organe où se prennent les décisions sur les évaluations et les études à mener au sein de l'IEA est l'assemblée générale qui est composée par les institutions membres de l'IEA. L'assemblée générale établit les orientations du programme de l'IEA. Elle admet des nouveaux membres, exclut les membres qui n'ont pas respecté leur devoir et désigne les comités chargés de mettre en œuvre les décisions. C'est à l'assemblée générale qu'il incombe de choisir le meilleur projet d'une étude et d'approuver son centre de coordination internationale. Par ailleurs, l'assemblée générale est le destinataire des rapports sur l'avancement des études et des enquêtes, sur les programmes de formation, sur les activités du centre de traitement des données, et sur les travaux des comités chargés d'accompagner chaque étude ou évaluation. Enfin, l'assemblée générale délibère sur les initiatives susceptibles d'améliorer la qualité des recherches de l'IEA. Elle recommande les démarches pour collecter les fonds nécessaires pour couvrir les frais des coûts internationaux de chaque étude, et elle approuve les comptes annuels de l'association. L'assemblée générale se réunit une fois par an, d'habitude pendant une semaine, au début de l'automne. Ont droit de vote au sein de l'assemblée générale uniquement les institutions membres de l'association. Dans le tableau 5 sont énumérés les lieux où se sont déroulées les assemblées générales de l'IEA depuis 1987 ainsi que le représentant officiel de la France au sein de cette assemblée. Il convient de rappeler que les membres de l'AG de l'IEA votent le lancement des études.

Tableau 5 : Assemblées Générales de l'IEA (dates et lieux)

Année	**Lieu**	**Délégué français**	**Institutions françaises responsables**
1987	New York, Etats-Unis	***Donnée manquante***	
1988	Frascati, Italie	***Donnée manquante***	
1989	Séoul, République de Corée	Daniel Robin	CIEP-Sèvre
1990	Pékin, Chine	Daniel Robin	CIEP-Sèvre
1991	Enschede, Pays-Bas	Alain Michel (observateur)	
1992	Monte Verità (Ascona), Suisse	Daniel Robin	CIEP-Sèvre
1993	El Escorial (Madrid), Espagne	Emilie Barrier	CIEP-Sèvre
1994	Djakarta, Indonésie	Alain Michel (observateur)	IGEN
1995	Riga, Lettonie	pas de repr. français	
1996	Vancouver, Canada	pas de repr. français	
1997	Sofia, Bulgarie	Gérard Bonnet	DEP-MEN
1998	Judean Hills, Israël	Gérard Bonnet	DEP-MEN
1999	Nesbru, Norvège	Gérard Bonnet	DEP-MEN
2000	Chiang Mai, Thaïlande	Gérard Bonnet	DEP-MEN
2001	pas de meeting	-	
2002	Marrakech, Maroc	Gérard Bonnet	DEP-MEN
2003	Lemesos, Chypre	Gérard Bonnet	DEP-MEN
2004	Taipei	Gérard Bonnet	DEP-MEN

1.2 - UNE ORGANISATION DECENTRALISEE DES PROJETS

Les membres de l'IEA ne sont que des institutions de recherche, aussi bien gouvernementales que non gouvernementales. Pour devenir membre de l'IEA, ces institutions doivent faire la preuve qu'elles ont les compétences et la capacité à conduire des enquêtes et qu'elles peuvent participer à des évaluations de ce genre ou qu'elles ont participé à de telles recherches ou qu'elles ont l'intention de participer à des études de l'IEA. L'assemblée générale de l'IEA est composée de représentants de ces institutions. L'assemblée élit, en rotation, sept membres du comité permanent, qui prend toutes les décisions entre les réunions annuelles de l'assemblée générale. Au début, le secrétariat de l'IEA exerçait les fonctions de centre de coordination des évaluations. A partir de la deuxième enquête internationale sur les mathématiques de 1982, chaque étude a eu son propre centre de coordination qui n'était pas localisé à la même place que le secrétariat. C'est donc en 1982 que l'IEA s'est, dans un certain sens, décentralisée et à partir de ce moment-là, que les études de l'IEA ont été dirigées par des centres qui ne coïncidaient pas avec le secrétariat de l'IEA. Par exemple, la deuxième enquête sur les mathématiques a été coordonnée par le centre localisé en Nouvelle-Zélande ; l'étude sur l'environnement de la classe, qui n'a pas abouti, avait son centre de coordination au Canada ; le centre de coordination de la première étude internationale sur la composition écrite a été tout d'abord aux Etats-Unis, puis s'est déplacé en Finlande ; le centre de coordination de la deuxième étude sur les sciences était en Australie, celui sur l'éducation préscolaire était aux Etats-Unis ; etc.

La conséquence de cette décentralisation dans l'organisation des enquêtes internationales au sein de l'IEA a eu comme effet une augmentation des coûts des enquêtes, car ce faisant, on multipliait le personnel en charge de la coordination internationale, chaque étude ayant le sien, y compris le personnel pour le traitement des données. Par ailleurs, toute communication entre les projets était rendue difficile en compliquant singulièrement l'échange de compétences et de connaissances méthodologiques mises au point en réalisant les études. Pour diminuer les coûts, ce n'est qu'à partir de l'étude sur la compréhension de la lecture en 1991 que l'IEA s'est dotée d'un unique centre de traitement des données. Enfin, à partir de 1997, au début donc du programme PISA, le secrétariat de l'IEA s'est établi définitivement à Amsterdam et a été structuré sur une base professionnelle. Avant cette année, le secrétariat bougeait d'un pays à l'autre jusqu'au début des années quatre-vingt-dix, lorsqu'il s'est établit à La Haye où il a toutefois fallu un certain nombre d'années avant de trouver une organisation stable, ce qui a été atteint avec le déplacement à Amsterdam.

Donc, un des problèmes qui a lesté l'IEA, et qui n'a pas été résolu par son assemblée générale pendant des années, a été la décentralisation de son organisation. Elle a empêché son secrétariat central de maîtriser aussi bien l'évolution scientifique que financière des enquêtes. La condition préalable d'un changement dans le sens d'une organisation centralisée était la constitution d'un secrétariat permanent stable. Cette décision a été prise relativement tard, après presque trente ans de fonctionnement, ce qui a été préjudiciable pour l'IEA. L'avenir nous dira s'il est possible de concevoir et de mener des études comparées axées sur des évaluations à grande échelle selon un modèle centralisé qui respecte la primauté de l'intérêt scientifique sur l'intérêt politique. A priori, cela n'est pas inconcevable.

La France n'a piloté aucune recherche d'IEA et n'a donc jamais été le siège d'un centre de coordination d'une enquête internationale. On peut se demander pourquoi la France n'a pas obtenu ou n'a pas demandé le pilotage d'une recherche de l'IEA. Vraisemblablement, l'Education nationale n'a jamais posé sa candidature pour assumer une telle responsabilité, cette attitude découlant automatiquement de l'absence de propositions françaises de recherche. On peut supposer que l'Education nationale n'a éprouvé aucun intérêt à promouvoir des études internationales comparées dans la mise en place de la DEP des évaluations internes, en sous-estimant de ce fait l'apport de connaissances sur son propre système provenant des études comparées au niveau international.

1.3 - FINANCEMENT DES PROJETS DE L'IEA

Comme on a déjà eu l'occasion de le dire, le problème qui a constitué une épine constante pour l'IEA a été celui du financement des enquêtes. Dès la première étude, l'IEA a été exposée à toutes sortes de déboires en ce qui concerne le financement et la recherche de fonds pour mener les enquêtes internationales à grande échelle. Toutes les études de l'IEA ont exigé des montages financiers compliqués. Dans la plupart des cas, ces fonds n'étaient pas gérés directement par l'IEA. Ils étaient versés aux Centres de coordination des différentes enquêtes internationales. Le secrétariat central de l'IEA n'était que partiellement informé par l'évolution financière des études, selon les situations plus ou moins critiques que les centres de coordination des études pouvaient rencontrer au cours de l'enquête. Si nous prenons par exemple l'étude sur l'éducation civique, celle-ci a été en grande partie financée par la « German Science Foundation » qui a versé directement l'argent à l'Université Humboldt de Berlin, le centre de coordination de l'étude. Cependant, d'autres contributions de moindre importance ont été aussi versées pour la même étude par de petites fondations privées américaines à l'Université de Maryland aux Etats-Unis, qui était le siège du comité directeur de l'étude. Le secrétariat central de l'IEA n'a que dans une partie minime participé aux coûts de cette étude.

Le problème du financement des enquêtes internationales a donc été une pierre d'achoppement constante pour l'IEA. Chaque institution de recherche participant à une étude, c'est-à-dire chaque système d'enseignement engagé dans une évaluation à grande échelle dans le cadre de l'IEA est supposé financer sa propre collecte des données, tandis que la partie internationale de l'étude est financée par un budget central. Au début de l'histoire de l'IEA, ce financement provenait presque exclusivement du gouvernement des Etats-Unis. La contribution américaine était par ailleurs complétée à chaque fois par des montants versés par des fondations privées comme la Ford Foundation, la Fondation Volkswagen, la Fondation de la Banque royale de Suède, ainsi que d'autres sources privées. L'IEA n'a jamais réussi à régler convenablement le financement de ses études, ce qui a en grande partie déterminé les problèmes que l'IEA a rencontrés pour assurer une production régulière de données selon un rythme stable. Elle n'a jamais trouvé non plus un module convaincant et stable pour financer les coûts internationaux des enquêtes : aurait-il fallu demander à chaque institution membre de verser une contribution annuelle pour financer ses coûts, ou échelonner cette contribution en fonction du nombre d'études auxquelles un pays ou une institution membre participait ? De même, l'IEA n'a que tardivement tranché sur les modalités de calcul des contributions des systèmes d'enseignement participant aux études. On verra qu'au contraire l'OCDE, qui est une institution des gouvernements, a d'emblée imposé une échelle des contributions selon des principes clairs. Il convient toutefois de rappeler que la différence entre l'OCDE et l'IEA est substantielle, étant donné que l'OCDE

est une organisation gouvernementale dont les fonds proviennent des budgets nationaux, tandis que l'IEA est une association de nature privée.

1.4 - SELECTION DES PROJETS DE L'IEA

L'identification et le choix des sujets des études internationales à grande échelle réalisées par l'IEA jusqu'au moment de l'entrée en scène de l'OCDE ne semblent pas répondre à une logique cohérente. Le choix des premières études a été effectué d'une manière informelle. Lorsque l'IEA était relativement petite et ne regroupait que 12 instituts comme c'était le cas au début des années soixante, il était plutôt facile de s'accorder sur les études à réaliser. Lorsque la taille de l'IEA gonfla jusqu'à dépasser la quarantaine d'institutions membres, il devint évident qu'il n'était plus possible de prendre une décision à l'amiable sur le programme des évaluations à grande échelle. Pour cette raison, l'assemblée générale décida de constituer un comité interne chargé de programmer les études et de trier les propositions sur la base d'un ensemble de critères mis au point par le comité et approuvés par l'assemblée générale. Ces critères sont les suivants :

1) Contribution probable du projet à l'amélioration des connaissances des systèmes d'enseignement, du fonctionnement des établissements, des stratégies d'apprentissage, de la théorie des programmes d'enseignement, et de l'amélioration des pratiques d'enseignement ;
2) Validité de la méthodologie proposée et tout particulièrement solidité des techniques d'enquêtes qui doivent être éprouvées et induire un niveau de risque d'erreurs minimales ;
3) Degré d'appropriation du projet par rapport à l'expertise et à la vocation de l'IEA, ce qui implique une vérification de la pertinence du projet par rapport aux ressources et aux prestations que l'IEA est susceptible d'offrir ;
4) Possibilité de financement du projet aussi bien au niveau national qu'international.

Les critères utilisés par l'IEA pour juger de l'opportunité d'engager une étude internationale à grande échelle écartent toute référence au rôle de l'éducation et de l'enseignement dans la société ou aux objectifs de l'éducation. Le premier critère fait référence aux paramètres d'ordre politique, mais aussi aux enjeux relatifs aux stratégies d'apprentissage. Le choix des enquêtes dépendra beaucoup de l'interprétation qui sera donnée à ce premier point.

Les propositions d'études sont examinées par le « future activities committee ». Si la proposition passe la rampe du comité, un deuxième comité ad hoc est constitué, composé par des spécialistes provenant de toutes les parties du monde dans le domaine de recherche propre au projet ainsi que par un représentant de l'IEA. Ce comité ad hoc a la responsabilité de rédiger une proposition de recherche détaillée. La rédaction de cette proposition demande en général une ou deux années de travail. Lorsque la proposition détaillée est prête, elle est distribuée aux membres de l'IEA, discutée dans le « future activities committee » et enfin elle est présentée à l'assemblée générale qui décide de l'accepter ou de la refuser.

La recherche des fonds pour financer l'étude commence après ce vote. Dans le cadre de l'IEA, cette opération demandait au moins deux ans. De ce fait, il fallait compter au sein de l'IEA sur un intervalle de temps d'au moins quatre-cinq ans entre le moment à partir duquel une esquisse de projet était présentée à l'association et le moment où on pouvait avoir la

certitude que l'étude pouvait être réalisée, car son financement était assuré sur la base des exigences posées par le déroulement et les spécifications méthodologiques de l'étude. Pour accélérer ce processus et réduire ce long délai de préparation, l'IEA a cédé à la tentation d'engager des études sans disposer d'une couverture complète de financement. Cette pratique a engendré une multitude de difficultés, car il a fallu courir derrière les financements potentiels pour couvrir les frais et pour empêcher le blocage d'un projet qui avait entre-temps démarré.

En conclusion, il y a un contraste évident entre le succès des études de l'IEA sur le plan international, scientifique et politique, et la fragilité foncière de l'IEA, déterminée par l'absence d'un programme de développement cohérent des études internationales à grande échelle, la précarité du dispositif de financement de ces études, et la décentralisation coûteuse de son organisation. Les enquêtes internationales de l'IEA semblent avoir été décidées au cas par cas, sur la base de quelques critères relativement simples, parmi lesquels celui d'ordre financier a souvent été déterminant. Il faut néanmoins reconnaître que le respect des critères méthodologiques a toujours été un critère important d'évaluation, ce qui a contribué à la renommée de ces enquêtes tout au moins au sein de la communauté scientifique.

Le processus de décision en vigueur au sein de l'IEA attribue un poids considérable aux membres de l'assemblée générale, ainsi qu'aux membres du comité ad hoc sur les activités futures. Les informations dont nous disposons ne nous renseignent pas sur le rôle des délégués français dans ces instances ainsi que sur les instructions qu'ils auraient reçues de l'Education nationale.

1.5 - LES COMITES DE L'IEA

Pour fonctionner, l'IEA a constitué une série de comités chargés de préparer et suivre les études pendant la période intercalée entre les assemblées générales annuelles. Ces comités chargés d'assurer et garantir le suivi et la qualité des études ainsi que le bon fonctionnement de l'ensemble de l'association ont eu une responsabilité considérable dans l'activité de l'IEA. Initialement, les comités suivants avaient été prévus :

– le comité technique, sur les enjeux relatifs à la méthodologie des études comparées et l'évaluation à grande échelle au niveau international, qui fonctionnera comme lieu d'expertise pour les nouveaux projets ;
– le comité du financement, chargé surtout de la recherche des fonds nécessaires pour financer les coûts internationaux des projets ;
– le comité du programme (future activities committee), pour l'examen et le tri des propositions d'étude (comité qui a disparu) ;
– le comité de la politique et des procédures, qui traite les enjeux et les retombées sur le plan politique des études effectuées et qui élabore des recommandations à l'intention de l'assemblée générale (comité qui a disparu) ;
– le comité des publications, qui détermine et supervise le programme des publications.

La France a été présente dans le comité permanent entre 1991 et 2000.

Tableau 6: Présences françaises dans les comités de l'IEA entre 1985 et 2004

Comités	1985-1990		1991-1995		1996-2000		2001-2004	
	Institutions membres	Délégués français	Composition	Délégués français	Composition	Délégués français	Composition	Délégués français
Comité permanent (Standing Committee)	Pas d'information	***Donnée manquante***	Nouvelle Zélande, Suède, Canada, Etats-Unis, **France**, Belgique, Hongrie, Corée, Espagne, Grèce, Norvège, Japon	***Donnée manquante***	Grèce, Nouvelle Zélande, Norvège, Etats-Unis, **France,** Belgique, Japon, Suisse, Israël, Hongrie, Philippines, Chypre, Pays-Bas	***Donnée manquante***	USA, Philippines, Belgique (Communauté flamande), Suisse, Lettonie, Pays-Bas, Japon, Slovenie, Finlande, Macedoine, Norvège	Aucun représentant français
Comité technique (Technical Committee)	***Donnée manquante***	***Donnée manquante***	***Donnée manquante***	***Donnée manquante***	Ray Adams, Keith Rust, WESTAT), Larry Hedges (Université de Chicago), Michael Martin (Boston College), Heiko Sibberns (DPC)	Aucun représentant français	Hans Wagemaker, Chair, Larry Hedges (University of Chicago), Pierre Foy (Statistics Canada), Eugene Johnson (AIR), Christian Monseur (Liege University), Jan-Eric Gustafson (Goteborg University), Michael Martin (Boston College, ex officio), Ina Mullis (Boston College, ex officio), Heiko Sibberns (DPC, ex officio)	Aucun représentant français
Comité des publications	Pas d'information	Pas d'information	Pas d'information	Pas d'information	Membres depuis 1996: Richard Wolf, (Université de Columbia), Geoff Masters (ACER), David Nevo (Université de Tel-Aviv), Hans Pelgrum (Université de Twente), Armin Gretler (Suisse)	Aucun représentant français	En 2004 Dick Wolf part. Il est remplacé par David Robitaille (Université de British Columbia) Reconstruction du Comité en cours.	
Comité de financement	***Donnée manquante***	***Donnée manquante***	***Donnée manquante***	***Donnée manquante***	***Donnée manquante***	***Donnée manquante***	***Donnée manquante***	

Source : Madame Barbara Malak, secrétariat IEA.
Note : Les données manquantes relatives à la France n'ont pas été communiquées aux auteurs.

1.6 - LE CAS DU PROJET TIMSS

Pour terminer cet excursus sur l'IEA, il nous paraît utile de fournir quelques indications sur le projet TIMSS dans le but de mieux appréhender la complexité de l'organisation aussi bien au niveau international que national pour réaliser un projet de grande envergure.

La planification de TIMSS a commencé en 1989. La première réunion des coordinateurs nationaux de la recherche s'est déroulée en 1990. La collecte des données s'est déroulée en 1994 dans l'hémisphère sud et au printemps 1995 dans l'hémisphère nord. Le premier rapport faisant état des résultats nationaux non désagrégés a été publié en 1996. L'opération TIMSS s'est donc étendue sur une période de sept ans, sans compter les années qui ont précédé la mise en œuvre du projet pendant laquelle la prise de décision a été préparée et la décision elle-même d'effectuer l'étude a été adoptée (1988).

Pour se rendre compte de l'envergure de cette recherche et donc de la taille de l'organisation nécessaire pour la réaliser aussi bien au niveau international qu'à l'intérieur des différents systèmes d'enseignement, il suffit de rappeler quelques données de base :

– Pays participants : 45 ;
– Niveaux d'étude testés : 5 (3^e, 4^e, 7^e et 8^e année ainsi que l'année terminale du secondaire) ;
– Population testée : plus d'un demi-million d'élèves ;
– Langues utilisées : les traductions ont été effectuées en plus de 30 langues ;
– Etablissements scolaires concernés ; 15'000
– Questionnaires élèves, enseignants et directeurs d'établissement scolaires : 1500 questions au total ;
– Personnes pour la passation des tests et le traitement des données : plus d'un million.

Ces dimensions ne peuvent pas être maîtrisées sans qu'il y ait : une organisation hautement structurée, un financement approprié, des organes décisionnels au sein desquels des compromis et des négociations sont réalisés pour parvenir à une entente sur le déroulement de l'enquête, son calendrier, les objectifs à atteindre, la manière de traiter et de présenter les résultats. Pour exercer une influence quelconque sur la conception, le déroulement, les analyses et les publications, il est indispensable de se doter d'une organisation appropriée. On peut se demander si le dispositif français est à la hauteur de l'expertise de la DEP, de la communauté française de recherche dans le domaine de l'enseignement des mathématiques et des sciences et dans le domaine psychométrique et enfin des conséquences politiques et pédagogiques que les résultats d'une enquête internationale de cette envergure peuvent provoquer.

2 - PLANIFICATION, CONSTRUCTION ET MISE EN ŒUVRE DU PROJET PISA DE L'OCDE DANS L'OPTIQUE DE LA PRESENCE FRANÇAISE

2.1 - DESCRIPTION DE L'ORGANISATION ET DE LA STRUCTURE DU PROGRAMME PISA DE L'OCDE

Nous allons maintenant nous concentrer sur le programme PISA de l'OCDE pour illustrer plus en détail sa structure, les modalités décisionnelles, les niveaux et processus de financement. Nous nous concentrons sur le programme PISA car tout d'abord il se démarque, comme nous l'avons déjà dit, des études de l'IEA par le fait même que l'organisme à l'origine du programme n'est pas une assemblée générale d'institutions de recherche, mais un organisme gouvernemental, dans ce cas précis le Comité de l'éducation de l'OCDE épaulé par le Comité directeur du Centre pour l'innovation et la recherche en éducation (CERI) de l'OCDE. D'autre part, la gestion et la structure organisationnelle du programme PISA de l'OCDE ont été conçues par des anciens de l'IEA. PISA a amplement profité de l'expérience et du travail réalisé au sein de l'IEA.

Comme on a déjà eu l'occasion de le dire, un système d'enseignement qui ne dispose pas d'un dispositif approprié d'intervention dans les différentes instances décisionnelles n'est pas en mesure d'exercer une influence sur la conception et l'évolution d'une recherche internationale et sera condamné à exécuter, voire à appliquer d'une manière passive les décisions prises par d'autres. La seule marge de manœuvre qu'un pays pourrait avoir dans ce cas de figure se résume à faire tout son possible pour respecter les normes de la comparabilité internationale afin de produire des données de grande qualité sur le plan national et ensuite d'interpréter ses propres résultats d'une manière libre, c'est-à-dire en les éclairant avec une panoplie de données contextuelles nationales que le niveau international n'est pas en mesure d'appréhender et donc encore moins d'intégrer dans ses analyses.

Le projet PISA a commencé à être planifié en 1997 après une phase relativement longue de gestation dans les coulisses du projet INES (International Indicators of Educational Systems), phase qui s'est grosso modo étalée entre 1994 et 1996. En avril 1997, le comité directeur du CERI-OCDE et le comité de l'éducation de l'OCDE prirent la décision de réaliser une enquête permanente à intervalle régulier sur les acquis et compétences des élèves, l'enquête PISA (acronyme de l'expression anglaise « Programme for International Student Assessment » – Programme pour l'évaluation internationale des élèves).

Après cette décision, qui donnait le feu vert pour mettre en œuvre le programme, la réalisation fut menée au pas de charge. Le 7 ou 8 mai 1997, au cours d'une réunion organisée à Budapest, à laquelle étaient présents les délégués de 34 pays (parmi lesquels la France) intéressés à participer à la réalisation du programme, furent prises les décisions suivantes portant sur :

– la structure de gestion du programme ainsi que le rôle opérationnel du Conseil des pays participants (Board of Participating Countries) chargé de piloter le programme ;
– l'ampleur et les domaines couverts par les instruments de l'étude ;
– les termes de référence pour lancer l'appel d'offre international ;
– la formule pour répartir les coûts internationaux de l'étude parmi les pays qui auraient adhéré au programme.

Au début d'octobre 1997 était rendues publiques les modalités de la procédure de l'appel d'offre international pour la collecte de données nécessaires à l'élaboration périodique d'indicateurs sur les résultats des élèves, et le 12 octobre 1997 l'appel d'offre international était adressé à un certain nombre de souscripteurs éventuels, identifiés par les pays membres de l'OCDE comme des instituts potentiellement en mesure de répondre à l'appel d'offre. Parmi les sept soumissionnaires potentiels était indiqué Monsieur Gérard Bonnet, chargé de mission à la direction de l'évaluation de la prospective du Ministère de l'éducation nationale[43].

La première réunion du Conseil des pays participants (BPC) eut lieu les 7 et 8 octobre 1997 à la veille du lancement de l'appel d'offre international. En même temps, le comité des pays participants a été invité à choisir cinq experts parmi une liste de spécialistes qui aurait dû constituer le groupe chargé d'examiner et d'évaluer d'un point de vue technique les dossiers qui auraient été soumis à la suite de l'appel d'offre international. La liste de ces experts répartis en quatre domaines de compétence se trouve dans le tableau 7. Il est intéressant de noter qu'il n'y a dans cette liste qu'un seul spécialiste français, Paul Dickes de l'Université de Nancy II. Plusieurs de ces experts signalés par les membres du BPC provenaient des cercles de spécialistes de l'IEA.

[43] Les autres soumissionnaires étaient les suivants : le Centre international d'étude TIMSS auprès du Boston College à Chestnut Hill, Massachusetts ; l'IEA ; l'institut hollandais pour le développement des tests SITO ; le centre hollandais OCTO ; le centre de recherche autrichien pour l'IEA à Salzbourg ; l'Australian Council for Educational Research (ACER).

Tableau 7 : Candidatures proposées par les pays membres au BPC (Conseil des pays participants) pour composer le jury du concours international pour la réalisation du PISA, répartis en quatre domaines de compétence, 1997

Domaine	Nom et prénom	Fonction/employeur	Pays
Contenu et mesure	Atash, Nadir	Conseiller indépendant	Etats-Unis
	Baumert, Jürgen	Max-Planck-Institut	Allemagne
	Bertrand, Richard	Laval University	Canada
	Binkley, Marilyn	National Center for Education Statistics	Etats-Unis
	Dickes, Paul	Université de Nancy II	France
Echantillonnage et mise en œuvre	Hambleton, Ron	University of Massachusetts, Amherst	Etats-Unis
	Lombardo, Giovanni	Université La Sapienza, Rome	Italie
	Posthlethwaite, Neville	Consultant indépendant	Royaume-Uni
	Ross, Kenneth	International Institute for Educational Planning, Paris	Australie/Unesco
	Trivellato, Paolo	Université La Sapienza, Rome	Italie
Gestion et financement	Elliott, Emerson	National Committee for Accrediting Teacher Education	Etats-Unis
	Gil, Guillermo	National Institute for Quality and Evaluation (INCE)	Espagne
	Konttinen, Raimo	Professor Emeritus, Research Institute of Education	Finlande
	Olkinuora, Errkki	University of Turku	Finlande
	Ryo, Watanabe	National Institute for Educational Research	Japon
Autres experts	Cordova, Fernando	Ministère de l'Education	Mexique
	Crocker, Robert	Expert indépendant	Canada
	Hawker, David	Curriculum and Assessment Division, Qualifications and Curriculum Authority	Royaume-Uni
	Hill, Peter	University of Melbourne	Australie
	van Hoesel, P.H.M.	Institute for Small Business Research and Consultancy	Pays-Bas
	Hutmacher, Walo	Consultant indépendant	Suisse
	Kettemann, Bernhard	Institut für Anglistik, Graz	Autriche
	de Landsheere, Gilbert	Université de Liège	Belgique
	Laroche, Léo	Expert indépendant	Canada
	Olkinovra, Erkki	Expert indépendant	Finlande
	Porter, Andrew	University of Wisconsin, Madison	Etats-Unis
	Reeff, Jean-Paul	Ministère de l'éducation nationale	Luxembourg
	Veenhuijs, J.B.M.	Consultant indépendant	Pays-Bas

2.3 - PRODUITS ATTENDUS DU PROGRAMME PISA

Dans les intentions de l'OCDE, du Comité de l'éducation ainsi que du Comité directeur du CERI, le programme PISA a été élaboré pour délivrer essentiellement quatre types de produits :

– un ensemble d'indicateurs de base susceptible de fournir aux décideurs un profil de référence de l'état des connaissances des élèves de leur pays ainsi qu'une description élaborée des caractéristiques des principaux sous-groupes d'élèves ;
– un ensemble d'indicateurs de contexte éclairant la manière avec laquelle les compétences et les connaissances des élèves peuvent être mises en relation avec des variables démographiques, sociales, économiques et éducatives ;
– des indicateurs de tendance qui seront disponibles grâce à la nature cyclique de l'étude qui permettra de collecter, selon un périodicité de trois ans, des données dans les mêmes domaines ;

– des connaissances approfondies sur le fonctionnement des systèmes d'enseignement qui pourront être utilisées pour élaborer des analyses poussées des politiques de l'éducation.

2.4 - FACTEURS DE VALIDITE D'UNE ETUDE INTERNATIONALE A GRANDE ECHELLE

L'analyse de l'organisation, du déroulement et des produits livrés par les études à grande échelle organisées depuis quatre décennies permet d'identifier cinq composantes déterminantes de réussite de ces études :

– les instruments d'enquête (tests, questionnaires, etc.) de haut niveau, structurés en fonction des objectifs d'étude ;
– les laboratoires qualifiés pour organiser l'enquête, effectuer les tests, recueillir les données, traiter les informations, fournir des rapports et livrer dans les délais les produits attendus ;
– une organisation internationale capable de payer la qualité de l'exécution de l'étude ;
– un dispositif de financement et de contrôle comptable indépendant et solide ;
– une stratégie d'information explicite.

Un pays ou un système d'enseignement sera plus ou moins influent dans la mesure dans laquelle il parviendra à placer des spécialistes dans les organes décisionnels de ces composantes. Il est évident qu'une opération de ce type ne mobilise pas uniquement des ressources scientifiques, mais aussi un réseau politique et administratif. Par ailleurs, les scientifiques engagés dans ces projets ne devraient pas opérer en vase clos, pour leur propre compte, mais garder des liaisons actives avec l'instance nationale qui pilote la participation du pays à l'enquête. Ces échanges sont vitaux aussi bien pour anticiper les choix stratégiques que pour former sur le plan interne une relève scientifique insérée dans le réseau mondial de recherche. Si ce volet ne fonctionne pas, un pays reste isolé et ne profite pas de ces travaux, tandis que tous les avantages sont empochés par les pays les plus puissants qui parviennent à former une élite scientifique aux frais des autres pays.

2.5 - FACTEURS DE SUCCES D'UNE ETUDE INTERNATIONALE A GRANDE ECHELLE - STRATEGIE ET OBJECTIFS A ATTEINDRE

L'exigence de disposer d'informations solides sur les produits d'un système d'enseignement afin de permettre tout d'abord aux décideurs de prendre des décisions en connaissance de cause et aux acteurs du système d'enseignement de rendre compte des ressources mises à leur disposition avant de demander des moyens supplémentaires pour la mise en œuvre de nouvelles réformes de l'enseignement est un facteur déterminant dans l'élaboration d'une stratégie de collecte, traitement et diffusion de données sur l'enseignement. Il n'est pas simple d'expliquer les écarts de résultats entre écoles qui opèrent dans un même système d'enseignement et qui reçoivent les mêmes ressources, une fois égalisées les différences entre élèves et entre enseignants, si on ne dispose pas de renseignements précis sur la structure, l'organisation, le fonctionnement d'un système d'enseignement. Les études internationales visant à produire des données comparables ont permis d'accomplir des avancées considérables dans la compréhension du fonctionnement des systèmes d'enseignement et l'identification des paramètres ayant une incidence significative sur les résultats scolaires.

Les progrès de ces dernières décennies dans les études sur l'évaluation des élèves ont mis en évidence l'inadéquation existante entre les pratiques encore en vigueur en matière de production d'informations sur les résultats d'un système d'enseignement aussi bien au niveau national qu'international et les exigences de connaissances pour piloter les systèmes d'enseignement. La stratégie qui détermine la conception d'une étude internationale à grande échelle doit tenir compte de cet état des choses et être élaborée de manière à combler les lacunes de type méthodologique repérées aussi bien au niveau national qu'international, mais pour ce faire il est indispensable de disposer de la « force de frappe » nécessaire afin de forcer les partenaires du projet à prendre en compte des variables ayant une valeur prédictive élevée mais dont l'inclusion dans les outils engendre des coûts supplémentaires ou des risques politiques pour certains décideurs.

L'avancée représentée par le programme PISA de l'OCDE consiste dans le fait que d'emblée l'OCDE s'est penchée sur la stratégie nécessaire pour dessiner le programme d'enquête. Cette réflexion a commencé bien avant que les travaux d'organisation de l'enquête ne débutent. Les premiers documents signalant la présence d'une réflexion sur la stratégie d'information remontent à 1994 et une première ébauche de stratégie d'information avait déjà été produite au cours de l'été 1995, après l'assemblée générale du projet « INES » qui s'était déroulée à Lahti en Finlande en juin 1995. Le document fondateur du programme PISA est un document stratégique élaboré par le réseau A du projet INES, c'est-à-dire par le réseau constitué par un ensemble de pays qui se sont volontairement regroupés afin de réfléchir sur la meilleure façon de produire des indicateurs sur les acquis et les compétences des élèves. C'est au sein de ce réseau, piloté par les Etats-Unis, qui à cet effet ont amplement financé les travaux de secrétariat de ce réseau, que la stratégie du programme PISA a été conçue et mise au point. Les conclusions de cette réflexion se trouvent dans le document DEELSA/ED/CERI/CD(97)4 du 28 mars 1997. Ce document, mis au point au sein du réseau A du projet INES de l'OCDE, a été discuté et approuvé par le Comité directeur du CERI et par le Comité de l'éducation de l'OCDE. Les représentants de la France dans les instances du programme PISA dès son début sont indiqués dans le tableau 8.

Tableau 8 : Délégué(e)s de la France dans les instances du programme PISA de l'OCDE, 1995-2000

Organes PISA	1995	1996	1997	1998	1999	2000
Groupe directeur de INES	Claude Thélot (DEP)	Claude Thélot (DEP)	Claude Thélot (DEP)	DM	DM	DM
Groupe des coordinateurs nationaux de INES	Claude Sauvageot (DEP)	Claude Sauvageot (DEP)	Claude Sauvageot (DEP)	DM	DM	DM
Réseau A de INES	Bernard Ernst (DEP)	Bernard Ernst (DEP)	Jacqueline Levasseur (DEP)	Jacqueline Levasseur	DM	DM
Comité des pays participants à PISA	DM	DM	DM	DM	DM	DM
Comité directeur du CERI	DM	DM	DM	DM	DM	DM
Comité de l'éducation	DRIC	DRIC	DRIC	DRIC	DRIC	DRIC

DRIC : Direction des relations internationales du Ministère de l'Education Nationale
DM : Données manquantes non communiquées aux auteurs.

Les instances du programme PISA citées dans le tableau 8 ont un poids considérable dans la définition de la stratégie du programme, dont les buts et le financement sont discutés et votés au sein de ces instances et souvent mis au point dans des discussions de couloir. Les

représentants français dans ces groupes ont donc participé à la détermination des orientations du programme PISA.

3 - DIMENSIONS CRITIQUES DU PROGRAMME PISA

La réalisation d'un programme international de recherche comporte un volet stratégique et un volet technique. Pour analyser l'engagement français, il est indispensable d'examiner d'une manière séparée ces deux dimensions, ce qui devait permettre d'esquisser les conditions à respecter pour une participation efficace, dynamique et rentable, aussi bien d'un point de vue scientifique qu'économique aux enquêtes internationales à grande échelle des acquis et des compétences des élèves ou des adultes.

3.1 - DIMENSIONS STRATEGIQUES

Pour mettre en œuvre une stratégie commune de collecte de données, il faut prendre en compte plusieurs paramètres et plusieurs dimensions qui convergent dans le processus décisionnel. Les pays participants doivent par exemple s'accorder sur un plan d'enquête, sur le type d'instruments à utiliser et à développer, sur la méthode de collecte de données. Sur tous ces points le débat est ouvert et l'accord n'est possible qu'à la suite de discussions prolongées rendant possible un compromis acceptable pour les pouvoirs publics engagés. Pour chacune de ces composantes, les négociations entre intervenants sont la règle. Elles se développent à tous les niveaux et de ce fait les pays participants doivent disposer des ressources nécessaires pour pouvoir être présents autour des différentes tables de discussions avec des représentants non seulement compétents dans les domaines traités mais ayant reçu des instructions uniformes émanant d'une stratégie nationale qui vise à atteindre des objectifs délibérés.

La recherche d'un compromis entre pays est inéluctable. Il n'est pas possible dans le cas des études internationales à grande échelle ayant une dimension comparée d'imposer à la plupart des pays un seul modèle ou une seule procédure. A ce propos, l'échec de l'expérience IAEP est éloquent et instructif. Nécessairement la méthode doit être souple pour s'accommoder des contraintes multiples et souvent divergentes auxquelles sont confrontés les pays participants. Le compromis consiste à négocier le respect des normes internationales de qualité sur le plan scientifique avec les exigences parfois particulièrement rigides en vigueur à l'intérieur d'un pays ou appliquées dans un système d'enseignement. Certains sujets, par exemple, ne peuvent pas être abordés dans les tests pour des raisons culturelles ou d'autres thèmes doivent être écartés pour des raisons juridiques relatives à la protection de la sphère privée.

3.2 - DIMENSIONS TECHNIQUES DU PROGRAMME PISA

Le volet technique comporte les éléments méthodologiques et opérationnels suivants :

3.2.1 - Méthodologie

Les parties constitutives de l'approche méthodologique peuvent être regroupées autour des composantes suivantes :

- domaines à étudier ;
- cycle d'enquêtes ;
- population visée ;
- l'échantillonnage ;
- questionnaires contextuels ;
- structure des instruments d'évaluation.

Nous allons ici passer rapidement en revue ces composantes pour mettre en évidence les choix opérés au sein du PISA.

(i) Domaines à étudier
Dans le programme PISA, il n'y a pas eu d'hésitation sur le choix des domaines à étudier. D'emblée, il a été décidé de se concentrer sur la lecture dans la langue de l'enseignement, les mathématiques et les sciences, parce que ces domaines intéressent les décideurs des pays de l'OCDE, parce qu'ils sont des domaines universels, et enfin parce qu'ils peuvent être mesurés partout.

(ii) Cycle d'enquêtes
Un autre choix caractéristique du programme PISA, comme nous l'avons par ailleurs déjà signalé, a été celui de mettre en œuvre un cycle d'enquêtes couvrant une période de 9 ans, une collecte de données étant prévue tous les trois ans pour évaluer une matière, différente chaque fois, considérée comme matière dominante, ainsi que deux autres matières considérées comme secondaires (qui consisteront en sous-ensembles d'items provenant des collectes de données antérieures sur la matière dominante). En recueillant les données tous les trois ans, on génère un flux ininterrompu de renseignements sur les acquis des élèves qui permet de mieux apprécier le rendement des systèmes d'enseignement à la fin de la scolarité obligatoire.

(iii) Population visée
Le programme PISA vise une population définie en fonction de l'âge et non en fonction du niveau scolaire fréquenté. Ce choix est une solution de commodité car au niveau international il est plus facile de construire un échantillon comparable d'élèves basé sur l'âge : pour obtenir des résultats comparables, tous les pays évalueront des élèves du même âge. Se différenciant des études internationales effectuées auparavant, le programme PISA a retenu comme âge de la population de l'enquête le plus élevé auquel la scolarisation est encore quasiment universelle dans les pays participants de l'OCDE. Cet âge se situe entre 15 et 16 ans. En général, dans la plupart des pays de l'OCDE, c'est à 15 ans que se termine l'obligation scolaire, ce qui permet de trouver les élèves auxquels faire passer les tests ou soumettre les questionnaires.

(iv) L'échantillonnage
Des négociations entre experts des systèmes d'enseignement sont inévitables pour parvenir à définir le plan d'échantillonnage qui détermine les élèves qui peuvent être exclus de l'enquête. Il s'agit par exemple d'élèves qui s'expriment dans une autre langue que celle du test ou qui fréquentent depuis moins de deux ans un établissement qui pratique la langue du

test, d'élèves des établissements de l'enseignement spécial, incapables de passer le test, ou d'élèves ayant des handicaps mais intégrés dans les classes normales. Les catégories d'élèves qu'on peut exclure du test, ainsi que la proportion de ces élèves, peut varier considérablement. Certains systèmes d'enseignement peuvent être tentés d'exclure le plus grand nombre possible d'élèves qu'ils considèrent faibles, ce qui les avantagerait par rapport à des systèmes d'enseignement qui filtrent moins leurs élèves. La délimitation de la population à inclure dans l'échantillonnage est une opération de la plus grande importance parce qu'elle a une incidence sur la comparabilité des résultats du test. En ce qui concerne la France, la responsabilité de l'échantillonnage a été assumée dans le cadre du programme PISA par la DEP[44].

