Vers un test adaptatif - criterié, combinant l'utilisation de l'IRT et de l'ASI, pour l'évaluation du socle commun de connaissances et de compétences

Antoine Bodin
& équipe de recherche "socle commun"

IREM d'AIX-MARSEILLE - Université de la Méditerranée

antoinebodin@mac.com

Actes 5eme colloque international de l'Analyse statistique implicative. Palerme – Italie – Novembre 2010

Résumé. La communication fera le point sur une recherche visant au développement d'une méthode de d'évaluation sur ordinateur, de nature adaptative, combinant l'utilisation de l'analyse des réponses à l'item (IRT) et de l'analyse statistique implicative (ASI). Nous montrerons comment l'ASI permet d'envisager des tests fiables sur le plan édumétrique tout en étant critériés. La nécessité de disposer d'un tel test pour l'évaluation du socle commun de connaissances et de compétences, du moins pour son volet mathématique, a constitué le point de départ de cette recherche qui allie la recherche sur le plan théorique et le développement sur le plan technique.

Après avoir évoqué la problématique concernant l'évaluation en général, l'évaluation certificative en particulier, et la difficulté d'obtenir des évaluations qui soient à la fois valides d'un point de vue didactique et fidèles d'un point de vue édumétrique, nous présenterons la problématique du socle commun français. Les grandes lignes de la construction d'un test seront présentées ainsi que certaines difficultés qui restent à résoudre, en particulier en ce qui concerne le logiciel CHIC.

Abstract. The communication will take stock of a research aimed to developing a Computerized Adaptive Testing method (CAT) combining the use of the Item Response Analysis (IRT) and of the Statistical Implicative Analysis (ASI). We will show how ASI allows to planning tests both edumetricaly trustworthy and criterion-referenced. The need to be able to rely on such a test to asses all French citizens in relation to the French official national common core of knowledge and skills ("socle commun"), at least for its mathematical component, was the starting point for this research which seeks to combine theoretical research and technical development.

After outlining some issues pertaining to assessment in general and certification in particular, as well as the difficulty to obtaining estimates that would be both didactically valid and edumetricaly reliable, we will present the French common core set of issues. The outline of the test building process will be presented as well as some problems which remain unresolved, particularly regarding the CHIC software.

1. Présentation de la communication

La recherche porte sur la possibilité de développer un test adapté à la certification d'un domaine particulier de connaissances. Dans notre cas précis, il s'agit du volet mathématique du socle commun de connaissances et de compétences tel qu'il est défini, en France, par le législateur, mais la problématique et les solutions proposées pourront être facilement étendues à d'autres domaines.

La communication mettra en évidence l'intérêt qu'il y aurait de pouvoir disposer, pour une telle évaluation, d'une banque de questions d'évaluation aussi riche que possible et d'un test adaptatif à support informatique associé à cette banque. Elle montrera l'utilité et les insuffisances, en ce qui concerne la construction d'un tel test, des méthodes issues de la psychométrie et comment le recours à l'analyse implicative peut aider à la résolution de certaines des difficultés rencontrées.

L'état de développement de notre test et les premiers résultats de l'expérimentation en cours illustreront les aspects théoriques présentés.

2. L'évaluation des connaissances à la croisée de deux paradigmes

Longtemps, l'évaluation des connaissances, du moins lorsqu'elle se voulait objective, s'est appuyée sur la théorie classique des tests, qui, on le sait, privilégie la fidélité par rapport à la validité et, de ce fait, ne permet qu'indirectement le contrôle de la qualité des items. Toute l'évaluation reposait alors sur l'idée qu'il y avait un objet à mesurer, que cet objet avait une vraie valeur, et que la mesure de cette valeur était possible. Dès lors, il était clair qu'en répétant la mesure un nombre de fois suffisant on pouvait approcher cette valeur d'aussi près que l'on voulait. Cette conception prévaut encore dans bien des cas. Les théories de l'analyse des réponses à l'item (IRT)[1], bien qu'issues de la psychométrie, comme cela est le cas de la théorie classique des tests, ont apporté un progrès notable en ce qui concerne l'analyse de la qualité des questions et leur calibrage, en permettant des comparaisons que la théorie classique ne permettait pas et, surtout, en permettant la construction d'échelles quasiment indépendantes des questions d'évaluation, et même des groupes de sujets, qui ont présidé à leur élaboration ; de plus, avec l'IRT, les questions d'évaluation et les sujets de l'évaluation peuvent être rapportés à une même échelle.

Toutefois, l'IRT reste sous l'emprise de la psychométrie et du mythe de la vraie note (Chevallard, Y. 1986). Les échelles construites dans ce modèle permettent sans doute de placer avec précision les sujets comme les questions les uns et par rapport aux autres mais elles n'autorisent ni l'analyse qualitative du questionnement évaluatif ni l'analyse des significations des résultats des sujets.

On sait combien, au cours des vingt dernières années, l'IRT a investi le champ des études évaluatives à grande échelle : d'abord les études internationales (TIMSS, PISA, PIRLS, ALL,...) et, de plus en plus, les études nationales à grande échelle. Dans ces études, ce qui compte le plus est que la position relative de deux pays ou de deux groupes de sujets sur l'échelle produite ne puisse pas être contestée sur le plan

[1] IRT : Item Response Theory

méthodologique. La signification de l'échelle obtenue est rarement interrogée (il s'agit en réalité d'une échelle artificielle, normalisée de moyenne 500 et d'écart-type 100) et moins encore n'est interrogée la signification d'un écart particulier sur cette échelle. Cela peut satisfaire les politiques et les media qui disposent ainsi d'un instrument simple pour commenter le complexe. Ainsi tout ce monde pourra s'entendre en évoquant, par exemple, une variation de 20 "points" sur l'échelle sans avoir aucune idée de la façon dont un "point" est défini. Les spécialistes ne sont pas dupes des limites du modèle et leurs interprétations sont en général beaucoup plus prudentes que celles des utilisateurs de seconde main. De fait, ils accompagnent le plus souvent leurs analyses quantitatives d'analyses qualitatives (analyses a priori et a posteriori des items, et de sous domaines de l'étude, utilisation de taxonomies pour distinguer les niveaux de complexité cognitive des items,...). Il n'en reste pas moins que c'est toujours le modèle psychométrique qui prédomine.

Une autre approche, que nous qualifions de didactique, peut être opposée à l'approche psychométrique. Elle consiste à prendre en compte la spécificité des connaissances à évaluer, à identifier les organisations possibles de ces connaissances et à chercher à placer les sujets évalués par rapport à ces organisations. Encore faut-il pour que l'on puisse parler d'évaluation que l'on ait l'intention d'attribuer des valeurs différenciées à ces organisations et que l'on puisse réellement le faire. Ensuite, la fiabilité de l'évaluation dépendra de la façon dont la validité et la fidélité seront assurées. Rappelons qu'une évaluation est dite valide si elle évalue bien ce qu'elle prétend évaluer ; elle est dite fidèle, si la répétition de l'évaluation n'est pas susceptible de modifier le jugement (du moins pas de façon notable). Une première condition pour assurer au minimum cette validité serait de pouvoir identifier des organisations de connaissances jouissant d'une certaine stabilité, c'est-à-dire qui ne dépende pas trop des groupes de sujets concernés ni des questions utilisées pour l'évaluation.

La recherche peut produire des données dont l'analyse peut justement suggérer de telles organisations et, sur ce point, l'analyse statistique implicative (ASI) a depuis longtemps fait ces preuves. Toutefois cette méthode à elle seule n'est pas en mesure de préciser les valeurs évoquées ci-dessus. Pour utiliser les concepts de la mesure, l'approche psychométrique assure la fidélité sans assurer la validité, tandis que l'approche didactique est susceptible d'assurer la validité, du moins pour des domaines de connaissance suffisamment réduits et bien structurés, mais cela sans grand souci pour la valeur proprement dite et sans pouvoir assurer la fidélité.

L'idée de chercher à associer les deux approches dans un même modèle n'est pas nouvelle, nous l'avons proposée dès les années 90 dans plusieurs articles (Bodin 1997) et nous l'avons souvent expérimentée pour l'analyse d'évaluations à grande échelle (Bodin ...). Dans la recherche en cours, nous cherchons à concrétiser cette idée en la mettant au service du développement d'un test (et plus généralement d'un système de "testing") échappant aux défauts des examens et des tests habituels.