Pour avoir une valeur représentative, la taille de l'échantillon a été d'emblée définie à raison d'environ 4500 élèves par pays. La taille réelle de l'échantillon requis pour chaque pays a été fixée par l'arbitre expert en échantillonnage désigné par le Conseil des pays participants parmi les noms proposés par le Consortium chargé de réaliser l'enquête. De ce fait, le représentant de la France dans le Conseil des pays participants a également eu un rôle dans le choix de l'arbitre expert en échantillonnage d'où dépend le jugement sur la taille et la validité de l'échantillon du système d'enseignement participant à l'enquête. Le nombre de niveaux d'échantillonnage et les méthodes de sélection des échantillons déterminent la composition de l'échantillon et les limites de l'exploitation des résultats. Dans le cas du PISA, l'Allemagne et la Suisse ont décidé d'avoir un double échantillonnage, l'un concernant la population de 15 ans comme requis par l'OCDE et un deuxième concernant la population de la neuvième année, indépendamment de l'âge[45]. Par ailleurs, pour faire en sorte que les mesures des acquis des élèves puissent être liées de manière fiable aux variables structurelles, l'échantillon d'élèves doit concerner 150 écoles au moins, ce qui permet de calculer la variance inter et intra-établissements des acquis des élèves. Par ailleurs, les élèves sélectionnés dans une école pour faire partie de l'échantillon peuvent être rassemblés dans une classe fictive le jour du test, mais on peut aussi décider de faire passer le test à la classe entière où se trouvent ces élèves. En Suisse, on a retenu cette deuxième option et choisi deux classes entières pour chacune des 150 écoles prises en considération, ce qui a d'un côté considérablement augmenté la taille de l'échantillon national avec une répercussion bien entendu sur le coût de participation à l'enquête, mais ce faisant on a pu réaliser des analyses fines sur l'effet-établissement et l'effet-classe sur les résultats, calculs que l'échantillon officiel minimal de l'OCDE ne rend pas possible. On voit ici le rôle important que le représentant d'un système d'enseignement peut avoir sur la définition de l'échantillon. Celui-ci n'est pas automatiquement imposé mais peut être négocié. Cependant, pour déterminer taille, configuration, composition de l'échantillon, il faut qu'à l'intérieur de chaque système d'enseignement il y ait un organisme de suivi qui dispose de l'autorité nécessaire pour donner des instructions à ses représentants dans les organes internationaux qui prennent des décisions de ce genre, faute de quoi l'échantillon est imposé de l'extérieur.

(v) Questionnaires contextuels

Pour appréhender les variables dont on suppose qu'elles exercent une influence sur les niveaux de compétences dans les disciplines de l'enquête, le programme PISA, comme par ailleurs toutes les enquêtes précédentes de l'IEA, a prévu une collecte d'informations contextuelles. Le programme PISA s'est borné a recueillir des informations au niveau des

[44] Les auteurs n'ont pas obtenu les noms du responsable de l'échantillonnage ni celui du service chargé de l'effectuer.

[45] L'échantillon helvétique lors de PISA 2003 comprenait 21.257 élèves et 398 écoles. L'échantillon de la Suisse romande (1,5 millions d'habitants) comprenait 9.561 élèves et 199 écoles. Les auteurs n'ont pas réussi à obtenir l'effectif de l'échantillon français.

élèves et au niveau de l'établissement et a renoncé à recueillir des informations au niveau des enseignants. Ce choix n'est pas le fruit d'un hasard, mais le résultat d'âpres discussions parmi les représentants des systèmes d'enseignement participant à l'enquête qui se sont déroulées au sein du Conseil des pays participants et qui ont amené à la constitution d'une majorité de pays contraire à l'inclusion dans le programme d'un questionnaire pour les enseignants, majorité appuyée d'ailleurs par l'OCDE. Dans les enquêtes de l'IEA, par exemple dans l'enquête PIRLS, au contraire on a prévu et mis au point un questionnaire à l'intention des enseignants.

(vi) Structure des instruments d'évaluation

La construction de ces questionnaires s'effectue au sein de groupes de travail spécialisés dans lesquels on discute l'architecture du questionnaire, sa structure, ainsi que la formulation des questions. Dans le cas du programme PISA on peut relever que le questionnaire sur les établissements soulève en général parmi les experts des perplexités qui ont néanmoins été balayées au moment de l'interprétation des données, car l'OCDE a établi des corrélations entre les réponses données par les directeurs des établissements dans le questionnaire sur les établissements et les résultats des élèves. Dans ce cas aussi, l'action des représentants ou des experts nationaux dans les groupes de travail qui mettent au point ces outils peut infléchir dans un sens ou dans un autre la nature de l'outil. Cependant, il va de soi que la force d'intervention dépend non seulement des compétences et capacités individuelles, mais aussi de l'habileté à négocier des alliances avec d'autres experts pour appuyer les positions qu'on voudrait faire accepter. En ce qui concerne le premier cycle d'enquêtes du programme PISA qui s'est déroulé en 2000, au questionnaire sur le contexte des élèves a été ajoutée en option facultative un questionnaire élaboré par le réseau A du projet INES sur la perception de soi. Les données recueillies grâce à cette option ont été par la suite amplement exploitée dans l'interprétation des résultats par l'OCDE car il s'est avéré que cette échelle était particulièrement fertile d'informations pour interpréter les résultats obtenus dans les tests.

3.2.2 - Opérationnalisation

L'efficacité de la participation au programme PISA dépend aussi de l'organisation interne du dispositif qu'un système d'enseignement développe pour réaliser l'enquête. La présence dans les différents organes de gestion du projet est un résultat de cette organisation. En ce qui concerne le programme PISA, un paramètre important à prendre en compte pour organiser le dispositif de suivi est son ancrage dans l'OCDE, car le programme PISA bénéficie de l'infrastructure juridique et financière de l'OCDE qui assure à l'ensemble du programme une solidité que les enquêtes de l'IEA n'ont jamais pu atteindre, mais cette caractéristique oblige les responsables des systèmes d'enseignement qui participent au programme à intégrer dans leur dispositif le réseau gouvernemental qui s'occupe des activités dans le domaine de l'éducation. Par ailleurs, le programme PISA s'insère dans une réflexion d'ensemble sur l'évolution des politiques de l'éducation des pays membres discutée au sein du Comité de l'éducation de l'OCDE qui est l'instance la plus élevée dans le domaine des politiques de l'éducation de l'organisation et du Comité directeur du Centre de l'OCDE sur l'innovation et la recherche en éducation. La complexité de ce modèle de gestion est par ailleurs amplifiée par le fait qu'au cours de toutes les étapes du programme, depuis la conception du projet jusqu'à sa mise en œuvre et à sa diffusion, les pouvoirs requis pour prendre les décisions nécessaires sont définis d'une manière précise. A chaque niveau de responsabilité correspond un niveau de décision équivalent.

En ce qui concerne le programme PISA, la structure et la gestion du projet, les rôles et les responsabilités des divers acteurs et la gestion des ressources mettent en jeu les instances suivantes :

– un comité national chargé de suivre l'enquête et de superviser son application au sein du pays ;
– le réseau A du projet INES ;
– le conseil des pays participants ;
– le groupe des experts fonctionnels ;
– le groupe consultatif et technique.

Passons maintenant en revue ces organismes au sein desquels la France est représentée pour en appréhender les compétences respectives.

(i) Le Comité national du programme PISA
Chaque pays a dû constituer un comité national composé de représentants des milieux nationaux de l'éducation, d'experts en éducation spécialisés dans l'évaluation des résultats scolaires et de décideurs politiques dans le domaine de l'éducation. Le rôle du comité national est de déterminer les modalités de participation au programme en donnant des instructions aux délégués nationaux qui représentent le système d'enseignement dans les différents organismes du programme. Ce comité peut formuler des appréciations sur la pertinence des instruments internationaux dans le contexte national, attirer l'attention sur les biais possibles de ces outils, mettre en évidence les problèmes de contrôle de qualité lors de la passation de l'enquête, discuter la composition de l'échantillon national, et déterminer la politique de diffusion des résultats à l'échelon du pays[46].

(ii) Le réseau A du projet INES :
Le réseau A du projet INES a été constitué au tout début du projet, en 1989 et depuis lors ce réseau est présidé par les Etats-Unis qui en ont assuré le secrétariat, ce qui signifie fournir un responsable du réseau, assumer le financement du fonctionnement du réseau et la responsabilité de la programmation des travaux. La fonction du réseau A est de définir, identifier et calculer des indicateurs sur les résultats de l'enseignement. La participation au réseau est facultative. Un pays décide de désigner un délégué pour ce réseau s'il a un intérêt à co-opérer avec d'autres pays de l'OCDE à la construction d'0un jeu d'indicateurs sur les résultats de l'enseignement. La participation à un réseau est financé par chaque pays participants. Les activités du réseau sont supervisées par le secrétariat de l'OCDE et les résultats des travaux du réseau sont présentés au groupe de direction du projet INES. Le réseau A a acquis une importance considérable avec le programme PISA qui a été conçu et élaboré au sein de ce réseau. La France a été présente dans ce réseau depuis le début avec un délégué désigné par la DEP. En 2003, le délégué français dans le réseau A était Thierry Rocher de la DEP.

(iii) Le conseil des pays participants
Le conseil des pays participants est « de facto » l'organe directeur du programme PISA. Chaque pays prenant part à un cycle d'enquêtes est représenté dans le conseil des pays participants. Les principales tâches de cet organe sont les suivantes :
- définir les objectifs d'action du projet d'évaluation et les matières à tester ;
- fixer les domaines prioritaires d'actions en ce qui concerne les indicateurs à élaborer ultérieurement pour être inclus dans le recueil d'indicateurs de l'OCDE « Regard sur

[46] Les auteurs n'ont pas reçu des informations sur la présence d'un comité national pour l'enquête PISA en France.

l'éducation », proposer l'élaboration des instruments nécessaires pour recueillir les informations dont on a besoin pour produire ces indicateurs et les types d'analyses à effectuer ;
- déterminer le champ de l'enquête, son étendue et approuver l'offre internationale pour sa réalisation ;
- inspirer les orientations des rapports d'analyse des résultats.

Les représentants auprès du conseil des pays participants sont désignés par les gouvernements des pays membres participants. On s'attend à ce que les membres de ce conseil aient une connaissance approfondie des projets d'évaluation des acquis des élèves et de leur relation avec la politique. C'est à ce conseil qu'incombe la responsabilité principale d'assurer la cohérence de l'ensemble du programme et des initiatives proposées pour le plan d'enquête. De ce fait, la désignation du représentant national au sein de ce conseil représente une étape importante dans la stratégie d'un pays qui souhaite agir activement sur le plan international pour infléchir le programme ou pour l'orienter dans une direction ou dans une autre. En 2003, le représentant français dans ce conseil était Gérard Bonnet de la DEP.

(iv) Les groupes d'experts fonctionnels

Les groupes d'experts ont la responsabilité de fournir à l'organisme ayant gagné l'appel d'offre pour la réalisation du programme les cadres théoriques et conceptuels nécessaires pour construire les instruments de l'enquête. Dans le cadre du programme PISA quatre groupes d'experts ont été constitués : un pour chaque matière testée et un chargé de la mise au point des questionnaires contextuels. La désignation s'opère sur la base de propositions émanant des systèmes d'enseignement participant à l'enquête. Par le truchement de ces groupes, tous les pays peuvent influencer les orientations du programme. Le mandataire chargé de la mise en œuvre du programme est responsable en dernier ressort de la réalisation de l'enquête, mais il est tenu d'écouter les pays participants et de leur donner une part active dans l'élaboration des instruments d'évaluation. Ceux-ci seront construits par le maître d'ouvrage après avoir trouvé un consensus avec les groupes d'experts respectifs. De ce fait, la désignation des experts est un enjeu scientifique, car il implique des choix théoriques qui vont façonner les outils d'enquête et un enjeu politique d'autant plus important que le nombre d'experts est limité, ce qui oblige les pays à trouver un accord sur quelques noms.

Dans le cadre du programme PISA, les groupes d'experts dans les trois disciplines sont composés de huit spécialistes tandis que le groupe d'experts s'occupant des questionnaires contextuels comprenait cinq experts. Pour PISA 2003, le groupe d'experts des mathématiques comprenait 10 membres. Le choix des membres de ces groupes est donc le fruit d'une alchimie particulière au niveau international. En effet, il incombe au maître d'ouvrage de proposer les nominations à l'OCDE qui désignera les experts en consultation avec le conseil des pays participants. Les possibilités pour les pays d'influencer la composition de ces groupes d'experts est plutôt limitée car, logiquement, les groupes d'experts comprendront surtout des spécialistes connus par le maître d'ouvrage, ayant sa confiance et celle du secrétariat de l'OCDE. Les orientations et la structure des questionnaires utilisés dans le programme PISA ont été en grande partie déterminées par les compromis atteints au sein des trois groupes d'experts fonctionnels dont la composition a été le fruit d'un ensemble de considérations dans lequel les facteurs géopolitiques et scientifiques ont été soigneusement calibrés en fonction d'autres paramètres dictés par des affinités culturelles et scientifiques.

Lors de la réunion des 27-29 janvier 1998, le groupe exécutif du Conseil des pays participants a réexaminé les propositions soumises par les pays participants et par ACER pour les désignations des experts des quatre groupes fonctionnels. Le choix final a été effectué par le président du conseil des pays participants (l'Américain Eugene Owen), en consultation avec le vice-président et le Secrétariat de l'OCDE. La composition finale de ces groupes a été établie par le président du conseil des pays participants et ACER, les pays membres du programme jouant un rôle plutôt anodin dans cette affaire.

La composition de ces quatre groupes était la suivante :

Tableau 9 : Composition des groupes d'experts fonctionnels en 1998 pour le cycle PISA 2000

Groupe Lecture		**Groupe Mathématiques**		**Groupe Science**		**Groupe Questionnaire de contexte**	
Irwin Kirsch, Président, ETS, Princeton	Etats-Unis	Jan de Lange, Président, Freudenthal Institute Université d'Utrecht	Pays-Bas	Wynne Harlen, Président, Scottish Council for Research in Education, Edinburgh	UK	Trevor Williams, président, Westat, Rockville	Etats-Unis
Marilyn Binkley, NCES, Washington D.C.	Etats-Unis	Raimondo Bolletta, CEDE, Frascati	Italie	Peter Fensham, Monash University, Melbourne	Australie	Roel Bosker, Université de Twente, Enschede	Pays-Bas
Judith Kadar-Fülop, Ministère de la culture et de l'éducation, Budapest	Hongrie	Sean Close, St Patricks College, Dublin	Irlande	Raoul Gagliardi, Université de Genève, Genève	Suisse	Aletta Grisay, Université de Liège, Liège	Belgique
John de Jong, Swets Language testing Unit, Arnhem	Pays- Bas	Maria Luisa Moreno Martinez, INCE, Madrid	Espagne	Mi-Young Hong, Korea Institute of Curriculum & Evaluation, Séoul	Corée	Stan Jones, Statistics Canada, Yarmouth Nova Scotia	Canada
Dominique Lafontaine, Université de Liège, Liège	Belgique	Mogens Niss, IMFUFA, Université de Roskilde, Roskilde	Danemark	Svein Lie, Université de Oslo	Norvège	Horst Lofgren, Lund University Malmö	Suède
Rainer Lehmann, Institut für Erziehungs-wissenschaften, Université Humboldt, Berlin	Allemagne	Kyung Mee Park, Chungbuk National University Séoul	Corée	Margarita Petrich Moreno, Ministère de l'éducation, Mexico	Mexique	Petr Mateju, Institut of Sociology, Académie des sciences Tchèques, Prague	République Tchèque
Pirjo Linnakylä, Université de Jyvaskyla, Jyvaskyla	Finlande	Thomas Romberg Université du Wisconsin, Madison	Etats-Unis	Senta Raizen, National Center for Improving Science Education, Washington D.C.	Etats-Unis	Erich Ramseier, Service de la recherche en éducation du Canton de Berne, Berne	Suisse
Peter Mosenthal, Université de Syracuse, Syracuse NY	Etats-Unis			Elizabeth Stage, Université de Californie, Oakland	Etats-Unis	Judith Torney-Purta, Université du Maryland, College Oark MD	Etats-Unis

Martine Rémond, INRP, Paris	France					Franz Weinert, Max-Plank Institute für Psychologische Forschung, Münich	Allemagne
Ryo Watanabe, National Institute for Educational Research, Tokyo	Japon					Douglas Willms, Université du New Brunswick, Fredericton	Canada

Ce tableau mérite les remarques suivantes :

En premier lieu, en ce qui concerne la répartition par pays, on note que sur 35 experts il y en a dix du continent nord américain (donc environ 1/3) ; ensuite, un seul expert français est inséré dans le groupe d'experts. Il s'agit de Madame Martine Remond de l'INRP de Paris[47] ; le groupe européen est consistant, car il comprend 19 experts, parmi ceux-ci trois Hollandais, trois Allemands, deux Belges, deux Suisses ; pour le continent asiatique, il y a quatre représentants, dont deux de la Corée, un du Japon et un de l'Australie. Compte tenu de la fonction des groupes d'experts, on peut s'interroger sur la procédure de sélection. Elle a été apparemment transparente, mais des pans entiers de la recherche en éducation ne sont pas représentés dans ces groupes.

Sur les 34 experts désignés par ACER et le Secrétariat de l'OCDE avec l'accord du conseil des pays participants il y a une seule Française. Si nous considérons le rôle important joué en France par la recherche sur l'enseignement des mathématiques, on peut être étonné qu'aucun expert français n'ait été désignée dans ce groupe[48]. La Suisse, qui n'a pas de tradition scientifique dans le domaine de l'évaluation internationale à grande échelle, a néanmoins réussi à imposer deux experts. Un résultat de ce genre n'est évidemment pas le fruit du hasard, sans rien enlever, ce disant, aux qualités intrinsèques de ces nominations.

(v) Le groupe consultatif technique

Le groupe consultatif technique assure la qualité technique du projet en supervisant le travail du maître d'ouvrage. Ce groupe constitue une espèce de forum dans lequel se retrouvent toutes les personnes qui exercent des fonctions opérationnelles importantes dans le déroulement du projet comme, par exemple, les principaux sous-traitants auxquels le maître d'ouvrage assigne des tâches techniques spécifiques. C'est au sein de ce groupe que les différents spécialistes ayant des compétences techniques se retrouvent et discutent les aspects méthodologiques du programme.

Les membres du groupe consultatif technique sont désignés par l'OCDE sur proposition du maître d'ouvrage. Le secrétariat de l'OCDE consulte le conseil des pays participants avant de désigner les membres du groupe consultatif technique. Encore une fois apparaît ici la place charnière de ce conseil et donc de ses membres, dont les relations avec le Secrétariat de l'OCDE constitue la clé de voûte qui détermine la force d'influence d'un pays.

[47] L'OCDE a aussi créé pour l'occasion un comité d'analyse des implications culturelles (cultural review panel) dans lequel la France était représentée par Pierre Vrignaud.

[48] Cette absence française a persisté par la suite, car le groupe d'experts en mathématiques qui a préparé le test de mathématiques pour PISA 2003, lorsque les maths étaient le domaine principal testé, ne comprenait aucun spécialiste français.

Il n'y a pas d'experts externes au projet qui font partie du groupe consultatif technique. La composition de ce groupe est restreinte aux experts qui participent au programme, car pour pouvoir intervenir en toute connaissance de cause dans les discussions et participer à l'élaboration des recommandations d'ordre technique et opérationnel, il est indispensable d'avoir une connaissance approfondie du programme. Cette hypothèse, en partie juste, présente néanmoins le risque d'engendrer un clan fermé de spécialistes maîtrisant les paramètres du programme.

Le conseil des pays participants peut toutefois faire appel à des compétences spécialisées d'appoint pour clarifier, le cas échéant, des enjeux méthodologiques controversés. Cependant, il est évident que le délégué d'un pays ou d'un système d'enseignement qui voudrait entreprendre une démarche de ce type doit disposer d'arguments considérables et trouver des alliés au sein du conseil des pays participants, ce qui n'est pas une opération particulièrement aisée ni du point de vue scientifique, ni du point de vue diplomatique.

Pour le groupe de conseillers techniques de PISA 2000 et PISA 2003 il n'y avait aucun expert français.

4 - L'ORGANISATION DU PROGRAMME PISA

Dans l'organigramme de PISA, deux composantes constituent les piliers du programme : les chefs de projets nationaux et le maître d'ouvrage.

4.1 - LES CHEFS DE PROJETS NATIONAUX

Les chefs de projets nationaux occupent une position clé dans la réalisation du programme PISA. Leur mission est celle de conduire les enquêtes dans le contexte national. De ce fait, ils sont en contact direct avec le maître d'ouvrage pour toutes les questions liées à la réalisation de l'enquête, comme par exemple la vérification des fichiers, le contrôle de l'échantillonnage, la vérification de la population exclue, etc. Les chefs de projets nationaux doivent être des spécialistes des évaluations et des enquêtes et doivent avoir la capacité d'organiser et de conduire l'enquête sur le plan national. Leur désignation est effectuée par les gouvernements des pays participants ou par les responsables des systèmes d'enseignement engagés dans le programme.

Il incombe notamment aux chefs de projet nationaux de vérifier la traduction des tests et des questionnaires, d'imprimer les outils selon le standard de présentation commun décidé par le maître d'ouvrage. Tous les documents préparés dans les pays sont examinés par un arbitre expert en collecte de données qui donne son avis par rapport au respect des critères de couleurs, de mise en page, d'ordre des questions, de forme des réponses et d'instructions.

De même, il incombe au responsable national du projet d'établir et de développer les contacts et la collaboration avec les établissements scolaires. Pour obtenir la collaboration des écoles et d'autres organismes administratifs il faut que leurs représentants soient informés des essais sur le terrain et de l'enquête principale ou invités à y prendre part. Chaque pays ou chaque système d'enseignement participant doit élaborer, fournir et diffuser des informations au sujet de l'étude. Pour aider les chefs de projets nationaux dans cette

tâche, le maître d'ouvrage prépare une brochure comprenant une explication des essais et fournissant les informations de base, mais chaque système d'enseignement est libre d'adapter ces informations selon la sensibilité de ses enseignants et de ses écoles. Tout autre document nécessaire pour solliciter l'accord et la collaboration des écoles doit être établi par les systèmes d'enseignement participants sous la direction ou la supervision du chef de projet national.

La constitution de l'échantillon national est également une responsabilité du chef du projet national qui doit prévoir un dispositif pour sélectionner les élèves dans chaque école conformément aux procédures mises au point par l'arbitre expert en échantillonnage. Les exclusions ne seront autorisées que pour des raisons convenues à l'avance et dans des proportions ne dépassant pas un pourcentage fixé à l'échelon international.

Il est prévu de vérifier attentivement les procédures d'échantillonnage pour éviter de s'apercevoir, après l'administration des tests, qu'un échantillonnage non comparable a été utilisé, ce qui exclurait les données du système d'enseignement en cause de l'analyse et de la publication.

Un pré-test est prévu sur le terrain un an avant l'enquête principale dans tous les systèmes d'enseignement participant au programme. Environ 800 élèves sont testés chaque fois. Cet exercice est conduit sous la responsabilité du chef de projet national qui assure la traduction des instruments et des manuels utilisés à cette occasion.

L'entente entre les différents chefs de projets nationaux est indispensable pour assurer des modalités d'administration des tests les plus identiques possibles dans chaque pays. Or, l'utilisation du personnel dans les écoles, l'organisation de la passation des tests elle-même, le rôle du personnel de chaque école dans cet exercice doivent respecter des critères les plus communs possible afin de garantir la comparabilité maximale du test. Il faut en premier lieu éviter que les élèves d'une école soient avantagés ou défavorisés par les modalités appliquées localement lors de la passation des épreuves.

Le chef du projet PISA 2000 en France a été Jean-Pierre Jeantheau (DEP), celui de PISA 2003 Anne-Laure Monnier (DEP).

4.2 - LA MAITRISE D'OUVRAGE

La réalisation du programme PISA ne peut pas être effectuée par le secrétariat de l'OCDE qui a d'autres responsabilités ni par un groupe ad hoc de pays qui se mettraient d'accord pour réunir les ressources nécessaires dans le but de réaliser le programme. Les exigences techniques à respecter pour faire en sorte que le test se déroule dans des conditions optimales le plus semblables possible dans les différents pays et dans plusieurs milliers d'écoles afin d'assurer un niveau de comparabilité élevé, exigent une organisation spécialisée. Par ailleurs, la quantité considérable de tests administrés (on dépasse ici les 300 000 élèves testés) exige la disponibilité d'un centre de calcul puissant en mesure de traiter rapidement les données, de fournir toutes les indications nécessaires concernant les résultats obtenus dans les différents systèmes d'enseignement, d'interagir avec les chefs de projets nationaux dans la vérification et le nettoyage de données avant de procéder à leur diffusion. Ces tâches sont donc mandatées à un maître d'ouvrage principal qui est chargé de l'exécution du contrat signé avec l'OCDE. Le choix du maître d'ouvrage est un exercice

complexe et délicat à la fois qui implique un appel d'offre international. L'organisme retenu est désigné à la suite d'une procédure d'évaluation. En ce qui concerne l'enquête PISA 1 en 2000 et l'enquête PISA 2 en 2003, le maître d'ouvrage principal a été l'« Australian Council for Educational Research » (ACER) qui a été choisi à la suite d'un appel d'offre international.

L'appel d'offre international pour la réalisation de la première enquête du programme PISA a été lancé le 9 octobre 1997. Dans un document du 12 octobre 1997, l'OCDE a envoyé une lettre d'invitation pour participer à l'appel d'offre international à sept institutions différentes et dont on peut supposer qu'elles avaient été mises au courant auparavant des grands traits de l'appel d'offre, car on ne pouvait pas imaginer qu'au cours de trente jours ces institutions auraient pu constituer des consortiums solides et répondre avec des soumissions élaborées à l'appel d'offre de l'OCDE. Les institutions invitées ont été les suivantes :

- le Centre international d'études TIMSS, situé au Boston College à Chestnut Hill (Massachusetts, Etats-Unis) ;
- l'IEA, notamment le secrétariat à Amsterdam qui était de toute façon le mandataire de TIMSS et le superviseur de l'enquête TIMSS coordonnée par le Centre international du Boston College ;
- l'Institut national hollandais pour le développement des tests (CITO) à Arnhem ;
- le Département d'éducation de l'Université de Twente à Enschede, Pays-Bas ;
- la Direction de l'évaluation et de la prospective du Ministère français de l'éducation nationale ;
- l'Institut autrichien des sciences de l'éducation qui était aussi le centre chargé de réaliser les enquêtes IEA en Autriche ;
- le Conseil australien de recherche en éducation (ACER).

Les commentaires suivants peuvent être formulés à propos de cette liste :

- l'IEA est également contactée mais il ne faut pas ignorer qu'à ce moment-là elle était présidée par un Hollandais, le prof. Tjeerd Plomb, enseignant à l 'université de Twente ;
- un centre national IEA, celui en Autriche, se trouve sur la liste, mais aucun autre centre national correspondant de l'IEA ;
- sur les sept invités, trois sont dans le giron de l'IEA et deux se retrouveront dans le consortium gagnant de l'appel d'offre (ACER et CITO) ;
- la DEP, qui est un organisme de l'administration, apparaît isolée dans un réseau composé par des institutions scientifiques plus ou moins en relation entre elles.

Ces institutions n'ont pas été choisies au hasard ; elles ont été repérées en fonction d'une stratégie concoctée par l'OCDE, mais qui est délicate à expliciter faute d'informations suffisantes. Nous n'avons pas d'éléments qui nous permettent de connaître les critères adoptés pour déterminer le cercle des candidats potentiels et de savoir si des consultations au préalable ont été menées, si des indications ont été adressées à l'OCDE ou si l'OCDE avait elle-même pris des contacts pour délimiter le cercle des candidats potentiels à la réalisation du programme PISA.

Trois concurrents ont répondu à l'appel d'offre international : l'Université de Bourgogne (France) (voir Annexe 5) ; le Boston College des Etats-Unis ; et l'Australian Council for Educational Research.

Le conseil des pays participants a adopté la procédure d'évaluation proposée par le secrétariat de l'OCDE qui prévoyait dans un premier temps une évaluation technique des propositions. A cet effet, le conseil des pays participants a choisi des experts internationaux pour effectuer cette première évaluation. Ce jury était composé des personnes suivantes :

– Nadir Atash (PARSA, Etats-Unis);
– Jürgen Baumert, du Max Planck Institute for Human Development and Education, Berlin, Allemagne ;
– Marilyn Binkley du National Center for Education Statistics des Etats-Unis ;
– David Hawker du Qualification and Curriculum Autority, Royaume-Uni ;
– Neville Postlethwaite, Consultant international britannique, ancien Président de l'IEA ;
– Jean-Paul Reeff, du Ministère de l'Education du Luxembourg.

Ces experts se sont réunis à Paris du 20 au 22 novembre 1997 avec les objectifs suivants :

– évaluer chaque dossier sur la base des critères d'évaluation établie dans le cahier des charges ;
– décrire les faiblesses et les points de force de chaque dossier toujours par rapport au cahier des charges ;
– préparer une description comparative des différentes approches proposées, des modèles de gestions et d'organisation de la proposition ;
– évaluer le rapport entre les coûts et la qualité technique de la proposition ;
– identifier les questions techniques nécessitant d'être clarifiées pour chacune des propositions.

Le conseil des pays participants a reçu le 26 novembre 1997 un rapport détaillé du groupe d'experts ainsi que les trois dossiers. Ces documents ont tous été rendus public à la fin de la compétition pour l'appel d'offres.

Prenant en compte la qualité technique du programme de travail proposé, la capacité d'organisation, les qualifications du personnel, les compétences de gestion ainsi que l'expérience antérieure et le devis soumis, le groupe d'experts a retenu la proposition soumise par l'Australian Council for Educational Research qui a obtenu 83 points sur 100 dans l'évaluation. A la deuxième place s'est située la proposition du Boston College avec 79 points, tandis que la proposition soumise par l'Université de Bourgogne a été considérée comme inadéquate sur plusieurs points importants du cahier des charges et de ce fait le groupe d'experts a suggéré de ne pas la prendre en considération. Cette proposition a obtenu 42 points au total.

Du point de vue des coûts, la proposition la plus chère était celle du consortium piloté par l'Université de Bourgogne dont le devis s'élevait à 54 millions de francs français ; la proposition la moins chère était celle du Boston College avec une demande de 43 millions de francs français (donc environ 10 millions de moins que la proposition française), et enfin la proposition de l'ACER était de 50 millions de francs français.

Les points forts de l'Australian Council for Educational Research identifiés par le jury international ont été les suivants :

– la proposition a démontré une compréhension claire des objectifs politiques du projet et aborde les tâches décrites dans le cahier des charges d'une manière correcte, mieux

encore, créative. Des principes clairs ont été établis pour le travail de développement qui prévoit un plan opérationnel détaillé ;

- la proposition offre une structure de gestion solide qui combine à la fois une gestion centralisée avec des mécanismes qui assurent une large collaboration internationale et une représentation de tous les pays dans les différentes composantes de la phase de développement et d'exécution du projet. En particulier, ACER a été en mesure de proposer cinq groupes d'experts fonctionnels de grande qualité qui constitueront le noyau intellectuel du projet donnant la garantie de la présence d'une expertise internationale à très haut niveau ;
- un élément caractéristique de la proposition de ACER est l'intégration, dans le cadre de développement, du domaine de la lecture comme domaine principal et des domaines des mathématiques et des sciences comme domaines secondaires. L'esquisse proposée par ACER a le mérite d'offrir une interaction entre ces trois domaines et de proposer un plan de développement consensuel et flexible. Par ailleurs, la proposition de l'ACER représente un pas en avant par rapport au projet TIMSS de l'IEA. Les solutions proposées par ACER sont à cet égard très convaincantes, car elles vont au-delà du cadre théorique de TIMSS pour aborder les questions sous un angle transdisciplinaire ;
- les pays membres ont exprimé plusieurs fois, pendant la phase de préparation du programme, leur souci de prêter une très grande attention à la vérification de la qualité de toute la procédure d'enquête, en tenant particulièrement compte des expériences de l'IEA et de l'étude IALS. Or, sur ce point la proposition de l'ACER se révèle la meilleure. En effet, elle comprend une procédure de contrôle de la qualité, y compris un programme étendu de visite des sites, aussi bien des centres nationaux que des écoles engagées dans l'étude aussi bien pendant le pré-test que pendant l'enquête principale, qui est susceptible d'améliorer considérablement la qualité de l'étude ;
- enfin, la proposition de l'ACER offre un dispositif statistique et des procédures d'analyses susceptibles d'assurer un compte rendu approprié des résultats de l'enquête. Les propositions d'amélioration des outils renforcent la qualité des mesures, ainsi que le potentiel analytique des instruments qui seront administrés aux élèves. L'ACER a par ailleurs bien compris qu'il était nécessaire d'atteindre un consensus dans le développement des outils et dans la mise au point des procédures pratiques de travail ;
- un dernier point positif pour l'ACER a été l'exploitation des nouvelles technologies pour concevoir, gérer, piloter le projet et traiter les données. De ce fait, le projet aurait été utilisé pour transférer des technologies aux pays membres et pour échanger les technologies entre maître d'ouvrage et systèmes d'enseignement, ce qui a été considéré comme une valeur ajoutée de la proposition.

Après l'évaluation du jury international, la procédure prévoyait une deuxième étape de nature politique. Celle-ci s'est déroulée les 15/16 décembre 1997, également à l'OCDE. Les buts de cette étape étaient les suivants :

- réviser les propositions de l'appel d'offres international à la lumière de l'examen technique effectué par le jury international ;
- fournir des indications pour sélectionner le maître d'ouvrage avec qui initier une négociation pour finaliser le contrat. Ce travail a été délégué au groupe exécutif du conseil des pays participants constitué par un nombre restreint des membres de ce conseil. Nous ne savons pas si la France a fait partie de ce groupe exécutif ;
- établir la procédure à suivre pour permettre au groupe exécutif de négocier le contrat.

Lors de cette réunion, le conseil des pays participants a écouté les représentants de l'ACER et le représentant du Boston College, qui avaient été invités à présenter leur projet, tandis que le troisième concurrent, l'Université de Bourgogne, avait été écarté. Après chaque présentation, les membres du conseil pouvaient poser des questions pour cibler ou clarifier la nature des propositions. Trois membres du jury étaient aussi présents à cette réunion pour expliquer la procédure qu'ils avaient suivie pour évaluer les propositions et donner leur avis.

A la suite de ces présentations, la proposition de l'ACER a reçu le meilleur accueil. De ce fait, le conseil des pays participants a décidé à l'unanimité de mandater son groupe exécutif à engager des négociations avec l'ACER dans le but de parvenir à l'établissement d'un contrat. Cependant, au préalable, l'ACER a été invité à clarifier un certain nombre de points de sa proposition, notamment :

– la mesure de la littératie et l'approche proposée pour évaluer les compétences en lecture. L'ACER a été invitée à fournir la garantie que son approche prenait en compte les tous derniers développements en la matière et en particulier les meilleures idées de la proposition de ce que l 'OCDE a appelé par la suite le « Consortium européen », mais qui n'était en fait que la proposition de l'Université de Bourgogne ;
– la manière pour assurer l'équivalence conceptuelle et psychométrique des outils à développer ;
– le développement équilibré de la composante mathématique et scientifique de l'évaluation dans un texte centré sur les compétences en lecture ;
– le recours aux méthodes et aux technologies les plus avancées pour le développement des échelles d'évaluation et pour assurer l'établissement de séries historiques solides ;
– les procédures de recours envisagées dans le cas où on ne serait pas parvenu à un consensus sur le plan international, en particulier à l'intérieur de la communauté scientifique représentée dans le groupe d'experts fonctionnel ;
– la transparence la plus complète à tout moment de la procédure des méthodes mises en œuvre, insistant sur la nécessité d'assurer le transfert de compétences entre pays ;
– la manière avec laquelle ACER envisageait de traiter les écarts constatés en matière de procédure d'administration des tests ainsi qu'en ce qui concerne leur impact sur la comparabilité des résultats. Le conseil des pays participants en particulier exigea une grande rigueur en ce qui concerne la composition de la population du test afin qu'elle soit la plus consistante parmi tous les pays participants ;
– l'inclusion d'options nationales dans le test qui ne devait pas se faire au détriment de la cohérence d'ensemble des instruments et de la comparabilité des résultats ;
– la classification des catégories socio-économiques qui devait être perfectionnée ;
– le profil des chefs de projet nationaux dont on souhaitait une formulation explicite ;
– la réduction de 10% du devis par rapport à celui présenté.

Les réponses de l'ACER aux points soulevés par le conseil des pays participants furent distribuées à tous les pays le 14 janvier 1998 pour examen et commentaire. Le groupe exécutif du conseil des pays participants se rencontra le 27 et 28 janvier 1998 à Washington pour finaliser les négociations avec l'ACER. A cette réunion, le groupe exécutif parvint à obtenir un accord satisfaisant avec l'ACER sur les points substantiels du contrat. Le 31 janvier 1998, le contrat fut distribué à tous les membres du conseil des pays participants, tandis que l'OCDE régla avec l'ACER les questions légales et administratives. Sur la base de cet accord entre l'OCDE et l'ACER, le contrat fut signé le 2 février 1998. L'ACER y figurait

comme maître d'ouvrage principal à la tête d'un consortium qui comprenait les organisations suivantes :

– l'Australian Council for Educational Research (ACER);
– the Netherlands National Institute for Educational Measurement (CITO);
– le Service de pédagogie expérimentale de l'Université de Liège (SPE) ;
– WESTAT (il s'agit d'une entreprise privée de recherche spécialisée dans les enquêtes statistiques qui travaille essentiellement pour le Gouvernement américain et dont le siège se trouve à Rockeville dans le Maryland où sont occupées environ 1700 personnes).

Pour conclure cette partie dédiée à la description de l'organisation du programme PISA, une remarque s'impose : la mise en œuvre de cette opération a été effectuée avec une rapidité extraordinaire. Tout a été mis sur pied en sept mois, entre juin 97 et janvier 98, y compris la préparation de l'appel d'offre, son lancement, l'évaluation des soumissions, le choix du maître d'ouvrage, la sélection des experts. Si l'on considère la taille du projet, on ne peut que rester étonnés face à la vitesse avec laquelle l'opération a été menée. On peut supposer que cela a été possible car des travaux préparatoires ont été accomplis ailleurs, en dehors du processus officiel décrit et présenté dans les documents OCDE. Le réseau A du projet INES était entré en matière depuis longtemps, mais ce n'est pas au sein du réseau A que les détails de la machine organisatrice ont été mis au point. Par ailleurs, le délai d'un mois laissé pour constituer un consortium international en mesure de soumettre une proposition pour conduire une étude est particulièrement court et à cet égard on peut émettre l'hypothèse que le consortium européen, piloté par l'Université de Bourgogne, n'a pas eu le temps suffisant pour élaborer la proposition ou n'a pas eu les informations ou les indications appropriées pour entamer une réflexion au préalable lui permettant d'être prêt avec une proposition élaborée et avec des sous-traitants identifiés au moment opportun, ou que la nature du consortium et les modalités de sa composition n'étaient pas compatibles avec la procédure et les temps imposés par l'OCDE qui supposaient comme interlocuteurs des pôles de recherche structurés autrement ou déjà existants.

On peut ainsi se demander si au-delà de la relative transparence du processus, l'opération qui a mobilisé une partie importante d'anciens collaborateurs de l'IEA n'ait pas été conçue et élaborée dans d'autres instances que celles officiellement reconnues. C'est un problème important de gestion de la recherche internationale qui mériterait d'être exploré ultérieurement.

5 - LE CAS DU CONSORTIUM EUROPEEN PILOTE PAR L'UNIVERSITE DE BOURGOGNE

(voir Annexe 5)

Il est appréciable de constater qu'avec ce premier appel d'offre international la France ait réagi de manière positive et qu'une proposition émanant de la France, celle présentée par l'Université de Bourgogne, ait été soumise à l'OCDE, bien qu'elle ait été assez rapidement écartée par le groupe d'experts chargé d'évaluer les propositions ainsi que par le conseil des pays participants. Cependant, il est nécessaire de noter ici que pour la première fois en plus de 35 d'études internationales à grande échelle en éducation, la France prenait l'initiative de

présenter une proposition, de la développer et de poser sa candidature pour effectuer une enquête internationale à grande échelle. Cette démarche représente une nouveauté qui mérite d'être soulignée. En effet, si nous prenons en considération l'expertise accumulée en France, en particulier auprès de la DEP, en matière d'évaluation à grande échelle des élèves, rien ne justifiait l'absence de la France sur le plan international en tant qu'acteur principal dans la mise en œuvre des études internationales à grande échelle et dans la détermination des politiques d'évaluation concordées sur le plan international. Avec le programme PISA, en 1997, la France rompt avec le passé et se présente sur la scène avec une proposition propre, quoique enrobée dans un consortium de partenaires européens. Le fait qu'elle ait été balayée peut se justifier en partie par l'absence de contacts internationaux, l'isolement relatif de la France au cours des trente premières années dans ce domaine, la faiblesse des relations établies par les chercheurs français au niveau international dans ce domaine et peut être aussi par des insuffisances de la recherche en éducation dans le domaine de l'évaluation à grande échelle en France, à l'exception des travaux menés par la DEP et quelques laboratoires universitaires.

CHAPITRE III - COUTS, FINANCEMENTS ET ENCADREMENT LEGAL

La grande différence entre les enquêtes IEA et le programme PISA de l'OCDE se situe très vraisemblablement au niveau de l'organisation du financement, de la comptabilité et du cadre législatif qui discipline la participation à l'étude, les droits et devoirs des pays engagés dans la recherche.

En ce qui concerne le programme PISA, les aspects financiers et légaux sont pris en charge par l'OCDE. C'est ainsi que les instruments légaux de l'OCDE ont été utilisés pour mettre au point le règlement de l'étude et que les services administratifs et financiers de l'OCDE ainsi que le service comptable de l'organisation se sont occupés du financement et ont déterminé les contributions des pays membres. De ce fait, le programme PISA se déroule dans un cadre budgétaire extrêmement solide qui engage les gouvernements des pays participants et qui impose aux pays des obligations précises, négociées et clarifiées à l'avance, selon des règles communes à l'ensemble des projets de l'OCDE.

Tableau 10 : Coûts internationaux des études de l'IEA (estimations)

Etudes internationales	**Coût international global annuel**	**Durée de l'étude en années**	**Coût international total**	**Contribution annuelle aux coûts internationaux demandée aux institutions participantes (ticket d'entrée)**
Compréhension de la lecture (Reading Literacy)	250 000 US$	4	2 millions US$	Pas de données
Première enquête internationales sur les nouvelles technologies à l'école (Computers in Education)	Pas de données	4 +1	Pas de données	Pas de données
L'éducation pré-scolaire (PPP : Pre Primary Education) Troisième phase (enquête longitudinale)	Pas de données (étude réalisée sous les auspices de l'IEA, mais sans recevoir des fonds de sa part)	17	Pas de données	Pas de données
Troisième enquête internationale sur les mathématiques et les sciences (TIMSS)	4000 000 US$ (?)	4 + 2	Pas de données	Pas de données
Troisième enquête internationale sur les mathématiques et les sciences – Réplique (TIMSS-R)	Pas de données	4	Pas de données	40,000 US$
Deuxième enquête internationale sur les nouvelles technologies à l'école (Computers in Education) (IEA-SITES)	Pas de données	3 + 3	Pas de données	4,000 US$
Education civique (CIVED)	Pas de données	3 + 4	Pas de données	5,000 US$
Progrès dans la compréhension de la lecture à 9 ans (CM1 (enquête PIRLS), 2001	Pas de données	5	Pas de données	20, 000 US$
Progrès dans la compréhension de la lecture à 9 ans (CM1 (enquête PIRLS) 2006	Budget en cours de préparation	5	Budget en cours de préparation	30, 000 USD

La démarche appliquée à l'OCDE pour le programme PISA a été tout autre que celle de l'IEA. D'emblée, il était clair que pour éviter les travers financiers de l'IEA il fallait assurer le financement du programme avec un cadre solide imposant une formule de répartition des frais internationaux de l'enquête basée sur l'échelle 1 des contributions nationales à la partie 1 du budget de l'OCDE, ce qui permettait d'éviter toute contestation sur les coûts et d'engendrer un risque de déficit ou d'insuffisances de moyens pour la réalisation de l'étude. Deux exceptions ont été reconnues :

– les coûts auraient été, logiquement, distribués uniquement parmi les pays qui auraient accepté de participer au projet ;
– un plafond des contributions aurait été adopté pour tenir sous contrôle les dépenses et empêcher un gonflement des dépenses, chemin faisant.

Pour le premier cycle d'enquête du programme PISA, le plafond des contributions nationales a été fixé de manière telle a être équivalent à 40.000 $ par an pour la durée du premier cycle qui a été fixée sur quatre années, en y incluant les études préliminaires et l'année finale de diffusion des résultats. Ceci signifie que les pays de l'OCDE ayant décidé de mettre en œuvre le programme PISA et d'y participer avaient d'emblée accepté de fournir 160.000 $ américains chacun pour la réalisation de l'enquête et de couvrir ce montant avec leurs ressources propres. Il a été aussi convenu que ce plafond aurait pu être ajusté en fonction de l'évolution du deuxième cycle d'enquête.

Bien évidemment, si un des pays membres de l'OCDE refuse de participer à l'étude, la contribution des autres pays est augmentée de manière proportionnelle pour compenser cette défaillance. Donc, la clé de base pour le financement du programme PISA est l'échelle de la participation des pays au budget commun de l'OCDE et la source de financement est publique, car l'argent est versé par les gouvernements.

En plus de la contribution aux coûts internationaux de l'étude, coûts en grande partie déterminés par les frais du maître d'ouvrage, chaque système d'enseignement engagé dans l'étude doit calculer ses propres coûts pour l'organisation et la réalisation de l'enquête sur le plan national, ce qu'on convient d'appeler les coûts nationaux de l'enquête. En ce qui concerne la France, les coûts nationaux pour PISA 2000 ont été de $450.000. A cette somme il convient d'ajouter $287.500 pour la contribution aux coûts internationaux. Les coûts totaux pour la France de la participation à PISA 2000 ont donc été de $750.000. Ces chiffres approximatifs donnent un ordre de grandeur acceptable qui permet de comprendre l'importance du volet financier dans l'organisation de la participation à ce projet. Les sommes en jeu sont importantes, ce qui devait obliger à inclure dans le dispositif national qui négocie le profil de l'enquête et qui la réalise sur le terrain, des experts compétents sur le plan financier et comptable, travaillant en étroite liaison avec les experts du domaine pour pouvoir apprécier les implications des propositions formulées par le maître d'ouvrage à l'OCDE[49].

Enfin, ces chiffres ne comprennent pas les coûts des options facultatives qui permettent d'ajouter au programme principal des études complémentaires. Nous n'avons pas à disposition un tableau exhaustif ou suffisamment complet des modules optionnels proposés lors du test PISA 2000 ni sur le nombre des systèmes d'enseignement qui ont profité de cette possibilité.

[49] Par exemple, pour PISA 2006, les pays participants ont reçu la proposition de financer l'inclusion dans la panoplie des instruments d'un questionnaire adressé aux parents dont le coût survient à environ €250.000 par pays.

Dans le contrat, établi entre l'OCDE et le mandant ACER pour la réalisation du premier cycle du programme PISA, on a proposé un contrat au mandant ACER de FF 49.753.878 (on était en 1998 et l'Euro n'avait pas encore été introduit) net des taxes avec une première tranche de FF 11.138.123 pour 1998.Il s'agit de sommes considérables à la hauteur d'un projet de recherche à l'échelle mondiale de grande envergure qui permet à un certain nombre d'institution faisant partie du consortium piloté par ACER de pouvoir se développer et se positionner dans l'espace éducatif mondial comme des centres ou des pôles spécialisés dans les évaluations à grande échelle et sur les études internationales comparées.

D'autre part, au passage, l'OCDE prélève également sa partie du budget international estimé à FF 6.400.000 environ pour la période 1998-2001.

CHAPITRE IV - LA FRANCE ET LES ENQUETES INTERNATIONALES

1 - HISTORIQUE DES ENQUETES SUR L'EVALUATION DES ELEVES EN FRANCE

Les enquêtes nationales sur les compétences et les acquis des élèves, ainsi que sur les déterminants de la réussite scolaire, ont précédé, en France comme dans d'autres pays, les enquêtes internationales. En France, on peut dater de telles recherches des années 1930. De par ses intérêts scientifiques et ses compétences, l'INETOP (Institut National d'Etude du Travail et d'Orientation Professionnelle, fondé en 1928 par le psychologue Henri Pièron ; sur ces points voir Vrignaud, à paraître) a été un des premiers laboratoire français à organiser des enquêtes sur de grands échantillons pour recueillir des données psychologiques, sociologiques et pédagogiques sur les enfants et adolescents d'âge scolaire. Ces travaux d'enquêtes ont permis de recueillir quantité d'informations scientifiques utiles pour mettre à l'épreuve des hypothèses sur le développement cognitif et personnel en relation avec les apprentissages et la vie scolaire.

1.1 - LA DOCIMOLOGIE

Pour Henri Piéron, les recherches concernant la notation des enseignants et les examens, étaient inséparables des recherches sur l'évaluation psychométrique. Il s'agissait de montrer le manque de fiabilité des premières par rapport à la meilleure fiabilité des secondes. Dès 1927 Piéron écrit : « la Psychotechnique doit comprendre une branche consacrée à l'étude critique des procédés classiques de sélection fondés sur les examens et concours d'ordre scolaire, à la 'docimologie' ».

Dans les premières études, à l'initiative de Pièron, à la fin des années 1920, les chercheurs observent un manque de fiabilité dans les évaluations scolaires, d'une part par l'observation concrètes de pratiques réelles d'évaluation (exemple : l'analyse de jurys d'examen), d'autre part par la mise en place d'expériences spécifiques (exemple : expériences de multi-corrections de copies). Des synthèses de ces recherches ont été publiées à plusieurs reprises par Piéron puis par Reuchlin et ses collaborateurs (en particulier en 1935, puis en 1958). En 1963, Pièron publie son ouvrage devenu classique : « Examens et docimologie ». Les principaux constats portent sur le manque de fiabilité des notations et mettent en évidence dans la notation de copies d'examens et de concours, trois biais principaux :

1° sur le niveau de sévérité,
2° sur l'utilisation de l'échelle de notation,
3°sur le classement des copies.

De plus, ces biais sont indépendants les uns des autres. Ils sont attribuables à la subjectivité de l'examinateur et mettent en évidence un défaut de fidélité et de validité de la notation.

Pour améliorer la fidélité, Piéron propose de donner aux examinateurs une formation à la pratique de l'évaluation. Par la suite, on s'est proposé de recourir à une docimologie positive.

La docimologie positive vise à améliorer les pratiques d'évaluation en s'appuyant en particulier sur la construction d'épreuves d'évaluation plus objectives : questionnaires de connaissances scolaires et épreuves psychopédagogiques. Ce programme de construction d'épreuves objectives a certainement joué un rôle d'initiateur dans les premiers travaux sur l'évaluation des acquis des élèves par le Ministère de l'Éducation Nationale, qui seront évoqués plus loin.

Il reste que le recours à des épreuves objectives pour les évaluations certificatives a été peu suivi en France. Nous avons discuté des raisons de ces réticences dans un article récent (Vrignaud, 2003). On retrouve dans la littérature internationale les arguments développés dans le contexte français à travers les lignes de tension existant entre objectivité et authenticité des épreuves. Au nom de l'authenticité, les pédagogues privilégient les évaluations à partir de productions des élèves jugées plus proches de la mise en œuvre de ce qui est enseigné. Notons cependant qu'à côté des QCM s'est développée une recherche d'élaboration d'items objectifs à réponse construite, pour faire pièce à cette prise de position des pédagogues.

L'objectivité obtenue par une recherche de standardisation du questionnement n'est pas suffisante en elle-même pour assurer la validité de l'évaluation : un accord des juges sur les contenus à évaluer est aussi nécessaire (cf supra : 3ème source de biais), ce qui amène à prendre en compte les objectifs pédagogiques, atteints ou non, par les élèves. Ces objectifs doivent être formulés de façon opérationnelle, i.e. en termes de performances de l'élève, identifiables de façon univoque. L'Inetop a initié la diffusion en France de cette approche de l'évaluation par les objectifs qui s'est concrétisée, par la suite, notamment dans le courant de l'évaluation formative (voir notamment Bacher, 1973 ; Bonora, 1972, 1973, 1996). Les attitudes des enseignants à l'égard des objectifs pédagogiques de leur discipline ont fait l'objet à l'Inetop d'une recherche qui a mis en évidence l'existence de divergences au sein de cette population, en ce qui concerne leurs options pédagogiques fondamentales, divergences qui suggèrent la nécessité de la concertation entre agents de l'éducation et, par suite, de l'évaluation (Bonora, 1988). Cette approche par les objectifs a été utilisée par le MEN à partir des années 80 pour construire les instruments de ses enquêtes nationales sur les connaissances des élèves, par un travail de concertation soutenue entre enseignants, IPR, IG et autres agents de l'éducation nationale sur les objectifs pédagogiques à prendre en compte, et leur traduction en termes de performances attendues de l'élève.

La docimologie a été l'objet d'un projet d'enquêtes internationales : le projet Carnegie. Pour renforcer les conclusions des recherches menées en France, Piéron indique que des observations sur le caractère subjectif de la notation ont été également observées dans d'autres pays (Suisse, Angleterre, Belgique, Inde, Etats-Unis…). Ce qui explique qu'en 1931, à l'initiative des Etats-Unis, une grande enquête internationale portant sur « Les conceptions, les méthodes, la technique et la portée pédagogique et sociale des examens et concours » fut lancé par la Carnegie Corporation (enquête connue sous le nom de Enquête Carnegie).

En France, les études ont été menées sur des épreuves réelles du Baccalauréat. 100 copies de chaque matière furent tirées au hasard dans les archives, les notes ont été relevées puis les copies furent recopiées afin de les confier à 5 examinateurs de baccalauréat.

1.2 - LES ENQUETES DE L'INETOP

On peut citer parmi les premiers travaux l'enquête de Laugier, Weinberg et Cassin (1939). Il faudrait également citer ici les enquêtes sur la notation réalisées par Piéron dans le cadre de ses études docimologiques sur lesquelles on aura l'occasion de revenir dans le dernier paragraphe.

Le niveau intellectuel des enfants d'âge scolaire a donné lieu à deux enquêtes avec la participation de l'INED. La première enquête qui utilisait le test « mosaïque » construit par Gille a eu lieu en 1944 et ses résultats ont été publiés en 1950 (Heuyer, G., Piéron, H., Piéron, Mme H. & Sauvy, A). La collecte des données pour une seconde enquête plus complète s'est déroulée en 1968 (INED & INETOP, 1969, 1971, 1973). Ces enquêtes sont sans doute uniques en France par la taille des échantillons, la quantité de données recueillies et les analyses effectuées. Ces données permettent aujourd'hui de faire des recherches sur l'évolution du niveau intellectuel des jeunes français. Ainsi, André Flieller de l'Université de Nancy 2, a pu comparer les performances au test « mosaïque » de Gille à quarante ans d'intervalle (Flieller, 1989). Il faut également signaler que, pour la seconde enquête, des questionnaires spécifiques avaient été construits par Pierre Benedetto. Les subtests utilisés étaient composés de manière telle que les aptitudes mesurées pouvaient être comparées du Cours Préparatoire à la cinquième.

Un second thème d'enquêtes réalisées par l'INETOP est l'orientation. La plus connue est sans doute celle portant sur les élèves à la fin du premier cycle secondaire (Reuchlin & Bacher, 1969). Cette enquête comprenait de nombreuses épreuves permettant de recueillir des informations sur les aptitudes intellectuelles, les acquis scolaires, les résultats scolaires, les projets d'études, et différentes variables se rapportant au niveau socio-économique des élèves. L'exploitation des résultats de cette enquête a donné lieu à de nombreuses publications. Elle a été accompagnée d'un suivi quatre années plus tard (Bacher & Isambert-Jamati, 1970).

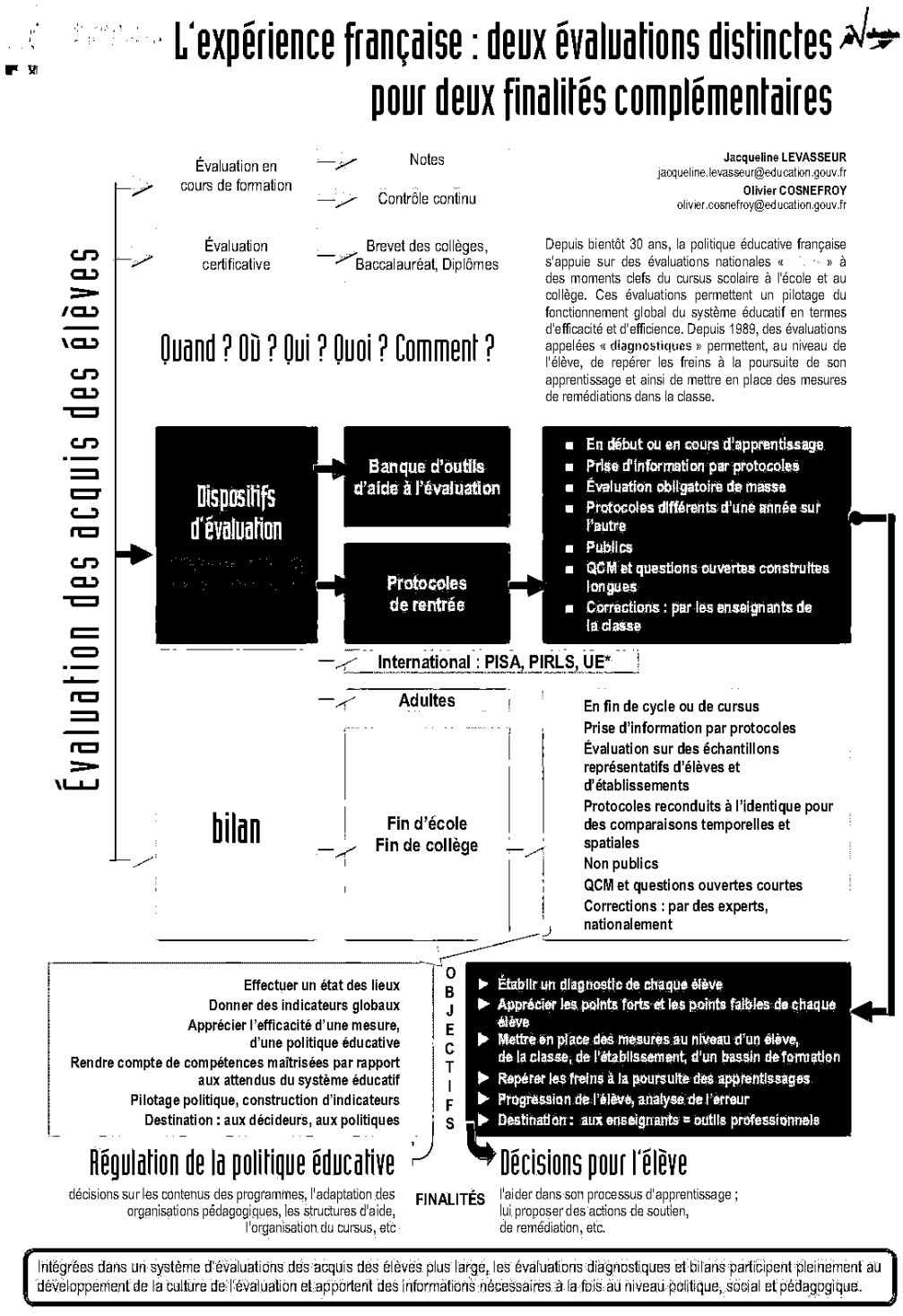
1.3 - LES SERVICES DU MEN

Bien que dès 1964, ait été mis en place au Ministère un service chargé de produire des indicateurs statistiques sur l'éducation et la formation, il faudra attendre le la fin des années 1970 pour assister à la mise en place des évaluations bilan sur les acquis des élèves. Par la suite, différentes enquêtes ont été conduites dans les années 1970 pour aboutir aux enquêtes nationales sur les acquis des élèves en primaire (fin 1970) et au collège (1980 pour les sixièmes, 1982 pour les cinquièmes). Pour ces travaux, le Ministère s'est appuyé sur l'aide méthodologique fournie principalement par l'INETOP et l'Université Nancy 2 (GRAPCO). Dans le prolongement de cette dernière enquête, a été mis en place, en 1984, le premier dispositif portant sur les acquis des élèves de 3éme par le Service de la Prévision, des Statistiques et de l'Evaluation (SPRESE). Comme nous l'avons fait remarquer dans notre rapport sur les évaluations bilan fin de troisième pour le Haut Conseil de l'évaluation de l'école (Salines & Vrignaud, 2001), le service chargé des évaluations a progressivement élargi ses missions pour se consacrer à des évaluations diagnostiques passées par tous les élèves à trois niveaux fondamentaux de leur cursus scolaire. La figure 1 est la présentation par Jacqueline Levasseur, Chef de Bureau à la DEP, mission de l'évaluation des élèves et des étudiants, et son collaborateur Olivier Cosnefroy, des différentes évaluations françaises,

pilotées par la DEP. Cette présentation met en évidence la diversité des évaluations (bilan, diagnostique) et la couverture des différents niveaux scolaires du primaire au collège. Elle montre aussi la place des évaluations internationales au sein de ce dispositif.

L'existence d'un dispositif d'évaluation bien rôdé et fournissant régulièrement des informations sur les acquis des élèves a eu au moins deux conséquences importantes par rapport à la position française quant aux évaluations internationales. D'abord, il faut remarquer que cet organisme n'a pas participé immédiatement aux évaluations internationales (celles de l'IEA) et a attendu une dizaine d'années pour s'investir dans ce domaine (avec l'enquête IALS). On peut donc faire remarquer que malgré l'expérience solide et reconnue de ce service (Dans son ouvrage sur le « Le pilotage des systèmes éducatifs », Gilbert de Landsheere (1994) présente le dispositif français comme « un développement exemplaire »), ce service n'avait pas la même expérience des évaluations internationales. Ainsi, les Modèles de Réponse à l'Item, largement utilisés pour présenter les résultats de ces enquêtes n'étaient pas directement connus et utilisés au sein de la DEP. Une seconde conséquence est que la France a regardé les résultats à la lumière de ses propres évaluations. Ce qui n'a pas toujours été le cas dans la mesure où de nombreux pays ne disposaient pas, eux, de services d'évaluation nationale des acquis des élèves et ont collecté ce type de résultats pour la première fois à l'occasion de PISA.

Figure 1. Le dispositif français d'évaluation (Levasseur et Cosnefroy, 2004)



*Bonnet, G. (éd) (2004), *The assessment of pupils' skills in English in eight European countries*. Bonnet, G. & al. (2001), *The use of national reading tests for international comparisons : ways of overcoming cultural bias*, European network of policy makers for the evaluation of education systems, http://eisa.ac.education.fr/reval

2 - LA FRANCE PROMOTEUR D'ENQUETES

2.1 - LE DOMAINE FRANCOPHONE

Dans le domaine des évaluations internationales des compétences des élèves, la France a lancé des enquêtes, en particulier dans le monde francophone. De nombreux pays francophones du Maghreb et d'Afrique ont créé leurs dispositifs d'évaluation des élèves avec l'appui d'experts français. Dans ce cadre a été élaboré un programme visant aux comparaisons entre pays d'Afrique de l'Ouest francophones (le programme PASEC : Programme d'Analyse des Systèmes Educatifs). Dans le cadre de la francophonie, on peut également citer une enquête sur la production de l'écrit dans quatre pays francophones (Québec, Belgique, Suisse et France, voir Colmant & Desclaux, 1996). Plus proches de perspectives internationales, Le Réseau européen des responsables des politiques d'évaluation des systèmes éducatifs a été le promoteur de plusieurs enquêtes internationales.

2.2 - LE DOMAINE EUROPEEN

Nous avons présenté *Le Réseau européen des responsables des politiques d'évaluation des systèmes éducatifs* (RERPESE), son rôle dans la création d'une meilleure connaissance mutuelle et l'organisation d'une réflexion commune des politiques et des services d'évaluation des élèves dans les différents pays européens. Nous avons également évoqué les différentes initiatives et les travaux d'enquêtes pilotées par ce réseau. Nous ne reviendrons donc pas ici sur ces éléments qui ont déjà été décrits et dont les éléments méthodologiques originaux seront abordés dans la partie dédiée à la méthodologie des enquêtes pilotées par le RERPESE.

3 - LA PSYCHOMETRIE EN FRANCE

Pour terminer ce tour d'horizon des travaux français dans le domaine de l'évaluation nationale et internationale des compétences des élèves, nous avons jugé utile de développer une question dont les conséquences se font déjà et risquent de se faire encore davantage sentir dans ce domaine : celui de la place de la psychométrie dans le paysage de la recherche fondamentale et appliquée française.

Le nom de la France est associé à la création de la méthode des tests puisque le premier test d'« intelligence » a été créé par le français Alfred Binet en 1904. Par contre, la France apparaît plutôt en retrait par rapport à de nombreux pays industrialisés en ce qui concerne la recherche psychométrique. On peut constater le même retard dans les nombreux domaines où on peut recourir à l'utilisation d'instruments psychométriques par exemple l'évaluation et le recrutement du personnel comme le montrent, entre autres, les travaux de Bruchon-Schweizer (Bruchon-Schweitzer & Ferrieux, 1991), sur les méthodes de recrutement en France et en Europe. Cette place marginale de la psychométrie dans le paysage français tient à une pluralité de causes (Vrignaud, 2000). On cite souvent le rejet de ces méthodes développée dans la ligne du mouvement de mille neuf cent soixante-huit. Il existe pourtant

en France une tradition psychométrique visible à travers l'édition de tests en français et l'existence de deux revues publiant dans ce domaine : la « Revue Européenne de Psychologie Appliquée » et « Psychologie et Psychométrie ». Cependant, le domaine est mal identifié dans les instances d'évaluation des laboratoires de recherche. Il existe peu de laboratoires présentant dans leur programme des thématiques explicitement centrées sur les méthodes psychométriques (on peut citer sans être exclusif l'INETOP, le GRAPCO à Nancy 2, Rennes). Et, le plus souvent, ces thématiques sont associées à la psychologie différentielle. Il s'agit donc moins de développer des recherches sur les méthodes psychométriques que de les appliquer à l'étude des différences interindividuelles dans différents domaines cognitifs et conatifs. Le développement et les recherches des méthodologiques psychométriques dans le domaine de l'éducation ont donné lieu à la constitution d'une discipline spécifique « l'édumétrie » qui bien identifiée dans la recherche internationale et dont les résultats sont publiés dans les revues consacrés à *« l'Educational Measurement »* parmi lesquelles on peut citer : *Educational and Psychological Measurement*, *Journal of Educational Measurement*, *Journal of Educational and Behavioral Statistics, Journal of Psychoeducational Assessment, Educational Measurement : Issues and Practic.* On a déjà noté l'absence des enseignants chercheurs français en Sciences de l'Education dans le domaine des enquêtes sur l'évaluation des compétences des élèves tant nationales qu'internationales.

Une association, l'ADMÉÉ (Association pour le Développement des Méthodologies d'Évaluation en Éducation-Europe) rassemble les chercheurs, principalement francophones, travaillant sur l'évaluation en éducation. Cette association est la branche européenne créée en 1985, d'une association canadienne francophone datant des années 1970 dont le sigle a une signification quelque peu différente puisqu'il s'agit de l'Association pour le Développement de la *Mesure* et de l'Évaluation en Éducation. Ce glissement terminologique nous semble particulièrement significatif d'une méfiance envers les méthodes quantitatives « dures ». On peut signaler, parmi les colloques organisés par l'ADMÉÉ-Europe celui de 1999, organisé par l'IREDU, qui portait sur « *L'évaluation des politiques d'éducation ».* L'activité de l'ADMEE est l'exception qui confirme la règle en ce qui concerne l'intérêt des chercheurs francophones pour l'évaluation mais il faut souligner que les chercheurs des pays francophones d'Europe y sont très actifs et qu'elle rassemble des chercheurs de plusieurs disciplines (économie, sociologie, psychologie et sciences de l'éducation).

Une autre raison de la méconnaissance de la psychométrie est son absence des formations supérieures en statistiques. Que ce soient dans les formations universitaires ou de grandes écoles (par exemple l'ENSAE), la psychométrie n'est pas enseignée. Hors, il ne suffit pas de maîtriser les statistiques même à un haut niveau pour appréhender les concepts et les approches de la mesure développées en psychométrie. Cette absence d'enseignement explique les difficultés rencontrées en France pour appréhender le domaine de l'évaluation des élèves et les problèmes méthodologiques posés par les enquêtes internationales. Il faut également souligner la part prédominante prise dans le traitement des enquêtes par des méthodes d'analyse des données « à la française », en particulier, l'analyse factorielle des correspondances. Ces méthodes ont pendant longtemps été des exceptions culturelles françaises (les analyses en facteurs communs et principaux, les modèles structuraux, les Modèles de Réponse à l'Item ou l'échelonnement multidimensionnel étant majoritairement utilisés dans le reste du monde). Elles ont en partie occulté dans le domaine des traitements d'enquêtes les approches anglo-saxonnes. Ceci a d'ailleurs pour conséquence qu'il a parfois été difficile pour les chercheurs français de publier dans des revues internationales en

traitant leurs données selon ces méthodes peu reconnues, voire ignorées, par les experts internationaux.

Pour souligner les conséquences de ces insuffisances françaises dans le domaine psychométrique, on peut faire remarquer que les deux éditeurs français de tests (les Editions et Applications Psychométriques et les Editions du Centre de Psychologie Appliquée) ont été rachetés par un éditeur de test américain (*Psychological Corporation*) qui les a fusionnés. A long terme, les conséquences de ce type de mondialisation peuvent être la disparition de l'édition de tests français au profit uniquement de l'adaptation de tests américains. C'est d'ailleurs déjà en partie le cas, les tests les plus utilisés étant des adaptations de tests américains : on teste l'intelligence des enfants français avec l'adaptation des tests américains (*Wechsler Adult Intelligence Scale for Children, Kaufman Assessment Battery for Children*) et la personnalité des cadres français avec des adaptations de questionnaires américains (16 PF, NEO-PI). Il ne s'agit pas, bien sûr, de préjuger de la qualité de la mesure : ces épreuves américaines et leurs adaptations sont majoritairement d'excellente qualité. Cependant, on peut se demander si cette situation ne mériterait pas d'être traitée comme dans le domaine culturel en cherchant à préserver spécificités et particularismes culturels.

Ces considérations sur l'édition de tests n'ont pas que des côtés anecdotiques. Du point de vue économique, derrière le marché des tests et de l'évaluation, il y a le marché de l'éducation. Pour montrer l'importance de ces enjeux, on peut prendre l'exemple du développement d'un organisme américain tel *qu'Educational Testing Service* (sur le développement d'ETS voir Lemann, 1999). Cet organisme a acquis aux Etats-Unis un rôle dominant sur le marché de l'évaluation en gérant les examens d'entrée dans les Universités de la côte Est des Etats-Unis grâce en particulier au fameux *Scholastic Aptitude Test* (SAT). Aujourd'hui, le rôle d'ETS est prépondérant par ses enquêtes au niveau fédéral sur les acquis des élèves et leur évolution (*National Assessment of Education in Progress*). Pour illustrer ce rôle sur le marché mondial de l'Education, il faut signaler qu'ETS a construit le *Test Of English as a Foreign Language* (TOEFL). Le TOEFL est aujourd'hui une référence en termes de compétences en anglais y compris pour les employeurs ce qui conduit de nombreux organismes de formation à construire leurs programmes en relation avec la préparation du TOEFL et ses contenus. Avec le développement de la formation et de la certification à distance supportée par Internet, le marché de l'Éducation devient mondial. Il est donc essentiel que la France garde ou, plutôt, développe dans ce domaine de l'évaluation un potentiel scientifique et technologique.

4 – LA PUBLICATION DES RESULTATS

Nous distinguerons trois types de diffusion des résultats des enquêtes internationales : 1) les publications de présentation aux responsables des systèmes éducatifs et aux enseignants, principalement celles du Ministère de l'Education Nationale, destinées principalement ; 2) les publications scientifiques visant à exploiter et à discuter des résultats ; 3) les publications de vulgarisation à destination du grand public par l'intermédiaire des media (journaux, radio, TV).

1) Les publications de présentation aux éducateurs

La DEP édite plusieurs publications destinées à faire connaître ses travaux : les Notes d'Information et les Notes d'Évaluation, la revue «Education et Formation » et des numéros thématiques : les Dossiers d'Éducation et Formation. A partir des années 1990, moment où la Dep a été responsable du pilotage des enquêtes pour la France, les résultats des principales enquêtes internationales ont été chacune l'objet d'une Note d'Information ou d'Évaluation : TIMSS (Jouvenceau, 1992 ; Servant & Murat, 1996 ; Servant, 1997) ; PISA 2000 (Bourny, Dupé, Robin, & Rocher, 2001), PIRLS (Colmant, & Mulliez, 2003) ; et dernièrement PISA 2003 (Bourny, Fumel, Monnier, Rocher, 2004). Les résultats de PISA 2000 ont été l'objet d'un numéro des Dossiers (Bourny, Braxmeyer, Dupé, Rémond, Robin, & Rocher, 2002). Ce numéro rassemble les contributions de plusieurs auteurs (spécialistes de la DEP et universitaires) qui présentent l'enquête, les résultats des élèves français et discutent plusieurs hypothèses concernant les performances des élèves français comparées à celles des élèves des autres pays participants. Duru-Bellat, Mons, & Suchaut (2004) ont publié des éléments de leur rapport cité plus bas dans Éducation et Formation.

Les travaux du RERPESE ont également été présentés dans ces publications en particulier les résultats de l'enquête sur les compétences en anglais des élèves de 15 à 16 ans. La première version de cette enquête en 1996 a été publiée dans un numéro des dossiers (Levasseur et Shu, 1997). Les résultats de l'enquête 2002/2003 portant sur les compétences en anglais dans huit pays européens ont été présentés dans une Note d'évaluation (Bonnet et Levasseur, 2004). Dans le cadre du RERPESE, il faut mentionner la lettre d'informations du Réseau (ÉVALUATION) qui paraît deux fois chaque année.

On peut également inclure dans ces publications dédiées plus particulièrement aux éducateurs, les rapports réalisés à la demande du HCéé qui comportent des éléments de présentation et de discussion des enquêtes internationales, en particulier : Salines et Vrignaud, 2001 ; Céard, Rémond, & Varier, 2003 ;

Et enfin, nous citerons quelques publications destinées à faire connaître les résultats à des publics enseignants spécialisés comme Dupé, & Olivier, 2002 ou Robin et Rocher, 2003.

2) Les publications scientifiques

L'activité de publication scientifique peut être répartie en articles et communications scientifiques (revues à comité de lecture, congrès internationaux) et en travaux universitaires (thèses). Par ailleurs, on peut intégrer dans cette catégorie les rapports réalisés à la demande de la DEP sur des points scientifiques ou méthodologiques des enquêtes.

L'activité de publication dans des revues à comité de lecture a été particulièrement abondante pour l'enquête de l'IEA sur les sciences. Le pilotage de cette enquête par l'INETOP (Institut National d'Etude du Travail et d'Orientation Professionnelle) a permis aux chercheurs de cet institut à en exploiter les résultats par des publications : Bonora (1972, 1974a et b, 1975). Excepté pour cette enquête, nous n'avons pas trouvé mention de publications exploitant les résultats pour les autres enquêtes de l'IEA auxquelles la France a participé avant les années 1990.

L'enquête IALS a donné lieu à de nombreux rapports sur les résultats français et leur fiabilité. On peut répartir ces travaux en trois phases : une première phase avant (ou peu

après) la publication des résultats par l'OCDE (Dickes & Vrignaud, 1995 ; Igersheim, Pluvinage, & Bonnin, 1995 ;) ; une seconde phase comprenant des études françaises publiées dans les années qui ont suivi la publication (Rémond, 1996 ; Blum & Guérin, 1997 ; Dickes & Flieller, 1997 ; Blanchard, & Vrignaud, 1996) et, enfin, une troisième phase suite à l'étude dans le cadre de la Commission européenne (Blum, Goldstein, & Guérin-Pace, 2001). Les auteurs des différents rapports ont souvent publié sur leurs analyses de IALS dans des revues et des ouvrages : Guérin-Pace, & Blum, 1999 ; Rémond, 2001 ; Vrignaud, 2001a et b ; Vrignaud & Chartier, 1999).

Des études complémentaires à partir des données françaises et internationales de l'enquête PISA ont été l'objet de différents rapports (Bautier, Crinon, Rayou, & Rochex, 2003 ; Bonnet, 2002 ; Duru-Bellat, Mons, & Suchaut, 2004a) et de plusieurs publications (Bonnet, 2002 ; Duru-Bellat, Mons, & Suchaut, 2003 et 2004b ; Meuret, 2003 ; Mons, 2004a).

Il faut également signaler des revues de questions qui ne portent pas une enquête spécifique mais sur la méthodologie des comparaisons internationales (Murat, & Rocher, 2004 ; Rocher, 2003 ; Vrignaud, 2002).

On peut enfin mentionner les communications et publications autour des enquêtes du RERPESE, principalement l'enquête portant sur l'évaluation de la littéracie dans quatre pays européens (Bonnet, 1998 ; 2004a et b ; Vrignaud & Rémond, 2002).

Parmi les travaux universitaires, la thèse de Nathalie Mons (2004b), dirigée par le Professeur Duru-Bellat, à l'IREDU, est consacrée à la perspective comparative en éducation. La partie expérimentale utilise les données de l'enquête PISA 2000 pour une recherche sur les effets des politiques de décentralisation en éducation. Cette thèse est, à notre connaissance, la seule française consacrée aux évaluations internationales. Il faut également signaler la création, en 2003, d'un séminaire à l'école pratique des hautes études en sciences sociales, animé par Jean-Richard Cytermann (2003), dont plusieurs séances ont porté sur les enquêtes internationales et leur impact sur les politiques éducatives dans différents pays européens.

Ce tour d'horizon, montre donc, sauf erreurs ou omissions de notre part, que, excepté l'enquête IEA sur les Sciences, les chercheurs français n'ont pas publié sur les enquêtes internationales avant les années 1990. Il existe davantage de publications exploitant les données françaises ou de publications plus générales sur les enquêtes internationales à partir de l'enquête IALS et PISA. Ces travaux présentent trois caractéristiques : 1) ils sont le fait de chercheurs ayant des « habitudes » de travail avec la DEP ; 2) ces chercheurs sont plutôt des psychologues et des sociologues ; 3) ces travaux adoptent souvent une position critique vis-à-vis des résultats des enquêtes.

3) Les publications destinées au grand public

Les articles présentant les résultats des enquêtes internationales sur les acquis des élèves sont peu documentés jusqu'à l'enquête IALS. Bien que nous n'ayons pu obtenir de certitude sur ce point, il semble qu'il n'y ait pas eu de présentation à la presse française des résultats des enquêtes antérieures à IALS. On peut noter de manière sporadique quelques présentations brèves des résultats français dans des revues spécialisées (Le Monde de l'Éducation) ou de vulgarisation scientifique (Sciences et Avenir pour les enquêtes sur les mathématiques).

La médiatisation des résultats des enquêtes internationales, à partir des années 1990, apparaît particulièrement orchestrée par l'OCDE pour l'enquête IALS. Il faut rappeler, ici, que cette enquête était au départ pilotée par *Statistics Canada*. Ce n'est qu'au cours du processus que l'OCDE a obtenu de prendre en charge la publication des résultats de l'enquête et a décidé de donner une large publicité médiatisée à ce travail.
Les enquêtes suivantes tant celles pilotées par l'OCDE (PISA) que celles pilotées par l'IEA (PIRLS) ont donné lieu à des conférences de presse systématiques. Les résultats ont été largement repris dans les principaux média comme le montre le dossier de presse publié par l'OCDE et téléchargeable sur le site PISA (OECD, 2002). On note qu'excepté des quotidiens comme Le Monde et le Figaro qui consacrent une surface importante à la couverture de cette enquête, la plupart des articles se contentent de citer le classement des élèves français.