Précisons encore que, en France, mais aussi dans nombre d'autres pays, l'évaluation courante des acquis des élèves aussi bien dans les classes que dans les examens (Brevet National des Collège ; Baccalauréats,...) procède d'un syncrétisme dans lequel on peut identifier des éléments susceptibles d'être rattachés, selon les

cas, aux conceptions psychométriques ou aux conceptions didactiques. Nous sommes là au niveau de l'habitus où les questions techniques de validité et de fidélité sont généralement ignorées. Il importe de préciser que notre recherche est centrée sur l'évaluation certificative, ce qui laisse totalement de côté la question de l'évaluation formative pour laquelle la problématique est totalement différente (Bodin...).
Ce rappel de questions générales concernant l'évaluation nous a semblé un préalable nécessaire à la présentation de notre propre problématique.

3. La problématique du socle commun et de son évaluation

Suite aux recommandations de la Commission Européenne (2005), la France s'est dotée d'un socle commun de connaissances et de compétences (2006) censé concerner tous les habitants de ce pays ; ce socle devant être acquis par tous à l'issue de la scolarité obligatoire. Plus précisément :
"*la scolarité obligatoire doit au moins garantir à chaque élève les moyens nécessaires à l'acquisition d'un socle commun constitué d'un ensemble de connaissances et de compétences qu'il est indispensable de maîtriser pour accomplir avec succès sa scolarité, poursuivre sa formation, construire son avenir personnel et professionnel et réussir sa vie en société*".
Le socle est organisé en sept piliers que nous ne décrirons pas ici. Des documents officiels détaillent les connaissances et les compétences de chaque pilier. Parmi ces piliers, seul le pilier 3 (*principaux éléments de mathématiques et la culture scientifique et technologique*) comporte un volet mathématique. Une originalité à signaler : les éléments du socle ne sont pas compensables. Par exemple une insuffisance dans la partie mathématique ne pourra pas être compensée par une excellence en langue vivante (pilier 2) :
"il ne peut ... y avoir de compensation entre les compétences requises qui composent un tout, à la manière des qualités de l'homme ou des droits et des devoirs du citoyen."
De la même façon, une compétence dans le domaine géométrique ne pourra pas compenser un manque de connaissance en statistiques ou en probabilité ; et réciproquement ! Il s'ensuit que les notes et les moyennes traditionnelles (du moins en France) se trouvent totalement disqualifiées pour l'évaluation du socle. Seule l'évaluation critériée est susceptible de faire l'affaire, mais outre qu'elle est largement étrangère aux habitudes des enseignants (et surtout des élèves et de leurs familles), elle est en contradiction avec le système d'évaluation qui continue à prévaloir dans les examens et même dans les conseils de classe. Pour le moment, les enseignants se trouvent aux prises avec des injonctions contradictoires avec le risque de devoir passer plus de temps à faire de l'évaluation certificative qu'à s'occuper de la formation de leurs élèves.
En ce qui concerne le volet mathématique du socle, l'évaluation critériée devrait porter sur l'ensemble des connaissances et des compétences spécifiées (lesquelles ne représentent qu'environ 1/12 de l'ensemble du socle commun). Il faudrait pour cela envisager une évaluation comportant plus de 150 questions et supposant un temps de passation de plusieurs heures (5 heures au minimum). Cela est d'autant moins réaliste pour l'évaluation certificative, qu'il faudrait alors construire un nouveau test pour chaque passation (comme cela

se fait pour les examens traditionnels avec tous les défauts qu'on leur connaît en matière de validité et de fidélité).

Il faut donc trouver autre chose. Des solutions partielles existent, telle le "testing" adaptatif utilisé aussi bien pour le TOEFL pour l'anglais, pour le SAT aux USA (SAT : Reasoning Achievement Test, qui a remplacé le "Scholastic Academic Test" qui était sans doute plus connu), ou encore le test de recrutement des développeurs de ...Microsoft ! Disons de suite que ce que nous avons vu de ces différents tests a pu nous donner des idées, mais que nous n'avons rien trouvé de totalement satisfaisant (même en faisant abstraction des contenus, que l'on pourrait toujours remplacer par un contenu adapté à nos besoins).

À notre avis, la problématique du socle a toute raison d'être prise au sérieux. Elle peut conduire à une rénovation en profondeur des enseignements dispensés dans le cadre de la scolarité obligatoire (et des pratiques d'évaluation), et en particulier, en ce qui concerne les mathématiques, contribuer à leur redonner du sens et à les rendre à nouveau "socialement désirables" (Chevallard, Y. 2007). Elle peut tout aussi bien se dissoudre dans l'insignifiance et la routine et générer des effets négatifs notables. Pour éviter cela, il faut que les enseignants puissent disposer d'outils efficaces, en particulier en ce qui concerne l'évaluation. C'est là une tâche qui ne peut laisser indifférent un institut de recherche sur l'enseignement des mathématiques, tâche dans laquelle notre équipe s'est investie depuis trois ans.

D'après les statistiques officielles, en France, c'est au moins 150 000 jeunes qui sortent chaque année du système éducatif sans posséder les connaissances et les compétences définies par le socle. De son côté, la Commission Européenne estime actuellement à 77 millions (!) le nombre d'européens qui n'ont pas le niveau requis pour le socle (Commission Européenne 2009).

Partant d'un travail effectué avec l'École de la deuxième chance de Marseille (jeunes de plus de 16 ans, bien représentatifs des 150 000 jeunes que nous venons d'évoquer), avec notre équipe de recherche de l'IREM d'Aix-Marseille, nous avons choisi de prendre la question du socle et de son évaluation par l'aval. La question devenant pour nous : quelles sont en fin de compte les connaissances et les compétences nécessaires à tous, indépendamment du cursus de formation que chacun pourra avoir suivi (ou non suivi !) ? Comment est-il possible de positionner des individus (encore dans le système scolaire ou sortis du système scolaire) par rapport aux demandes du socle ? Précisons en particulier que les écoles de la deuxième chance ainsi que d'autres structures de formation hors système scolaire obligatoire sont placées dans l'obligation légale de faire ce positionnement.

4. Connaissances, compétences et évaluation

Puisque, pour utiliser des termes à la fois courants et officiels nous parlons de connaissances et de compétences, précisons le sens que nous donnons à ces termes. Pour nous :

- *Avoir des connaissances*, signifie pour nous : connaître des faits, des définitions, des règles, des procédures. Il est entendu ici que « connaître » suppose compréhension et capacité à reconnaître ou

appliquer dans les cas ne demandant pas une mobilisation personnelle. Par exemple, connaître une procédure suppose de savoir la mettre en œuvre dans les cas triviaux.

- *Avoir des compétences*, signifie pour nous : avoir des connaissances ET être capable de mobiliser ces connaissances dans des situations qui ne les appellent pas directement. Par exemple, organiser son espace de vie, y prévoir la place du mobilier, met en jeu des connaissances de nature géométrique mais ne les appelle pas directement. Résoudre un « vrai » problème de mathématiques - c'est-à-dire un problème qui pose question sans préciser ni la démarche à suivre, ni les outils à utiliser, relève du niveau des compétences.

  On a souvent opposé les compétences dites transversales et les compétences disciplinaires. Pour le moment, ce n'est pas notre propos. Nous nous intéressons aux compétences de nature mathématique susceptibles d'opérer dans le cadre mathématique ou dans un cadre interdisciplinaire.

Cela dit, ces définitions (*a minima*) ne cherchent pas à régler la question, mais, simplement, à réduire, au sein de notre équipe et dans nos relations avec l'extérieur, le flou, pour ne pas dire la confusion générale qui les entoure. Elles se veulent opératoires en matière d'évaluation et c'est dans l'opérationnalisation que l'on verra si les différences suggérées sont, ou non, pertinentes.

En ce qui concerne les connaissances et les compétences attendues, nous sommes restés dans le cadre des instructions officielles, nous réservant le droit de faire, à l'issue de notre recherche, des observations sur d'éventuels excès ou manques.

Avec les définitions précisées ci-dessus, on pourrait penser que l'évaluation du socle commun de connaissances et de compétences pourrait se limiter à celle des compétences qu'il définit (mal). Cependant, outre la difficulté qu'il y a à évaluer directement les compétences, plusieurs raisons nous ont convaincus de la nécessité d'évaluer séparément les connaissances et les compétences :

- Cela correspond aux attentes de la société.
- Cela correspond aux attentes des jeunes adultes, qui, ayant été en échec à ce niveau, éprouvent le besoin d'être rassurés.
- Les personnes ayant quitté le système scolaire auront à faire face à des tests ou entretiens d'embauche, éventuellement à des tests d'entrée dans des formations complémentaires, qui porteront largement sur les connaissances. Cela évolue lentement mais on reste loin d'une évaluation qui serait centrée sur les compétences.
- En n'évaluant qu'au niveau des compétences on perd de vue les connaissances manquantes ou insuffisantes et il est alors difficile de certifier une maîtrise suffisante de l'ensemble du référentiel, ou encore de pouvoir proposer les compléments de formation nécessaires.

Ajoutons que cette distinction connaissances-compétences correspond à une observation que nous avons faite de façon récurrente sur de très nombreuses évaluations à grande échelle (EVAPM, PISA,..) :

- Dans un même domaine des mathématiques, la corrélation entre les niveaux de réussite observés sur les connaissances d'une part et sur les compétences d'autre part est assez faible : en tout cas, plus faible que ce que l'on pourrait penser a priori.

Avant d'aller plus loin nous devons insister sur le fait que ce que nous voulons évaluer, ce sont bien des connaissances et des compétences. Toutefois, pour cela, nous sommes bien obligés de passer par des questions supposées les opérationnaliser (c'est-à-dire les opérationnalisant plus ou moins bien !).

Une première difficulté provient du fait que, sauf - peut-être - dans les cas triviaux de simple rappel de connaissance, quel que soit le niveau de précision utilisé dans la description d'une connaissance ou d'une compétence, les questions utilisées pour l'opérationnaliser (pour l'évaluer) apparaîtront comme étant de difficultés différentes.

Ainsi, pour prendre un cas simpliste, supposons la connaissance « savoir additionner deux nombres entiers naturels », et soit 3 questions supposées évaluer cette connaissance (c'est-à-dire trois additions à effectuer). Rangeons ces trois questions par ordre de taux de réussite croissants : $a \leq b \leq c$.

L'idéal serait évidemment que l'on ait $a = b = c$, que ce soit les mêmes élèves qui réussissent (ou qui échouent) à a, b et c, et que, de plus, on puisse étendre cette égalité à un ensemble assez important d'autres additions. Non seulement cela n'arrive jamais, mais de plus en revenant au cas général, on observe facilement que :

1 - Étant donné deux questions a et b, susceptibles, a priori, d'opérationnaliser la même connaissance ou la même compétence, il est rare que a soit substituable à b, c'est-à-dire que le remplacement de a par b dans l'évaluation soit sans effet sur les résultats des élèves.

2 - Dans les mêmes conditions, dans le cas où a est moins bien réussie que b (on dira alors que a est plus difficile que b), il est exceptionnel que, parmi les élèves réussissant a, la proportion de ceux réussissant b soit supérieure à 0,8 (voir exemple plus loin).

3 - Toujours dans ces conditions, l'indice classique d'implication statistique n'est pas de grande utilité. En effet, il est loin de permettre de prédire avec une marge d'erreur acceptable que lorsque l'on réussit a, on réussit aussi b. Cet indice reste utile pour une première structuration d'un ensemble de questions en terme de tendance (ou de propension) mais on ne peut guère lui demander davantage.

4 - Ajoutons ici, que pour compliquer les choses, ces mêmes études (EVAPM, PISA,…) montrent que les corrélations entre les divers domaines des mathématiques (i.e. algèbre - géométrie) sont elles mêmes assez faibles.

Cette relative indépendance aussi bien des questions que des domaines et sous-domaines mathématiques rend problématique l'idée de vouloir passer du questionnement évaluatif à l'évaluation des connaissance et des compétences.

Nous prenons comme hypothèses de recherche les points suivants :

1- Il est possible d'organiser l'ensemble des connaissances et des compétences qu'il s'agit de contrôler (d'évaluer) (ensemble que nous appelons le référentiel) en un ensemble d'unités porteuses de sens. Par exemple :

- savoir faire à la main des opérations courantes (connaissance) ;
- savoir modéliser des situations simples conduisant à des additions, exécuter les opérations à la main ou à la machine et addition, et interpréter les résultats (compétence).

2 - Il est possible que le nombre de ces unités soit suffisamment réduit : une vingtaine au maximum, pour que chaque élève (ou sujet) puisse être positionné de façon lisible et communicable par rapport à ce référentiel réduit.

3 - Il est possible d'associer à chacune de ces unités un ensemble de questions l'opérationnalisant de la façon suivante :

- ces questions peuvent être de difficultés différentes, mais, dans cet ensemble, les questions de même niveau de difficulté sont interchangeables (dans un sens à définir).
- on est à peu près assurés que les sujets qui réussissent les questions (de cet ensemble) d'un certain niveau de difficulté réussiraient les questions de difficulté moindre du même ensemble.

Cela revient à définir des niveaux de maîtrise pour chacune des unités du référentiel réduit.

Nous avons employé plus haut les expressions « faire des opérations courantes », « modéliser des situations simples ». Il est clair que cela demande à être précisé. En fait, la démarche est dialectique et récursive : les questions d'évaluation du point 3 explicitent les unités du point 2, mais une question qui ne satisfait aux critères du point 3 est soit exclue du groupe de questions opérationnalisant l'unité, soit oblige à une nouvelle définition de l'unité.

Tout cela peut paraître très lourd et hors de portée, mais il faut rappeler ici que nous voulons faire un test adaptatif (voir plus loin) utilisable à grande échelle et sur une longue durée, dans des conditions de sécurité contrôlée (ce qui suppose une banque de questions très importante). Cela ne veut pas dire que le test sera intangible, bien au contraire : une fois mis en service sa révision devra être permanente pour assurer que les hypothèses ci dessus continuent à être vérifiées (et bien sûr pour tenir compte de l'évolution de la définition du socle commun). De plus nous voulons un test qui informe le candidat non seulement sur son niveau global, notion qui on le sait n'a aucun sens intrinsèque, mais aussi sur ses points forts et sur ses manques.

Pour la réalisation de ce projet, nous faisons encore une quatrième hypothèse :

4 - L'analyse statistique implicative (ASI) peut aider là structurer le référentiel et la banque de questions de façon à satisfaire les 3 hypothèses précédentes.

Avant de voir la façon dont nous comptons utiliser l'ASI, disons quelques mots sur la façon dont nous envisageons l'évaluation adaptative.

5. Organisation du test adaptatif

Étant donné un domaine à évaluer, appelons R (R = $\{r_1 , r_2 ,…., r_r \}$ le référentiel réduit dont il est question plus haut. Ce référentiel doit recouvrir le domaine à évaluer (ici le socle) sous ses différents aspects (connaissances, compétences, niveaux de complexité,...) et soit Q une banque de questions d'évaluation associée à R. Supposons encore qu'à chaque élément r_i du référentiel R soit associé un ensemble de questions destinées à l'opérationnaliser. Soit Q_i cet ensemble. $Q_i = \{ q_{i,n} \}$; n = 1 ; 2 ; 3 ;… ; n_i. $Q_i \subset Q$.

Un test adaptatif sur R consiste à faire passer à une personne p un sous ensemble Q' de questions de Q et à estimer à partir de ses réponses à Q' le score qu'il obtiendrait sur Q tout entier.

Plusieurs questions se posent alors :

A1 - Comment choisir la première question ($q_{1,p}$) posée à la personne p ?

A2 - La question $q_{n,p}$ d'ordre n ayant été posée à la personne p, comment choisir la question suivante : $q_{n+1,p}$, d'ordre n + 1 ?

A3 - Pour cette personne p, à quel ordre N (question $q_{N,p}$) faut-il arrêter la suite des questions ?

A4 - Quelle méthode utiliser pour estimer à partir de ses réponses au sous ensemble Q' le score θ qu'aurait la personne p sur l'ensemble de questions Q ?