On peut enfin inclure dans cette catégorie des ouvrages grands publics sur les enquêtes internationales sur la littéracie comme celui de Blum et Guérin (2000), ainsi que des articles dans des revues spécialisées comme (par exemple l'article de Pacteau et Vrignaud, 2001 dans Sciences Humaines).

Synthèse : on peut dire que la psychométrie et l'édumétrie sont des domaines peu développés dans le paysage de la recherche française. Il en est de même pour les applications de ces disciplines (construction de tests, enquêtes). Cette méconnaissance a sans doute joué sur le peu d'intérêt manifesté pour les enquêtes internationales. Les effets principaux continuent à s'en faire sentir aujourd'hui avec la disparition de l'édition de tests française et la rareté des experts français pouvant intervenir dans ce domaine tant au niveau national qu'international.

CHAPITRE V - METHODOLOGIE DE L'EVALUATION DES COMPETENCES

Il est impossible de présenter et surtout de discuter des enquêtes internationales sans aborder de manière approfondie les méthodes psychométriques. Nous dirons que la psychométrie et ses méthodes sont les fondations du bâtiment. Cette présentation est, évidemment très technique. Les lecteurs peuvent, bien sûr, aborder directement les chapitres suivants qui traitent de la méthodologie des enquêtes internationales proprement dites quitte à revenir ensuite sur la présentation de certains concepts développés dans ce premier chapitre.

J'aborderai ces questions de méthodes à partir de la psychologie car les sciences de l'éducation sont une discipline jeune par rapport à la précédente et, du fait que la plupart des pionniers en sciences de l'éducation avaient reçu une formation en psychologie, ils ont adopté les méthodes de cette discipline. Par ailleurs, les premières recherches sur les performances scolaires faisant appel à la mesure (on peut citer les travaux du psychologue anglais Spearman qui s'était intéressé en 1904 aux écoliers londoniens). Je montrerai que la majeure partie des méthodes utilisées ont été élaborées au sein de la psychologie ou plutôt de la psychométrie. On parle aujourd'hui de l'édumétrie pour définir un champ équivalent à celui de la psychométrie en sciences de l'éducation, mais les méthodes et modèles sont largement similaires et bien souvent les chercheurs de ce domaine travaillent et publient dans les deux champs.

La mesure, c'est à dire l'assignation de grandeurs à des objets en respectant certaines propriétés de ceux-ci, a posé en psychologie des problèmes particuliers qui ont abouti au développement de solutions originales au sein de cette discipline. Ces méthodes se sont trouvées rassemblées dans la psychométrie qui définit les méthodes à mettre en oeuvre, depuis les dispositifs de collecte des données jusqu'à la définition de normes de fiabilité (pour une présentation des théories et méthodes psychométriques, on se reportera, en français, à des ouvrages comme ceux de Dickes, Tournois, Flieller et Kop, 1994 ou de Laveault et Grégoire, 2002). La démarche de validation de la mesure en psychométrie repose sur le principe selon lequel toute mesure est un construit. La validation de ce construit nécessite, comme condition première, de pouvoir émettre des hypothèses sur sa nature et son fonctionnement. S'assurer de la qualité de la mesure nécessitera de juger de l'adéquation du modèle de construction en testant les hypothèses que ce modèle permet d'énoncer sur des données recueillies. On parlera ici d'un modèle de mesure, et la démarche hypothético-déductive consiste à tester l'adéquation de ce modèle de mesure aux données. Plusieurs approches peuvent être mises en œuvre pour tester cette adéquation (on en trouvera une présentation dans les ouvrages cités supra). Les trois approches les plus généralement utilisées sont l'approche classique (formalisée par Lord et Novick, 1969), les Modèles de Réponse à l'Item et les modèles structuraux.

1 - LES PRINCIPAUX MODELES DE MESURE

Il ne s'agit pas de faire ici un exposé systématique de ces questions mais d'en donner les lignes directrices et les principaux concepts afin de discuter des problèmes posés par les

mesures comparatives qui sont l'objet des enquêtes internationales, de présenter les solutions adoptées dans les différentes enquêtes et enfin, d'évoquer les insuffisances éventuelles de ces solutions et les points sur lesquels des recherches et des développements sont nécessaires. Cet exposé est évidemment technique mais c'est justement un des problèmes cruciaux de ces enquêtes que la compréhension des résultats et surtout de leurs limites est liée à des questions méthodologiques complexes.

En effet, la mesure étant un construit, on va inférer à partir de ce construit. On mesure la performance et on infère sur les compétences. C'est à dire qu'on passe de « tel élève ou tel établissement, tel pays obtient un score de tant au test de littéracie… », c'est sa performance, à « donc les compétences de tel élève, tel pays, se situent à tel et tel niveau… ». Cette inférence sur les compétences est valide dans la mesure où l'épreuve mesure bien ce qu'elle est censée mesurée. Certes, les outils statistiques permettent d'étayer ce passage de la performance à la compétence. Les études de validation, la prise en considération de l'erreur de mesure permettent d'éviter des biais mais certaines questions fondamentales restent ouvertes. Reprenons ici l'exemple de la mesure de l'intelligence. On a longtemps et à juste titre reproché aux tests d'intelligence de prendre en compte une partie seulement des comportements que l'on peut inclure comme rendant compte de l'intelligence.

Si la question de l'échantillonnage des sujets est un thème bien connu des enquêtes, la question de l'échantillonnage des tâches et par voie de conséquence des items opérationnalisant ces tâches est bien moins connue des non spécialistes voire, parfois, des spécialistes. Il s'agit de s'assurer que les items sont représentatifs de l'univers évalué et surtout que certains aspects de cet univers n'ont pas été négligés. Ainsi, les tests d'intelligence, ont souvent été critiqués comme survalorisant des intelligences scolaires, verbales correspondant aux connaissances d'adultes vivant en milieu urbain. Des auteurs comme Sternberg et surtout Gardner (1993/1997) qui dans sa théorie des intelligences multiples propose près d'une dizaine de formes d'intelligence, ont montré que l'échantillonnage du domaine pouvait être souvent réducteur. Ceci est d'autant plus réducteur qu'on a pu avoir tendance à privilégier un facteur général d'intelligence. La position des personnes et surtout des groupes sociaux ou ethniques aux Etats-Unis a conduit à écrire les pires pages de l'histoire de la psychométrie avec les travaux de l'équipe de Terman et ses propositions d'eugénisme basées sur les performances intellectuelles. On trouvera une critique de ces travaux dans différents ouvrages spécialisés ainsi que dans l'ouvrage de Jay Gould « la mal mesure de l'homme ». La publication de l'ouvrage « *the bell curve* » par Hernbstein et Murray en 1994, ouvrage basé sur la comparaison des groupes ethniques aux tests d'intelligence générale et stigmatisant le groupe des « *afro-américain* » pour leurs résultats en moyenne inférieurs à ceux des autres groupes ethniques nord américains est malheureusement un exemple que ce type de travaux socialement et politiquement dangereux resurgissent régulièrement.

1.1 - LA THEORIE CLASSIQUE DES TESTS

Un des principaux objectifs de la psychométrie est de réduire l'erreur de mesure. De manière habituelle, on rappelle qu'une des manières de formuler l'objectif de la psychométrie est :

Score observé = Score vrai + Erreur de mesure

On cherche à distinguer performance (les résultats observés) et compétence (l'aptitude, le trait qui a produit cette performance et que l'on cherche à évaluer). On observe une performance et on en infère sur la compétence. Le but est de rendre cette inférence le plus fiable possible. Il faudrait en fait parler de nombreuses sources d'erreurs qui influent sur la performance et que l'on va chercher à prendre en considération et à réduire. La standardisation est un des moyens de réduction des conditions d'erreur. L'appréciation des sources d'erreur sur la qualité du construit est l'objet de l'étude de la fidélité interne. Il s'agit de s'assurer que le passage des items à la variable évaluée est fiable. Cela permet de s'assurer que le calcul d'un score à partir des items, en général en faisant la somme des points accordés pour des réponses correctes est fiable. Ce qui ne serait pas le cas, par exemple, dans le cas où les items mesureraient des compétences différentes. C'est pourquoi on parle ici d'homogénéité ou de consistance interne. L'analyse interne se fait à deux niveaux : local celui des items et global celui du score.

Au niveau des items, on s'intéresse principalement à deux de leurs caractéristiques : leur difficulté et leur discrimination.

1.1.1 - Indice de difficulté de l'item

Dans le cas d'un score dichotomique (bonne ou mauvaise réponse), la difficulté de l'item est souvent estimée par la proportion d'élèves de l'échantillon qui donnent une réponse correcte. Le score moyen est une variante pour des items polytomiques (réponse multiple ordonnée). Dans le cadre de la théorie psychométrique, on sait que cette proportion observée est entachée d'erreur. Ainsi dans un autre échantillon, on n'observera sans doute pas la même proportion de réussite.

L'interprétation de l'absence de réponse doit être prise en compte. Elle est fonction de la position de l'item : pour les items intermédiaires, elle signale une déclaration d'ignorance et/ou une absence de prise de risque ; pour les items terminaux un manque de temps. La distinction entre ces deux types de non réponse est importante car elle permet de déterminer si le test est un test de puissance ou de vitesse. Le codage des non réponse comme échec ou comme item non examiné est fondamentale pour l'estimation de la difficulté des items (en particulier dans le cadre des modèles de réponse à l'item). Le codage des omissions terminales comme des items non examinés postule qu'ils n'informent pas sur la connaissance des élèves qui n'y ont pas répondu. La proportion de réussite sera estimée à partir des seuls élèves de l'échantillon qui ont répondu à l'item. Cette procédure évite d'interpréter comme absence de maîtrise du domaine ce qui dépend en fait de la vitesse de travail et du temps de passation. Différentes méthodes pour pallier ces absences de réponses existent, les enquêtes internationales ont souvent employé des approches très sophistiquées dans la mesure où, de plus, le plan de collecte des données est un plan incomplet par construction (tous les élèves ne passent pas tous les items). Ce point sera développé plus loin.

1.1.2 - Indice de discrimination de l'item

L'indice se fonde sur la corrélation entre l'item et le critère ou sur les différences des indices de difficulté pour des élèves appartenant à des groupes différents relativement à un critère. La valeur numérique de l'indice dépend de l'échantillon d'élèves examinés et de la variable critère. Le critère interne ou externe devrait être un indicateur parfaitement valide de la

variable à mesurer. Un critère interne tel que le score total corrigé est une fonction de la performance sur tous les items moins celui qui est examiné.

Si les réponses à l'item ne sont pas dichotomiques mais graduées, on peut calculer des coefficients polysériaux qui sont des généralisations des coefficients présentés.

Cet indicateur peut être calculé sur les distracteurs des questions à choix multiples. Il convient de s'assurer qu'ils discriminent bien sur la variable étudiée et non pas sur une variable parasite.

Ce type d'analyse peut s'appliquer si l'on ne dispose que d'un nombre relativement restreint de protocoles. L'un de ses inconvénients majeurs réside dans la dépendance des statistiques de l'item : elles sont calculées à partir des réponses du groupe d'élèves examinés et ne valent que dans la mesure où les sujets sont aussi proches que possible de la population cible. Il convient donc de garder les caractéristiques de chaque passation, de bien préciser sur quel échantillon et à quelle phase de l'apprentissage ils ont été obtenus, de définir la population à laquelle la procédure paraît applicable et les conditions d'observation qui autorisent une généralisation de la mesure.

La prise en compte de l'indice de discrimination est importante pour s'assurer de la fiabilité des items de l'épreuve (suppression des items peu discriminants donc peu informatifs). La discrimination de l'item renseigne sur la qualité et la quantité d'information apportée par l'item pour déterminer la compétence du sujet. Un item au pouvoir discriminant élevé apporte beaucoup d'information sur la compétence du sujet, un item peu discriminant renseigne peu sur la compétence du sujet. Lors des pré-expérimentations d'épreuves, on retirera les items ayant une discrimination faible ou nulle car ces items n'apportent aucune information utile voire, dans certains cas, peuvent introduire un bruit inutile.

Pour bien comprendre ce qu'est la discrimination, il est utile de se représenter que l'administration d'une épreuve a pour but de recueillir de l'information sur la compétence dans un domaine donné d'un sujet ou d'un groupe de sujets. Une bonne épreuve est une épreuve qui va apporter une information de qualité sur la compétence en satisfaisant un critère d'économie qui est le temps d'administration, la quantité d'exercices que le sujet doit effectuer. La qualité de l'information doit être appréciée en relation avec la difficulté de l'épreuve. C'est la question que se posent de manière classique les enseignants en proposant à leurs élèves des épreuves correspondant à leur niveau. Si l'épreuve est trop facile, elle sera réussie par presque tous les élèves (ce que les psychométriciens appellent un effet plancher) et ne fera que séparer les quelques élèves les plus faibles des moyens. Si l'épreuve est très difficile, elle sera échouée par la majorité des élèves (ce que les psychométriciens appellent l'effet plafond), elle ne fera que séparer les meilleurs des moyens. Si l'épreuve est de difficulté moyenne, elle sera plus ou moins réussie selon le niveau de compétence de la plupart des élèves, elle permettra de classer les élèves moyens mais elle renseignera peu sur les élèves les plus faibles qui échoueront majoritairement une épreuve trop difficile pour eux, de même, elle renseignera mal sur les élèves les plus compétents qui réussiront majoritairement une épreuve trop facile pour eux.

Le pouvoir discriminant d'un item doit donc être apprécié en relation avec sa difficulté. La difficulté de l'item indique la zone de compétence sur laquelle il informe. Un item facile renseigne sur les faibles niveaux qui seuls échoueront, un item difficile sur les niveaux

élevés qui seuls réussiront, etc. Son pouvoir discriminant indique dans quelle mesure sa réussite sépare bien le niveau de compétence correspondant à sa difficulté des autres.

1.1.3 - Au niveau global

De la même manière qu'on s'est intéressé à la validité des items, on va étudier la fiabilité de l'épreuve au niveau global. On parle d'homogénéité ou de consistance interne. Dans la théorie classique des tests, celle-ci est estimée par le coefficient alpha de Cronbach. Cet indicateur répond à la question « l'ensemble des items est-il suffisamment homogène pour que le calcul d'un score soit valide ». La valeur de l'alpha dépend à la fois de l'homogénéité des items (appréciée à partir de leurs intercorrélations) et de leur nombre. A homogénéité donnée, on peut augmenter la consistance interne du test en augmentant sa longueur. Ce point est important dans la mesure où les épreuves pour les évaluations bilan sont en général plutôt longues.

Synthèse : Les qualités d'une épreuve sont étroitement dépendantes des qualités des items. Le niveau de difficulté permet de déterminer le niveau de compétence qu'ils permettent de mesurer de manière fiable. Leur pouvoir discriminant permet de déterminer la qualité de l'information que chacun d'eux apporte pour mesurer le niveau de compétence correspondant à sa difficulté. Une épreuve du type évaluation bilan destinée à évaluer les compétences des élèves d'un niveau donné devra donc comprendre des items efficaces pour la plage de compétences que l'on cherche à évaluer.

1.2 - LES MODELES DE REPONSE A L'ITEM (MRI)

1.2.1 - Présentation

Ces modèles regroupés sous l'appellation générique de Modèles de Réponse à l'Item : MRI (en anglais 'Item Response Modeling': IRM[50]) ont été créés il y a une trentaine d'années (voir pour une présentation Hambleton & Swaminathan, 1985). Il faut signaler qu'ils ont été « inventés » à peu près simultanément et de manière indépendante au Danemark par le mathématicien Georg Rasch qui cherchait un modèle permettant de comparer des compétences d'élèves en lecture à plusieurs années d'intervalle et, aux Etats-Unis, par le statisticien Allan Birnbaum qui cherchait à améliorer les modèles de mesure en psychométrie. Ces modèles ont profondément renouvelé l'approche psychométrique car d'une part ils offrent un cadre unitaire pour penser l'ensemble des concepts psychométriques (exposés supra à propos du modèle classique) et d'autre part, ils offrent un nouveau cadre d'interprétation des résultats aux tests en situant la performance des sujets par rapport à des tâches et non plus par rapport à la compétence d'autres sujets. Ces modèles sont probabilistes. On postule que la probabilité qu'un sujet j donne une réponse correcte à un item i est fonction de la compétence (θ_j) du sujet et de la difficulté de l'item (d_i).

$Pr(X=1) = f(d_i , \theta_j)$ (1)

[50] En anglais, le terme de « Item Response Theory », en abrégé IRT, est plus largement utilisé. Le terme de modèle paraît plus approprié dans la mesure où il s'agit de rendre compte du comportement du sujet répondant à un item plutôt que de construire une théorie psychologique du comportement du sujet comme le font remarquer Goldstein & Wood (1989).

Les modèles MRI sont basés sur la recherche d'un modèle mathématique du fonctionnement de l'item permettant de représenter la relation entre difficulté de l'item et compétence. On utilise en général la fonction logistique. Le modèle le plus général comprend trois paramètres[51] pour modéliser le fonctionnement de l'item :

$$pr(x_{ij}=1) = c_i + (1-c_i)\frac{EXP(a_i(\theta_j - b_i))}{1+EXP(a_i(\theta_j - b_i))} \quad (2)$$

avec :

b_i. la difficulté de l'item,
a_i. la pente (discrimination de l'item),
c_i. le paramètre de réponse au hasard,
θ_j la compétence du sujet.

Les modèles MRI fournissent un cadre conceptuel unitaire pour l'ensemble des concepts de la psychométrie. Chacun des paramètres renseigne sur le fonctionnement de l'item. On peut les rapprocher des paramètres classiques : b_i., la difficulté de l'item, de la fréquence de réussite[52], a_i., la pente (discrimination de l'item), de la corrélation item/test, c_i., de l'étude des distracteurs. Le paramètre de compétence θ_j est une estimation de la mesure vraie de sa compétence. L'explication de la compétence et de la difficulté de l'item par une même variable latente justifie explicitement la comparaison entre items et entre sujets. Les paramètres de difficulté vont permettre de comparer les items entre eux. Les paramètres de compétences autorisent la comparaison des sujets et des groupes de sujets. Toutes les opérations de construction de tests et d'interprétation des résultats demandant d'assurer l'équivalence des items et des tests ou la comparaison de différentes populations vont se trouver ainsi facilitées.

La question du nombre de paramètres du modèle a été souvent discutée. Les tenants du modèle à deux ou trois paramètres ont eu tendance à présenter le modèle de Rasch (à un seul paramètre) comme un cas particulier du modèle général à plusieurs paramètres. Les tenants du modèle de Rasch ont mis en avant que leur modèle possédait des propriétés particulières rendant l'estimation des paramètres plus robuste et plus simple à interpréter. En particulier, l'approche des items polytomiques est plus facile à réaliser (dont la cotation se fait selon plusieurs niveaux échelonnés).

Ces options ont un retentissement sur les méthodes préconisées pour le traitement des enquêtes internationales. Ainsi, ACER utilise un modèle dérivé du modèle de Rasch implanté dans son logiciel CONQUEST alors qu'ETS s'appuie sur un modèle à deux paramètres avec des algorithmes d'estimation implantés dans le logiciel BILOG.

[51] Dans le cas où la prise en compte d'une probabilité de réponse au hasard n'est pas pertinente, le paramètre c_i.ne sera pas intégré au modèle. Si on peut faire l'hypothèse d'une égale discrimination des items, le paramètre de pente a_i. aura la même valeur pour tous les items (modèle de Rasch).
[52] En particulier dans le cas du modèle à un paramètre.

1.2.2 - L'estimation des paramètres

L'estimation des paramètres des MRI nécessite l'utilisation de logiciels[53]. La mise en oeuvre de cette estimation n'est pas une opération anodine. L'appréciation de l'adéquation des modèles MRI se pose aux différentes étapes de l'estimation des paramètres de difficulté des items et de compétence des sujets. En amont, les modèles MRI reposent sur des conditions de validité nombreuses : unidimensionnalité, indépendance conditionnelle des items[54], et, pour le modèle de Rasch, égal pouvoir discriminant des différents items. Ces conditions sont parfois difficiles à tenir et à vérifier. Ainsi Hambleton, Swaminathan & Rogers (1991) recensent une vingtaine de procédures à mettre en oeuvre pour s'assurer de la possibilité d'application du modèle aux données. Signalons également l'ensemble de travaux menés par l'équipe de Stout (Bolt & Stout, 1996; Shealy & Stout, 1993a et b) à l'Université de Chicago qui a permis de trouver des cadres conceptuels plus performants pour tester certaines hypothèses (unidimensionalité, indépendance conditionnelle, fonctionnement différentiel des items). En aval, une fois les paramètres estimés, il est nécessaire de juger de l'adéquation du modèle aux données au niveau global et au niveau de chaque item.

A chacune de ces étapes, la (bonne) estimation des paramètres requiert de disposer de populations numériquement importantes (on entend par là plusieurs centaines voire plutôt plusieurs milliers de sujets). Ces exigences en font des méthodes peu économiques à mettre en oeuvre. Les conditions de validité sont rarement remplies d'emblée. Bien souvent il est nécessaire d'éliminer des items à l'issue d'un prétest, et d'en éliminer encore sur les données définitives. Il est donc nécessaire de disposer d'un ensemble important d'items.

Il est nécessaire de spécifier les objectifs des MRI. En général, les MRI sont présentés comme une solution intéressante pour l'évaluation d'une compétence à un niveau plutôt global voire au niveau d'un système. En effet, bien que les items et les sujets jouent un rôle symétrique, on s'est en général intéressé à la mesure de la compétence du sujet (thêta). Rappelons que cette variable est considérée comme une variable sans erreur. L'estimation de la compétence par cette variable est une solution aux problèmes classique posés par la séparation du score vrai et l'erreur de mesure.

1.2.3 - L'interprétation

Les modèles MRI ont été présentés par leurs avocats comme renouvelant la théorie de la mesure. Rasch argumentait que l'estimation de la difficulté des items et de la compétence des sujets étaient indépendantes, ce qui fondait le concept d'objectivité spécifique. Quels que soient les items passés par un sujet, on obtiendra une même estimation de sa compétence. Quels que soient les groupes de sujets auxquels l'item a été administré, on obtiendra une même estimation de sa difficulté. Cette idée a été souvent considérée comme peu « réaliste » et semble d'ailleurs ne pas avoir donné lieu à de nombreuses études comme le constate Andersen (1995) dans un ouvrage de synthèse sur les développements du modèle de Rasch (Fischer & Molenaar, 1995). Pourtant, il existe une abondante littérature sur les modèles MRI. On trouve dans cette littérature, d'une part, des travaux méthodologiques portant sur l'estimation des paramètres et sur l'extension du modèle à des

[53] Il existe de nombreux logiciels (BILOG et BILOG-MG, MULTILOG, ConQUEST, etc.). Les logiciels continuent à évoluer du fait que les procédures d'estimation des paramètres, en particulier dans le modèle à plusieurs paramètres sont toujours l'objet d'améliorations et de discussions.

[54] L'indépendance conditionnelle présuppose que pour un niveau de compétence donné, la réussite à un item quelconque est indépendante de la réussite aux autres items.

situations plus variées que les items dichotomiques (par exemple des items à réponses ordonnées). Nous ne traiterons pas ici de ces développements et renverrons aux revues de question sur ce thème (par exemple Drasgow & Hulin, 1990 ; Goldstein & Wood, 1989). On trouve, d'autre part, des applications des modèles MRI à de nombreux problèmes rencontrés dans les situations d'évaluation des personnes. Nous nous intéresserons à ce dernier type de travaux à travers l'interprétation des résultats. Le mérite des modèles MRI est certainement de fournir un cadre conceptuel unitaire. L'explication de la compétence et de la difficulté de l'item par une même variable latente justifie explicitement la comparaison entre items et entre sujets. Les paramètres de difficulté vont permettre de comparer les items entre eux. Les paramètres de compétences autorisent la comparaison des sujets et des groupes de sujets. Toutes les opérations de construction de tests et d'interprétation des résultats demandant d'assurer l'équivalence des items et des tests ou la comparaison de différentes populations vont se trouver ainsi facilitées.

Dès que les paramètres de l'item sont connus par estimation, on peut déterminer la compétence du sujet (valeur thêta) sur la même échelle que celle de l'item. Ainsi on peut mettre en regard les sujets selon leur compétence et les items selon leur difficulté. Cette manière de procéder nous montre comment les modèles MRI conduisent à évaluer la compétence d'une sujet différemment de la procédure classique de l'étalonnage (situer la performance du sujet dans la distribution des performances de sa population de référence) ou de la procédure d'évaluation critériée (situer la performance du sujet en référence à un critère de maîtrise). Les modèles MRI définissent la compétence du sujet comme sa probabilité de résoudre des items d'une difficulté donnée. La compétence se définit donc par rapport à des tâches et non par rapport à d'autres sujets. Le paramètre de compétence du sujet définit sa zone de compétence qui peut être mise en relation avec les paramètres de difficulté des items. La définition de la zone de compétence nécessite de décider du seuil de probabilité de réussite retenu pour définir une maîtrise de l'item. Peut on considérer qu'un seuil supérieur à 50 % montre que l'item peut être résolu par le sujet ou vaut-il mieux considérer qu'un seuil proche de 100 % reflète mieux la compétence du sujet ? Par exemple dans les évaluations éducatives aux États-Unis, le seuil de 80 % est généralement retenu (Kirsch, 1995). Ce seuil a l'avantage de garantir une probabilité quasi certaine de réussite, mais sa sévérité peut être trompeuse quant aux réussites réelles des sujets. En effet, les probabilités sont fortes pour que les sujets réussissent d'autres items de difficulté plus grande que celle définie par leur zone de compétence. Un second problème est celui de la définition de la compétence en fonction du contenu des items. Dire qu'un sujet est capable de résoudre des items d'une difficulté donnée renvoie à la définition opérationnelle de ces items. Cette définition peut paraître simple quand le contenu des items s'y prête : par exemple la complexité d'opérations arithmétiques, le nombre d'inférences à effectuer pour conduire un raisonnement. Ce type d'analyse apparaît souvent simplificatrice au regard des modèles de résolution proposés par la psychologie cognitive. Nous rediscuterons de ces points à propos de la définition des compétences dans les enquêtes internationales.

1.2.4 - Limites des modèles MRI

Deux critiques importantes ont été faites à l'encontre des modèles MRI d'une part la difficulté de juger de leur adéquation aux données, d'autre part le caractère globalisant de leur interprétation. L'appréciation de l'adéquation des modèles MRI se pose aux différentes étapes de l'estimation des paramètres de difficulté des items et de compétence des sujets. D'abord, l'estimation des paramètres des modèles MRI repose sur des conditions de validité nombreuses : unidimensionnalité, indépendance conditionnelle des items, égal pouvoir

discriminant pour le modèle de Rasch. Ces conditions sont parfois difficiles à tenir et à vérifier. Ainsi Hambleton (1991) recense les procédures à mettre en oeuvre pour s'assurer de la possibilité d'application du modèle aux données. Ensuite, bien qu'il existe plusieurs logiciels pour la mise en oeuvre des MRI, les nombreuses procédures d'estimation des paramètres en particulier dans le modèle à plusieurs paramètres sont toujours l'objet d'améliorations et de discussions. Enfin, une fois les paramètres estimés, il est nécessaire de juger de l'écart entre les fréquences observées et les probabilités calculées. Des procédures pour juger de l'adéquation sont mises en oeuvre pour chaque item et pour l'ensemble du test. Le nombre de procédures, de tests de signification élaborés (pour une revue de questions voir Flieller, 1995) montre l'embarras que suscitent ces démarches. A chacune de ces étapes, la mise en oeuvre de ces procédures requiert de disposer de populations numériquement importantes (on entend par là plusieurs centaines voire plutôt plusieurs milliers de sujets). Il est permis de penser que ces éléments devraient rencontrer des solutions plus satisfaisantes dans les années qui viennent et que l'on verra mieux quelles sont les situations d'évaluation pour lesquelles l'application des modèles MRI est pertinente. La seconde critique porte sur la réalité psychologique du modèle. Ainsi Reuchlin (1997) conteste le caractère continu du modèle qui présuppose qu'un sujet peut toujours réussir un item. La réponse à un item a un caractère discret. La réussite à un item difficile n'est pas peu probable pour un sujet peu compétent, elle est tout simplement impossible. Une contestation moins radicale porte sur l'interprétation psychologique du modèle.

L'unidimensionnalité de la variable latente présuppose que les différences interindividuelles ne sont que des différences de puissance, que les différences de difficulté entre items ne sont que des différences quantitatives. On présuppose ainsi que quel que soit le niveau de compétence des sujets, ceux-ci mettent en oeuvre des processus et des stratégies similaires pour répondre aux items. Cette critique a déjà été souvent portée à l'encontre des scores dont le caractère globalisant n'informe pas sur les processus sous-jacents (voir par exemple Huteau et Lautrey, 1978). Deux directions de recherches ont tenté d'intégrer aux MRI des hypothèses sur les processus. Embretson (1985) a proposé un modèle pour les cas où il est possible de décomposer la résolution de l'item en plusieurs étapes hiérarchisables. Ces différentes étapes donnant chacune lieu à l'estimation de paramètres spécifiques, l'auteur est parvenu par cette méthode à identifier les processus employés par les sujets (Embretson, 1995). Une autre direction consiste à prendre en compte les caractéristiques des populations dans l'estimation des paramètres. Par exemple Mislevy a proposé des modèles prenant en compte des différences entre populations en particulier quand on fait l'hypothèse que les sujets emploient des stratégies différentes (Mislevy & Verhelst, 1990 ; Sheehan & Mislevy, 1990). Ces derniers modèles ont été adaptés pour le traitement des données des enquêtes internationales.

Synthèse : les modèles de réponse à l'item ont apporté un renouvellement profond de la psychométrie depuis une cinquantaine d'années. Leur capacité à traiter de manière globale l'ensemble des informations sur les items et les sujets facilite fortement l'interprétation des résultats d'épreuves d'évaluation des compétences. Leurs inconvénients est de requérir des conditions fortes (unidimensionnalité, indépendance locale) pour être appliqués de manière valide. Les enquêtes internationales ont privilégié ces modèles car ils permettent également de traiter de l'identification des biais et de la comparaison des résultats de différentes populations. De ce fait, on a privilégié une approche unidimensionnelle et une définition de la compétence largement basée sur la psychométrie.

2 - LES CONDITIONS DE VALIDITE

Les modèles de mesure requièrent différentes conditions pour être valides. Nous allons aborder deux parmi les plus importantes d'entre elles : l'une est relativement connue : l'unidimensionnalité, la seconde est moins connue mais tout aussi fondamentale : l'indépendance locale ou conditionnelle des items. Ces points peuvent paraître techniques et éloignés des préoccupations pratiques de l'évaluation des compétences des élèves. Mais ils sont essentiels pour comprendre les problèmes posés pour assurer la comparabilité de la mesure, et les moyens mis en œuvre pour les résoudre, ainsi que les conséquences que peuvent avoir les solutions adoptées sur la fiabilité et l'interprétation des résultats.

2.1 - L'UNIDIMENSIONNALITE

Le nombre de variables à introduire dans un modèle pour rendre compte d'un ensemble de comportements est une question classique en psychologie. Dans le cadre des modèles de mesure, si on se situe au niveau des items et non plus des variables, cette question peut être formulée en termes de dimensionnalité d'un ensemble d'items.

La question centrale posée est celle de la prise en compte de différentes dimensions et par conséquent de plusieurs compétences expliquant la performance des sujets aux items. Si l'on considère par exemple trois échelles, les relations entre leurs scores peuvent se situer entre deux situations extrêmes : 1) il n'existe aucune relation entre elles ; 2) la relation entre les dimensions est tellement forte qu'il n'y a pas lieu de les distinguer, elles mesurent la même chose. Dans le cas 1, les dimensions sont orthogonales (les corrélations sont nulles), il faut présenter et interpréter les résultats de chacune des échelles séparément. Dans le cas 2, les corrélations sont proches de 1, il n'y a pas lieu d'interpréter séparément les dimensions., les compétences mesurées sont complètement redondantes et ne se distinguent que par un artefact sémantique qui consiste à les nommer différemment. La plupart du temps, les données se situent entre ces deux pôles. La question porte, alors, sur à partir de quel seuil la liaison entre les dimensions peut-elle être estimée comme suffisamment faible pour considérer que les dimensions mesurées correspondent à des compétences différentes. Cette question a été au coeur de la plupart des débats autour des modèles psychologiques des aptitudes. On sait qu'un modèle unifactoriel, celui du psychologue anglais Spearman, était opposé à un modèle pluraliste défendu en particulier par le psychologue et mathématicien américain Thurstone. On sait également que ces deux modèles avaient été mis en relation avec les régimes politiques dans lesquels vivaient les chercheurs : le pouvoir centralisateur pour Spearman et son facteur unique, le régime fédéral américain pour Thurstone et sa théorie multifactorielle. On sait également que ces questions du nombre de facteur ont été le moteur des évolutions de l'analyse factorielle anglo-saxonne et dans son étape actuelle des modèles d'équations structurales et de l'analyse factorielle confirmatoire. En fait, les modèles hiérarchiques ont permis de montrer l'équivalence de ces modèles. Un premier niveau de facteurs rend compte des performances dans des tests et un second niveau explique les performances communes à l'ensemble des aptitudes de premier niveau. Le choix portera alors sur l'intérêt de mesures spécifiques mais en partie redondantes ou d'une seule mesure générale, plus éloignée des contextes.

Cette question de la dimensionnalité des compétences s'inscrit dans ce débat. On cherche à savoir si les résultats peuvent être présentés sur une ou plusieurs échelles. Cependant, ce

débat apparaît faussé car pour des raisons de fiabilité de la mesure, on s'attache au fait que les épreuves soient fortement unidimensionnelles car cette condition est requise par le modèle de mesure employé : le MRI. L'unidimensionnalité est à la fois la structure recherchée et la condition (l'hypothèse au sens de *assumption*) de différents modèles : du modèle unifactoriel bien sûr, mais aussi des MRI. En effet, les modèles de base des MRI nécessitent la condition d'unidimensionnalité : on doit rendre compte des relations entre items (estimés par leurs paramètres) et entre les sujets ainsi qu'entre items et sujets par une seule variable latente. Il existe des MRI à plusieurs dimensions (voir pour les développements récents dans ce domaine, Fischer et Molenaar, 1995, ainsi que van der Linden et Hambleton, 1997), on en trouvera plusieurs exemples dans les traités récents sur les avancées des MRI, mais ils sont encore peu utilisés et je ne les aborderai pas dans le cadre de cette présentation. Je voudrai ici m'attarder sur les moyens de vérifier cette unidimensionnalité. On peut bien sûr vérifier l'unidimensionnalité à partir de l'adéquation des modèles aux données. L'utilisation des modèles d'équations structurales à cet effet a été présentée plus haut. Je voudrais, ici, m'intéresser aux MRI. Les algorithmes d'estimation des paramètres des MRI travaillent itérativement en minimisant l'écart entre les fréquences de réussite (ou les patrons de réponses) observés et estimés (à partir des estimations des paramètres du modèle). Lorsqu'un minimum est atteint, on calcule un indicateur d'adéquation entre les données et les estimations. Si l'hypothèse de l'existence d'un écart ne peut être rejetée, alors on considère que les estimations, donc le modèle, rendent compte des données. Cela peut permettre de considérer que les données sont unidimensionnelles dans la mesure où un modèle unidimensionnel en rend compte. Mais, ce raisonnement peut apparaître tautologique. De plus, la décision de rejet de l'hypothèse nulle n'est pas toujours simple du fait de la trop grande sensibilité des indicateurs à la taille des échantillons. D'autres procédures s'appuyant sur le rapport de la taille de la première valeur propre d'une analyse en facteurs principaux et la taille de la suivante ont parfois été préconisés (voir pour un exemple de mise en œuvre Bonora et Vrignaud, 1997a et 1997b). Les approches des modèles structuraux apparaissent plus prometteuses dans la mesure où la réponse en termes d'indicateurs est plus claire. Encore faut-il utiliser des statistiques appropriées. En effet, on insiste sur le fait que les indicateurs d'adéquation des modèles structuraux sont valides lorsque la matrice à ajuster est la matrice des covariances de variables dont les distributions sont multinormales. Dans de nombreux cas, en psychométrie, les réponses des sujets sont dichotomiques (succès/échec) ou ordinales. Des solutions ont été proposées, par exemple l'utilisation des matrices de corrélations polychoriques dans LISREL ou dans M-PLUS de Muthèn, ou dans d'autres logiciels le recours à des indicateurs insensibles à la forme des distributions estimés par la méthode du *Bootstrap*. D'autres approches de l'unidimensionnalité ont été proposées en confrontant les données observées à des données simulées dont on connaît l'écart à la dimensionnalité (Nandakumar, 1994).

On peut penser que cette inflation méthodologique éloigne des questions d'évaluation. Ce n'est pas mon opinion, en effet, il est indispensable pour raisonner sur un construit psychologique, de le valider en testant les hypothèses auxquels il donne lieu. L'hypothèse d'unidimensionnalité justifie l'utilisation de la variable mesurée pour classer les sujets sur un continuum. Il faut signaler que Harvey Goldstein Goldstein, 2004 ; Goldstein, Bonnet, Rocher, *submitted*) a montré, en appliquant les modèles d'équations structurales aux données anglaises et françaises que les données n'étaient pas unidimensionnelles, mais à tout le moins bidimensionnelles. L'écart à l'unidimensionnalité est révélateur de failles dans le dispositif de mesure et doit recevoir une interprétation psychologique. Je montrerai les implications de ces éléments à propos des biais d'items ou plutôt du Fonctionnement

Différentiel des Items (FDI) qui sera présenté et discuté à propos des comparaisons interculturelles.

Synthèse : On peut donc dire que la plupart des enquêtes internationales privilégient l'unidimensionnalité. Cette option est en accord complet avec la théorie psychométrique et les modèles de mesure utilisés. Ceci a des conséquences importantes sur la présentation des résultats qui privilégieront l'ordre des performances des pays donc leur classement.

2.2 - L'INDEPENDANCE LOCALE

2.2.1 - Définition

Je commencerai par rappeler brièvement ce qu'est l'indépendance locale. La plupart des modèles psychométriques nécessitent de faire l'hypothèse que la réponse d'un sujet à un item ne dépend pas de ses réponses aux autres items de l'instrument[55]. En termes psychométriques, on considère que la réussite d'un sujet à un item ne dépend que de sa compétence sur le trait latent mesuré par l'item et de rien d'autre (en particulier pas de ses réponses aux items antérieurs). En termes de probabilités, la probabilité de réussite à deux items quelconques d'un test est le produit des probabilités de réussite à chacun des items. Si on ne peut pas retenir l'hypothèse d'indépendance conditionnelle, alors il faudrait introduire un troisième terme représentant la dépendance conditionnelle entre ces deux items comme la probabilité particulière de réussite à ces deux items, leur interaction. Il est souvent difficile de tester l'hypothèse d'indépendance conditionnelle, il est, par contre, aisé d'identifier des situations où, par construction, cette hypothèse est violée : par exemple le cas des items en cascade quand on cote la réponse et sa justification. Je tiens à préciser que l'indépendance locale est une condition nécessaire pour l'ensemble des modèles de mesure. Elle est explicitement exigée dans les MRI qui représentent les compétences des sujets et les difficultés des items en termes de probabilités : la cote de réussite d'un item versus la réussite à un autre item est estimée par le rapport entre les paramètres de difficulté de ces items. Cette limite des MRI aurait pu être dépassée, comme le suggérait le statisticien anglais Harvey Goldstein, en introduisant des paramètres supplémentaires qui prendraient en compte l'indépendance locale (Goldstein, 1980). Par exemple Bradlow, Wainer et Wang (1998) proposent un MRI incluant des paramètres représentant la dépendance locale et élaborent un algorithme permettant l'estimation de ces paramètres. Elle peut, en revanche, être traitée explicitement dans les modèles structuraux en l'introduisant dans le modèle sous forme de covariances d'erreur.