Ici, $Q' = Q'_p = (q_{1,p} , q_{2,p} , q_{3,p}..... q_{N,p})$

A5 - Quelle information le test retourne t-il au candidat ?

Pour le point A1, on fixe *a priori* un niveau de départ θ_0 qui peut ou non être le même pour tous les candidats. Les autres questions sont résolues par l'utilisation conjointe de l'analyse des réponses aux items (IRT) et de l'analyse statistique implicative (ASI).

On trouvera en annexe un schéma illustrant le fonctionnement possible d'un test adaptatif.

6. Apports de l'analyse implicative

Il s'agit donc de structurer la banque Q et, corrélativement, de définir et de structurer le référentiel réduit R.

Nous l'avons dit plus haut, l'indice classique d'implication ne permet pas de faire ce travail avec une fiabilité suffisante (respect des 3 premières hypothèses).

Nous avons longtemps résisté à utiliser l'implication entropique dans la mesure où la structuration qu'elle permettait de proposer nous semblait trop pauvre. En effet, pour chacun des nombreux essais que nous avons faits, sur de nombreuses évaluations de types divers, les groupements et organisations que nous avons pu faire en nous appuyant sur cet indice étaient toujours très éclatés et semblaient peu utilisables.

Bien sûr, nous avons essayé d'autres pistes pour essayer d'atteindre le même but, en particulier les analyses de type factorielles, mais aucune d'elles ne s'est montrée capable de nous rapprocher d'une solution acceptable. Finalement, il a fallu nous rendre à l'évidence : l'implication entropique était difficile à utiliser simplement parce que la structure sous-jacente aux évaluations était pauvre. Pour l'évaluation, le fait de savoir que la manifestation d'un comportement a (réussite à une question de type a) rend le comportement b

plus probable que le comportement b est intéressant mais bien insuffisant ; ce que nous avons besoin de savoir, c'est si, dans ces conditions, le comportement b est, ou non, très probable.

Notons alors p(b/a) la probabilité de réalisation de b lorsque a est réalisé. Le niveau de probabilité acceptable reste arbitraire mais, a priori, ne devrait pas être inférieur à 0,8. Disant cela nous sommes conscient que, malgré tout ce qui a pu être dit par d'autres et par nous mêmes (!), nous sommes en quelque sorte restés attachés à une conception « inclusion » de l'implication statistique.

Reprenons donc le problème avec, pour l'illustrer, 3 questions extraite d'une étude EVAPM portant sur 2150 sujets : questions a, b et c. Ces questions sont choisies de façon à relever du même sous domaine (calcul numérique), du même type de tâches (exécution de procédures routinières) et de même registre (nombres positifs décimaux). Elles seraient donc de bonnes candidates pour figurer dans un dans un même groupe Q_i illustrant un même élément r_i du référentiel R. Pour être précis, on peut consulter ces questions en annexe, mais ce n'est pas ce qui importe ; l'expérience montre en effet que l'on obtient un résultat semblable avec la plupart des triplets de questions de même type. Voyons ce résultat.

D'abord les taux de réussite :

Question	Q_1	Q_2	Q_3
Taux de réussite	42 %	54%	63%

Ensuite les probabilités conditionnelles (fig 1):

a \ b	Q_1	Q_2	Q_3
Q_1		0,63	0,75
Q_2			0,70
Q_3			

p (b/a)

Puis les valeurs de l'indice d'implication classique (Poisson) (fig 2) :

a \ b	Q_1	Q_2	Q_3
Q_1		0,99	1
Q_2			0,99
Q_3			

$\varphi(a ; b)$

Et, finalement, celles de l'indice d'implication entropique (Poisson) (fig. 3):

a \ b	Q_1	Q_2	Q_3
Q_1		0,60	0,74
Q_2			0,54
Q_3			

$\varphi_e(a ; b)$

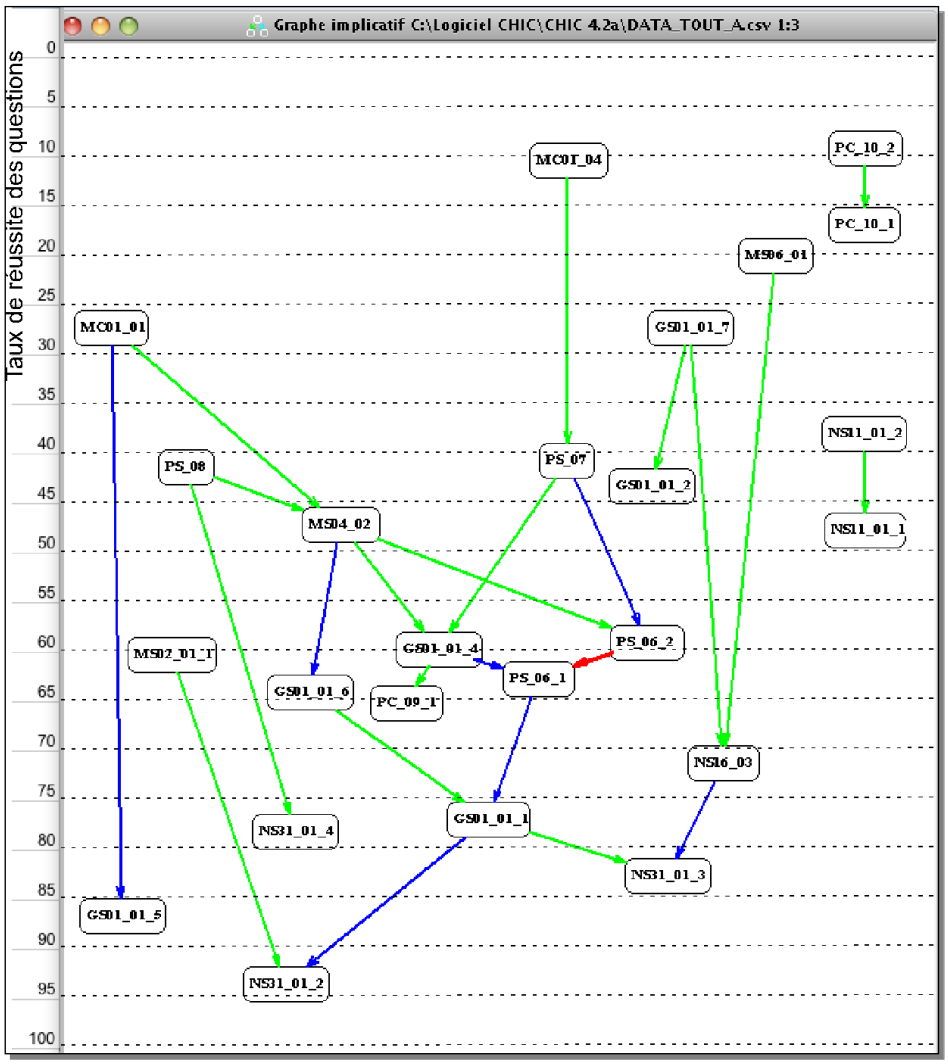
On le voit les valeurs de l'indice entropique ne sont pas très éloignées des probabilités conditionnelles.

Toutefois, contrairement à l'indice d'implication entropique, l'indice p(a/b) n'est sensible ni aux tailles respectives de a, de b, et de la population parente, ni à l'information apportée par la réalisation de b sachant que a est réalisée (information au sens de Shannon). L'utilisation des probabilités conditionnelles reste utile pour contrôler le sens de nos implication, mais l'indice d'implication entropique est maintenant l'instrument que nous privilégions pour tenter de structurer Q et R.

Nous appelons carte (ou graphe) des compétences le résultat de cette structuration de R ; nous noterons GR cette carte.

Plus précisément, dans un premier temps, l'analyse a priori du référentiel et des questions associées, ainsi que la pré-expérimentation de ces questions, nous permet de proposer un état GR_0 de cette structuration, qui maintenant comporte donc une structuration en unités, unités elles mêmes munies de niveaux, le tout devant vérifier les conditions des hypothèse 2 et 3.

Le graphe suivant présente une analyse implicative réalisée avec l'indice classique (Poisson). Le graphe est construit de façon à rendre compte des intensités implicatives d'une part et des taux de réussite d'autre part.



Graphe implicatif des questions (indice classique) - Fig. 4

Dans ce graphe, les premières lettres des étiquettes (N, M, P, G, I) désignent les sous-thèmes du référentiel

N - Nombres et Calcul ; M - Grandeurs et mesure ; P - Proportionnalité

G - Géométrie ; I - Incertitude (Statistiques et probabilités)

La seconde lettre, S ou C, désigne respectivement des connaissances ou des compétences (au sens défini plus haut). Le désordre qui apparaît est habituel ; il est bien connu des didacticiens comme de tous ceux qui s'intéressent à l'évaluation des connaissances. L'idéal d'une organisation en chaînes bien ordonnées qui conduirait à des échelles de Guttman est, d'une certaine façon, sous-entendue par l'édumétrie classique, mais n'est pratiquement jamais vérifié. L'évaluation courante, non critériée, par le jeu des notes et des moyennes, comme les évaluations à grande échelle lorsqu'elles s'appuient sur les modèles des réponses à l'item entretiennent cette illusion d'unidimensionnalité d'une façon dommageable pour la qualité de l'évaluation (faible validité, perte d'information,...) et donc pour la qualité des jugements et des décisions qui sont prises en son nom.

Plutôt que de fermer les yeux sur ce désordre, nous en prenons acte et cherchons à en tirer profit.

Le problème de l'évaluation n'étant pas de placer un candidat par rapport à un ensemble de tâches (les questions de l'évaluation), mais plutôt par rapport à un ensemble de types de tâches, il convient cependant de passer de la carte des questions à une carte des items du référentiel. C'est ce qui nous amène à vouloir organiser et structurer le référentiel R. Pour le moment, notre référentiel de départ comporte une centaine d'items, mais nous avons dit que nous voulions en déduire un référentiel réduit à une vingtaine d'items au maximum. Nous comptons faire ce travail dans l'esprit des remarques précédentes (essentiellement respect de notre hypothèse 3). Le référentiel de départ (100 pages) est consultable sur notre site et un référentiel réduit est en cours de construction. Ce qui doit conduire à un état 0 de ce référentiel et de ses graphes associés. Pour aller plus loin nous devons poursuivre l'expérimentation sur des populations suffisamment nombreuses et variées.

Lors de la passation d'un test, c'est en effet par rapport à ce référentiel qu'il conviendra de positionner les individus (et non simplement par rapport à des questions réussies ou non)..

Mais, nous l'avons dit, le processus de construction d'un test adaptatif ne peut qu'être évolutif. Dans un premier temps, nous définissons les états 0 de R et de Q (R_0 et Q_0) en utilisant les données dont nous disposons et les analyses a priori des éléments du référentiel non réduit et des questions de la banque Q.

Partant de ces états R_0 et Q_0 un algorithme (en cours de mise au point) utilisera les résultats des passations successives pour affiner les organisations de Q et de R, donnant ainsi R_t et Q_t à tout instant t.

Un problème ne manquera pas alors de se poser qui concerne la fidélité du test. L'évolution dans le temps de Q et de R fera que les scores d'un même individu mesurés à des temps t_1 et t_2 différents ne seront pas nécessairement égaux. Toutefois, le suivi du résultat du test dans le temps permettra de contrôler cette évolution.

Le passage des idées émises ici à une pratique fiable suppose une masse importante de données recueillies dans des conditions standardisées dans le cadre d'une expérimentation à grande échelle des questions de la banque Q, ce qui serait difficilement réalisable si ces questions n'étaient pas déjà proposées dans un test, lui même passé à grande échelle. Pour sortir de ce cercle vicieux, la solution que nous avons retenue consiste à partir d'un test T_0 , aussi fiable que possible compte tenu des connaissances dont nous disposons (données et analyses) et de laisser ce test s'améliorer au fur et à mesure des passations.

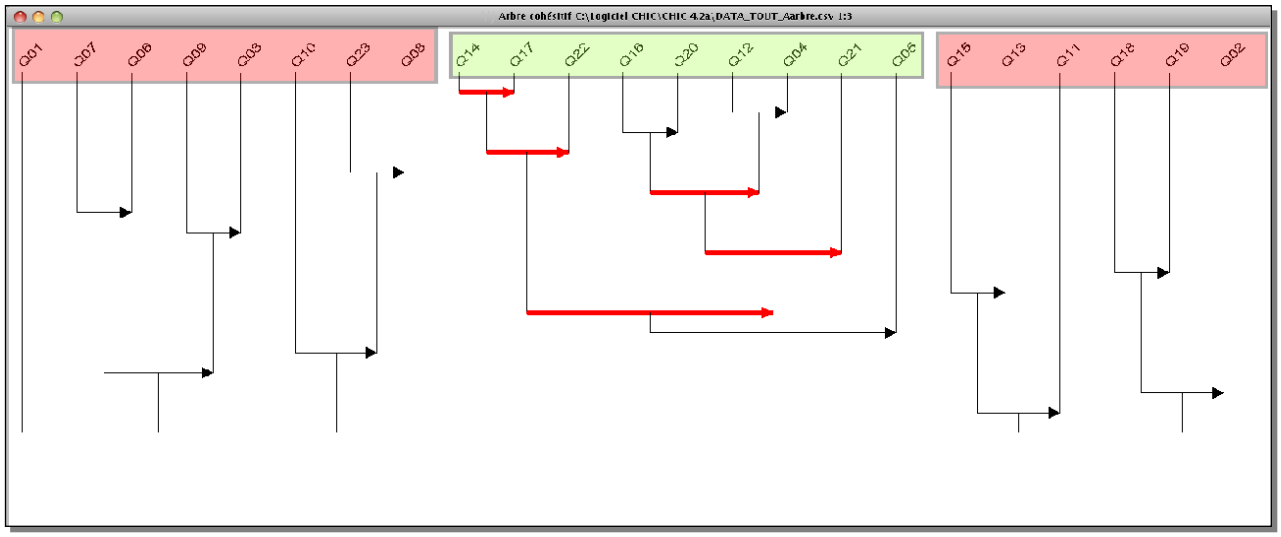
Au moment où nous écrivons ce texte, nous avons commencé à expérimenter deux tests de ce type avec des professeurs et des élèves des classes de collège (quatrième et troisième). Ces tests et les résultats obtenus sont consultables sur notre site. Il est prévu que ces expérimentations se poursuivront au cours de l'année 2010-2011.

7. L'analyse implicative cohésive comme aide à l'organisation des compétences

L'analyse implicative et cohésive (avec indice classique ou indice entropique) nous fournit un autre moyen de contrôler la pertinence des structurations proposées.

L'arbre suivant présente l'organisation implicative cohésive d'un ensemble de questions.

Les questions Q14, Q17 et Q22 par exemple, pourront peut-être être considérées comme opérationnalisant la même unité de compétence, éventuellement à des niveaux différents, et cela sous réserve de vérification de notre hypothèse 3. Il n'en sera pas de même pour les questions Q01, et Q10 (qui d'ailleurs, ici, n'ont pas été construites pour évaluer la même connaissance ou la même compétence).



Arbre implicatif cohésif - Fig 5

Nous l'avons déjà signalé, la notion de compétence garde, dans la littérature pédagogique un caractère flou important. Une question de recherche concerne la possibilité même de définir une compétence. Si l'on admet comme cela est généralement le cas, qu'une compétence est constituée de connaissances, de capacités et d'attitudes (et bien que les mots utilisés ici posent aussi des problèmes) susceptibles d'être mobilisés par le sujet pour la réalisation d'une tâche donnée, l'analyse implicative cohésive peut aider à identifier de telles compétences et à en rejeter d'autres supposées. Après analyse des contenus, des groupes tels que celui qui est mis sur fond vert pourront éventuellement être considérés comme relevant de la même compétence. Il n'en sera sans doute pas de même pour les groupes placés sur fond rouge.

Notons encore que l'analyse des courbes de réponses associées à l'IRT peut aider à affiner l'identification et la caractérisation des compétences.