2.2.2 - Des conséquences mésestimées

La condition d'indépendance locale est essentielle pour la fiabilité de la mise en œuvre des MRI puisque les paramètres des items sont estimés et valides sous cette condition d'indépendance locale. Et, pourtant on cherche rarement à vérifier que les réponses des

[55] La réponse d'un sujet ne dépend pas, non plus, des réponses des autres sujets à cet item. Les réponses à un test ou à un questionnaire psychométrique se présentent sous la forme d'une matrice comprenant les sujets en ligne et les items en colonne. L'indépendance conditionnelle joue autant entre les colonnes (facette items) qu'entre les sujets (facette ligne). L'hypothèse d'indépendance locale est également requise en ce qui concerne les sujets. L'indépendance locale postule que la probabilité que deux sujets réussissent le même item ne dépend que de leur compétence, de leur position sur le trait latent. Pour la clarté de cet exposé notionnel, je me limiterai à la facette items.

sujets sont bien indépendantes ou plutôt qu'on ne peut pas mettre en évidence de dépendance entre les réponses aux items. Cela tient peut être moins à l'ignorance de cette condition qu'à la difficulté de la vérifier. J'ai mis en en œuvre plusieurs approches possibles pour vérifier cette condition dans une étude portant sur l'évolution des compétences en mathématiques des élèves de collège à dix ans d'intervalle (Bonora et Vrignaud, 1997a et 1997b).

L'absence d'information sur le respect de cette condition est regrettable mais il est plus grave de tenir cette hypothèse pour vraie alors qu'elle ne l'est pas par construction. On peut trouver de nombreuses situations de *testing* où, par construction, la condition d'indépendance locale n'est pas respectée (Vrignaud, 2003). Ainsi, dans l'évaluation de la lecture ou plutôt de la littéracie, on demande souvent de répondre à plusieurs questions posées sur le même texte. Cette manière de procéder se justifie par le fait que l'investissement du sujet, tant cognitif que temporel, pour s'approprier des objets complexes, ici un texte, doit être rentabilisé au mieux. On utilise en anglais l'expression de *testlet* pour de tels exercices comprenant plusieurs items. J'ai malheureusement pu constater que les psychométriciens n'ont jamais évoqué ni tenu compte des biais induits par cette dépendance dans le traitement des résultats des enquêtes internationales sur la littéracie que j'ai été conduit à expertiser (Dickes & Vrignaud, 1995). Ces bais ont pourtant des effets non négligeables comme l'ont montré les quelques recherches réalisées sur les *testlets* (par exemple Wainer & Thissen, 1996). Les indicateurs psychométriques classiques tels que l'alpha de Cronbach sont biaisés dans le sens d'une surestimation. Actuellement, on cherche à construire des modèles prenant en compte la dépendance locale.

Synthèse : La condition d'indépendance locale est requise pour la validité des modèles de mesure psychométriques, en particulier les MRI. Cette condition devrait être vérifiée avant l'application de ces modèles. Elle est parfois violée par construction par exemple dans le cas de questions portant sur un même texte. Ce point n'introduit pas forcément des biais majeurs, mais, on peut être surpris de ne pas toujours voir ce point abordé dans des traitements d'enquêtes présentant sur d'autres points méthodologiques un haut degré de sophistication.

CHAPITRE VI - METHODOLOGIE DES COMPARAISONS INTERNATIONALES

1 - RAPPEL HISTORIQUE

Les questions d'équivalence de la mesure selon les caractéristiques des sujets se sont posées à partir du constat de l'existence de différences entre des groupes identifiés selon leur contexte socio-culturel. Cette question s'est développée, depuis le début du siècle dernier, de manière complémentaire, dans deux domaines de la psychologie : la psychologie interculturelle et la psychométrie. L'objectif des deux disciplines est similaire : assurer l'équivalence de la mesure, mais la position épistémologique de l'utilisateur du dispositif de mesure est d'une certaine manière différente. En psychologie interculturelle, on se situe d'emblée dans une perspective comparative. En psychométrie, il s'agit plutôt de vérification, d'analyses post hoc, visant à garantir la fiabilité de la mesure. Si le cadre d'interprétation diffère, les méthodes utilisées sont dans la plupart des cas les mêmes. Cette situation nous conduit à poser la question : « A quelle notion de culture fait-on référence lorsqu'on qualifie les biais de culturels ? »

Au début du siècle, les recherches comparatives menées sur des sociétés « éloignées » des sociétés occidentales, étaient l'objet de la psychologie interculturelle. Devant les difficultés que pose la définition de sociétés et de groupes sociaux qui vont servir à établir les termes de la comparaison, la psychologie interculturelle a redéfini progressivement les sujets de ses études comparatives. Les recherches de la psychologie interculturelle s'étendent aujourd'hui aussi bien de petits groupes homogènes - les clans, les ethnies - que de plus vastes groupes hétérogènes - les nations ou les sous-groupes identifiés dans la notion de pluralité culturelle de nos sociétés actuelles. Ainsi la comparaison des performances à des tests d'aptitude entre des groupes de sujets issus de différentes classes sociales, de différents groupes ethniques tels qu'ils se voient définis aux Etats Unis ressort de la même démarche, des mêmes méthodes que la comparaison entre des sociétés différentes (les ethnies des anthropologues, les nations). Dans de nombreux cas, ces groupes se distinguent aussi par leurs langues. La nécessité de traduire ou plutôt d'adapter les dispositifs utilisés pose des problèmes particuliers.

Sur le plan historique, il peut être utile de rappeler qu'un mouvement de remise en question des tests s'est particulièrement développé aux Etats-Unis. Une des conséquences du constat de l'existence de différences notables entre groupes sociaux plus d'un écart type et de l'ampleur variable de cette différence selon les domaines cognitifs, a été la construction de tests culture free, c'est à dire censés minimiser les écarts entre groupes sociaux (sur les tests culture free voir la synthèse de Demangeon, 1976). Après le quasi abandon de l'idée de tests culture free, les constructeurs de tests ont plutôt cherché à développer et à perfectionner des méthodes d'identification des biais. On cherche à assurer l'équité entre les individus, dans les situations d'évaluation, quelles que soient les caractéristiques des individus (sur les débats aux États Unis, voir Bacher, 1982, et pour des travaux récents, Benson et Hutchinson, 1997). Le terme d'équité se laisse définir en référence à égalité ou plutôt à inégalité. L'équité vise à réduire ou à compenser les situations d'inégalité entre groupes, engendrées par les systèmes sociaux. Dans le cadre de la théorie générale de la validité définie par Messick (1989), un test est considéré comme équitable s'il aboutit, pour le

psychologue, à prendre des décisions identiques pour des sujets ayant des compétences identiques, quelles que soient, par ailleurs, les caractéristiques de ces sujets. Dans l'optique de l'équité, l'étude des biais est nécessaire lorsqu'on peut faire l'hypothèse que la culture d'un groupe social dominant est introduite implicitement dans le dispositif de mesure. On s'intéresse, en général, dans ce but, aux principales variables souvent invoquées comme pouvant freiner le parcours scolaire et professionnel des personnes : genre, origine sociale des parents, origine culturelle, ethnicité.

Ces problèmes sont en partie à l'origine de la recherche de tests *culture free* qui pourraient évaluer les performances de sujets en éliminant les biais induits par les différences de contexte culturel. Ces tests *culture free* étaient d'ailleurs autant sinon plus défendus pour l'évaluation de groupes sociaux « défavorisés » de nos sociétés que pour les recherches de la psychologie interculturelle. Bien souvent, on faisait l'hypothèse qu'il suffisait qu'un test ne fasse appel au langage ni pour la compréhension des consignes ni pour la résolution des items pour être *culture free*. Les tests non verbaux comme les matrices de Raven ont longtemps eu cette réputation et l'ont gardée comme en témoignent la comparaison des résultats collectés dans différents pays présentés par John Raven ou la défense, dans un récent colloque sur l'adaptation des tests, de l'utilisation d'un test non verbal comme alternative aux procédures de traduction et d'adaptation par Bracken, Naglieri et Bardos (1999).
Dans les travaux sur des sociétés où les individus sont a priori peu familiarisés avec les situations impliquées par les procédures de test, on a essayé également de pallier cette situation par une procédure de familiarisation ou d'apprentissage. On peut citer à ce propos les recherches du britannique Dempster (1954), d'Ombredane et collaborateurs (Ombredane, Robaye, & Plumail, 1956) dans des populations africaines.

Synthèse : Il est important de souligner que ce type de recherche repose sur une perspective comparatiste unidimensionnelle avec une représentativité limitée des comportements évalués. Les choix faits dans le domaine des enquêtes internationales qui ont majoritairement privilégiés l'unidimensionnalité et un type d'échantillonnage des tâches. On ne prend pas en compte les mêmes compétences en partant comme dans les enquêtes de l'IEA des curricula ou d'une définition d'une compétence générale à valeur universelle comme dans les enquêtes de type IALS et PISA.

2 - ASSURER L'EQUIVALENCE

Les comparaisons ne peuvent être fidèles que si on peut démontrer qu'il existe une équivalence des mesures dans les différents pays. Le fait d'administrer une même épreuve traduite dans les langues des pays considérés ne suffit pas pour considérer que les scores obtenus par les élèves de différents pays peuvent être comparés entre eux. On définit trois niveaux d'équivalence de la mesure :

- l'équivalence de construit : la conception de l'instrument, l'arrière plan théorique permettent de faire l'hypothèse que les instruments évaluent la même aptitude. C'est le niveau le plus bas de l'équivalence. Il doit être démontré par exemple à l'aide d'un réseau nomothétique comportant une étude de la validité convergente dans les différents pays ;

- l'équivalence au niveau de la mesure : par exemple les échelles Fahrenheit et Celsius mesurent bien le même construit, mais elles ne possèdent pas la propriété d'équivalence, pour cela il faut procéder à une transformation linéaire ;
- l'équivalence scalaire : les échelles Celsius et Kelvin ont une équivalence de mesure mais pas l'équivalence scalaire, il faut pour cela fixer une origine commune aux deux échelles.

Procéder à des comparaisons internationales exige donc que les instruments possèdent la propriété d'équivalence scalaire. Après avoir démontré l'équivalence ou plus pragmatiquement réduit les sources potentielles de biais, il faut mettre en place une procédure de parallélisation (*equating*) pour placer les résultats nationaux sur une échelle commune. Les procédures d'*equating* reposent sur la construction d'une méthode de conversion à partir d'épreuves communes qui servent d'ancrage. Plusieurs méthodes peuvent être utilisées : des méthodes de régression, la méthode d'équipercentile et les méthodes basées sur les Modèles de Réponse à l'Item (pour une revue récente voir Kolen, 2004). Ces dernières sont aujourd'hui les plus utilisées dans la mesure où les données des enquêtes sont traitées par des MRI. D'abord, nous allons présenter les démarches permettant d'identifier et de réduire les biais culturels qui sont susceptibles de se manifester lorsqu'on administre une même épreuve à différents groupes culturels et/ou linguistiques. Ces méthodes ont été élaborées depuis largement un siècle dans le cadre de la psychométrie afin d'assurer l'équité dans l'évaluation par les tests et de la psychologie interculturelle afin de s'assurer de l'équivalence de la mesure lorsqu'on compare des populations. C'est en s'appuyant sur ce cadre méthodologique qu'ETS a élaboré des procédures principalement basées sur les MRI utilisées dans les enquêtes NAEP (Johnson, 1992) permettant de faire des comparaisons temporelles (entre les différentes cohortes ayant passé les épreuves) et fédérales (entre les différents états des USA). Ces procédures ont été ensuite étendues à la situation des enquêtes internationales. Nous présenterons ensuite les procédures d'equating.

2.1 - L'IDENTIFICATION ET LA REDUCTION DES BIAIS CULTURELS

2.1.1 - Définition

On dit qu'une mesure est biaisée dès lors qu'elle ne mesure pas ou, imparfaitement, ce qu'elle est censée mesurée. Sur l'analyse des biais en général, on peut se reporter, pour des publications en français, à notre revue de questions (Vrignaud, 2002) ainsi qu'à des ouvrages généraux sur la psychométrie (Dickes et coll., 1994 ; Laveault et Grégoire, 2002). On est en présence d'un biais culturel dès lors que la nature de la variable mesurée est modifiée en fonction des caractéristiques des sujets. Le biais culturel n'est pas dans le dispositif de mesure, il s'agit d'un effet d'interaction entre le dispositif de mesure et les caractéristiques des sujets. Les biais, les méthodes pour les identifier peuvent être classés en fonction des différents niveaux et aspects du dispositif de mesure où ils se manifestent : le construit, la situation d'administration, les items. Il est commode d'établir une taxonomie des biais, comme le proposent en particulier Van de Vijver et Tanzer (1997), en fonction des éléments précédents : biais de construit, de méthode, d'item. Le concept de biais est inséparable du concept d'équivalence. L'équivalence est la condition indispensable pour établir des comparaisons. On ne peut assumer l'équivalence en se fondant sur des démarches intuitives. Il est nécessaire de la démontrer, ce qui revient à s'assurer que le

dispositif de mesure est dépourvu de biais ou, de manière plus modeste, que les méthodes mises en œuvre n'en identifient pas.

L'étude des biais s'est développée principalement et de manière complémentaire, dans deux domaines de la psychologie : la psychologie interculturelle et la psychométrie. L'objectif des deux disciplines est similaire : assurer l'équivalence de la mesure, mais la position épistémologique de l'utilisateur du dispositif de mesure est d'une certaine manière différente. En psychologie interculturelle, on se situe d'emblée dans une perspective comparative. En psychométrie, il s'agit plutôt de vérification, d'analyses *post hoc*, visant à garantir la fiabilité de la mesure.

Un mouvement de remise en question des tests s'est particulièrement développé aux Etats-Unis du fait des contestations sociales, souvent traduites en actions judiciaires, liées à l'éventuelle présence de biais, à l'encontre de certains groupes sociaux, dans des épreuves utilisées pour la sélection professionnelle ou pour l'entrée à l'université. Après le quasi-abandon de l'idée de tests « *culture free* », les constructeurs de tests ont plutôt cherché à développé et perfectionner des méthodes d'identification des biais culturels. On cherche à assurer l'équité entre les individus, dans les situations d'évaluation, quelles que soient leurs caractéristiques (sur le développement des méthodes aux Etats-Unis voir Benson et Hutchinson, 1997 ; on trouvera un exposé historique des débats politiques et sociaux sur la question de l'équité dans l'ouvrage de Lemann (1999) sur le SAT - *Scholastic Aptitude Test* - et les procédures de sélection à l'entrée dans les universités américaines). Le terme d'équité se laisse définir en référence à égalité ou plutôt à inégalité. L'équité vise à réduire ou à compenser les situations d'inégalité entre groupes engendrées par les systèmes sociaux. Dans le cadre de la théorie générale de la validité définie par Messick (1989), un test est considéré comme équitable s'il aboutit pour le psychologue à prendre des décisions identiques pour des sujets ayant des compétences identiques quelles que soient, par ailleurs, les caractéristiques de ces sujets. Dans l'optique de l'équité, l'étude des biais est nécessaire lorsqu'on peut faire l'hypothèse que la culture d'un groupe social dominant est introduite implicitement dans le dispositif de mesure. On s'intéresse en général dans ce but aux principales variables souvent invoquées comme pouvant freiner le parcours scolaire et professionnel des personnes : genre, origine sociale des parents, origine culturelle, ethnicité.

Dans les enquêtes internationales l'étude des biais vise à vérifier que l'épreuve utilisée n'est pas biaisée dans les différents groupes comparés qui sont définis comme les nations ou au sein d'un même pays des groupes identifiés par l'usage d'une langue commune. Ce point a une importance méthodologique capitale dans la mesure où l'étude des biais se fera sur des versions linguistiques différentes. Nous reviendrons plus longuement sur ce point après avoir exposé les méthodes d'identification des biais d'item.

Nous exposerons d'abord les biais de construit et de méthode. Nous présenterons ensuite l'identification des biais d'items en développant les principales méthodes utilisées. Cette dernière partie est plus longue que les précédentes pour au moins quatre raisons : 1) les items étant les briques à partir desquelles sont bâtis les construits de niveau supérieur (test, questionnaire), l'existence de biais à ce niveau implique qu'ils se retrouveront aux autres niveaux ; 2) les méthodes utilisées pour identifier les biais d'items nous paraissent avoir une valeur pédagogique exemplaire pour exposer ce qui se joue dans les biais ; 3) un soin particulier a été apporté à la détection des biais d'items dans les enquêtes internationales ; 4) les biais d'items nous semblent être le type de biais le moins connu, du moins dans les

milieux francophones, alors qu'ils ont donné lieu à nombre de développements méthodologiques très importants au cours des vingt dernières années.

2.1.2 - Les biais de construit

Les biais de construit invalident le dispositif de mesure puisqu'ils mettent en évidence l'inconsistance de la variable mesurée entre les groupes. Rappelons que ce qui est observé est la performance à un test. Pour le psychométricien, cette performance dépend d'une variable non observée qualifiée de latente. Lorsque l'instrument ne mesure pas la même variable latente selon les groupes en présence, on est en présence d'un biais de construit.

Les mesures des aptitudes dans des sociétés différentes sont parmi les exemples les biais de construit les plus connus. En effet, le concept d'intelligence à la base des tests n'est pas représentatif des comportements considérés comme intelligents dans de nombreuses sociétés (ce que montre bien Gardner (1993/1997) dans sa théorie des intelligences multiples). On peut rattacher ces biais de construits à des positions d'absolutisme culturel (voir la définition de ces termes dans le paragraphe consacré à la référence) où l'on impose aux différentes sociétés le cadre théorique construit dans une autre. Une des manières de pallier les biais de construit est d'adopter un cadre de relativisme culturel modéré, en définissant le concept à partir des points de vue des différentes sociétés étudiées, par exemple, en échantillonnant les items à partir des définitions de l'intelligence produites dans les différentes sociétés.

Dans le cas des enquêtes internationales, la question des biais de construit se pose de manière différente selon la façon dont la compétence mesurée a été définie et construite. Dans le cas des enquêtes de l'IEA où on a cherché à établir un noyau commun à partir des curricula nationaux, on s'est ainsi assuré de la pertinence du construit dans les différents pays. Dans le cas des approches de type IALS ou PISA, on se retrouve dans une situation proche de celle de la mesure de l'intelligence, on s'appuie sur la psychométrie pour définir le construit. On pourrait de manière humoristique paraphraser la boutade attribuée à Binet « l'Intelligence ? C'est ce que mesure mon test… », en disant « La littéracie ? C'est ce que mesure notre épreuve… ». Il faudra démontrer que la compétence évaluée est bien pertinente dans les différents pays. Nous discuterons ce point dans le paragraphe traitant de la compétence.

2.1.3 - Les biais de méthode

Les biais de méthode concernent moins l'instrument que ses conditions d'administration et de passation. Un biais de méthode bien connu est celui de l'inégale familiarisation avec le matériel ou les procédures. Ainsi, beaucoup de travaux sur la comparaison des performances aux tests d'intelligence auprès de populations traditionnelles sont biaisés car le matériel n'était pas familier à des populations pas ou peu scolarisées. D'autres biais peuvent être induits par les différences culturelles entre l'administrateur du test et le répondant. On observe également des biais de méthode produits par l'incomparabilité des échantillons. Par exemple, les élèves de certains pays peuvent être davantage motivés pour participer à une enquête internationale.

La composition de l'échantillon est un problème essentiel dans les enquêtes internationales. La définition des populations entrant dans la comparaison (quels élèves peuvent être exclus ?) doit être très claire pour éviter des biais induits par des différences de composition

des populations. On a également souvent évoqué les biais liés aux différences d'exposition à l'enseignement et aux programmes (*opportunity to learn*).

A titre d'exemple de biais induits par la composition de l'échantillon, on peut rappeler un aspect des traitements de l'enquête IEA de 1970 qui portait sur l'enseignement des sciences dans 19 pays (cf. Comber & Keeves, 1973). Les variables indépendantes avaient pu être regroupées en 4 « blocs « ordonnés chronologiquement, ce qui permettait d'inférer le sens de la causalité, et d'appliquer une analyse de régression « pas-à-pas « (stepwise), tout en se donnant de bonnes chances d'épuiser l'univers des variables les plus pertinentes. Or, à l'issue de l'analyse inter-élèves (population IV : fin de scolarité secondaire), on a constaté que les variables familiales ont un poids nettement plus important aux USA que dans les autres pays en moyenne (9% de v. expliquée contre 2%). Or, ce phénomène ne tient pas tant à un effet plus important ici que là de l'environnement familial, mais plutôt au fait que le taux de scolarisation à ce niveau scolaire était beaucoup plus élevé aux USA. Dans les autres pays, la sélection s'étant exercée plus tôt dans le cursus, avait entraîné une homogénéisation du milieu familial, dont l'effet ne pouvait plus, de ce fait, s'exprimer en fin de cursus.

2.1.4 - Les biais d'items

Traditionnellement qualifiés de biais item/test, les biais d'items sont des nuisances pour la qualité de la mesure. En effet, on peut montrer qu'un item biaisé mesure une autre variable que la variable qu'il est censé mesurer et que cette variable « parasite » favorise ou défavorise un des groupes étudiés. Une nuisance est ainsi introduite dans la mesure. Un exemple historique de biais d'item a été fourni par un test de vocabulaire suédois qui comportait une question sur le vocabulaire de la pêche à la ligne. Cet item était moins bien réussi par les filles que par les garçons du fait que l'activité à laquelle il faisait référence était une activité davantage pratiquée par les garçons que par les filles. Cet item ne mesurait pas la maîtrise de la langue (variable mesurée) mais des connaissances spécialisées dépendantes du genre (variable parasite). L'étude des biais item/test s'inscrit maintenant dans la perspective plus vaste du concept de Fonctionnement différentiel de l'item (en abrégé, FDI), formulation plus neutre qui permet de considérer le FDI comme une méthode heuristique d'étude des différences entre groupes au niveau des items et non plus seulement comme une nuisance psychométrique.

2.1.4.1 - Définition

Il est important de bien distinguer deux notions qui rendent compte d'éventuelles différences de performance entre des groupes : 1° l'impact du test, 2° le FDI. L'impact est un effet général de l'appartenance à un groupe sur la variable mesurée. Par exemple, les scores moyens des hommes sont en général inférieurs aux scores moyens des femmes dans les tests verbaux. Cet effet du genre sur la performance générale aux épreuves verbales est l'impact. En présence d'impact la réussite à chacun des items devrait être plus élevée (ou plus basse) dans un des groupes pour tous les items.

Le FDI porte, lui, sur les différences de « réussite » aux items indépendamment de l'effet éventuel de l'impact. Pour éliminer l'effet de l'impact, il faut comparer la réussite des sujets de chacun des groupes aux items suspects de FDI en tenant la performance égale. Le FDI peut porter sur chacune des caractéristiques de l'item : 1° sa difficulté ; 2° sa discrimination.

Lorsque la différence de réussite à l'item est de même sens en faveur ou en défaveur du même groupe dans toutes les classes de sujets le FDI est dit « uniforme ». Le FDI uniforme porte uniquement sur la difficulté de l'item. Il existe un écart en faveur du même groupe à tous les niveaux de compétence. Lorsque la différence de réussite change de sens selon le niveau de performance des sujets (par exemple la différence est en faveur d'un groupe pour les classes de performance faibles et en défaveur du même groupe pour les classes de performance élevée) on parle de FDI « croisé ». Le FDI croisé porte sur la discrimination de l'item celui ci est plus discriminant dans un groupe que dans l'autre. Si on se représente aisément la signification psychologique d'un FDI uniforme, celle d'un FDI croisé peut être plus délicate.

2.1.4.2 - Identification du FDI

Différentes méthodes ont été développées pour identifier le FDI. Pour un ouvrage méthodologique introductif on peut se référer à Camilli et Shepard, 1994, ou, en français, à Vrignaud, 2002 ; Laveault et Grégoire, 2002 ; l'ouvrage consacré au symposium conduit sur ce sujet à *Educational Testing Service* par Holland et Wainer, 1993 fournit des exposés méthodologiques détaillés et aborde également les aspects sociaux et juridiques du FDI ; signalons, enfin, la revue de Millsap et Everston (1993) qui replace les différentes procédures dans un cadre conceptuel particulièrement clair.

Précisons d'emblée que comparer directement les fréquences de réussite des items dans les groupes étudiés ne permet pas d'identifier correctement le FDI, car cette approche ne distingue pas entre l'impact et le FDI. Pour identifier le FDI, il est nécessaire de comparer la réussite des sujets aux items à compétence égale pour éliminer l'effet éventuel de l'impact. Les méthodes se distinguent selon que, pour constituer des classes de sujets de compétence homogène, elles effectuent l'appariement des sujets sur la variable observée (c'est à dire le score) ou sur une variable latente (c'est à dire les valeurs de compétence des modèles de réponse à l'item). On peut également distinguer les méthodes selon leurs caractères paramétriques ou non.

Il existe des méthodes travaillant à partir du score observé : la méthode de Mantel Haenszel (non paramétrique) et la régression logistique (méthode paramétrique), d'autres méthodes travaillant à partir de variables latentes : l'approche par les modèles de réponse à l'item (que nous abrégerons désormais en MRI), méthode paramétrique et, une méthode non paramétrique, l'indicateur de biais simultané. Nous présentons ces méthodes en annexe.

2.1.4.3 - Portée et limite de l'étude du FDI dans les enquêtes internationales

Comme nous l'avons écrit supra, l'étude des FDI selon une ou plusieurs des méthodes disponibles est indispensable pour vérifier l'équivalence entre les différentes versions linguistiques de l'épreuve. Nous insistons à nouveau sur le fait que la comparaison des fréquences de réussite n'informe en aucune manière sur l'existence éventuelle de biais. Dans le plupart des enquêtes, des items suspects de FDI ont été éliminés ou ont été l'objet d'un traitement particulier. Par exemple on calcule des spécifiques pour les paramètres de l'item dans le ou les pays où un FDI est identifié ; ces valeurs spécifiques remplaceront la valeur commune utilisé pour l'ensemble des pays (cette méthode a été utilisée dans l'enquête IALS).

On peut donc dire que l'identification des FDI est faite de manière systématique et soigneuse. Il faut néanmoins mentionner trois problèmes pour lesquelles les solutions

actuelles ne sont pas complètement satisfaisantes et qui expliquent que certains biais peuvent ne pas avoir été identifiés correctement par les méthodes employées. Ces trois principaux problèmes sont celui de la traduction, de la référence et des questions méthodologiques.

Le choix de(s) la référence(s)

Dans la recherche de FDI entre deux groupes, un des groupes est considéré comme la référence, l'autre comme le groupe cible. Si ce choix a peu d'importance dans le cas de deux groupes, il devient plus crucial dans le cas de plusieurs groupes. Dans le cadre d'une comparaison internationale entre plusieurs pays prendra-t-on un des pays comme référence ou procédera-t-on à toutes les comparaisons entre les pays deux à deux ? Ce choix renvoie à une approche plutôt interne (centrée sur chacune des populations étudiée) ou externe (centrée sur une population considérée comme référence). Effectuer toutes les comparaisons possibles est coûteux et peut induire des artefacts[56]. On peut également prendre comme référence l'ensemble formé par l'agrégation de tous les pays (en laissant en dehors le pays étudié) mais dans ce cas le risque est que cette agrégation peut elle-même contenir de nombreux items suspects qui vont biaiser la détection du FDI. Il faudrait procéder en plusieurs étapes et en éliminant les items suspects à chacune des étapes. Ce processus est très coûteux et là aussi les résultats ne sont pas garantis (voir sur ce point Millsap et Everston, 1993).

En général dans les enquêtes internationales, on a procédé en utilisant un des pays comme référence. Ce choix est loin d'être neutre. Car conceptuellement, le FDI est un écart à l'unidimensionnalité. L'item est sensible à plus d'une dimension : celle qu'il est censé mesuré et une autre dimension parasite. Comme nous l'avons exposé à propos de l'unidimensionnalité, l'ensemble des items n'est jamais complètement expliqué par une dimension. Choisir un pays de référence revient donc à fixer la dimension en l'alignant sur les réponses, les performances d'un pays donné. Ce choix a donc des implications importantes d'abord sur la construction de la dimension et, ensuite, sur les écarts à la dimension qui pourront être interprété comme des FDI. On peut dire que si l'on avait adopté une solution différente, d'autres items auraient été identifiés comme biaisés pour d'autres populations.

Bien sûr, les méthodes utilisées pour l'identification du FDI sont relativement robustes et on peut considérer que les cas de FDI les plus flagrants sont correctement détectés. On souhaiterait cependant que ce choix de la référence soit davantage discuté, que des contre-validations soient effectuées en changeant la référence.

Les items traduits sont-ils identiques ?

Les méthodes d'identification du FDI ont été mises au point pour comparer des groupes qui ont passé les *mêmes* items : par exemple les hommes et les femmes. Dans le cadre des comparaisons internationales, on compare le fonctionnement d'items qui ne sont plus absolument identiques puisqu'ils ont été traduits. On a pu montrer qu'il est possible que les items soient de difficulté différentes dans deux langues sans que le FDI soit identifié (Sireci,

[56] On sait que les seuils de signification sont biaisés dans le cas de comparaisons non orthogonales (faux positifs). Ainsi si on compare dix pays, on devra faire 45 comparaisons. On peut s'attendre à trouver deux comparaisons significatives à .05 du seul fait du hasard.

1997 ; 1999). L'idée défendue par Sireci est que, s'il existe un biais sur l'ensemble des items d'une version linguistique donnée, alors ce biais sera indétectable puisqu'il sera confondu avec l'impact. La fiabilité des procédures statistiques repose sur l'hypothèse d'équivalence des traductions. En effet, un item peut être plus facile dans une langue que dans l'autre du fait de l'utilisation de termes similaires dans la question et la réponse pour une langue mais pas pour l'autre. Des procédures peuvent être préconisées (jugements d'experts sur l'équivalence de la difficulté à partir de critères linguistiques, recueil de données sur des sujets bilingues). Mais on voit que ces remèdes sont loin de régler la question et que l'équivalence d'épreuves traduites ne peut jamais être complètement démontrée par le recours unique à des méthodes psychométriques. Le minimum serait de réaliser des analyses complémentaires de FDI sur des populations ayant passé la même version linguistique (par exemple la version française administrée en Belgique, en France et en Suisse).

Choix des approches méthodologiques

Nous avons présenté quatre approches possibles pour l'identification des FDI. Il est certain que l'utilisation des indices de biais fournis par les MRI est cohérente avec l'utilisation de ces derniers comme modèle de mesure pour l'ensemble de l'enquête. Cependant, il serait utile de compléter l'étude par les résultats d'autres approches complémentaires. Nous avons déjà évoqué l'intérêt de l'approche de Stout et collaborateurs. Cette approche serait particulièrement pertinente dans la mesure où elle permet de tester des fonctionnements différentiels de regroupements d'items. Justement, dans le cas, en particulier, de la littéracie, le test est constitué d'exercices comprenant un document et plusieurs items. Il serait donc logique de tester la présence éventuelle de biais sur l'ensemble des items rattachés à un document et non pas sur les items isolés. Cette approche à partir des regroupements serait méthodologiquement plus conforme à la dépendance existant entre items rattachés à un même document. De plus, l'approche de Stout comprend également des méthodes permettant de vérifier l'unidimensionnalité d'un ensemble d'items ce qui apporterait des informations complémentaires sur ce point crucial pour le traitement des données. Il nous semble que cette approche alternative n'a pas été utilisée non, pour son manque de pertinence, mais, du fait de la valorisation des approches implémentées dans les logiciels diffusés par les organismes en charge du traitement des données (par exemple le logiciel BILOG-MG édité par ETS, le logiciel Conquest édité par ACER).

FDI uniforme ou croisé

Nous avons vu que les MRI sont l'approche la plus utilisée dans les enquêtes internationales pour l'identification du FDI. En général, on s'intéresse seulement au paramètre de difficulté (FDI uniforme). Pourtant lorsqu'on utilise le modèle à deux ou trois paramètres, l'étude du FDI devrait porter également sur le paramètre de discrimination (FDI croisé). L'étude des biais liés à différences de discrimination de l'item est techniquement plus délicate surtout dans le cadre des MRI. Elle est plus aisée à réaliser dans le cadre des modèles d'équations structurales. Certes, il n'est pas forcément souhaitable d'augmenter indéfiniment le nombre et la complexité des traitements mais cette absence de discussion apparaît paradoxale eu égard à la valorisation des méthodes et techniques employées dans les rapports sur le traitement des données. Il semble qu'on suive pour l'ensemble des traitements la pente la plus facile eu égard aux logiciels utilisés.

L'appariement

Comme nous l'avons exposé plus haut, on est en présence de FDI lorsque la réussite diffère entre les groupes pour des sujets de compétence égale. Par exemple si on étudie le FDI selon le sexe, on va constituer des classes de filles et de garçons de compétence égale. La méthode la plus simple pour constituer cet appariement est d'utiliser le score total au test. Dans le cas des MRI, l'appariement est fait sur la variable latente, les paramètres q. Cette manière de procéder peut poser problème car le ou les items biaisés vont entrer dans le calcul du score et par conséquent influer sur l'estimation de la compétence. Comme il est difficile de disposer par ailleurs d'une mesure de la compétence[57] dont on sache qu'elle n'est pas biaisée, on conseille en général de faire une première analyse en utilisant tous les items pour calculer le score. Si un FDI est identifié sur certains items, on reprendra l'analyse en appariant les sujets sur un score recalculé en excluant les items biaisés. Cette procédure est évidemment lourde à mettre en œuvre et de ce fait, elle est rarement employée. Elle n'a pas été utilisée à notre connaissance dans les enquêtes internationales.

La question de la robustesse des procédures

La littérature est riche en discussions souvent à partir d'études sur des données simulées pour produire ou non des FDI plus ou moins importants sur les avantages et inconvénients des différentes méthodes. Des questions comme l'utilisation du score observé ou d'une variable latente, l'intégration ou non de(s) l'item(s) biaisé(s) dans le calcul de la variable, la détermination de seuils pour les indicateurs descriptifs, la sensibilité (parfois trop grande) des tests de signification sont de celles qui devraient recevoir de meilleures réponses. Schématiquement, on sait que la taille de l'échantillon entre dans le calcul des tests utilisés pour identifier le FDI. On connaît bien en statistiques ce problème lié à la dépendance du test à la taille de l'échantillon. Cette dépendance augmente le risque de seconde espèce, ici, le risque de conclure à la présence d'un FDI alors qu'il n'en existe pas. En général, pour se prémunir contre cette trop grande sensibilité de la méthode, on choisit de retenir des valeurs élevées pour les tests et leur seuil de signification ou alors de travailler sur un sous-échantillon de taille plus réduite tiré de façon aléatoire de l'échantillon principal.

On peut également s'appuyer sur des études de simulation. Ainsi, la procédure employée par Flieller (1999), pour l'étude du FDI dans un test lexical, nous paraît tout à fait recommandable. On met en oeuvre la méthode d'identification selon une répartition des sujets dans des groupes aléatoires ; dans cette condition, aucun FDI ne devrait se manifester. Ceci permet d'estimer la sensibilité de la méthode utilisée et sa tendance à détecter de « faux positifs ». Flieller (1999) a pu constater qu'aucun FDI n'était détecté en comparant le groupe des sujets de rang pair et le groupe des sujets de rang impair. Les logiciels construits par Stout et collaborateur offrent également un ensemble de procédures pour produire des données simulées permettant d'étudier le fonctionnement et la robustesse des indices de FDI selon différentes conditions. Ce type de procédures n'a pas été rapporté à propos du traitement des enquêtes internationales.

L'interprétation

L'interprétation du FDI pose problème puisqu'il est fonction de la comparaison entre deux groupes d'items communs : ceux qui fonctionnent de manière similaire, et les autres, taxés de FDI. Cette procédure conduit à écarter les items qui ne s'inscrivent pas dans le processus unidimensionnel. Elle peut, dans ce cadre, également se révéler douteuse si le nombre

[57] Par exemple les résultats à un autre test évaluant la même compétence dont on ait fait la preuve qu'il est exempt de FDI.

d'items ne présentant pas de FDI est équivalent (voire inférieur) au nombre d'items présentant du FDI : la dimension unique prise en compte par le modèle est-elle bien, alors, la dimension la plus pertinente ? Dans les enquêtes internationales utilisant les modèles MRI, les items présentant un FDI seront soit éliminés, soit intégrés dans l'analyse comme items spécifiques : les valeurs des paramètres de ces items seront estimées séparément pour chaque population. Par exemple dans l'enquête IALS, sur 114 item, 13 ont été éliminés et 53 ont reçu un paramètre spécifique pour au moins un des 10 pays étudiés (tableau 10.8 p. 171). Le FDI est considéré comme une source de nuisance dont les causes ressortent aux défauts de construction du test (traduction, différences de contexte) et sont effectivement des biais. Mais d'autres facteurs peuvent intervenir : modifications dans l'environnement socio-éducatif, modification de l'enseignement des domaines spécifiques du programme. Ceci laisse penser qu'on pourrait adopter une position différente en considérant que les causes du FDI apportent, à leur manière, des informations intéressantes sur les différences entre les pays étudiés.

Synthèse : La détection des biais culturels est la pierre de touche des enquêtes internationales dans la mesure où elle permet de s'assurer de l'équivalence de la mesure dans les différentes populations. Les biais de construit sont peu discutés car il semble aller de soi que la compétence mesurée est bien la même dans les différentes populations. Un soin particulier est apporté au contrôle des différents points susceptibles d'induire des biais de méthode : comparabilité des échantillons, des conditions d'administration, de cotation. Les biais d'items ou plutôt les FDI font l'objet d'études statistiques systématiques. Cependant, les méthodes employées peuvent être discutées. D'abord car on connaît mal leur capacité à prendre en charge des items traduits. Ensuite, car elles pourraient être complétées par d'autres approches statistiques.

3 - LA MISE EN PARALLELE (EQUATING)

3.1 - DEFINITION

La comparaison directe des scores observés n'est possible que si on peut démontrer l'équivalence scalaire entre les différentes versions linguistiques de l'épreuve a pu être démontrée. Ce qui n'est jamais le cas. Il est donc nécessaire de la construire à partir des données. Ces procédures souvent désignées sous le nom d'*« equating »*, de *« scaling »* ou de *« linking »* sont plus ou moins complexes selon le plan de recueil de données. On est dans la situation d'un plan d'ancrage par les items communs pour des groupes non équivalents lorsque deux groupes (désignés g1 et g2) ont passé chacun un test différent (X pour g1, Y pour g2), chacun des deux tests comprenant des items communs considérés comme un sous-test V.

Dans le cas des enquêtes internationales, on considére que les différentes versions linguistiques de l'épreuve sont constituées d'items identiques et que tous les items sont utilisés pour la procédure d'*equating*. Le plan d'expérience qui est celui des enquêtes internationales ne présuppose pas l'identité des populations parentes (groupes non équivalents). Cette situation correspond à un plan avec ancrage par les items communs pour des groupes non équivalents. On parlerait ici d'equating entre les différentes versions d'un même test en désignant ces versions par X_A, X_B, X_C,... X_Z. On dispose de trois types de

procédures 1) la régression linéaire, 2) l'équipercentile, 3) les MRI. Nous allons les exposer en partant du cas le plus général : deux tests X et Y comprenant un noyau commun V. Une simple transposition permet de se représenter comment les procédures s'appliquent au cas des enquêtes dans lesquelles l'ensemble des items d'un même test traduit a été administré aux différentes populations.

3.2 - LES DIFFERENTES APPROCHES

3.2.1 - Equating par la régression linéaire

On va s'appuyer sur les relations entre les tests X et Y et le sous-test V commun aux deux groupes pour calculer les équations de régression entre les tests X et Y pour l'ensemble de l'échantillon. On pourra ainsi estimer les scores des sujets du groupe g1 au test Y qu'ils n'ont pas passé, réciproquement on estimera les scores des sujets du groupe g2 au test X qu'ils n'ont pas passé. Cette procédure a été développée par Angoff (cf Kolen & Brennan, 1995, chapitre 4). Il existe plusieurs méthodes pour calculer ces équations de régression. Elles varient en fonction des hypothèses sur la nature des tests (sont-ils congénériques) et leurs conditions de validité. En général, les résultats des différentes méthodes sont largement équivalents quand les données sont bien adaptées au traitement. C'est à dire quand les tests sont de même longueur, et quand les items communs remplissent les conditions d'équivalence : représentativité du contenu, similarité des emplacements.

Par exemple avec la méthode de Tucker et Levine, on fait l'hypothèse que les deux groupes constituent une seule population synthétique pour laquelle on va estimer les moyennes et les variances des deux tests pour l'ensemble des sujets. Les équations de régression seront construites à partir des paramètres de la population synthétique qui prennent en compte les différences de moyenne et de variance des deux groupes au sous-test constitué des items communs. Les résultats de la population synthétique sont constitués des résultats observés du groupe A et des résultats estimés du groupe B qui n'a pas passé ce test. On peut alors calculer les scores réduits pour les deux groupes au test passé par un seul groupe.

3.2.2 - Equating par la méthode des équipercentiles

Le rationnel de cette méthode consiste à utiliser les distributions conditionnelles des tests X et Y conditionnellement au sous-test V. On construit ainsi la table de contingence croisant les distributions (en percentiles) des scores au test X et au sous-test V. Les cases de cette table de contingence présentent pour chaque score au sous-test V la distribution des scores au test X. On construit de la même manière une table de contingence pour le test Y et le sous-test V. On s'appuie sur le principe que les distributions aux tests X et Y devraient être équivalentes conditionnellement au sous-test V. On reconstruira la distribution des scores au test Y des sujets qui ont passé le test X à partir des fréquences marginales de ces tables de contingence.

3.2.3 - Equating par les MRI

Les modèles MRI offrent plusieurs possibilités pour la mise en relation des items des deux tests et des paramètres de compétence des populations. Rappelons que lors de l'estimation des paramètres de difficulté et de compétence, chaque échelle est indéterminée. Pour lever

cette indétermination, on fixe arbitrairement la distribution d'un des paramètres de difficulté ou de compétence. En général les paramètres de la distribution des items (ou de la population) sont fixés à 0 pour la moyenne et 1 pour l'écart type. Ceci implique :

1. que des paramètres obtenus sur des données différentes ne sont pas directement comparables ;
2. que les paramètres sont invariants à une transformation linéaire près. On s'appuie sur cette propriété pour mettre sur une même échelle, difficulté des items et compétences des sujets.

Plusieurs procédures peuvent être utilisées pour un plan comprenant des groupes non équivalents avec un ancrage par des items communs. Les deux solutions généralement utilisées sont : soit la transformation linéaire à partir d'une équation calculée sur les items communs, soit l'estimation commune des paramètres des items.

Dans cette dernière procédure, tous les items sont considérés comme appartenant à un seul test. Les items communs servent d'ancrage et les items non communs sont considérés comme n'ayant pas été présentés au groupe qui a passé le test dont ces items ne font pas partie. On pourra ainsi estimer la difficulté de l'ensemble des items sur une même échelle. Une distribution différente sera estimée pour chacun des groupes, l'un des groupes servant de référence. La compétence des sujets des deux groupes sera ainsi placée sur une même échelle. Cette procédure est mise en oeuvre avec plus ou moins de sophistication par de nombreux logiciels (par exemple BILOG-MG).

Synthèse : Il n'est pas possible de comparer directement les scores à une même épreuve passée par des groupes (pays) différents. Il faut appliquer une procédure d'equating pour assurer le passage entre les scores à une version linguistique et les scores à une autre version. Ces procédures ne sont pas simplement la transformation de chacune des variables séparément mais une transformation intégrant les relations existant entre les deux variables. On dispose de trois procédures d'equating. Dans les enquêtes internationales, l'utilisation des MRI fournit une procédure d'equating. Cependant, il pourrait être souhaitable de contrôler la robustesse de cette procédure en mettant en œuvre une des autres procédures classiques.

CHAPITRE VII - LES ENQUETES INTERNATIONALES SONT-ELLES BIAISEES ?

Différentes critiques ont été évoquées par les organismes et laboratoires français et étrangers qui se sont impliqués dans une réflexion sur la validité de l'enquête IALS. Ces réflexions sont évoquées dans les rapports d'expertise en particulier ceux du projet européen sur les résultats de l'enquête IALS piloté par l'ONS anglais (Goldstein). Sur cette dernière, on trouve également les rapports commandés par la DEP : celui du GRAPCO de Nancy2 de Dickes et Flieller ; celui de l'INED par Guérin et Blum ; celui de Martine Rémond. Nous ne ferons pas ici une présentation détaillée et exhaustive des critiques évoquées par ces experts. Certaines critiques ne valent que pour certains points de l'enquête IALS qui était une enquête très particulière puisqu'elle visait à évaluer les compétences des adultes par une collecte des données à domicile auprès des ménages. Néanmoins, certaines des interrogations soulevées ont posé des problèmes plus généraux sur lesquels on peut s'interroger pour la plupart des enquêtes internationales,en particulier, celles qui cherchent à évaluer des compétences plutôt que la maîtrise de contenus d'enseignements. Nous allons présenter quelques points conceptuels et méthodologiques qui peuvent poser problème en essayant de montrer quelle peut être la portée de ces critiques.

1 - DEFINITION ET SIGNIFICATION DE LA COMPETENCE EVALUEE : L'EXEMPLE DE LA LITTERACIE

Nous avons vu que pour assurer l'équivalence, il fallait valider la nature du construit et son interprétation dans les différents pays. La contestation la plus fondamentale porte donc sur la validité des enquêtes internationales basées sur les compétences (enquête de type PISA) et non sur celles basées sur les curricula (enquêtes de type IEA). Les questions centrales sont : « que mesurent ces enquêtes ? » et « cette mesure est-elle pertinente pour l'ensemble des pays en général et la France en particulier ? ». Pour discuter ces questions, nous prendrons l'exemple de la littéracie car c'est la compétence évaluée dans les enquêtes internationales qui a été le plus discutée et commentée.

Nous avons vu que les enquêtes de l'IEA étaient basées sur les curricula. On cherchait à construire une sorte de dénominateur commun des acquis des élèves dans les différents pays. Cette méthode avait bien sûr des inconvénients : coûteuse en temps, le noyau commun n'était pas forcément représentatif des programmes de tous les pays participants. Elle avait malgré tout un mérite celle d'établir une relation directe entre les input pédagogiques et les outputs en termes de maîtrise et de compétence. L'interprétation des scores s'en trouvait d'une certaine manière facilitée.

Le choix fait par des enquêtes de type PISA d'évaluer des compétences n'est pas exempt de tout questionnement scientifique et idéologique. En effet, on se souvient des débats sur la mesure de l'intelligence et de la boutade de Binet. On court ici le risque de déclarer la compétence c'est ce que mesure notre test. En effet, comment définir la population parente de toutes les compétences nécessaires pour vivre et réussir dans notre monde moderne ? Et même, si l'on réduit ce choix à une compétence comme la littéracie, comment être sûr que l'on échantillonne les items (les tâches) de manière à réellement balayer le domaine ? Ne

court-on pas le risque comme dans les tests d'intelligence de sur-représenter voire de ne représenter que les tâches en relation avec les apprentissages scolaires et le milieu culturel dominant (par exemple aux Etats-Unis le WASP) et d'assister aux terribles dérives induites par les travaux de Terman comme l'évoquent Blum et Guérin (2000) ? Ce danger des dérives justifiées par une idéologie supportée par une psychométrie détournée de ses fins scientifiques est réel lorsqu'on assiste à une résurgence de la comparaison des performances des groupes ethniques à un ensemble de tests, comparaison justifiant la place de ces groupes dans le système social et économique (Hernnstein et Murray, 1994). Il y a un risque de dérive idéologique important à considérer ces compétences comme dotées d'une réalité autonome et objective alors qu'elles sont étroitement dépendantes de l'adaptation dans un type de société donné qui les valorise dans une optique économique.

On va justifier la compétence par la validité de la mesure mais sur quelle validité s'appuie-t-on ? Celle du modèle de mesure en l'occurrence les MRI. Nous avons évoqué à propos des MRI qu'un de leur apport intéressant était de placer sur la même variable latente la difficulté des items et la compétence des sujets. La compétence d'un sujet peut donc être interprétée en relation avec les items qui sont dans sa zone de compétence. Réciproquement, les items peuvent être interprétés en relation avec les caractéristiques et la compétence des sujets qui sont capables de les réussir. La construction de l'échelle de compétence dans les enquêtes utilisant les MRI sera donc largement basée sur les regroupements d'items à partir de leurs indices de difficulté. Ainsi, dans la plupart des enquêtes internationales on définit plusieurs niveaux (en général cinq) de compétences. L'interprétation de chacun de ces niveaux est ensuite enrichie par l'analyse cognitive des items classés dans ce niveau. Ce système de définition d'une compétence est essentiellement psychométrique même s'il reçoit un habillage de psychologie cognitive. Un tel système a été particulièrement développé par Kirsch et collaborateurs dans les enquêtes NAEP puis IALS et PISA (voir par exemple Kirsch, Jungeblut, & Mosenthal, 1998). Cette approche présente deux inconvénients majeurs.

Le premier est d'être partiellement tautologique : cet item est facile puisqu'il est réussi par un grand nombre de sujets donc il correspond à des opérations de niveau faible. Il faudrait enrichir davantage l'interprétation en utilisant les procédures psychométriques qui doivent suivre la validation du modèle de mesure : les validités critériées convergentes et divergentes. Cela n'a pas toujours été fait, voire cela a parfois conduit à mettre en doute la fiabilité de la mesure. Que penser par exemple du constat montrant en France qu'une proportion importante de diplômés de l'enseignement supérieur n'atteignent que le niveau le plus bas (extraire une information simple d'un texte) de littéracie ?

Un second problème est la manière dont on peut déterminer le niveau auquel appartient un item. En effet, on prend en compte le paramètre de difficulté, non pas en lui-même, mais en recherchant quel niveau de compétence est nécessaire pour maîtriser un item de ce niveau de difficulté. Un item sera donc classé dans la catégorie correspondant au niveau de compétence permettant d'avoir une probabilité (en général 75 ou 80% de chances) de le réussir. Mais, les sujets qui ont un niveau de compétence inférieur ont encore une probabilité élevée de le réussir si leurs compétences sont proches de la coupure séparant les classes de niveau. La qualité de cette séparation peut être appréciée à partir du pouvoir discriminant des items (paramètre a du modèle à deux ou trois paramètres ; paramètre de discrimination identique dans le cas du modèle de Rasch ou de ses dérivés). L'information donnée par ces niveaux apparaît donc relativement floue et imprécise dans la mesure où les coupures sont relativement arbitraires : le fait d'être classé dans un niveau de compétence ne veut en

aucun cas dire que le sujet n'est pas capable de fonctionner à des niveaux de compétence plus élevées. L'interprétation des niveaux n'est pas toujours facile car certains niveaux possèdent parfois peu d'items (en général les niveaux supérieurs). Et, surtout, l'interprétation en termes de fonctionnement cognitif n'est pas fondée sur l'analyse des tâches et des processus mais plutôt sur le modèle de mesure psychométrique.

Si l'on choisit une approche des compétences, alors, il est nécessaire de définir les compétences en termes de domaines, opération qui seule pourra valider l'interprétation de la mesure psychométrique puisqu'elle permettra de vérifier la couverture du domaine de la compétence par les épreuves construites. Cette approche a été l'objet d'une enquête internationale pilotée par l'OCDE : le programme DESECO (1999). Nous présentons en annexe un résumé schématique des résultats de ce programme. Il s'agissait de demander à différents experts : philosophes (Canto-Sperber & Dupuy, 1999), ethnologue (Goody, 1999), psychologue (Haste, 1999), économistes (Levy & Murnane, 1999), spécialistes des sciences de l'éducation (Perrenoud, 1999) comment on pourrait définir les compétences nécessaires pour vivre et réussir dans le monde moderne. Ce type de travaux pourrait permettre de définir les compétences évaluées sur des bases théoriques et non uniquement psychométriques. La validité du construit et son interprétation s'en trouveraient réellement validées. Il ne semble pas que les résultats de DESECO aient été injectés dans les réflexions sur les enquêtes internationales d'évaluation des compétences.

2 - LA QUESTION DE LA REFERENCE

La psychologie interculturelle a remis en question ses modes d'approche qui tendaient à imposer des représentations scientifiques postulées universelles à d'autres sociétés. Le fait que la psychologie se soit développées dans les sociétés occidentales, rendait cette discipline peu apte à rendre compte des différences culturelles.

Un premier courant de recherche développé depuis les années 1930 consistait à administrer à des groupes culturels différents les tests et les questionnaires utilisés pour les recherches psychologiques en Amérique du Nord et en Europe. Les différences de résultats établies entre les populations occidentales et les populations d'autres cultures étaient censées refléter des différences dans le fonctionnement psychologique des individus selon leur culture d'appartenance. On sait combien ces travaux ont été contestés en particulier dans le champs des aptitudes intellectuelles. La forte contestation de la validité de ces résultats a initié une réflexion sur la construction d'épreuves adaptées ou *« fair »* ou *« culture free »*. Ce courant a également rencontré d'importantes difficultés. Les recherches sur les questions de l'équivalence, du fonctionnement différentiel des tests et des items sont en partie l'héritage de ce courant de recherche.

2.1 - LES DEUX POINTS DE VUE

Parmi les apports de la psychologie interculturelle qui permettent de mieux construire la problématique des enquêtes comparatives, le concept de référence est particulièrement pertinent. J'ai argumenté dans plusieurs publications (Vrignaud, 2001 ; Vrignaud et Bonora, 1998) sur l'apport que ce concept pourrait constituer pour mieux définir les problématiques comparatives de ces enquêtes. La question de la référence se pose dès la définition des

groupes observés. On parle de type de référence externe ou interne en suivant, de manière métaphorique, les idées développées par le linguiste Pike dans les années 1930. L'utilisation d'une référence externe est souvent qualifiée d'approche « étique », en référence à la phonétique qui cherche à décrire tous les phonèmes possibles de toutes les langues – point de vue scientifique à valeur universelle – et l'utilisation d'une référence interne d'« émique », en référence à la phonémique qui s'attache à la description des phonèmes dans une langue donnée – point de vue développé à l'intérieur du système culturel étudié. Les résultats de recherches interculturelles conduites d'un point de vue « émique » peuvent aboutir à construire un point de vue étique. L'universel est bâti sur la part commune aux résultats émiques observés. On parlera ici d'étique dérivé (Berry, 1993) – ou de relativisme culturel modéré – pour le distinguer de l'étique imposé (absolutisme culturel intégral). Ces concepts sont largement développés dans des traités de psychologie interculturelle comme celui de Berry, Poortinga, Segall et Dasen, 1992.

Bien sûr ce point de vue peut paraître plus pertinent dans le domaine psychologique que dans le domaine de l'évaluation éducative. En éducation, on peut se demander si la différence entre les deux points de vue « etic » et « emic » est pertinente. Les apprentissages de base ne sont-ils pas des universaux tellement similaires dans l'ensemble des systèmes scolaires que la point de vue « emic » serait dénué de pertinence ?

Même si nous répondons par l'affirmative à cette question, il nous semble qu'au moins une partie du domaine des compétences des élèves devrait être abordé d'un point de vue « emic » : le domaine des compétences de base de la vie quotidienne *(« basic life skills/survival skills »)*. En effet, ces compétences peuvent largement varier d'une culture à l'autre voire à l'intérieur d'une même culture selon l'environnement et l'histoire des groupes et des individus. A moins de privilégier l'hypothèse de la globalisation et de l'universalité des valeurs et des modes de vie à l'aune d'un système socio-économique considéré comme la référence.

Par ailleurs, il n'est pas certain qu'un point de vue « imposed etic » soit complètement défendable pour la partie directement liée aux curricula scolaires (« 3Rs »). Des traitements secondaires sur les résultats d'enquêtes internationales (INED sur IALS) dans le domaine de la littéracie ont montré que des différences interculturelles entre pays avaient été minimisées. Par exemple, les programmes scolaires peuvent attacher une plus grande importance à certains types d'exercices, d'approches pédagogiques. On donne un poids plus important à l'étude de textes « littéraires » dans les pays européens que dans les pays nord américains, dans certains pays, l'utilisation systématique pour les évaluations scolaires de QCM crée une plus grande familiarité avec ce type d'instruments.

2.2 - POIDS DE LA REFERENCE EXTERNE

Le choix d'une référence, selon cette optique de la psychologie interculturelle, n'a pas toujours été suffisamment explicité dans les enquêtes internationales sur les compétences des élèves. Certaines de ces enquêtes se sont développées à partir d'une référence étique : par exemple les modèles et épreuves utilisées dans l'enquête IALS étaient directement issus, parfois sans aucune modification, de l'enquête américaine sur la littéracie chez les adultes, elle-même reprenant le matériel de l'enquête sur la littéracie chez les jeunes aux Etats-Unis (NALS). Le choix d'une référence a également des conséquences importantes sur la pertinence des méthodes visant à identifier les biais de mesure.

Nous lançons cependant ici une discussion qui dépasse le cadre de notre exposé sur la méthodologie de traitement. Il nous paraissait indispensable d'exposer cette problématique qui distingue bien une approche dans laquelle les différences interculturelles sont gommées au profit d'un point de vue postulé universel et une approche respectant les différences interculturelles. Sur le plan méthodologique, la plupart des méthodes statistiques de traitement des enquêtes d'évaluation sont en harmonie avec une perspective «imposed etic» et la renforcent. La recherche des biais aboutit à faire disparaître les éléments de différenciation interculturelles. L'unidimensionnalité de la variable (pour le calcul des scores et les modèles MRI) ne permet d'identifier que des différences quantitatives. On retrouve cette distinction dans la recherche comparative en éducation comme l'a souligné Mons (2003) d'un côté certains auteurs ont privilégié le particularisme et d'autre les invariants.

L'enquête OCDE sur la littératie (IALS) illustre la contamination par une telle approche. Dans l'enquête NAEP citée plus haut, était incluse une épreuve d'évaluation des compétences en lectures pour la population des adolescents et jeunes adultes (Young Adult Literacy Survey). Dans un second temps une enquête menée auprès d'une population d'adultes « National Adult Literacy Survey « dont les tests étaient construits pour être mis en relation avec les tests de la YALS, a permis de faire des comparaisons entre les générations. Dans un troisième temps des tests IALS permettant la mise en relation avec ceux de la NALS ont été construits et utilisés pour l'enquête l'OCDE (enquête « International Adult Literacy Survey « (1995)), assurant la comparaison entre les résultats des dix pays. On procède comme si la définition de la littéracie correspondant au construit de la première enquête (NAEP) avait une valeur universelle.

Pour assurer l'homogénéité du construit nécessaire au modèle de mesure, nous avons vu que les données une fois collectées, on procède d'une part à l'élimination des items présentant un FDI d'autre part à l'ancrage sur une population afin d'assurer l'equating entre les populations. L'élimination des items suspects de FDI implique de réduire le fonctionnement des sujets à un même dénominateur qui est dépendant du groupe de référence. Il est d'usage de considérer que l'ancrage ne joue que sur l'origine de l'échelle. De manière optimiste, cette opération est présentée comme anodine. On cite la métaphore de la fixation du zéro des échelles de température (Celsius, Fahrenheit, Kelvin). Cette métaphore ne nous paraît pas refléter la réalité de la mesure en psychologie. La mesure n'est pas seulement la fixation d'une échelle, elle définit le construit. Dans l'exemple, de la IALS, le choix d'utiliser comme référence la NALS (et pour celle ci la YALS) introduit comme référence à l'homogénéité les items communs de ces échelles. Le construit se trouve donc déterminé par les items de l'échelle destinée à l'ancrage. On peut considérer que dans le cas d'échelle unidimensionnelle, cela revient, en partie, à aligner la dimension en se servant des items de l'épreuve d'ancrage.

Même les enquêtes qui ont construit leur épreuve d'un point de vue emic dérivé comme par exemple PISA où les pays participants étaient invités à proposer des exercices et des items ne sont pas exempts de telles critiques. En effet, parmi les items proposés par les Etats-Unis, on retrouve une partie des items utilisés dans IALS et les enquêtes antérieures comme on peut le constater dans le tableau 3.1.

Tableau 3.1. Codes des items de IALS repris dans PISA

IALS ID	PISA ID and unit name
COREQ1S1	r232q01 unicef
B1Q5S1	r 233q01 diapers
B1Q6S1	r 233q02 diapers
B1Q10S1	r234q01 personnel
B1Q11S1	r234q02 personnel
B2Q1S1	r235q01 impatiens
B2Q3S1	r235q02 impatiens
B2Q6S1	r236q01 newrules
B2Q7S1	r236q02 newrules
B3Q7S1	r237q01 job interview
B3Q8S1	r237q02 job interview
B3Q9S1	r237q03 job interview
B3Q11S1	r238 q01 bicycle
B3Q12S1	r238 q02 bicycle
B3Q13S1	r239q01 allergies
B3Q15S1	r239q02 allergies
B4Q1S1	r240q01 painreliever
B4Q2S1	r240q02 painreliever
B4Q6S1	r241q01 warranty hotpoint
B4Q7S1	r241q02 warranty hotpoint
B5Q1S1	r242q01 marathon
B5Q2S1	r242q02 marathon
B5Q3S1	r243q01 childseat
B5Q4S1	r243q02 childseat
B5Q5S1	r243q03 childseat
B5Q6S1	r243q04 childseat
B6Q1S1	r244q01 scrambled
B6Q7S1	r245q01 movie summaries
B6Q8S1	r245q02 movie summaries
B7Q10S1	r246q01 contact employer
B7Q11S1	r246q02 contact employer
B7Q13S1	r247q01 firesafety
B7Q14S1	r247q02 firesafety
B7Q15S1	r247q03 firesafety

Synthèse : Les recherches interculturelles insistent sur la nécessité de distinguer une approche externe (« etic ») et une approche interne (« emic ») aux cultures étudiées. Certaines enquêtes privilégient une approche « etic ». Il semble souhaitable de travailler avec des références internes à chacun des univers culturels sans imposer au départ une référence culturelle externe, au moins dans le domaine des compétences de base de la vie quotidienne. Si on souhaite que les différences interculturelles puissent se manifester pleinement, il est nécessaire d'adopter une perspective d'abord « emic » permettant de passer à une perspective « derived etic ».

3 - LA TRADUCTION

L'importante diversité des langues a de quoi interroger non seulement le linguiste et l'anthropologue, mais aussi le psychologue (Fuchs, 2002). Il existe un consensus pour chiffrer le nombre de langues actuellement parlées dans le monde entre 4000 et 5000 ! Peu d'hypothèses ont été avancées pour tenter d'expliquer les raisons, l'intérêt – par exemple en termes d'évolution – d'une telle diversité. Les psychologues semblent également peu sensibilisés aux implications de cette diversité. On peut dire sans grand risque de se tromper que plus de 90% des résultats des recherches en psychologie ont été établis sur des sujets parlant des langues indo-européennes (le *Standard Average European* de Whorf).

La traduction ou plutôt l'adaptation des tests et des questionnaires psychométriques a donné lieu au développement d'une méthodologie spécifique. Une simple traduction accompagnée d'une rétro-traduction n'est pas suffisante pour s'assurer de l'équivalence. On trouvera la description de l'ensemble des procédures dans des synthèses comme celle de Behling et Law (2000). La Commission Internationale des Tests a développé et publié des recommandations sur les normes à respecter (Hambleton, 1994).

Il n'entre pas dans le cadre de ce rapport de discuter ces hypothèses, je ferai simplement remarquer que dans les enquêtes internationales, les concepteurs se placent implicitement dans une perspective universaliste en considérant la variabilité linguistique comme périphérique dans le traitement langagier. Pourtant la variabilité linguistique intervient au niveau sémantique. Par exemple, dans les épreuves de vocabulaire du WISC III, les termes retenus dans chacun des pays ne sont pas les mêmes car les prétests ont montré que la difficulté des mots n'était pas la même selon les langues (voir par exemple, Sarrazzin, 1999). L'hypothèse de la grammaire universelle a été également questionnée par les travaux sur l'acquisition du langage (Kail, 2000 ; Hickman, 2002). Les recherches visent à mettre en relation l'acquisition plus ou moins précoce de concepts aussi fondamentaux que le temps et l'espace avec la variabilité de la structuration de ces concepts entre les langues. On peut faire l'hypothèse que certaines inférences, certaines constructions, sont plus difficiles à faire dans certaines langues que dans d'autres. On peut à ce propos souligner la différence de longueur du même texte en français et en anglais, ce qui peut jouer sur le temps de lecture. Des analyses secondaires ont mis en évidence de tels biais cognitifs dans des enquêtes comme IALS ou PISA (Rémond, 1996 ; 2001). Encore une fois, il ne s'agit pas de se perdre dans des arguties méthodologiques mais de montrer en quoi des choix méthodologiques mêmes très sophistiqués ont une incidence sur le construit et sur l'interprétation des résultats.

Synthèse : La traduction de l'épreuve dans les différentes langues des pays participants a donné lieu à de nombreuses vérifications. Ce point mériterait néanmoins davantage de recherches et d'explicitations. D'abord, sur un plan théorique, on aimerait savoir dans quelle mesure la variabilité linguistique peut influer sur la performance à des épreuves utilisant le langage comme la littéracie. Sur un plan méthodologique, il faudrait s'assurer que les méthodes utilisées pour identifier les biais culturels, élaborées pour travailler sur des items identiques, sont complètement efficaces dans le cas des items traduits.

4 - LA METHODE DES PLANS EQUILIBRES INCOMPLETS PAR BLOCS

Il apparaît difficile de concilier deux exigences : celle de présenter un grand nombre d'exercices pour mieux assurer la représentativité des compétences évaluées et celle de ne pas accroître de manière excessive le temps et la charge de travail des sujets. Pour concilier ces deux exigences, la solution consiste à ne pas administrer tous les exercices à tous les sujets. Il faut néanmoins disposer d'un lien (ancrage) entre tous les exercices si l'on veut pouvoir les placer sur une même échelle. Il suffit pour cela de disposer d'informations sur toutes les paires d'items. On va répartir les exercices (items) en plusieurs blocs de longueur (temps de passation) à peu près égale. Chaque sujet ne passera qu'un nombre de blocs correspondant au temps de passation choisi. La question devient de construire des combinaisons des blocs de telle manière que chacun des blocs soit combiné à chacun des blocs restant, c'est à dire que toutes les paires de blocs soit présentes dans le dispositif expérimental. On va donc construire des cahiers de passation comprenant au moins deux blocs. Il s'agit alors de réduire le nombre de combinaisons des paires de cahiers pour maîtriser l'explosion combinatoire que pourrait engendrer la nécessité de construire toutes les combinaisons de paires de blocs. En général, on a choisi de construire des cahiers comprenant trois blocs pour s'appuyer sur une méthode de construction des plans expérimentaux bien connue : celle des triades. Pour neutraliser les effets liés à l'apprentissage et à la fatigabilité, on va contrôler l'ordre de passation des blocs en les contrebalançant. Chaque bloc apparaîtra au moins une fois dans les différentes positions de l'ordre de passation d'où le nom de « cahiers tournants » sous lequel ce dispositif expérimental est souvent désigné en français. Ainsi, dans le cas de trois blocs, chacun des blocs apparaîtra au moins une fois en position initiale, intermédiaire et finale.

Cette question de la réduction du nombre de combinaisons des conditions expérimentales est une question classique dans la construction des protocoles expérimentaux. Parmi les travaux qui ont systématisé cette approche, on peut citer ceux de Cochran et Cox (1950). La méthode des triades est un mode de recueil de données qui permet d'obtenir des informations sur toutes les paires d'objets inclus dans le dispositif expérimental en réduisant le nombre de présentations. Cette situation se rencontre fréquemment en sciences humaines lorsqu'on étudie les jugements de similarité, nous l'avons, nous-mêmes, employé pour l'étude des processus de catégorisation (Vrignaud, 1999). Une organisation des triades pour réduire le nombre de blocs à présenter aux sujets, en perdant un minimum d'information, a été proposée par Burton et Nerlove (voir Burton & Nerlove, 1976 ainsi que Weller & Romney, 1988). Cette présentation dite en plan équilibré par blocs incomplets (*balanced incomplete blocks design*, en abrégé BIB) est particulièrement intéressante car elle permet de réduire drastiquement le nombre de blocs à présenter aux sujets pour recueillir des informations sur toutes les paires d'objets. Par exemple l'évaluation des similarités entre 7 objets nécessite de présenter 21 paires. A partir de 7 objets, on peut constituer 35 triades

($\binom{n}{3} = \frac{n!}{3!(n-3)!}$).

A l'intérieur de ces 35 triades, chaque paire apparaît cinq fois. On peut donc en utilisant seulement sept triades recueillir les jugements de similarité entre les 21 paires. Si l'on souhaite avoir davantage de robustesse, on peut utiliser deux ensembles différents de 7 triades soit 14 objets. Cette présentation se révèle particulièrement efficace face à l'explosion combinatoire quand le nombre d'objets étudiés augmente.

La méthode peut être généralisée à un nombre quelconque d'objets (n), de taille de blocs (k : triades, quadruplets, etc.) de nombre d'occurrences de paires (λ). Il faut néanmoins signaler qu'il n'existe pas de solutions pour certaines combinaisons de n et de k. Une méthode de construction des blocs a été proposée par Cochran et Cox (1950).

Un plan expérimental comprenant sept blocs d'items et aboutissant à la construction de sept cahiers différents a été utilisé dans le cadre des enquêtes NAEP (Johnson, 1992). Les blocs ont été disposés à l'intérieur des cahiers de telle manière que l'ordre est parfaitement contrebalancé pour chacun des blocs. Nous présentons dans le tableau 3.2. le plan expérimental utilisé dans la NAEP (Johnson, 1992) pour sept blocs d'items (A,B,C,D,E,F) selon la méthode des triades et le contre-balancement. On peut constater que toutes les paires de blocs se retrouvent dans au moins un cahier et que chaque bloc se trouve une fois dans une des trois positions (initiale, intermédiaire et finale).

Tableau 3.2. Construction de sept cahiers à partir de sept blocs selon la méthode des triades

Cahier	Blocs		
1	A	B	D
2	B	C	E
3	C	D	F
4	D	E	G
5	E	F	A
6	F	G	B
7	G	A	C

Portée et limite de la méthode des cahiers tournants

Les principaux avantages de la méthode des cahiers tournants est comme nous l'avons vu de disposer d'informations sur de nombreux exercices ou items en limitant le temps de passation. La méthode des triades permet de construire une présentation systématique des blocs conforme aux exigences de fiabilité des méthodes d'analyse des données (connaître les liaisons – corrélations – entre toutes les paires d'items, disposer d'un plan équilibré et contrebalancé).

Les inconvénients découlent du nombre limité de méthodes d'analyse des données capables de prendre en charge un tel plan. En effet, ce plan d'expérience a été élaboré dans le cadre des protocoles expérimentaux destinés à étudier l'effet de différents facteurs expérimentaux. Il s'agit de pouvoir multiplier le nombre de facteurs sans multiplier le nombre de groupes expérimentaux. On utilise l'analyse de variance pour analyser les données. De même, l'extension de ce type de plan à l'étude des jugements de similitudes conduit à recueillir l'information sur toutes les paires d'objets chez un même sujet. Les données produites par chaque sujet sont ensuite agrégées au niveau du groupe et de nombreuses méthodes d'analyse des données travaillant à partir d'une matrice de distance peuvent ensuite être appliquées (par exemple l'échelonnement multidimensionnel). Dans ces deux situations, le caractère incomplet du plan ne pose pas de problème particulier. Ce n'est pas le cas de la collecte des données par la méthode des cahiers tournants. Dans cette dernière situation, les cahiers étant été administrés à des groupes indépendants, on dispose certes de données sur les mêmes sujets mettant en relation les paires de blocs mais incomplètes par construction. Le tableau 3.3. présente la présence et l'absence de données résultant de la

passation des cahiers par des groupes indépendants. On voit que chaque paire de blocs est présente pour un groupe et un seul. Pour chacun des groupes, on dispose donc de données pour trois blocs et trois paires de blocs. Par contre, on peut considérer que les données pour les quatre blocs restants et les 18 paires de blocs correspondantes sont manquantes.

Tableau 3.3. Données présentes et manquantes entre blocs pour chaque groupe/cahier

	Bloc						
Groupe/cahier	A	B	C	D	E	F	G
1	*	*		*			
2		*	*		*		
3			*	*		*	
4				*	*		*
5	*				*	*	
6		*				*	*
7	*		*				*

La présence de données manquantes systématiquement par construction nécessite de recourir à des méthodes permettant de traiter ce type de données. On s'appuie sur les réflexions théoriques particulièrement pertinentes par leur caractère de généralisation théorique et leur profondeur de vue qui ont été développées par Rubin (1987) à partir de la recherche de solution au problème des données manquantes dans les tests. Il s'agit d'imputer au sujet sa compétence à partir d'un ensemble de valeurs incomplètes. De manière générale, Rubin en vient à considérer que pour l'estimation de la compétence vraie d'un sujet est, selon la théorie psychométrique, une situation où les données sont manquantes. En effet, la compétence n'est connue que conditionnellement aux réponses du sujet à un nombre réduit de questions : celles qui sont incluses dans le test qu'il a passé y compris dans le cas où il a répondu à toutes les questions du test. Dans le cadre des MRI, cette formulation le conduit à repenser son algorithme d'estimation des paramètres en deux étapes (algorithme EM, Rubin, 1991). Mislevy a poursuivi cette réflexion et son application à l'estimation des paramètres des modèles MRI dans le cadre d'enquêtes dont les données ont été collectées dans un plan expérimental utilisant les cahiers tournants. Mislevy et collaborateurs introduisent dans le modèle les données descriptives du contexte du sujet (*background variables*) afin de rendre l'estimation du paramètre de compétence des sujets plus robuste. Il s'agit d'estimer la compétence des sujets conditionnellement aux réponses qu'ils ont donné aux items auxquels ils ont répondu (donc sans inclure les items manquant par construction et les non réponses aux items présentés) et conditionnellement aux variables décrivant leur contexte socio-économique. On obtiendra ainsi une estimation de l'ensemble des paramètres des items et des paramètres de compétence des sujets sur une échelle unique même si les sujets n'ont pas passé tous les items et si les items n'ont pas été administrés à tous les groupes. Les procédures d'*equating* seront ensuite étendues, dans le cas des enquêtes internationales à l'ensemble des groupes nationaux entrant dans la comparaison. D'après les publications des auteurs sur ce sujet, la théorie réalise un apport majeur à la réflexion psychométrique et les procédures semblent donner des résultats robustes dans la situation des données incomplètes. Il est, d'ailleurs, à noter que cette procédure élaborée par les chercheurs d'ETS a été ensuite implantée dans le logiciel Conquest édité par ACER (Wu, Adams, & Wilson, 1997) lorsque ce groupe a été chargé du traitement des données PISA.

L'inconvénient de cette procédure est de ne pas pouvoir appliquer d'autres approches que les MRI et que ce type d'algorithme au traitement des données. L'*equating* ne peut être réalisé par d'autre méthode que les MRI (par exemple l'équipercentile ou la régression). D'autres modèles de mesure comme les modèles structuraux, ne peuvent être appliqués à l'ensemble des données ce qui aurait permis de recouper les informations obtenues. On introduit ici une dépendance totale aux MRI, et même plus à un algorithme d'estimation spécifique implanté uniquement dans les logiciels diffusés par deux organismes. Ceci est d'autant plus regrettable qu'il aurait suffi d'introduire un bloc commun administré à tous les sujets pour pouvoir mettre en œuvre de nombreuses autres approches que ce soit pour l'estimation des compétences ou l'identification des biais d'items. Ainsi, dans la dernière évaluation-bilan sur les compétences des élèves en fin de troisième, la DEP a utilisé la méthode des cahiers tournants mais en ajoutant un bloc commun ce qui permet pour l'analyse des données de mettre en œuvre d'autres approches que les MRI (Dauphin & Trosseille, 2004).

Synthèse : L'utilisation d'un plan de collecte des données permettant de ne pas administrer la totalité des items à l'ensemble des sujets est souvent utilisé dans les enquêtes internationales. Ce plan élaboré pour l'enquête américaine NAEP présente l'avantage de recueillir des informations pour de nombreux items, ce qui facilitera l'interprétation de l'échelle de compétence sans alourdir la tâche et le temps de passation des sujets. Il présente par contre l'inconvénient de ne pouvoir être traité que par une seule méthode psychométrique (les MRI). Il n'est pas possible d'effectuer des analyses complémentaires par d'autres approches pour vérifier la fiabilité de l'ensemble des données.

5 - METHODOLOGIES ALTERNATIVES DEVELOPPEES DANS DES PROJETS EUROPEENS ET FRANÇAIS

Les travaux du RERPESE ont été présentés dans le chapitre 2. Il s'agit ici d'entrer dans le détail de la méthodologie alternative proposée dans le projet sur l'évaluation des compétences en littéracie sans recourir à des épreuves traduites. Il ne s'agit pas de vanter cette méthode comme un palliant les manques des méthodologies utilisées dans les enquêtes internationales, mais de montrer que d'autres approches sont possibles et qu'elles apportent d'autres éclairages pour les comparaisons internationales des compétences des élèves. Nous terminerons ce chapitre en présentant brièvement une autre innovation méthodologique élaborée par des organismes français pour réaliser des enquêtes à domicile sur l'évaluation des compétences des adultes.

5.1 - COMPARAISON DES PROGRAMMES EN LITTERATIE SANS EPREUVES COMMUNES

Je discuterai davantage ce point méthodologique car j'en ai été responsable et il fournit un bon exemple des alternatives existant à l'approche décrite précédemment. La difficulté méthodologique tenait à la nécessité de trouver un moyen d'établir un ancrage malgré l'absence d'épreuve commune. Deux moyens de mettre en relation les résultats dans les différents pays nous ont paru offrir une solution : 1) un ancrage par le recours à une épreuve commune dont l'adaptation avait déjà été réalisée dans chacun des pays, 2) un ancrage par

des sujets bilingues passant les épreuves dans leurs deux langues de compétence, 3) un analyse des données multidimensionnelle.

5.1.1. Recours à une épreuve commune

La manière habituelle d'établir l'équivalence entre populations ayant passé des épreuves différentes est d'utiliser une épreuve d'ancrage. Cette épreuve identique pour toutes les populations permet de mettre en place des procédures statistiques *(equating)* de passage entre les épreuves : comparaison de la difficulté des épreuves, comparaison des scores des sujets. En toute rigueur, une épreuve d'ancrage doit être constituée d'items représentatifs du contenu des épreuves dont on veut s'assurer de l'équivalence. La méthodologie établie pour les comparaisons longitudinales et transversales consiste à utiliser des items communs – ou test noyau – qui passés par les différentes populations permettront de construire une transformation pour mettre tous les paramètres sur une échelle commune. Cette méthode peut être jugée pertinente dans le cadre d'enquêtes longitudinales ou de la méthode des « cahiers tournants » mais elle présente les inconvénients que nous avons soulignés plus haut lorsqu'on a affaire à des épreuves traduites constituées d'items traduits qui ne sont donc pas identiques d'une épreuve linguistique à une autre. La recherche d'une méthode d'ancrage au niveau des épreuves nous a donc paru contradictoire avec notre objectif.

Elle nous a également posé problème car dans les méthodologies établies, les modèles et traitements utilisés reposent sur le postulat d'unidimensionnalité. Ce qui implique l'unidimensionnalité – ou au moins l'existence d'une homogénéité forte – des items des différentes épreuves et bien sur de ceux de l'épreuve commune. Il aurait fallu construire une épreuve commune répondant à ces critères. Ce qui était un des résultats attendus de l'étude qui devrait permettre 1) de voir dans quelle mesure ce qui est évalué et les épreuves utilisées appartiennent à un domaine commun et homogène ; 2) d'identifier les éléments théoriques et pratiques permettant de construire une épreuve commune.

Pourtant l'utilisation d'une épreuve d'ancrage même imparfaite sur le plan théorique nous semblait nécessaire pour mener au mieux l'objectif précédent en offrant la possibilité d'établir une équivalence entre les épreuves à partir d'indicateurs statistiques en sus de l'équivalence a priori et de l'ancrage par les sujets bilingues présentée ci-dessous. Etant donnés notre problématique et les délais imposés, il nous a paru préférable de construire cette épreuve d'ancrage à partir d'un matériel existant déjà adapté dans les quatre pays.