Finalement, une évaluation, pour être fiable, devrait évaluer de façon séparée les divers sous-domaines (numérique, géométrique, grandeurs, statistiques,...) et de façon séparée les connaissances et les compétences de chacun des sous-domaines, et cela à différents niveaux de complexité cognitive. Cela mènerait rapidement

à une inflation évaluative peu acceptable. Le but de notre recherche est de montrer que l'on peut obtenir ce résultat d'une façon nettement plus économique en temps d'évaluation et sans rien sacrifier de la rigueur nécessaire.

Pour échapper au schéma traditionnel des examens français, où les épreuves sont toujours nouvelles mais jamais évaluées en matière de validité et de fidélité, nous sommes convaincus qu'il convient d'utiliser la technique des tests adaptatifs appuyée sur une banque informatisée de questions d'évaluation aussi riche et variée que possible.

Cette technique aujourd'hui largement utilisée dans le monde, permet, pour un thème donné, de disposer d'un test à la fois unique pour le prescripteur et toujours différent pour l'utilisateur, ce qui permet l'individualisation totale de l'évaluation. En effet, dans la mesure où la banque de questions est suffisamment riche, il n'y a aucune nécessité de chercher à garder les questions secrètes. Les personnes peuvent alors passer plusieurs fois le même test sans pour autant avoir à répondre aux mêmes questions ; elles peuvent aussi bien passer le test au même moment (sans pour autant avoir à répondre aux mêmes questions) qu'à des moments différents.

Ainsi, dans le cas de notre test en construction (mathématiques - socle commun), il est prévu que notre banque de questions d'évaluation comportera un millier de questions, tandis qu'un candidat particulier n'aura à répondre qu'à une trentaine de questions.

8. Utilisation de la théorie des réponses à l'item (IRT)

Pour le point A4 du paragraphe 5, l'IRT, dans sa fonction d'estimation probabiliste du score qu'un sujet obtiendrait sur un ensemble Q de questions en se contentant de lui poser un sous ensemble Q' de questions de Q, est de nature à résoudre la question. Nous ne développerons pas ici ce point qui est bien connu en édumétrie (Hambleton, 1985). Nous proposerons ultérieurement une annexe à ce document présentant la méthode.

L'utilisation conjointe de la démarche d'évaluation adaptative et de l'IRT permet de limiter la taille de Q' au strict minimum pour une estimation restant dans un intervalle de confiance d'amplitude prédéfinie. Cela résout partiellement le point A3 : on fixe a priori un intervalle de confiance et on arrête le test lorsque l'amplitude de l'intervalle de confiance du score estimé devient inférieure à l'intervalle de confiance choisi.

Reste que faisant ainsi, on demeure dans la conception psychométrique dont les insuffisances ont déjà été signalées. On fait alors comme si l'ensemble des connaissances et des compétences s'organisait en un continuum par rapport auquel on pourrait repérer aussi bien les sujets que les questions. Or la moindre étude didactique montre que cette unidimensionnalité globale ne correspond à aucune réalité.

Les autres apports de l'IRT concernent sa fonction de contrôle de la qualité des items (en particulier contrôle de divers biais : culturels, sexistes, liés à la formulation, ...) et sa fonction de contrôle de l'équivalence pour l'évaluation de plusieurs items *a priori* différents.

9. Arrêt du test

Dans le cas de l'utilisation stricte de l'IRT, la suite des questions posées à une personne p est arrêtée (point A4 du § 5) lorsque simultanément :

1 – l'amplitude de l'intervalle de confiance du score θ_n , estimé pour le sujet p, après la n° question du test au seuil de confiance prédéfini τ, devient inférieure à une valeur prédéfinie α. Plus précisément, lorsque :

$$\text{Prob}\ (\theta \in]\ \theta_n - \alpha\ ;\ \theta_n - \alpha\ [)\ > \tau$$

2 – Le recouvrement du graphe implicatif (GR) du référentiel d'évaluation atteint un niveau prédéfini.

10.Résultat du test

Après le passage du test, le sujet p se voit alors attribuer alors simultanément :

1 – Un score ….

Dans un premier temps, pour le calcul des scores, nous utiliserons une démarche classique et pragmatique en utilisant un indice de score qui prenne en compte tant le niveau de difficulté atteint dans la passation des questions, que des question non réussies en deçà de ce niveau. Ce score étant pondéré par les indice de difficulté des questions passées (réussies ou non). Précisons que, dans le cas du testing adaptatif, un candidat n'est pas confronté à des questions pour lesquelles sa probabilité de réussite est trop faible.

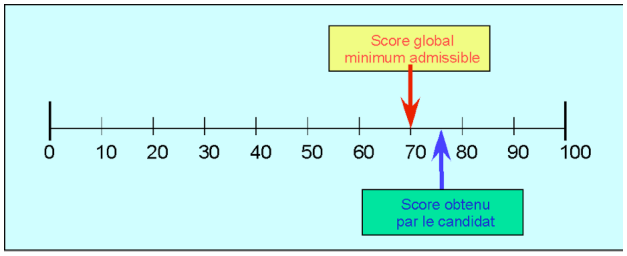
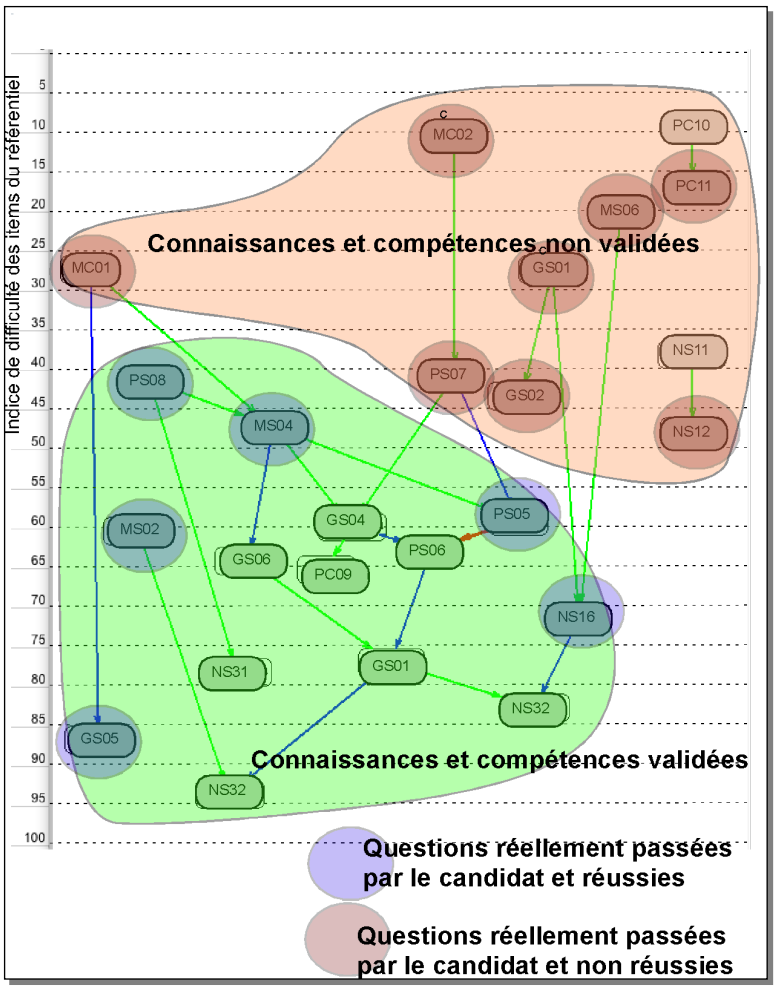


Figure 6

L'expérimentation sur des données issues d'études à grande échelle a fait la preuve de la fiabilité de l'algorithme envisagé. En particulier de sa compatibilité, tant avec la méthode classique de calcul des (nombre d'items réussis divisé par nombre d'items de l'épreuve, voir avec utilisation de pondérations), qu'avec les scores générés par l'IRT. Cette dernière compatibilité s'estimant en termes de corrélations, dans la mesure où les scores produits par l'IRT ne sont définis qu'à une transformation affine près.

D'autre part, cette démarche présente l'avantage d'être facilement compréhensible et acceptable par les candidats, ce qui n'est pas cas pour les scores issus de l'IRT.

2 - Une position (ou profil) estimée par rapport à la carte GR.



Positionnement d'un candidat (illustration fictive) - Fig. 7

3 - Le repérage du minimum à atteindre en terme de score et de position sur la carte GR, pour être admis (ici pour que le candidat puisse se voir valider le niveau socle).

Dans le cas présenté ci-dessus, le score global estimé du candidat (θ) est supérieur à un minimum fixé (encore à définir). Cependant l'analyse de sa position sur le graphe GR et donc de ses manques dans des sous-domaines particuliers pourra conduire à subordonner la certification à un complément de formation et d'entraînement sur certains points particuliers et à une nouvelle passation du test.

On aura ainsi construit un test qui sera à la fois un test critérié (test dont le résultat précise les points forts et les points faibles du sujet) et un test valide d'un point de vue édumétrique. De plus la passation de ce test ne demandera qu'un temps très réduit par rapport à qu'il faudrait investir pour une évaluation classique. La disposition d'une banque de taille importante (nous avons parlé d'un millier de questions) permettra de proposer des entraînements libres sans que cela soit de nature à nuire à la sécurité du test.

11. Questions techniques

Si l'on désigne par P l'ensemble des personnes ayant passé le test, les résultats obtenus lors diverses passations génèrent une matrice D sur laquelle opèrent tant l'IRT que l'ASI.

Le test T s'appuie donc sur :

- Le référentiel R de connaissances et de compétences à évaluer.
- La banque Q de questions d'évaluation censées opérationnaliser R (susceptibles d'évaluer les items de R).
- La population P des personnes ayant passé le test T.
- La base D des résultats des passations.

Toutefois ces différents composants doivent être actualisés au fur et à mesure des passations et des analyses et leurs articulations doivent être prévues pour que cette actualisation se fasse en temps réel. Pour garder cette contrainte à l'esprit, nous noterons T(t), R(t), Q(t), P(t) et D(t), les états de T, R, Q, T, D, à un instant t.

Dans la mesure où chaque personne ne passe qu'un nombre (très) réduit des questions de Q(t), la matrice D(t) comporte des trous. Cela n'empêche pas le calcul direct des indices d'implication et de cohésion de l'ASI, mais, pour le moment, le logiciel CHIC ne permet pas de traiter ce type de matrice, ni, en conséquence de produire les graphes souhaités. Cela peut se faire "manuellement", mais dans ce cas, il ne serait plus question d'actualisation en temps réel.

De plus la matrice D(t) comportera un millier de colonnes et, à terme, des milliers de lignes. CHIC n'est pas, pour l'instant capable de traiter des matrices de cette taille. Là aussi, on peut effectuer le travail par morceaux, mais cela complique fortement la tâche.

Nous comptons sur des améliorations de CHIC pour résoudre ces problèmes.

12. L'organisation de la recherche et du développement

Le schéma ci-contre précise l'organisation de la recherche.

La recherche est en cours, des questions théoriques comme des questions pratiques restent à résoudre, la banque Q est encore en construction (sous WIMS) et les questions disponibles sont en cours d'expérimentation. Un informaticien s'est joint à l'équipe pour nous aider à résoudre au cours de l'été les

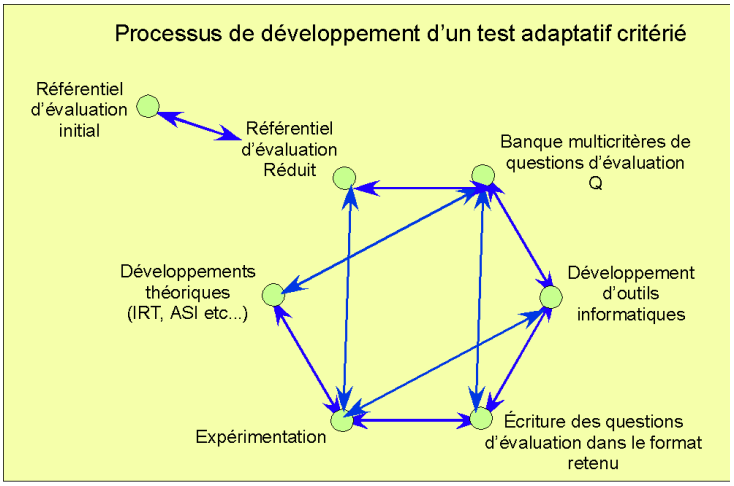


Fig 8

problèmes que posent encore la gestion des flux entre les différents éléments (R, Q, D) et les divers traitements issus de l'IRT et de l'ASI. Une première version du test adaptatif-critérié est prévu pour le mois d'octobre 2010 et pourra être présentée lors du colloque.

Pour le reste, notre équipe compte sur diverses collaborations dans le cadre de l'IREM d'Aix-Marseille et de l'Université de la Méditerranée (informaticien, didacticiens, statisticiens,..). Le chantier est vaste mais nous sommes convaincus de l'intérêt qu'il y aurait à le mener aussi loin que possible.

13. Conclusion

Le travail est en cours et je suis bien conscient que ce texte présente un chemin, mais montre en même temps que le but est loin d'être atteint. Nous espérons seulement avoir éclairé un chemin possible et contribué à le rendre souhaitable.

Enfin, je me dois de remercier les membres de l'équipe « socle et École de la deuxième chance » de l'IREM d'Aix-Marseille, dont les échanges m'ont stimulé et beaucoup apporté au cours de ces trois dernière années ». Je remercie aussi Régis Gras dont les encouragements ont été constants, ainsi que Christian Mauduit, directeur de l'IREM sans lequel ce travail n'aurait sans doute pas été amorcé. Mais je remercie aussi François Couturier et François Pétiard (Université de Franche Comté) qui ont accompagné ma réflexion et mon travail au cours des … 25 dernières années et grâce auxquels la banque EVAPMIB a pu être développée avec une qualité graphique exceptionnelle (LaTex) qui, en particulier, profite aux travaux actuels, ainsi que les membres successifs des équipes EVAPM (voir les sites internet correspondants).

14. Références

Bodin, A. (1997). Modèles sous jacents à l'analyse implicative et outils complémentaires. In A. Bodin, R. Gras, J. B. Lagrange : *implication statistique*, prépublication 97-32 - IRMAR de RENNES..

Bodin, A. (1997). L'évaluation du savoir mathématique - Questions et méthodes. *Recherches en Didactique des Mathématiques,* Éditions La Pensée Sauvage, Grenoble.

Bodin, A. (2002). Classification des questions d'évaluation et cadre de référence des études PISA

Bodin, A. (2006). Ce qui est vraiment évalué par PISA en mathématiques. Ce qui ne l'est pas. Un point de vue Français. Bulletin de l'APMEP Num. 463. pp. 240-265.

Bodin, A. (2006). Les mathématiques face aux évaluations nationales et internationales. De la première étude menée en 1960 aux études TIMSS et PISA … en passant par les études de la DEP et d'EVAPM. Communication séminaire de l'EHESS. Repères IREM, N°65, octobre 2006.

Bodin, A. (2007). Dissonances et convergences évaluatives - De l'évaluation dans la classe aux évaluations internationales : quelle cohérence ? Bulletin de l'APMEP N° 474 pp 47-79.

Bodin, A. (2007). What does PISA really assess? What it doesn't? A French view." In S. T. Hopmann (ed) : *PISA zufolge PISA / PISA According to PISA* - Wien

Bodin, A. (2008). Lecture et utilisation de PISA pour les enseignants. Petit x ; n° 78, pp. 53-78, IREM de Grenoble.

Bodin, A. (2009). L'étude PISA pour les mathématiques. Résultats français et réactions. *Gazette des mathématiciens* N°120 (Société Mathématique de France).

Chevallard, Y. (2007). Les mathématiques à l'école : pour une révolution épistémologique et didactique - Bulletin de l'APMEP. Num. 471. p. 439-461.

Chevallard.Y, Feldmann. S, (1986). Pour une analyse didactique de l'évaluation - IREM d'AIX MARSEILLE

Commission des communautés européennes (2009) : communication de la commission au parlement européen, au conseil, au comité économique et social européen et au comité des régions : les compétences clés dans un monde en mutation.

EURYDICE (2002). Compétences clés ; Un concept en développement dans l'enseignement général obligatoire. Commission Européenne.