Les items du subtest vocabulaire du WISC-3 *(Wechsler Intelligence Scale for Children)* nous ont semblé répondre de manière optimale à nos exigences pour trois raisons :

– le WISC-3 est un test bien connu, qui a été l'objet de plusieurs révisions et adaptations dans les quatre pays, chaque version nationale présentant une bonne fiabilité psychométrique ;
– le subtest « vocabulaire « est un bon indicateur des compétences verbales. ce subtest montre également une bonne validité concourante et prédictive avec les performances scolaires. On peut donc considérer que cela a un sens de le mettre en relation avec des épreuves d'évaluation de la lecture.
– Le fait que ce test a été construit pour être passé oralement dans une situation de face-à-face entre le psychologue et le sujet testé, soulève une difficulté importante dans le cadre de notre étude. Etant donnée la taille des échantillons interrogés (entre 1 000 et 2 000

sujets), il n'est pas possible de procéder ainsi. Nous avons donc décidé de faire passer ce test à l'écrit.

Il ne s'agit pas d'effectuer une opération d'*equating* pour situer les résultats nationaux sur une même échelle, mais de pouvoir mettre en relation des résultats à des épreuves différentes par des méthodes appropriées. L'analyse des résultats a permis de juger de la portée et des limites que l'on peut attendre de l'utilisation de l'épreuve de vocabulaire comme épreuve pivot entre les différentes épreuves nationales. L'utilisation du subtest vocabulaire du WISC-3 nous a paru offrir la possibilité de disposer d'une épreuve commune ne présentant pas les inconvénients d'une épreuve *ad hoc*. Cette épreuve est déjà adaptée aux pays concernés. Son utilisation apporte des éléments complémentaires pour la procédure de passage à partir des sujets bilingues.

5.1.2 - Le recours à des sujets bilingues

L'ancrage par les sujets bilingues fait aujourd'hui partie des méthodes préconisées pour établir l'équivalence des épreuves en psychologie interculturelle (Sireci, 2000). Elle a été surtout employée pour comparer des variables de comportement ou d'attitude (questionnaires de personnalité). Elle n'avait jamais été employée à notre connaissance pour évaluer les compétences en rapport avec la maîtrise de la langue elle-même. Les sujets passant deux épreuves dans des langues différents, on dispose de scores appariés pour ces deux épreuves, ce qui permet, si on fait l'hypothèse – forte et difficile à démontrer – que les sujets sont également compétents dans les deux langues de construire une méthode de calcul pour l'equating entre ces deux épreuves. Il est certain que l'hypothèse de compétences équivalentes des sujets bilingues dans les deux langues, est peu plausible et difficile à démontrer pour au moins deux types de raisons :

- sur le plan psycholinguistique, les recherches ont montré que les mécanismes psycholinguistiques des bilingues sont, d'une part, différents dans leurs deux langues et, d'autre part, différents de ceux des sujets monolingues ;
- sur le plan de l'équivalence de compétence des sujets bilingues et des sujets monolingues : en général, les sujets bilingues se situent à des niveaux de compétence extrêmes par rapport aux monolingues, soit vers le haut (enfants de parents cadres investis dans des carrières internationales) soit vers le bas (enfants de travailleurs migrants).

Il apparaît donc que le recours à des groupes de sujets bilingues pour établir l'équivalence n'est pas une solution parfaite. Cependant, elle pallie à sa manière les manques des méthodes comparant des versions linguistiques différentes dans des populations différentes.

5.1.3 - Une approche multidimensionnelle

J'ai présenté comme une limite des enquêtes internationales l'approche unidimensionnelle adoptée. La recherche que je viens d'évoquer m'a fourni l'occasion d'éprouver l'intérêt des approches multidimensionnelles pour ce type d'enquêtes.
Le plan expérimental était complexe : 4 épreuves nationales différentes ; 4 échantillons unilingues et 7 échantillons bilingues. Et, surtout, les données étaient structurellement incomplètes puisque les sujets « monolingues » n'avaient passé qu'une seule épreuve sur les quatre utilisées et les bilingues deux. Une méthodes multidimensionnelle m'a semblé

pouvoir répondre aux questions de comparaison en respectant les contraintes des données (plan incomplet) : l'utilisation de la régression PLS (Partial Least Squares) qui permet de gérer de manière efficace les données manquantes structurelles présentées par le plan. La régression PLS se distingue de la régression par les moindres carrés ordinaires (Ordinary Least Squares) par trois éléments principaux : l'utilisation de variables latentes (composantes), l'algorithme d'estimation, la procédure de validation. Elle est peu connue des psychologues mais est largement utilisée dans les domaines de la chimie et de l'économétrie. L'approche PLS consiste à construire des variables latentes (composantes) à partir d'un ou plusieurs blocs de variables pouvant avoir des relations de dépendance. L'approche a été mise au point par Hermann Wold (1982), pour modéliser les relations entre plusieurs blocs de variables indépendantes et dépendantes (soft modeling) et développée par son fils Svante Wold (S. Wold, Martens, et H. Wold, 1983) en se limitant au cas de deux blocs de variables. H. Wold a proposé pour estimer les valeurs des paramètres un algorithme NIPALS qui présente l'avantage de travailler sur chaque paire de variables et non sur l'ensemble des variables sous forme d'une matrice. Cette manière de procéder permet d'utiliser pour chaque paire de variables les observations présentant des données. Ces observations peuvent, de ce fait, être différentes d'une paire de variables à une autre. Ceci évite les inconvénients liés à la gestion des valeurs manquantes : éliminer les sujets présentant des valeurs manquantes ce qui dans cette enquête aboutirait à éliminer tous les sujets puisqu'ils n'ont passé qu'une seule épreuve ; ou remplacer les valeurs manquantes par des valeurs probables ce qui aurait peu de fiabilité dans notre cas étant donné l'impossibilité d'inférer les réponses sur plusieurs blocs de données. La mise en œuvre pour notre recherche a consisté à établir des relations entre épreuves nationales à partir des données des sujets bilingues. Comme il n'y avait pas lieu de distinguer dans l'analyse des blocs de variables dépendantes ou indépendantes, ce 'est pas une analyse de régression à laquelle on aboutit. La méthode PLS revient dans ce cas à effectuer une analyse en composantes principales (ACP). On l'utilise pour son algorithme dont l'avantage est, comme il vient d'être énoncé, de pouvoir travailler sur un ensemble de variables comportant de nombreuses données manquantes. Les données des sujets bilingues présentes pour plusieurs paires d'épreuves permettent à l'algorithme d'établir un lien entre les différentes épreuves nationales.

Les indicateurs montrent qu'on peut retenir deux facteurs. Ce phénomène montre que l'univers des items n'est pas unidimensionnel. Ces résultats et leur intérêt pour élargir les méthodologies des enquêtes comparatives sur les compétences des élèves ont été développés dans les rapports européens et dans plusieurs communications.

Nous sommes entrés dans de nombreux détails à propos de ce projet car il représente une étude de faisabilité d'une approche alternative par rapport aux enquêtes internationales utilisant une même épreuve traduite. On apporte ici des éléments de réponses à ce que peut être une approche de relativisme culturel modéré : assurer la mesure et respecter les cultures !

5.2 - INFORMATION ET VIE QUOTIDIENNE (IVQ)

Ce rapide tour d'horizon serait incomplet si on ne mentionnait pas l'enquête « Information et Vie Quotidienne » portant sur les compétences à l'utilisation d'information chez l'adulte. La première phase de ce projet regroupait la DEP, l'INSEE, l'INED, l'INETOP et le LASMAS. Il s'agissait d'évaluer la fiabilité de méthodes pour évaluer les compétences en littéracie chez

l'adulte. Bien que cette enquête n'avait pas de visée comparative internationale[58], elle s'inscrit néanmoins dans cette perspective dans la mesure où son mode de passation à domicile auprès d'adultes s'inscrit dans le prolongement des études complémentaires sur l'enquête IALS. Les trois apports les plus originaux d'IVQ étaient 1) l'utilisation d'un logiciel pour assister à la saisie des réponses par l'enquêteur, 2) la prise en compte de la motivation dans l'organisation des épreuves, 3) l'exploration des difficultés de lecture.

L'INSEE dispose d'un réseau d'enquêteurs à domicile qu'elle mobilise pour ses travaux. Les enquêteurs sont équipés d'un ordinateur portable sur lequel est implanté un logiciel permettant la saisie des réponses des enquêtés. Les enquêteurs de l'INSEE n'avaient jamais travaillé sur l'évaluation des compétences. Il était donc important de voir quelle serait la réception d'une telle enquête par les enquêteurs et les enquêtés. De plus, il fallait programmer le logiciel pour une tâche de nature différente de la saisie de réponses à des questions factuelles. Par exemple, le logiciel permettait de mesurer le temps mis par le sujet pour lire le texte et répondre aux différentes questions.

Le test fonctionne selon le principe du test adaptatif : un premier module dit « orientation » permet de séparer les sujets qui ont des difficultés en lecture, orientés vers un module diagnostic, et les sujets ayant acquis la lecture, orientés vers un module, dénommé « haut », comportant la lecture de textes. C'est ce module haut qui nous intéresse ici. En effet, le groupe de pilotage avait fait l'hypothèse que la motivation, définie comme l'intérêt pour le contenu des textes proposés, aurait un effet sur les performances en lecture. Cet intérêt implique également, selon nous, une familiarisation avec le vocabulaire, le style et le contenu des textes proposés. Pour tester cette hypothèse, le protocole comportait quatre thèmes : Famille et Société (FS) ; Cinéma, Spectacle et Loisirs (CSL) ; Sports (S) ; Littérature (L). Chacun des thèmes comprenait deux exercices de 5/6 items chacun. Chaque exercice comportait un texte sur la page de gauche et des questions se rapportant au texte sur la page de droite. Les textes retenus étaient des extraits d'articles de journaux pour les trois premiers thèmes, par exemple un compte rendu de match de football pour le thème S, une critique de film pour le thème CSL ; le thème L comportait deux textes courts d'auteurs « classiques » (Chamfort et Hugo). Le sujet après avoir lu chaque texte et les cinq ou six questions portant sur le texte, répondait oralement aux questions écrites, l'enquêteur saisissant immédiatement la réponse sur ordinateur. Dans une première phase, les sujets choisissaient un des quatre thèmes proposés (condition « choix »). Après avoir répondu à l'ensemble des questions pour les deux textes, l'enquêteur administrait, en le tirant au hasard, un des trois modules restant (condition « aléatoire »). Si la motivation a un effet sur la performance en lecture, on s'attend à ce que les thèmes soient mieux réussis dans la condition choix que dans la condition aléatoire. L'étude de faisabilité (Vallet et al., 2002) ayant montré l'intérêt de l'utilisation de la méthode de collecte des données et de l'exploration des difficultés en littéracie. L'utilisation d'épreuves à contenu varié apparaissait moins pertinente. Il a donc été décidé de constituer de nouvelles épreuves. Un accent particulier a été mis sur l'épreuve destinée aux bas niveaux de compétence. Ce nouveau dispositif a été utilisé pour une enquête conduite par l'INSEE, la DEP, le Commissariat au plan, le Ministère des Affaires Sociales auprès d'un échantillon représentatif de 10 000 français. Les résultats concernant les bas niveaux ont été présentés lors de *l'International colloquium* (conference) *of the ANLCI* (Agence Nationale de Lutte Contre l'Illettrisme) : « Assessment of low literacy levels ». Lyon, 5-7 November, 2003.

[58] La participation de l'ONS à une partie des travaux avait cependant permis d'envisager une enquête comparative Angleterre/France.

Synthèse : Alors que la position française en ce qui concerne les enquêtes pilotées par des organismes internationaux apparaît plutôt passive voire critique, la France, à travers le RERPESE, s'est fortement impliquée pour lancer et piloter des enquêtes entre pays européens. Les approches utilisées sont souvent originales.

Peut-on répondre à la question qui sert de titre à ce chapitre ? Nous avons insisté sur plusieurs éléments qui tendent à montrer que les enquêtes internationales ne sont pas exemptes de biais. Il faut distinguer les biais qui sont liés à des choix méthodologiques qui pourraient être pensés différemment et les biais qui sont inhérents à ce type d'enquêtes et qui nécessiteront, pour être mieux pris en compte, davantage de travaux scientifiques.

Le premier type de biais ont souvent été introduits par des choix qui auraient pu être différents. Ils apparaissent principalement dans les enquêtes où la on évalue une compétence plutôt que des acquis scolaires. Ainsi, la référence aux enquêtes américaines, le choix des « cahiers tournants », auraient pu être évités. On a privilégié une approche déjà développée dans le cadre américain qui présente des qualités certaines dans un cadre fédéral mais n'était peut être pas exportable telle quelle dans un cadre international. Cette approche a continué à être privilégiée dans les enquêtes suivantes dans la mesure où les organismes retenus par les appels d'offre avaient adopté des options similaires.

Le second type de biais, principalement le problème de l'équivalence des items traduits apparaît plus délicat et nécessite des recherches. Nous avons présenté des exemples d'enquêtes visant à adopter une approche alternative à l'utilisation d'épreuves traduites.

En conclusion, il est évident que les résultats des enquêtes internationales ne doivent pas être rejetés systématiquement mais être communiqués et réfléchis… avec néanmoins un minimum de prudence. La mesure de la compétence des élèves n'est pas le système métrique. Toute mesure est un construit, avons-nous rappelé dans le chapitre 1, et le danger d'un construit outre les biais éventuels est d'induire une réification. Il faut prendre garde à ce point lors de la communication des résultats et de leur utilisation dans des analyses secondaires.

REFERENCES

Andrieux, V. ; Bonnet, G. ; Chollet, P. ; Hornet, S. ; Levasseur, J. ; Maurer, M-P. ; Regnier, C. ; Stern, L. ; Thaurel-Richard, M. ; White, J. ; Wirth, A. (2002). *Enseignement et compétences en lecture à la fin de la scolarité primaire dans trois pays européens.* Paris : ministère de l'éducation nationale, DPD Edition. 105 pages.

Bacher, F. (1973). La docimologie, in M. Reuchlin (éd.), *Traité De Psychologie Appliquée, tome VI*, 27-87, Paris, P.U.F.

Bacher, F., Isambert-Jamati & coll. (1970). Enquête sur l'orientation à la fin du premier cycle secondaire : résultats d'une post-enquête effectuée quatre ans plus tard. *Bulletin de l'Institut National d'Orientation Professionnelle*, *26*, n° spécial.

Ball, S. (1998). Big Policies/Small World. An Introduction to International Perspectives in Education Policy, in: *Comparative Education* 34(2). pp. 130-199.

Ball, S., Van Zanten A. (1998). Logiques de marché et ethniques contextualisées dans les systèmes français et britanniques. In : *Education et société*, 1, 1998, pp. 47-71.

Bautier, E., Crinon, J., Rayou, P., & Rochex, J.-Y. (2003). Socialisation scolaire et non scolaire chez des élèves. Présupposés et mobilisés chez les jeunes évalués. Equipe ESCOL. Université Paris 8.

Baye, A. (2004). La gestion des spécificités linguistiques et culturelles dans les évaluations internationales de la lecture. In : *Politiques d'éducation et de formation*, n° 11, pp. 55-70.

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Smith, T.A. & Kelly, D.L. (1996a). *Science Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Boston: Boston College.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Kelly, D.L. & Smith, T.A. (1996b). *Mathematics Achievement in the Middle School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Boston: Boston College.

Benavot, A. (2004). A Global Study of Intended Instructional Time and Official School Curricula, 1980-2000. Manuscrit non publié, BIE-Genève.

Benson, J., & Hutchinson, S. (1997). The state of the art in bias research in the United States. *European Review of Applied Psychology*, *47*, 281-294.

Bernstein, B. (1971, 1975, 1977, 1990). Class, Codes and Control. Vol. I-IV, Routledge & Kegan Paul, London.

Berry, J.W., Poortinga, Y.H., Segall, M.H., & Dasen P.R. (1992). *Cross-cultural psychology. Research and applications*. Cambrige: Cambridge University Press.

Binkley, M.R., & Pignal, J.R. (1998). An analysis of items with different parameters across countries. In T. S. Murray, I.S. Kirsch & L.B. Jenkins (Eds), *Adult Literacy in OECD countries. Technical report on the first international adult literacy survey* (pp. 143-160). Washington DC: U.S. Department of Education. National center for Education Statistics.

Blanchard, S., & Vrignaud, P. (1996). *Étude préliminaire en vue de l'élaboration de nouveaux outils d'évaluation des compétences des adultes pour les enquêtes internationales dans le cadre de l'OCDE.* Rapport pour le Ministère de l'Éducation Nationale. Direction de l'Évaluation et de la Prospective.

Bloom, E.D., Cohen, E.J. (2002). Education for ALL: An Unfinished Revolution. In: *Daedalus*, Summer 2002, pp. 84-94; American Academy of Arts and Sciences (2003): Project on Universal Basic and Secondary Education.

Blum, A., & Guérin-Pace, F. (1997). *Analyse des disparités culturelles des réponses à l'enquête IALS*. INED/MEN.

Blum, A., & Guérin-Pace, F. (2000). *Des lettres et des chiffres.* Paris : Fayard.

Blum, A., Goldstein, H. & Guérin-Pace, F. (2001). International Adult Litteracy Survey (IALS): An analysis of international comparisons of adult litteracy. *Assessment in Education, 8,* 225-256.

Bonora, D. (1972). Les facteurs de la réussite scolaire dans les disciplines scientifiques : une enquête internationale. *L'Orientation Scolaire et Professionnelle*, *1*, 385-410.

Bonora, D. (1973). Les buts de l'éducation, in M. REUCHLIN (éd.), *Traité De Psychologie Appliquée, tome VI*, 139-191, Paris, P.U.F.

Bonora, D. (1974a). Un exemple de recherche en éducation comparée : les travaux de l'I.E.A. étudiés à la lumière des résultats parus. *Bulletin De L'association Francophone d'Education Comparée*.

Bonora, D. (1974b). Educational aims and curriculum in France: an I.E.A. survey. *Comparative Education Review*, *18*, 217-227.

Bonora, D. (1975). Enquête internationale sur l'enseignement des sciences : quelques résultats. *L'Orientation Scolaire et Professionnelle, 4*, 137-168.

Bonora, D. (1988). Prémisse à l'évaluation : les préférences des professeurs à l'égard des objectifs pédagogiques. *L'Orientation Scolaire et Professionnelle, 17*, 323-342.

Bonora, D. (1996). Les modalités de l'évaluation. *Revue Internationale d'Education, Sèvres, 11*, n° spécial : l'évaluation des élèves, 69-85.

Bonnet, G. (2004a). Vers une méthodologie alternative pour l'évaluation de la lecture dans les enquêtes internationales. *Education et Formations, 70,* 109-121.

Bonnet, G. (2004a). Evaluation of Education in the European Union: Policy and Methodology. *Assessment in Education: Principles, Policy and Practice, 11*, 179-191.

Bonnet, G. (Ed) (2004b). *The assessment of pupils' skills in English in eight European countries.* European network of policy makers for the evaluation of education systems, MEN/DEP Édition, Paris. 210 pages.

Bonnet, G., et al. (2004c). *Culturally Balanced Assessment of Reading (C-BAR).* European network of policy makers for the evaluation of education systems. Paris MEN/DEP Édition. Site **http://cisad.adc.education.fr/reva/**

Bonnet, G. (2002). « Reflections in a critical eye: on the pitfalls of international assessment »; Review Essay of « Knowledge and Skills for Life ; First Results from PISA 2000 ». *Assessment in Education: Principles, Policy and Practice, 9,* 385-397.

Bonnet, G. ; Braxmeyer, N. ; Horner, S. ; Lappalainen, H-P. ; Levasseur, J. ; Nardi, E. ; Rémond, M. ; Vrignaud, P. ; White, J., (2001). *The use of national reading tests for international comparisons: ways of overcoming cultural bias.* Ministère de l'éducation nationale. DPD Edition diffusion. Paris.

Bonnet, G. (Ed.) (1998). *The effectiveness of the teaching of English in the European Union, Report of the colloquium and Background documents*, 210 p., Ministère de l'éducation nationale, DPD Edition diffusion, Paris.

Bonnet, G., & Levasseur, J. (2004). Evaluation des compétences en anglais des élèves de 15 à 16 ans dans sept pays européens. *Note d'évaluation, 00-04*.

Bottani, N., Pegoraro, R. (2004). La situation de l'enseignement secondaire dans les pays de l'OCDE. Document préparé à l'occasion du séminaire « *L'enseignement secondaire à l'échelle mondiale : bilans et perspectives »*, organisé avec le BIE, Genève.

Bourny, G., Braxmeyer, N., Dupé, C., Rémond, M., Robin, I., & Rocher, T. (2002). Les compétences des élèves français à l'épreuve d'une évaluation internationale. Premiers résultats de l'enquête PISA 2000. *Les Dossiers, N° 137*.

Bourny, G., Fumel, S., Monnier, A.-L., Rocher, T. (2004). Les élèves de 15 ans. Premiers résultats de l'évaluation internationale PISA 2003. *Note évaluation ; 04 12*.

Bourny, G., Dupé, C., Robin, I., & Rocher, T. (2001). Les élèves de 15 ans. – Premiers résultats d'une évaluation internationale des acquis des élèves. *Note d'Information* ; *01.52.*

Bruchon-Schweitzer, M., & Ferrieux, D. (1991). Les méthodes d'évaluation du personnel utilisées pour le recrutement en France. *L'Orientation Scolaire et Professionnelle, 20*, 56-77.

Burton, M.L., & Nerlove, S.B. (1976). Balanced designs for triads tests: Two examples from English. *Social Science Research, 5*, 247-267.

Cacouault, M., Orivel, F. (ed.) (1993). L'évaluation des formations : points de vue comparatistes. Actes du 15e Congrès de l'Association européenne d'éducation comparée, Dijon, juin 1992. IREDU-CNRS-Université de Bourgogne, France.

Canto-Sperber, M. & Dupuy, J.-P. (1999). *Competencies for the good life and the good society.* Neuchâtel: OFS, NCES, OECD.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*, Thousand Oaks, CA: Sage.

Céard, M.-T., Rémond, M., & Varier, M. (2003). *L'appréciation des compétences des élèves et des jeunes en lecture et en écriture et l'évolution de ces compétences dans le temps*. Les rapports établis à la demande du Haut Conseil de l'Évaluation de l'École, 11. 147 pages.

Cochran, W.G., & Cox, G.M. (1950). *Experimental designs*. New York: John Wiley & Sons.

Colmant, M., & Desclaux, A. (1996). Savoir écrire en français en fin de scolarité obligatoire dans quatre communautés francophones. *Note d'information ; 96.20*.

Colmant, M., & Mulliez, A. (2003). Les élèves de CM1. Premiers résultats d'une évaluation internationale en lecture (PIRLS). *Note d'Information, 22*.

Comber, L.C., & Keeves, J.P. (1973). *Science Education in Nineteen Countries*. Stockholm: Almquist & Wiksell; New York: John Wiley & Sons.

CRESSWELL, M. and GUBB, J. (1987). *The Second International Mathematics Study in England and Wales* (International Studies in Pupil Performance Series). Windsor: NFER-NELSON.

Cytermann, J.-R., (2003). Les enjeux internationaux et européens de l'éducation. Séance N°1, Le système éducatif français à travers les comparaisons internationales. Document du séminaire EHESS.

Dauphin, L., & Trosseille, B. (2004). Les compétences générales des élèves en fin de collège. *Note d'Information. DEP.*

de Landsheere, G. (1994). *Le pilotage des systèmes d'éducation*. Bruxelles : De Boeck.

DeSeCo (1999a). *Definition and selection of competencies: Theoretical and conceptual foundations. Background paper.* Neuchâtel: OFS, NCES, OECD.

DeSeCo (1999b). *Comments on the DeSeCo expert opinions*. Neuchâtel : OFS, NCES, OECD.

Dickes, P., & Flieller, A. (1997). *Analyse secondaire des données françaises de la première enquête internationale sur la littéracie des adultes (enquête IALS).* Rapport pour le Ministère de l'Éducation Nationale. Université de Nancy, Laboratoire de Psychologie, équipe GRAPCO EA 1129.

Dickes, P., & Vrignaud, P. (1995). *Rapport sur les traitements des données françaises de l'enquête internationale sur la littéracie.* Rapport pour le Ministère de l'Éducation Nationale. Direction de l'Évaluation et de la Prospective.

Dickes, P., Tournois, J., Flieller, A., & Kop, J.L. (1994). *Psychométrie.* Paris : PUF.

Dupé, C. & Olivier, Y. (2002). L'évaluation PISA. *Le Bulletin vert de l'Association des Professeurs de Mathématiques*, *439*.

Duru-Bellat, M., Mons, N., & Suchaut, B. (2003). Contextes nationaux, organisation des systèmes éducatifs et inégalités entre élèves : l'éclairage de l'enquête PISA 2000. *Politique d'Education et de Formations*, *9*, 95-108.

Duru-Bellat, M., Mons, N., & Suchaut, B. (2004a). Caractéristiques des systèmes éducatifs et compétences des jeunes de 15 ans : l'éclairage des comparaisons entre pays. Rapport pour le M.E.N., DIJON : IREDU-CNRS.

Duru-Bellat, M., Mons, N., & Suchaut, B. (2004b). Organisation scolaire et inégalités sociales de performances : les enseignements de l'enquête PISA. *Éducation & formations*, *70*, 123-131.

European consortium for the assessment of pupils' achievements (1997). *Project for the evaluation of achievements and competencies in education. PEACE I.* Tender N° OECD/DEELSA/SID/BPC (97.1)7. Dijon : Université de Bourgogne /IREDU.

Flieller, A. (1999). Étude d'un test lexical (définitions lacunaires) par des Modèles de Réponse à l'Item. *Psychologie et Psychométrie*, *20*, 65-84.

Flieller, A. (1989). Application du modèle de Rasch à un problème de comparaison de générations. *Bulletin de Psychologie* ; *XLII* ; 86-91.

Foxman, D. (1992). *Learning Mathematics and Science: the Second International Assessment of Educational Progress in England*. Slough: NFER.

Fuchs, C. (2002). Place et rôle de la variabilité dans les théories linguistiques. In J. Lautrey, B. Mazoyer & P. van Geert. *Invariants et variabilités dans les sciences cognitives* (pp. 157-174). Paris : Éditions de la Maison des Sciences de l'Homme.

Gardner, H. (1993/1997). *Frames of mind [1993], trad. fr. Les formes de l'intelligence*, Paris, Editions Odile Jacob.

Goldstein, H. (2004). International comparisons of student attainment: some issues arising from the PISA study. *Assessment in Education, 11,* 319-330.

Goldstein, H. (1980). Dimensionality, bias, independance and measurement scale problems in latent trait test score models. *The British Journal of Mathematical and Statistical Psychology*, *33*, 234-246.

Goldstein H., & Wood R. (1989). Five decades of item response modelling. *The British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Goldstein, H., Bonnet, G., & Rocher, T. (submitted). Multilevel multidimensionnal structural equation models for the analysis of comparative data on educational performance.

Goody, J. (1999). *Education and competences*. Neuchâtel: OFS, NCES, OECD.

Gould, S.J. (1996/1981). *The mismeasure of man*. New York: W.W. Norton & Company [traduction française : *La mal-mesure de l'homme* (1997). Paris : Editions Odile Jacob].

Guérin-Pace, F., & Blum, A. (1999). L'illusion comparative. Les logiques d'élaboration et d'utilisation d'une enquête internationale sur l'illettrisme. *Population, 54,* 271-302.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessement*, *10*, 229-244.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage publications.

Harris, S., Keys, W., and Fernandes, C. (1997). *Third International Mathematics and Science Study: Second National Report Part 1*. Slough: NFER.

Haste, H. (1999). *Competencies; Psychological realities*. Neuchâtel: OFS, NCES, OECD.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve. Intelligence and class structure in american life*. New York: The Free Press.

Heuyer, G., Piéron, H., & Sauvy, A. (1950). *Le niveau intellectuel des enfants d'âge scolaire*. Paris : PUF.

Hickman, M. (2002). Espace, langage et catégorisation : le problème de la variabilité interlangues. In J. Lautrey, B. Mazoyer & P. van Geert. *Invariants et variabilités dans les sciences cognitives* (pp. 225-242). Paris : Éditions de la Maison des Sciences de l'Homme.

Holland, P.W., & Wainer, H. (Eds) (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

House, E. (1998). Les mécanismes institutionnels de l'évaluation. In: *Perspectives*. Revue trimestrielle d'éducation comparée, BIE, Volume 28 n°105, pp. 123-131.

House, E. (2004). Aspects politiques des évaluations à grande échelle aux Etats-Unis. In : Politiques d'éducation et de formation, Analyses et comparaisons internationales n° 11.

Hulin, C.L. (1987). A psychometric theory of evaluations of item and scale translations. *Journal of Cross-Cultural Psychology* ; *18* ; 115-142.

Husén, T. (1967). *International Study of Achievement in Mathematics: a Comparison of Twelve Countries, Volumes I and II*. London: John Wiley.

Husén, T., Postlethwaite, N. (1996). A Brief History of the International Association for the Evaluation of Educational Achievement (IEA). In: Assessment in Education, vol. 3, n° 2, pp. 129-141.

Igersheim, J., Pluvinage, F., & Bonnin, J.-M. (1995). *Expertise des résultats français de l'enquête IALS*. *Rapport pour le Ministère de l'Éducation Nationale. Direction de l'Évaluation et de la Prospective*. Strasbourg : Université Louis Pasteur.

International Association for the Evaluation of Educational Achievement (1988). *Science Achievement in Seventeen Countries: a Preliminary Report*. Oxford: Pergamon Press.

Institut National d'Études Démographiques & Institut National d'Études du Travail et d'Orientation Professionnelle (1969, 1973, 1978). *Enquête nationale sur le niveau intellectuel des enfants d'âge scolaire. Cahiers de l'INED, N° 54, 64, 83.* Paris : PUF.

Johnson, E.G. (1992). The design of the National Assessment of Educational Progress. *Journal of Educational Measurement,* 29, 95-110.

Jones, L.V., Okin, I. (2004). The Nation's Report Card. Evolution and Perspectives. Phi Delta Kappa Educational Foundation, Bloomington.

Jouvenceau, P. (1992). Evaluation internationale en mathématiques et en sciences, élèves de 13 ans. Mars 1991. *Note d'information ; 92.31*.

Kail, M. (2000). Acquisition syntaxique et diversité linguistique. In M. Kail & M. Fayol (Eds), *L'acquisition du langage. Le langage en développement. Au-delà de trois ans* (pp. 9-44). Paris : Presses Universitaires de France.

Kallaghan, T. (1996). IEA Studies and Educational Policy. In: Assessment in Education, vol. 3, n°2, pp. 143-160.

Kirsch, I.S., Jungeblut, A., & Mosenthal, P.B. (1998). The measurement of adult literacy. In T. S. Murray, I.S. Kirsch & L.B. Jenkins (Eds), *Adult Literacy in OECD countries. Technical report on the first international adult literacy survey* (pp. 105-134). Washington, DC: U.S. Department of Education. National Center for Education Statistics.

Kolen, M. (2004). Linking assessments: Concept and history. *Applied Psychological Measurement*, *28*, 219-226.

Kolen, M.J., & Brennan, R.L. (1995). *Test Equating. Methods and practices*. New York: Springer Verlag.

Lautenschlager, G.J., & Park, D.-G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement* ; *12* ; 365-376.

Laveault, D., & Grégoire, J. (2002). Introduction aux théories des tests en sciences humaines. 2e édition. Bruxelles : De Boeck Université.

Lemann, N. (1999). *The big test: The secret history of the american meritocracy.* New York: Farrar, Strauss and Giroux.

Levasseur, J. & Shu, L. (1997). « Espagne, France Suède : Évaluation des connaissances et compétences en anglais des élèves de 15 à 16 ans ». *Les Dossiers d'Éducation et formations*, *92*.

Levy, F., & Murnane, R.J. (1999). *Are there key competencies critical to economic success ?* Neuchâtel: OFS, NCES, OECD.

Li, H., & Stout, W. (1996). A new procedure for detecting crossing DIF. *Psychometrika* ; *61* ; 647-677.

Lord, F., Novick, M.R. (Eds) (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Messick, S. (1998). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. *ETS Research Report. RR98-48*. Princeton, NJ: Educational Testing Service.

Messick, S. (1994). Foundations of validity: Meaning and consequences in psychological assessment. *European Journal of Psychological Assessement, 10*, 1-9.

Meuret, D. (2003). Pourquoi les jeunes de 15 ans ont-ils à 15 ans des compétences inférieures à celles des jeunes d'autres pays. *Revue Française de Pédagogie, 142,* 89-104.

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*; *17*; 297-334.

Mislevy, (Robert), & Verhelst, (Norman). — Modeling item responses when different subjects employ different solution strategies, *Psychometrika*, *55*, 1990, p. 195-215.

Mislevy, R.J., Sheehan, K.M., & Wingersky, M. (1993). How to equate tests with little or no data. *Journal of Educational Measurement*, 30, 55-78.

Mislevy, R.J., Beaton, A.E., Kaplan, B., & Sheehan, K.M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133-161.

Mons, N. (2004a). Politiques de décentralisation en éducation : diversité internationale, légitimations théoriques et justifications empiriques. *Revue Française de Pédagogie, 146,* 41-52.

Mons, N. (2004b). *De l'école unifiée aux écoles plurielles. Evaluation internationale des politiques de différenciation et de diversification de l'offre éducative*. Thèse en Sciences de l'Education : Université de Bourgogne.

Moskowitz, J., Garet, M., et al. (1997). Implementing the Data Strategy: a plan for the analysis and presentation of outcomes indicators. Pelavia Research Institute, Washington.

Murat, F., & Rocher, T. (2004). The method used for international assessment of educational competencies, translated by Jason Tarsh. In *Comparing Learning Outcomes – International assessment of education policy*. pp. 190-214 London: Routledge Falmer.

Nandakumar, R. (1994). Asessing dimensionality of a set of item responses-Comparison of different approaches. *Journal of Educational Measurement*, *31*, 17-35.

Novoa, A., Lawn, M. (2002). Fabricating Europe. Kluwer Academic Publishers.

Oakland, T., & Lane, H.B. (2004). Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests. *International Journal of Testing*, *4*, 239-252.

OCDE (1999). Mesurer les connaissances et competences des élèves. Un nouveau cadre d'évaluation. Paris. 1999.

OECD (2002). *Knowledge and skills for life. First results from PISA 2000. PISA in the news in France Dec 2001-Jan 2002*. Author.

OECD (2004). What Makes School Systems Perform ? Seeing School Systems Through the Prism of PISA. Paris.

Pacteau, C. & Vrignaud, P. (2001). Éducation : les limites des évaluations internationales. *Sciences Humaines*, *114*, 8-9.

Paulston, R.G. (1988). Comparative and International Education: Paradigms, Theories, and Debates. In: Education: The Complete Encyclopaedia 1.1. New York, Elsevier Science.

Perrenoud, P. (1999). The key to social fields: Essay on the competences of an autonomous actor. Neuchâtel: OFS, NCES, OECD.

Piéron, H. et Reuchlin, M. (Ed.) (1958). Etudes docimologiques. De l'enseignement primaire à l'enseignement secondaire, BINOP, 2e série, 14e année, numéro spécial.

Piéron, H. (1963). Examens et docimologie. Paris : PUF.

Ramirez, F., Meyer, J.W. (2002). National curricula: World models and national historical legacies. Manuscrit non publié. Département de sociologie, Université de Stanford.

Reuchlin, M. (1997). La psychologie différentielle. Nouvelle édition entièrement refondue. Paris : PUF.

Reuchlin, M., & Bacher, F. (1969). L'orientation à la fin du premier cycle secondaire. Paris : PUF.

Rémond, M. (2001). Adapter n'est pas traduire : Adaptation dans différents contextes culturels d'épreuves d'évaluation de la littéracie. In C. Sabatier & P. Dasen (Ed.). Cultures, développements et éducation. Autres enfants, autres écoles (pp. 198-205). Paris : L'Harmattan.

Rémond, M. (1996). Complexity and ambiguity of items measuring competence: some reflections about the prose and document items which invite further questioning. Communication à l'O.C.D.E. de l'expertise de l'enquête internationale sur l'alphabétisation des adultes.

Robin I. & Rocher T. (2002). La compétence en lecture des jeunes de 15 ans : une comparaison internationale. In *Données sociales*, *2002*, INSEE.

Rocher T. (2003). La méthodologie des évaluations internationales de compétences, *Psychologie et Psychométrie ; Numéro spécial : Mesure et Education. 24* ; 117-146.

Rubin, D. (1987). *Multiple imputation for nonresponse in survevs.* New York: Wiley.

Rychen, D., Salganik, L. (2003). Key Competencies for a Successful Life and a Well-Functioning Society. Hogrefe & Huber Publ.: Seattle.

Salganik, L.H., Rychen, D.S., Moser, U., & Konstant, J.W. (1999). Projects on competencies in the OECD context: Analysis of theoretical and conceptual foundations. Neuchâtel: OFS, NCES, OECD.

Salganik, L., Rychen, D. (Ed.) (2001). Defining and Selecting Key Competencies, Hogrefe & Huber Publ.: Seattle.

Salganik, L.H., Provasnik, J.S. (2004). Defining Quality Education for Universal Basic and Secondary Education (UBASE).

Salines, M., & Vrignaud, P. (2001). Apprécier et certifier les acquis des élèves en fin de collège : diplôme et évaluations-bilans. *Les rapports établis à la demande du Haut Conseil de l'Evaluation de l'Ecole*, 2.

Sarrazin, G. (1999). The French-Canadian adaptation of the WISC-III. International Conference on Test Adaptation. Washington, DC : May 1999.

Schmidt, H.W. et al. (1996). Characterising Pedagogical Flow: An Investigation of Mathematics and Science Teaching in Six Countries. Dordrecht: Kluwer Academic Publishers.

Schriewer, J. (2000). Discourse formation in comparative education. Berne: P. Lang.

Schriewer, J. (2004). L'internationalisation des discours sur l'éducation : adoption d'une idéologie mondiale ou persistance du style de réflexion systémique spécifiquement nationale ? dans *Revue française de pédagogie*, *n° 146,* janv-févr-mars.

Servant, A. & Murat, F. (1996). Les connaissances des élèves en mathématiques et sciences en terminale. *Note d'information ; 96.49*.

Servant, A. (1997). Evaluation internationale en mathématiques et en sciences des élèves de cinquième et de quatrième. *Note d'information ; 97.06*.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detect test bias/DTF as well as item bias/DIF. *Psychometrika* ; *58* ; 159-194.

Shealy, R.T., & Stout, W.F. (1993). An item response theory model for test bias and differential test functioning. In P.W.Holland & H. Wainer (Eds). *Differential item functioning* (pp. 197-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sireci, S.G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice*, *16*, 12-19.

Sireci, S.C. (1999). Statistical methods for determining problematic items in tests adaptations. *Workshop delivered at the international conference on adapting tests for use in multiple languages and cultures*. May 1999 : Washington DC.

Stanat, P. & Baumert, J. (2001). Introduction. *European Journal of Psychology of Education*, *16*, 331-333.

Stout, W. & al. (1999). *Dimensionality-based DIF/DBF package: SIBTEST, POLY-SIBTEST, CROSSING SIBTEST, DIFSIM, DIFCOMP*. Urbana, IL: The William Stout Institute for Measurement.

Stout, W., & Roussos, L. (1996). *SIBTEST manual*. Urbana-Champaign, IL: University of Illinois.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameter of item response models. In P.W. Holland & H. Wainer (Eds), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tjeldvall, A. (2000). Torsten Husén: Conversations in Comparative Education. Phi Delta Kappa. Educational Foundation, Bloomington.

Vallet, A., Bonnet, G., Emin, J.-C., Levasseur, J., Rocher, T., Blum, A., Guérin-Pace, F., Vrignaud, P., d'Haultfoeuille, X., Murat, F., Verger, D., & Zamora P. (2002). *Enquête méthodologique « Information et Vie Quotidienne »*. Paris : Institut National de la Statistique et des Études Économiques.

Van de Vijver, F., & Tanzer, N. (1997). Bias and equivalence in cross cultural assessment: An overview, *European Review of Applied Psychology*, *47*, 26-33.

Vrignaud, P. (à paraître, 2005). L'INETOP : 75 années dans l'histoire de l'évaluation psychologique et pédagogique en France. *L'Orientaion Scolaire et Professionnelle*.

Vrignaud, P. (2003). Objectivité et authenticité dans l'évaluation. Avantages et inconvénients des Questions à Choix Multiples et des Questions à Réponses Complexes : importance du format de réponse pour l'évaluation des compétences verbales. *Psychologie et Psychométrie, 24 (2/3),* 147-188.

Vrignaud, P. (2002). Les biais de mesure : savoir les identifier pour y remédier. *Bulletin de Psychologie*, *55(6)*, 625-634.

Vrignaud, P. (2001a). Le fonctionnement différentiel des items : méthodologie du contrôle des biais dans l'adaptation des épreuves pour les enquêtes internationales. In C. Sabatier & P. Dasen (Ed.). *Cultures, développements et éducation. Autres enfants, autres écoles* (pp. 185-197). Paris : L'Harmattan.

Vrignaud, P. (2001b). Evaluations sans frontières : comparaisons interculturelles dans le domaine de la cognition. In M. Huteau (Ed.). *Les figures de l'intelligence*, pp. 79-115. Paris : Editions et Applications Psychologiques.

Vrignaud, P. (2000). Psychological Assessment: An Overview of French-Language Theory and Methods. In M.R. Rozenzweig & K. Pawlik (Eds). *The International Handbook of Psychology* (pp. 387-392). London: Sage.

Vrignaud, P. (1999). Using PLS to study individual differences in the recall of knowledge from semantic memory. In M. Tenenhaus & A. Morineau (Eds). *Les méthodes PLS*. Montreuil : Centre International de Statistiques et d'Informatique Appliquées.

Vrignaud, P. (1996). Les tests au XXIe siècle. Que peut-on attendre des évolutions méthodologiques et technologiques dans le domaine de l'évaluation psychologique des personnes ? *Pratiques Psychologiques, 2*, 5-28.

Vrignaud, P., & Rémond, M. (2002, July). *The use of national reading tests for international comparisons: Results from a feasibility study.* XXVth International Congress of Applied Psychology. Singapore.

Vrignaud, P., & Bonora, D. (1998). Literacy assessment and international comparisons. In Dan Wagner. *Literacy assessment for out-of-school youth and adults*. Philadelphia: UNESCO/International Literacy Institute.

Vrignaud, P. & Chartier, P. (1999). Quand les modèles de mesure deviennent réducteurs : apports et limites des Modèles de Réponse à l'Item pour les comparaisons internationales. *Communication au colloque de l'ADMEE : « Évaluation des politiques d'éducation »*. IREDU - Dijon, 15-17 septembre 1999.

Vrignaud, P., Rémond, M. (2002). The use of national reading tests for international comparisons: Results from a feasibility study. *Communication to the XXVth International Congress of Applied Psychology*. Singapore, 7-12 July 2002.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores ? What is the effect of local dependence on reliability ? *Educational Measurement: Issues and Practice*, *15*, 22-29.

Weinert, F. E. (1999). Concepts of competence. Neuchâtel: OFS, NCES, OECD.

Weller, S.C., & Romney, A.K. (1988). *Systematic data collection*. Newbury Park, CA: Sage.

Wu, M.L., Adams, R.J., Wilson, M.R. (1997). *ConQuest. Generalized item response modelling software.* Hawthorn, Australia: ACER.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG. Multiple-Group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

ANNEXE 1: PRINCIPAUX RAPPORTS DES ENQUETES DE L'IEA

Etude pilote en douze pays (1959-1962)

Foshay, A.W., Thorndike, R.L., Hotyat, F., Pidgeon, D.A., & Walker, D.A. (1962). *Educational Achievement of Thirteen-Year-Olds in Twelve Countries.* Hamburg: UNESCO Institute for Education.

Keeves, J.P. (1995). *The World of School Learning: Selected Key Findings from 35 Years of IEA Research*. The Hague: IEA.

Première enquête internationale sur les mathématiques (FIMS) (1961-1965)

Husén, T. (Ed.). (1967). *A Comparison of Twelve Countries: International Study of Achievement in Mathematics* (Vols. 1–2). Stockholm: Almquist & Wiksell.

Keeves, J.P. (1995). *The World of School Learning: Selected Key Findings from 35 Years of IEA Research*. The Hague: IEA.

Postlethwaite, T.N. (Ed.). (1967). *School Organization and Student Achievement: A Study Based on Achievement in Mathematics in Twelve Countries.* Stockholm: Almquist & Wiksell; New York: John Wiley.

Enquêtes sur six matières (Six Subjects Study) :

- **Première enquête internationale sur les sciences (FISS) (1966-1975)**

Comber, L.C., & Keeves, J. P. (1973). *Science Education in Nineteen Countries*. Stockholm: Almquist & Wiksell; New York: John Wiley & Sons.

Keeves, J.P. (1995). *The World of School Learning: Selected Key Findings from 35 Years of IEA Research*. The Hague: IEA.

Walker, D.A. (1976). *The IEA Six-Subject Survey: An Empirical Study of Education in Twenty-One Countries*. Stockholm: Almquist & Wiksell; New York: John Wiley & Sons.

- **Compréhension de la lecture (1967-1973)**

Keeves, J.P. (1995). *The World of School Learning: Selected Key Findings from 35 Years of IEA Research*. The Hague: IEA.

Thorndike, R.L. (1973). *Reading Comprehension Education in Fifteen Countries: An Empirical Study.* Stockholm: Almquist & Wiksell.

Walker, D.A. (1976). *The IEA Six-Subject Survey: An Empirical Study of Education in Twenty-One Countries*. Stockholm: Almquist & Wiksell; New York: John Wiley & Sons.

- **Littérature (1966-1973)**

Purves, A.C. (1973). *Literature Education in Ten Countries*. Stockholm: Almquist & Wiksell; New York: John Wiley & Sons.

- **Anglais comme langue étrangère (1968-1975)**

Lewis, E.G., & Massad, C.E. (1975). *The Teaching of English as a Foreign Language in Ten Countries*. Stockholm: Almquist & Wiksell; New York: John Wiley.

- **Français comme langue étrangère (1968-1975)**

Carrol, J.B. (1975). *The Teaching of French as a Foreign Language in Eight Countries*. Stockholm: Almquist & Wiksell; New York: John Wiley & Sons.

- **Education civique (1967-1976)**

Oppenheim, A.N., & Torney, J.V. (1974). *The Measurement of Children's Civic Attitudes in Different Nations*. Stockholm: Almquist & Wiksell; New York: John Wiley & Sons.

Torney, J.V., Oppenheim, A.N., & Farnen, R.F. (1976). *Civic Education in Ten Countries: An Empirical Study*. Stockholm: Almquist & Wiksell; New York: John Wiley & Sons.

Deuxième enquête internationale sur les mathématiques (1976-1989)

Travers, K.J., & Westbury, I. (Eds.). (1989). *The IEA Study of Mathematics I: Analysis of Mathematics Curricula*. Oxford: Pergamon Press.

Robitaille, D.F., & Garden. R.A. (Eds.). (1989). *The IEA Study of Mathematics II: Context and Outcomes of School Mathematics*. Oxford: Pergamon Press.

Burstein, L. (Ed.). (1992). *The IEA Study of Mathematics III: Student Growth and Classroom Processes*. Oxford: Pergamon Press.

Enquête sur l'enseignement en classe (1978-1989)

Anderson, L.W., Ryan, D.W., & Shapiro, B.J. (Eds.). (1989). *The IEA Classroom Environmental Study*. Oxford: Pergamon Press.

Deuxième enquête internationale sur les sciences (SISS) (1979-1991)

Rosier, M.J., & Keeves, J.P. (1991). *Science Education and Curricula in Twenty-Three Countries: The IEA Study of Science I.* Oxford: Pergamon Press.

Postlethwaite, T.N., & Wiley, D.E. (Eds.). (1992). *Science Achievement in Twenty-Three Countries: The IEA Study of Science II.* Oxford: Pergamon Press.

Keeves, J.P. (Ed.). (1992). *The IEA Science Study III: Changes in Science Education and Achievement: 1970 to 1984.* Oxford: Pergamon Press.

Enquête sur la composition écrite (1980-1988)

Gorman, T.P., Purves, A,, & Degenhart, R.E. (Eds.). (1988). *The IEA Study of Written Composition I: The International Writing Tasks and Scoring Scales.* Oxford: Pergamon Press.

Purves, A.C. (Ed.). (1992). *The IEA Study of Written Composition II: Education and Performance in Fourteen Countries.* Oxford: Pergamon Press.

Enquête sur la compréhension de la lecture (1988-1994)

Elley, W.B. (Ed.). (1992). *How in the World do Students Read?* Hamburg: Grindeldruck GMBH.

Lundberg, I., & Linnakylä, P. (1993). *Teaching Reading Around the World: IEA Study of Reading Literacy.* Hamburg: IEA.
Papanastasiou, C., & Froese, V. (Eds.). (2002). *Reading Literacy in 14 Countries*. Cyprus: Cyprus University Press.

Postlethwaite, T.N., & Ross, K.N. (1992). *Effective Schools in Reading: Implications for Educational Planners: An Exploratory Study. The IEA Study of Reading Literacy II.* Hamburg: IEA.

Elley, W.B. (Ed.). (1994). *The IEA Study of Reading Literacy: Achievement and Instruction in Thirty-two School Systems.* Oxford: Pergamon Press.

Wagemaker, H. (Ed.). (1996). *Are Girls Better Readers? Gender Differences in Reading Literacy in 32 Countries.* The Hague: IEA.

Première enquête internationale sur les nouvelles technologies à l'école (1985-1993)

Pelgrum, W.J., & Plomp, T. (Ed.). (1991). *The Use of Computers in Education Worldwide: Results from the IEA Computers in Education Survey in 19 Education Systems*. Oxford: Pergamon Press.

Pelgrum, W.J., Janssen Reinen, I.A.M., & Plomp, T. (Eds.). (1993). *Schools, Teachers, Students, and Computers: A Cross-National Perspective.* The Hague: IEA.

Pelgrum, W.J., & Plomp, T. (Ed.). (1993). *The IEA Study of Computers in Education: Implementation of an Innovation in 21 Education Systems*. Oxford: Pergamon Press.

Plomp, T., Anderson, R.E., & Kontogiannopoulou-Polydorides, G. (Eds.). (1996). *Cross-National Policies and Practices on Computers in Education.* Dordrecht, Boston, London: Kluwer Academic Publishers.

Troisième enquête internationale sur les mathématiques et les sciences (TIMSS) (1995)

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzales, E.J., Smith, T.A., & Kelly, D.L. (1996). *Science Achievement in the Middle School Years: IEA's TIMSS.* Chestnut Hill, MA: Boston College.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzales, E.J., Kelly, D.L., & Smith, T.A. (1996). *Mathematics Achievement in the Middle School Years: IEA's TIMSS.* Chestnut Hill, MA: Boston College.

Harmon, M., Smyth, T.A., Martin, M.O., Kelly, D.L., Beaton, A.E., Mullis, I.V.S., Gonzales, E.J., & Orpwood, G. (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Boston College.

Howson, G. (1995). *Mathematics Textbooks: A Comparative Study of Grade 8 Texts. TIMSS Monograph No 3.* Vancouver BC: Pacific Educational Press.

Martin, M.O., & Kelly, D.L. (Eds.). (1996). *TIMSS Technical Report: Volume I Design and Development.* Chestnut Hill, MA: Boston College.

Martin, M.O., & Kelly, D.L. (Eds.). (1997). *TIMSS Technical Report. Volume II. Implementation and Analysis, Primary and Middle School Years.* Chestnut Hill, MA: Boston College.

Martin, M.O., & Kelly, D.L. (Eds.). (1998). *TIMSS Technical Report: Volume III. Implementation and Analysis, Final Year of Secondary School.* Chestnut Hill, MA: Boston College.

Martin, M. O., & Mullis, I.V.S. (Eds). (1996). *TIMSS: Quality Assurance in Data Collection.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzales, E.J., Smith, T.A., & Kelly, D.L. (1997). *Science Achievement in the Primary School Years: IEA's TIMSS.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gonzales, E.J., Smith, T.A., & Kelly, D.L. (1999). *School Context for Learning and Instruction in IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gregory, K.D., Hoyle, C., & Shen, C. (2001). *Effective Schools in Science and Mathematics.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzales, E.J., Kelly, D.L., & Smith, T.A. (1998). *Mathematics and Science Achievement in the Final Year of Secondary School: IEA's TIMSS.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Fierros, E.G., Goldberg, A.L., & Stemler, S.E. (2000). *Gender Differences in Achievement: IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Boston College.

Robitaille, D.F. (Ed.). (1997). *National Context for Mathematics and Science Education: An Encyclopedia of the Educational Systems Participating in TIMSS*. Vancouver, BC: Pacific Educational Press.

Robitaille, D.F., & Beaton, A.E. (Eds.). (2002). *Secondary Analysis of the TIMSS Data.* Dordrecht, Boston, London: Kluwer Academic Publishers.

Robitaille, D.F., Beaton, A.E., & Plomp, T. (Eds.). (2000). *The Impact of TIMSS on the Teaching & Learning of Mathematics & Science.* Vancouver, BC: Pacific Educational Press.

Robitaille, D.F., & Garden, R.A. (Eds.). (1996). *Research Questions and Study Design: TIMSS Monograph 2.* Vancouver, BC: Pacific Educational Press.

Schmidt, W.H., Jorde, D., Cogan, L.S., Barrier, E., Gonzalo, I., Moser, U., Shimizu, K., Sawada, T., Valverde, G.A., McKnight, C., Prawat, R.S., Wiley, D.E., Raizen, S.A., Brotton, E.D., & Wolfe, R.G. (1996). *Characterizing Pedagogical Flow: An Investigation of Mathematics and Science Teaching in Six Countries.* Dordrecht, Boston, London: Kluwer Academic Publishers.

Schmidt, W.H., McKnight, C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (1996). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Mathematics.* Dordrecht, Boston, London: Kluwer Academic Publishers.

Schmidt, W.H., Raizen, S.A., Brotton, E.D., Bianchi, L.J., & Wolfe, R.G. (1996). *Many Visions, Many Aims: A Cross-National Investigation of Curricular Intentions in School Science.* Dordrecht, Boston, London: Kluwer Academic Publishers.

L'éducation préscolaire (1986-2002)

Phase 1
Olmsted, P.P., & Weikart, D.P. (1989). *How Nations Serve Young Children: Profiles of Child Care and Education in 14 Countries.* Ypsilanti, MI: High/Scope Press.
Olmsted, P.P., & Weikart, D.P. (1994). *Families Speak: Early Care and Education in 11 Countries.* Ypsilanti, MI: High/Scope Press.

Phase 2
Olmsted, P.P., & Montie, J. (2001). *Early Childhood Settings in 15 Countries: What Are Their Structural Characteristics?* Ypsilanti, MI: High/Scope Press.
Weikart, D. P. (1999). *What Should Young Children Learn? Teacher and Parent Views in 15 Countries*. Ypsilanti: MI: High/Scope Press.
Weikart, D.P., Olmsted, P.P., & Montie, J. (2003). *World of Preschool Experience: Observation in 15 Countries*. Ypsilanti, MI: High/Scope Press.

Phase 3
High/Scope Educational Research Foundation. (2003). *Sights and Sounds of Children: 14 Countries.* (DVD and various videotape formats.) Ypsilanti, MI: High/Scope Press.
Weikart, D.P, Montie, J., & Xiang, Z. (in press). Preschool *Experience and Age 7 Child, Outcomes: Findings from 10 Countries*. Ypsilanti, MI: High/Scope Press.

Troisième enquête internationale sur les mathématiques et les sciences : réplique (TIMSS-R) (1999)

Gonzalez, E. J., & Miles, J.A. (Eds.). (2001). *TIMSS 1999 User Guide for the International Database: IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill, MA: Boston College.

Martin, M.O., Gregory, K.D., & Stemler, S.E. (Eds.). (2000). *TIMSS 1999 Technical Report: IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gonzales, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 International Science Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzales, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 International Mathematics Report: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Chestnut Hill, MA: Boston College.

Deuxième enquête internationale sur les nouvelles technologies à l'école (SITES) (1999-2002)

Pelgrum, W. J., & Anderson, R. E. (Eds.). (1999, 2001). *ICT and the Emerging Paradigm for Life Long Learning: An IEA Educational Assessment of Infrastructure, Goals, and Practices in Twenty-six Countries.* Amsterdam: IEA

Kozma, R.B. (Ed.). (2003). *Technology, Innovation, and Educational Change: A Global Perspective.* Eugene, OR: ISTE

Tjeerd Plomp, Ronald E. Anderson, Nancy Law and Andreas Quale (Eds.). (2003). *Cross-national Policies and Practices on Information and Communication Technology in Education.* Greenwich: Information Age Publishing

Enquête sur l'éducation civique (CIVED) (1994-2002)

Phase 1
Steiner-Khamsi, G., Torney-Purta, J., & Schwille, J. (Eds.). (2002). *New Paradigms and Recurring Paradoxes in Education for Citizenship*. Oxford: Elsevier Science Ltd.

Torney-Purta, J., Schwille, J., & Amadeo, J.-A. (Eds.). (1999). *Civic Education Across Countries: Twenty-Four National Case Studies for the IEA Civic Education Project*. Delft: IEA.

Phase 2
Amadeo, J.-A., Torney-Purta, J., Lehmann, R., Husfeldt, V., & Nikolova, R. (2002). *Civic Knowledge and Engagement: An IEA Study of Upper Secondary Students in Sixteen Countries.* Amsterdam: IEA.

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and Education in Twenty-eight Countries: Civic Knowledge and Engagement at Age Fourteen*. Delft: IEA.

Enquête sur les progrès dans la compréhension de la lecture (PIRLS) (2001)

Campbell, J.R., Kelly, D.L., Mullis, I.V.S., Martin, M.O., & Sainsbury, M. (2001). *Framework and Specifications for PIRLS Assessment 2001* (2nd ed.). Chestnut Hill, MA: Boston College.

Gonzalez, E.J., & Kennedy, A.M. (2003). *PIRLS 2001 User Guide for the International Database.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & Kennedy, A.M. (2003). *Trends in Children's Reading Literacy Achievement 1991–2001: IEA's Repeat in Nine Countries of the 1991 Reading Literacy Study.* Chestnut Hill, MA: Boston College.

Martin, M.O., Mullis, I.V.S., & Kennedy, A.M. (Eds.). (2003). *PIRLS 2001 Technical Report*. Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Kennedy, A.M. (2003). *PIRLS 2001 International Report: IEA's Study of Reading Literacy Achievement in Primary School.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Kennedy, A.M., & Flaherty, C.L. (2002). PIRLS *2001 Encyclopedia: A Reference Guide to Reading Education in the Countries Participating in IEA's Progress in International Reading Literacy Study (PIRLS).* Chestnut Hill, MA: Boston College.

ANNEXE 2 : GASTON MIALARET

Gaston MIALARET est né le 10 octobre 1918 à Paris. Après ses études secondaires, il prépare la licence et le D.E.S. de mathématiques auprès de l'Université de Toulouse et, à Paris, la licence de psychologie, le diplôme de l'Institut de psychologie de l'Université de Paris, le Professorat de psycho-pédagogie des E.N.N.A. et le Certificat d'aptitude à l'Inspection des écoles primaires et à la Direction des Ecoles normales.

D'abord instituteur puis maître de classe d'application, il est chargé d'organiser la première classe de « sixième nouvelle » au lycée d'Albi avant de devenir professeur de mathématiques au Collège moderne et technique de la même ville.

Après son passage à l'Ecole normale supérieure de Saint-Cloud, il est chargé d'organiser le premier laboratoire de psycho-pédagogie de l'E.N.S. de Saint-Cloud tandis qu'il est Chargé d'enseignement à la Sorbonne, à l'Institut de psychologie, à l'E.N.N.A. de Paris.

En 1957, il soutient ses deux thèses en vue du Doctorat ès Lettres sur « L'apprentissage des mathématiques et Sélection » et « La formation des professeurs de mathématiques ».

Il est ensuite chargé d'organiser la licence de psychologie de l'Université de Caen où il est nommé en 1953. C'est dans cette université que se déroulera toute sa carrière : chef de travaux, maître de conférences, professeur sans chaire, professeur titulaire. Au sein de l'Université de Caen il développe les études de psychologie puis il est chargé, en 1967, d'organiser les études en vue de la licence et de la maîtrise en sciences de l'éducation. C'est à cette époque qu'il demande que l'intitulé de sa chaire soit modifié pour devenir « chaire de sciences de l'éducation ». Dans le cadre de ses activités caennaises, il obtient la création du troisième centre français médico-psycho-pédagogique et, après de nombreuses démarches, la possibilité de travailler dans une école de la ville de Caen qui sera considérée comme une école expérimentale. Puis il se bat pour obtenir la création d'un Institut de formation des conseillers d'orientation scolaire et professionnelle à Caen.

Après avoir pris sa retraite en 1984, il est appelé par l'UNESCO pour assurer la direction du Bureau international de l'Education à Genève (1987-1988).

ANNEXE 3 : L'AVIS DU NATIONAL RESEARCH COUNCIL AMÉRICAIN

Aux Etats-Unis, deux comités se sont penchés au début de l'année 2000 sur les avancées méthodologiques des enquêtes internationales sur vaste échelle : le comité sur les études comparées de l'éducation au niveau internationale, présidé par Emerson Eliott du Conseil national pour l'accréditation de la formation des enseignants et le Comité sur les tests et l'évaluation présidé par Mme Eva Baker, directrice du CRESST (Centre pour les recherches sur l'évaluation) de l'UCLA (Université de Los Angeles).

Le rapport conjoint de ces deux comités a été édité par A.C.Porter et A.Gamoran : *Methodological Advances in Cross-national Surveys of Educational Achievement*. National Research Council, National Academy Press, Washington DC 2002.

Les conclusions de ces deux comités sur la qualité de ces études sont fondamentalement positives, tout en reconnaissant que de nombreux problèmes subsistent et méritent d'être approfondis et résolus :

- La qualité des enquêtes est élevée
- Les résultats sont crédibles et peuvent faire autorité
- La méthodologie a fait des progrès considérables en 40 ans :
    - Meilleurs tests
    - Meilleurs échantillonnages
    - Meilleure documentation sur les tests
    - Meilleures analyses statistiques
- Les Etats-Unis doivent continuer à participer à ces enquêtes et doivent promouvoir le déroulement d'enquêtes internationales régulières sur vaste échelle sur les acquis des élèves
- Les deux comités ne se sont pas accordés sur la population qui devrait être échantillonée dans ces études (par âge ou par degré)

PRINCIPAUX POINTS EN SUSPENS (SELON LE NATIONAL RESEARCH COUNCIL)

- ***DEVELOPPER UNE MEILLEURE APPRECIATION DES DIFFERENCES DE CONTEXTES CULTURELS ET SOCIAUX DE L'ENSEIGNEMENT***
- ***COMPRENDRE LA PART DE L'INFLUENCE DES DIFFERENCES DE CONTEXTE SUR LES RESULTATS DES TESTS***

Remarques à propos de la nature des tests utilisés dans les enquêtes internationales comparées sur vaste échelle sur les acquis des élèves

- La conception de tests standardisés au niveau international est ambivalente : on hésite toujours entre l'approfondissement d'un domaine et l'exploration élargie du champ des apprentissages

- Le choix de traiter un large éventail de sujets oblige à préférer les questions à choix multiples plutôt que les questions ouvertes, ce qui limite la possibilité de tester les compétences complexes
- Il y a eu des progrès dans la création d'échelles communes pour un même degré d'une enquête à l'autre; mais il y a un échec dans la création d'échelles communes à plusieurs degrés dans la même enquête. Exemple : en TIMMS-R il y a impossibilité de mesurer les gains en passant du 4ème degré au 8ème, parce que les test n'ont pas été construits pour être sur la même échelle
- La conception de l'échantillonnage pour la POP3 de TIMSS est mauvaise (la dernière année du postobligatoire), car la proportion de la cohorte qui se trouve à l'école dans cette année varie énormément d'un pays à l'autre. Cet échantillonnage rend impossible toute comparaison

ANNEXE 4 : ENQUETES INTERNATIONALES SUR LES CONNAISSANCES EN MATHEMATIQUES

Les enquêtes TIMSS 2003 et PISA 2003 ont été les dernières d'une série d'évaluations internationales commencées en 1964. Les études antérieures dans ce domaine ont été les suivantes :

Année	Evaluation	Références
1964	First International Mathematics Study (FIMS)	Husen (1967) Pidgeon (1967)
1980–82	Second International Mathematics Study (SIMS)	Robitaille and Garden (1989) Cresswell and Gubb (1987)
1988	The first study carried out by the International Association for the Evaluation of Educational Progress (IAEP1)	Travers and Westbury (1989) Lapointe *et al.* (1989) Keys and Foxman (1989)
1991	The second study carried out by the International Association for the Evaluation of Educational Progress (IAEP2)	Lapointe *et al.* (1992a) Foxman (1992)
1994–5	The Third International Mathematics and Science Study (TIMSS)	Beaton *et al.* (1996a) Mullis *et al.* (1997) Keys, Harris and Fernandes (1996) (1997a) Harris, Keys and Fernandes(1997) Keys, Harris and Fernandes(1997b)
1998–9	The Third International Mathematics and Science Study Repeat (TIMSS-R)	Mullis *et al.* (2000) Ruddock (2000)

ANNEXE 5 : LE CONSORTIUM « UNIVERSITE DE BOURGOGNE » DANS PISA 2000

L'appel d'offre pour l'enquête PISA a été l'occasion pour la DEP et le RERPESE de rassembler des organismes et des laboratoires français et européens intéressés par l'évaluation (*European consortium for the assessment of pupils' achievements*, 1997). Pour des raisons de visibilité internationale et de facilité administrative, le projet a été identifié comme celui de l'Université de Bourgogne, mais en fait, la DEP en était l'initiateur. Les principaux organismes participant étaient :

Prime contractor :
IREDU (Université de Bourgogne)

Membres :
Office for National Statistics (Angleterre)
Direction de l'Evaluation et de la Prospective, MEN (France)
Skolverket (Suéde)
OCTO (Center for Applied Research in Education) (Netherland)
Centro Europeo dell'Educazione (Italia)

Partenaires :
Department of Educational and Psychological Research (Lund University, Sweden)
Department of Educational Measurement and Data Analysis (University of Twente, Netherland)
INRP (France)
INETOP/CNAM (France).

Il faut donc souligner que ce projet regroupait des organismes réputés dans le domaine de l'évaluation des compétences des élèves. Il a montré que des organismes français pouvaient avoir une attitude constructive dans le domaine et ne se cantonnaient pas seulement à la critique de l'enquête IALS.

Il est intéressant de noter que ce projet a eu deux retombées importantes. La première a été sa continuation dans un programme européen Socrates dédié à l'évaluation de la littéracie (ce programme fait l'objet du prochain paragraphe). La seconde est la reprise de la grille de compétences de lecture élaborée par Martine Rémond (INRP) pour ce projet dans le cadre de PISA.

ANNEXE 6 : LES METHODES D'IDENTIFICATION DU FDI

Avant de présenter les méthodes, il est utile de définir quelques termes et concepts employés. Le concept de cote est un concept central dans l'identification du FDI. La cote est le rapport entre la probabilité que l'événement se réalise ici la réussite à l'item et la probabilité que l'événement contraire se réalise ici l'échec à l'item. Pour identifier les FDI, on va comparer la cote des items dans les deux groupes. En effet, on peut faire l'hypothèse que si les chances de réussite dans les deux groupes sont les mêmes, à compétence égale, le rapport entre les cotes est proche de 1. Un écart important à cette valeur traduira un FDI. Dans la recherche de FDI entre deux groupes, un des groupes est considéré comme le groupe de référence, l'autre comme le groupe cible. En général, le groupe représentant la culture majoritaire sert de groupe de référence, et le groupe pour lequel on soupçonne un FDI le groupe cible. II est d'usage de désigner les items pour lesquels on a des raisons de faire l'hypothèse de la présence d'un FDI comme les items suspects, les items pour lesquels on a pu faire la preuve qu'ils n'étaient pas entachés de FDI comme les items valides.

La statistique de Mantel Haenszel

La statistique de Mantel Haenszel a été adaptée pour l'étude du FDI par Holland et Thayer (voir Dorans et Holland, 1993). Elle s'appuie sur le principe selon lequel le FDI se manifeste lorsque la fréquence de réussite de deux populations à un item diffère à compétence constante. Pour vérifier la présence de FDI selon cette approche, il est nécessaire de constituer des groupes de sujets appariés selon la compétence afin de comparer leurs fréquences de réussite. On constitue des classes de compétence homogène à partir du score au test étudié (en général, autant de groupes que de scores observés) On construit alors des tables de contingence à trois dimensions : niveau, groupe, réussite/échec, soit pour chaque niveau, une table de contingence croisant groupe et réussite/échec. L'avantage de la méthode de Mantel-Haenszel est la relative simplicité conceptuelle et la facilité de sa mise en œuvre. Elle est souvent employée dans les enquêtes internationales en complément des méthodes basées sur les MRI.

La régression logistique

La régression logistique permet d'étudier l'influence de différentes variables sur la réalisation d'un événement. La variable dépendante est dichotomique : réussite ou échec, mais on va travailler sur la cote. La variable dépendante est alors la cote transformée en logarithme. L'idée d'appliquer la régression logistique est due à Swaminathan et Rogers (1990). Les variables indépendantes sont le score total, une variable en général dichotomique codant le groupe et l'interaction entre la variable groupe et le score total. On fait l'hypothèse que, en l'absence de FDI, seul le coefficient de régression représentant l'effet du score total (la compétence) est significatif. Si le coefficient représentant l'effet du groupe est significatif, alors on est en présence de FDI uniforme et si le coefficient représentant l'effet de l'interaction entre le groupe et le score est significatif, alors on est en présence d'un FDI croisé.

L'approche par les MRI

Nous avons présenté longuement les MRI puisque c'est le modèle de mesure le plus utilisé dans les enquêtes internationales. Nous allons cependant rappeler rapidement les

principaux concepts des MRI pour introduire l'étude du FDI selon cette méthode. Les MRI sont fondés sur la recherche d'un modèle mathématique permettant de représenter la relation entre un item et un sujet. On utilise en général la fonction logistique. Le modèle le plus courant comprend deux paramètres pour modéliser le fonctionnement de l'item : bi, la difficulté de l'item, ai, la pente (discrimination de l'item), et un paramétre pour rendre compte de la compétence du sujet, θj. Ces paramètres permettent de représenter chaque item par le graphe de sa fonction (courbe caractéristique des items, CCI en abrégé).

L'étude du FDI sur le seul paramètre de difficulté des items (FDI uniforme) est le cas le plus simple, soit que l'on utilise le modèle de Rasch qui rend compte du fonctionnement des items à l'aide du seul paramètre de difficulté, soit que l'on postule que les paramètres de discrimination sont identiques dans les groupes étudiés. On procédera en testant, pour chaque item, l'hypothèse d'égalité du paramétre de difficulté estimé dans chacun des groupes. Pour tester cette hypothèse, on estime les paramètres de difficulté dans chacun des groupes pour lesquels on suspecte la présence de FDI, puis, on ramène à une même origine les deux distributions, ce qui élimine l'effet de l'impact. Le rapport entre la différence, subsistant éventuellement, entre l'estimation des items dans les groupes et l'erreur de mesure de cette différence fournit un indicateur de biais standardisé (*Standardized index of bias*).

L'utilisation des MRI pour l'identification des FDI s'intègre particulièrement bien dans la démarche de validation du dispositif de mesure lorsque le test est construit et calibré à l'aide de ces modèles ce qui est le cas des enquêtes internationales. Dans ces enquêtes, l'étude des biais est une étape de la mise en œuvre des MRI selon des approches basées sur le calcul d'indicateurs de biais.

L'approche de Stout

Stout et collaborateurs (Shealy et Stout, 1993) ont réalisé un ensemble de travaux sur le FDI en s'appuyant sur la dimensionnalité d'un ensemble d'items puisque la présence de FDI est le signe que l'on évalue plus d'une dimension. Cette approche est dénommée indicateur de biais simultané (Simultaneous index of bias). Ils recommandent, dans un premier temps, d'identifier l'ensemble des items valides (exempts de FDI) et l'item ou le sousensemble des items suspects par une analyse de la (des) dimensions) présentes) dans l'ensemble des items. Ils font, ensuite, l'hypothèse que les items suspects peuvent dépendre de deux variables latentes de compétence (la variable de compétence et une variable parasite), et que les deux groupes ne possèdent pas un niveau égal dans la compétence correspondant à cette variable parasite, ce qui implique un FDI sur les items qui y sont sensibles. Bien que leur cadre conceptuel soit celui des MRI, la procédure ne nécessite pas d'estimer les variables latentes pour effectuer la comparaison à compétence égale. Les auteurs utilisent une variable construite dans le cadre de la théorie classique des tests, un score vrai estimé à partir d'une régression sur le score observé, régression corrigeant l'erreur de mesure. L'appariement des sujets étant réalisé sur ce score vrai, on va comparer la réussite à l'item pour chacune des classes de compétences. L'indicateur de FDI, nommé bêta, est la somme des différences de réussite standardisées pour l'ensemble des classes.

Cette approche est particulièrement intéressante car elle permet d'étudier le FDI en prenant en considération plusieurs items à la fois, d'où sa dénomination de « simultanée ». En effet, un FDI, même faible, présent sur plusieurs items va se renforcer et peut, en fin de compte, avoir un effet très important sur le test dans son ensemble ce que Stout désigne comme un

« fonctionnement différentiel du test ». Dans ce cas, bêta est calculé sur 1e score moyen au bloc des items suspects, pour chaque classe de compétence sur les items valides. Cette approche peut de même être appliquée aux items polytomiques et généralisée pour l'identification du FDI croisé (Li et Stout, 1996).

ANNEXE 7

LE PROJET DE L'OCDE : DEFINITION ET SELECTION DES COMPETENCES FONDEMENTS THEORIQUES ET CONCEPTUELS (DESECO)

(www.portal-stat.admin.ch/deseco/index.htm)

1. PRESENTATION DU PROJET

1.1. PRESENTATION ET JUSTIFICATION DU PROJET :

Ce projet, qui s'est déroulé jusqu'à la fin 2001, a été lancé fin 1997, sur l'initiative de l'Office Fédéral Suisse de la statistique, dans le cadre du projet sur les indicateurs de l'OCDE. Le travail a été entrepris en étroite collaboration avec le réseau A du projet INES et avec ILSS. La Suisse a présidé et assuré la gestion du projet avec l'appui organisationnel des USA et du Secrétariat de l'OCDE.

La rapidité des changements survenant dans la vie économique, sociale et politique, l'avènement des nouvelles technologies, la globalisation de l'économie font prendre conscience de l'importance qu'ont les connaissances et les compétences pour assurer le bien-être dans l'avenir. Ces facteurs conduisent les décideurs à rechercher de l'information sur les compétences de la population ainsi que sur les effets qu'ont, sur leur développement, l'éducation, la formation, et l'apprentissage informel.

Pour informer ceux qui ont en charge la politique éducative, l'OCDE élabore des indicateurs de compétences utilisables dans des comparaisons internationales portant sur les « produits » des systèmes éducatifs et sur la distribution des compétences dans la population. Les compétences visées dans les enquêtes favorisent le bien-être individuel, social et économique mais, jusqu'alors, leur sélection et leur définition ne se sont pas appuyées sur des fondements théoriques solides. Le projet DESECO a cherché à développer un cadre conceptuel théoriquement fondé afin de mieux comprendre ce que sont les compétences nécessaires pour mener une vie personnellement et socialement valable dans un Etat moderne et démocratique. Il a été développé pour fournir un cadre de référence pour les travaux théoriques ultérieurs et pour la mesure des compétences.

1.2. OBJECTIFS :

Le projet avait trois objectifs :

1.2.1. Progresser dans la détermination des fondements théoriques des compétences clés.

La méthodologie adoptée est celle d'une approche scientifique à la fois interdisciplinaire et internationale : il s'agit de dégager un cadre théorique commun permettant d'identifier les compétences clés. Ces dernières sont nécessaires pour que les individus mènent une vie pleinement réussie et responsable et pour que la société soit en mesure de faire face aux défis du présent et de l'avenir.

1.2.2. Servir de point de référence pour développer des indicateurs et pour interpréter les résultats empiriques.

Le projet pourrait servir de point de référence pour développer et valider des indicateurs de compétence qui constitueraient une base plus solide pour l'interprétation des résultats des enquêtes. Les comparaisons entre les types de compétences exigées dans diverses situations et l'information recueillie sur ce que les individus apprennent dans différents environnements éducatifs fourniront des feed-back utiles à la politique éducative. Des critères pertinents pourront être sélectionnés à partir d'un cadre théoriquement fondé en vue d'une évaluation des systèmes éducatifs.

1.2.3. Encourager les allers et retours entre les approches conceptuelles et les travaux empiriques.

Dans ce projet, la notion de compétence a une extension large, recouvre tout un ensemble d'activités attendues d'un individu vivant dans une société moderne, complexe et démocratique : participation satisfaisante au marché du travail, à la politique, aux réseaux sociaux, aux relations interpersonnelles, satisfaction personnelle. Des effets positifs sont attendus pour la société en matière de compétition économique, de capacité de répondre aux défis sociaux et économiques, de bon fonctionnement des réseaux sociaux et de bien-être social général.

Une hypothèse fondamentale était sous-jacente au développement des indicateurs et aux tentatives effectuées pour mesurer les compétences des élèves et des adultes au cours des années récentes : les compétences clés existent et jouent un rôle particulièrement significatif dans la capacité de gérer sa vie. C'est cette hypothèse qui doit être soigneusement examinée.

Les connaissances transmises par les programmes d'études dans les institutions éducatives n'épuisent pas les aspects à considérer dans la détermination des compétences : les pratiques de recrutement comme les théories scientifiques attestent l'importance des qualifications, des compétences, des conduites, des attitudes et des valeurs qui ne relèvent pas des programmes d'études classiques.

Les compétences font l'objet de multiples approches conceptuelles. Les différentes perspectives tiennent d'une part aux disciplines engagées dans la définition des compétences clés et d'autre part aux champs sociaux dans lesquels ces compétences sont prises en considération. La définition et la sélection des compétences sont l'objet d'une négociation entre politiques et ne sont pas simplement un objet de réflexion scientifique.

2. PLAN D'ACTION

Le projet s'est déroulé selon trois phases principales.

2.1. PHASE 1 :

2.1.1. Analyse du travail antérieur relatif au développement des indicateurs effectué dans le cadre des projets de l'OCDE.

Une analyse théorique et conceptuelle des projets antérieurs traitant des compétences (CCC, IALS, HCIP), ainsi que des projets en cours, a été entreprise à partir de deux sources : les documents finalisant les projets et des entretiens semi-structurés avec des acteurs clés de ces projets. Elle vise à faire le point sur les théories scientifiques et sur les postulats culturels et normatifs qui ont servi de base à ces projets, sur les intentions originelles et sur les évolutions constatées lors du développement et de la réalisation des enquêtes.

2.1.2. Analyse des concepts de compétences.

Dans un rapport, réalisé par Weinert, les types de compétences ont été catégorisés et leurs fondements épistémologiques et conceptuels ont été analysés. Les différences et les convergences entre les concepts ont été identifiées. Le rapport contient un glossaire, des points de référence théoriques et des recommandations sur le cadre théorique le mieux susceptible de convenir pour catégoriser les approches existantes des concepts.

2.2. PHASE 2 :

2.2.1. Opinions des experts sur la compétence.

Des spécialistes de la psychologie, de la sociologie, de l'économie, de la philosophie et de l'anthropologie ont été invités à rédiger des contributions au sujet de la compétence.

2.2.2. Premier symposium international.

Afin de progresser dans les fondements théoriques qui sous-tendent l'identification des compétences pertinentes et afin de développer un cadre théorique commun, un symposium a été organisé à Neuchâtel (Suisse) en 1998, consacré à l'examen de l'analyse du concept de compétence, à l'examen et à la discussion des rapports des experts et à des commentaires de ces rapports par des représentants des divers champs sociaux.

2.2.3. Résultats de la phase 2.

Deux publications distinctes ont été consacrées aux contributions préparées pour ce symposium. La première a été centrée sur les politiques relatives au développement et à la mesure des compétences; la seconde, d'ordre scientifique, visait les personnes intéressées par le rapport complet.

2.3. PHASE 3 :

2.3.1. Processus de commentaire par pays et second symposium international.

Les pays participants ont pu fournir un feed-back sur les productions et sur les résultats obtenus lors des deux premières phases. Un second symposium s'est tenu à Genève début 2001 pour conclure le projet avec des recommandations destinées au milieu politique.

2.3.2. Résultats.

Tous les documents produits au cours de ce projet sont accessibles en ligne sur le site **www.portal-stat.admin.ch/deseco/index.htm**