Gras, R., Régnier J-C., Guillet, F. (2009). RNTI E16 Analyse statistique implicative – Cépadues ed. Toulouse

Gras, R. (1992). L'analyse des données: une méthodologie de traitement de questions de didactique, *Recherches en Didactique des Mathématiques*, Vol. 12-1.

Gras, R.(ed.) (2008). Statistical Implicative Analysis, Theory and Applications", Springer

Hambleton, R.K. ; Swaminathan, H. (1985). Item Response Theory - Principles and Applications - Kluwer Nijhoff Publishing.

Journal officiel de l'Union Européenne (2006). Recommandation du parlement européen et du conseil du 18 décembre 2006 sur les compétences clés pour l'éducation et la formation tout au long de la vie.

Journal Officiel de la république française. Décret n° 2006-830 du 11 juillet 2006 relatif au socle commun de connaissances et de compétences et modifiant le code de l'éducation.

Les documents produits par l'équipe de l'IREM d'Aix-Marseille, et plusieurs autres sont téléchargeables depuis la page "groupe socle & E2C" de l'IREM d'Aix-Marseille :

http://www.irem.univ-mrs.fr/spip.php?article177

À cette adresse on trouvera en particulier les documents suivants :

- Une communication faite lors du colloque Inter-Irem "Les mathématiciens et l'enseignement de leur discipline en France" - mars 2010 CIRM (Marseille Luminy)
- Une communication faite lors du Colloque international "De la culture commune au socle commun : enjeux, tensions, réinterprétations, déplacements". INRP novembre 2009.
- Grilles socle : présentation des textes réglementaires adaptée à l'évaluation.
- Référentiel socle <date mise à jour> : Notre référentiel d'évaluation et questions d'évaluation, commenté.
- Socle commun et évaluation <date mise à jour> : présentation de la problématique.

  Ce document regroupe plusieurs textes :

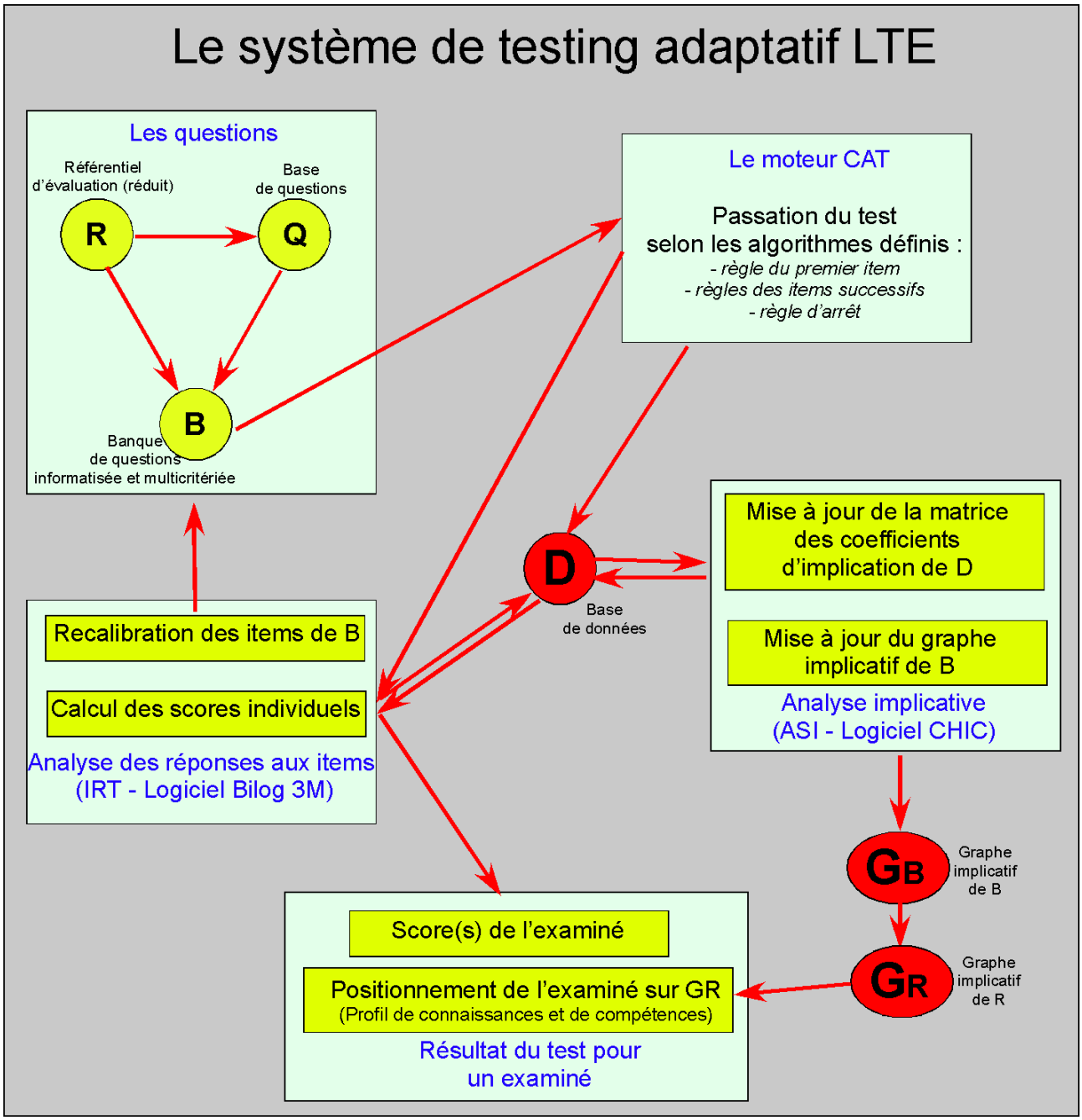
- Document complémentaire aux grilles Excel préparées pour l'évaluation du socle (cf. classeur Grilles socle) ; grilles issues du décret de 1996 et de l'opérationnalisation faite par la DEGESCO et l'IGEN.
- Évaluation certificative des acquis du socle commun de connaissances et de compétences.
- Le socle commun de connaissances et de compétences et mathématiques vu dans une perspective non scolaire (texte publié dans les cahiers pédagogiques - dossier numérique mars 2010).
- La question de la complexité des situations - Rapports entre complexité et compétences - Le cas des mathématiques - Quelques notes rapides sur la question.
- Le socle après la scolarité obligatoire (texte écrit pour l'école de la deuxième chance de Marseille).
- Les TIC en mathématiques et leur évaluation dans le socle commun de connaissances et de compétences (texte publié dans Mathematice, mars 2010).

❖ Épreuves expérimentées et grilles de recueil des résultats :
❖ Expérimentation - Épreuve A
❖ Expérimentation - Épreuve B
❖ Classification PISA 2003 : Cadre de référence commenté.
❖ Taxonomie de la complexité cognitive : Version remaniée de la taxonomie de Gras, R. révision A. Bodin 2010.

On trouvera de plus de nombreux documents concernant l'évaluation, le socle et les études internationales sur le site de l'auteur :

http://web.mac.com/antoinebodin/pro

Annexe 1 : illustration d'un fonctionnement possible d'un test adaptatif



Annexe 2 : présentation des 3 questions utilisées au §6.

Dans la division de 7 956 par 48 :				
a	Le quotient entier est 16 et le reste 276.	V	F	Jnsp
b	Le quotient entier est 1 657 et le reste 24.	V	F	Jnsp
c	Le quotient entier est 165 et le reste 36.	V	F	Jnsp
d	Le quotient entier est 36 et le reste 165.	V	F	Jnsp

Question Q1

Vrai ou Faux ?				
a	$3{,}7 = \frac{37}{10}$	V	F	Jnsp
b	$3{,}7 = \frac{0{,}37}{10}$	V	F	Jnsp
c	$0{,}03 = \frac{3}{7}$	V	F	Jnsp
d	$0{,}03 = \frac{3}{100}$	V	F	Jnsp

Question Q2

Vrai ou Faux ?				
a	$103{,}5 < 110{,}51$	V	F	Jnsp
b	$17{,}23 < 13{,}8$	V	F	Jnsp
c	$16{,}18 < 16{,}108$	V	F	Jnsp
d	$0{,}029 < 0{,}002\ 9$	V	F	Jnsp

Question Q3