Les apports de la recherche en psychologie aux tests d'intelligence : quelle conséquence pour la pratique?

How psychological research contributes to the tests of intelligence. What are the consequences for practical experience?

Philippe Chartier

Electronic version

URL: <https://journals.openedition.org/osp/4937>

DOI: 10.4000/osp.4937

ISSN: 2104-3795

Publisher

Institut national d'étude du travail et d'orientation professionnelle (INETOP)

Electronic reference

Philippe Chartier, "Les apports de la recherche en psychologie aux tests d'intelligence : quelle conséquence pour la pratique?", *L'orientation scolaire et professionnelle* [Online], 30/4 | 2001, Online since 15 December 2004, connection on 08 March 2023. URL: <http://journals.openedition.org/osp/4937> ; DOI: <https://doi.org/10.4000/osp.4937>

This text was automatically generated on 8 March 2023.

All rights reserved

Les apports de la recherche en psychologie aux tests d'intelligence : quelle conséquence pour la pratique?

How psychological research contributes to the tests of intelligence. What are the consequences for practical experience?

Philippe Chartier

L'objectif de cet article est de présenter quelques applications concrètes de recherches en psychologie en termes d'outils disponibles et/ou de procédures applicables dans le domaine des tests d'intelligence. Nous nous limiterons ici à l'approche « classique » de l'intelligence (une intelligence abstraite, académique...), sans aborder les autres aspects tels que l'intelligence sociale (une forme d'intelligence orientée vers les personnes) ou la créativité ¹.

La grande majorité des tests d'intelligence utilisés aujourd'hui repose toujours sur le même principe, qui date du début de la psychométrie (début du 20^e siècle) : différencier les sujets sur leur niveau de performance. Bien que les principales épreuves utilisées soient régulièrement révisées, le principe de ces épreuves date donc de près d'un siècle (Lautrey, 2001). L'une des principales critiques adressées à cette approche, c'est d'être peu informative sur les conditions de réalisation de cette performance.

Avec les nombreux travaux menés depuis les années 1970, en particulier ceux relevant de la psychologie cognitive, on pouvait s'attendre, d'une part, à un nouvel éclairage des tests classiques et, d'autre part, à un renouvellement en profondeur des épreuves utilisées avec l'apparition de tests originaux. Mais le bilan des applications reste limité • malgré les espoirs exprimés dans les années 80 pour mieux intégrer les apports de la psychologie cognitive dans l'élaboration de nouvelles épreuves d'intelligence (voir par exemple : Huteau & Lautrey, 1978) le constat, près de 20 ans plus tard, de l'un des

auteurs est assez mitigé : « *il n'y a pas eu de renouvellement profond des tests* » (Huteau, 1995, p. 405).

Des synthèses de recherches dans ce domaine sont régulièrement publiées (Huteau & Lautrey, 1978; Lautrey & Huteau, 1990; Huteau, 1990; Huteau, 1995 ; Vrignaud 1996 ; Huteau, 2001...) mais ces analyses théoriques présentent le plus souvent l'état des recherches (éventuellement accompagnées de courtes descriptions des procédures expérimentales utilisées) et beaucoup plus rarement des procédures concrètes d'évaluation réellement utilisables.

De plus, dans ces revues de questions, les conclusions des auteurs sont souvent teintées de pessimisme quant au passage des résultats de la recherche vers la pratique.

À défaut d'un *renouvellement profond*, des liens existent pourtant entre recherche et pratique : des procédures récentes d'analyse de tests anciens, des rénovations d'outils et même des créations de nouvelles épreuves, témoignent de l'application réelle de certains résultats de recherche.

Nous aborderons, dans une première partie, le cadre théorique général avant de présenter, dans une seconde partie, quelques exemples d'applications et de procédures d'évaluation effectivement utilisables aujourd'hui.

Les apports de la recherche

Nous ne présenterons ici que trois axes de recherche : les apports de la psychologie cognitive, le courant de l'évaluation dynamique et les apports méthodologiques (nous avons retenu ce découpage pour faciliter la présentation de notre analyse, mais il n'y a pas réellement, selon nous, de délimitation aussi précise entre ces trois domaines).

Les apports de la psychologie cognitive à la compréhension des tests

À partir des années 60, le courant de la psychologie cognitive, centré sur la compréhension et l'explication des activités mentales, va souvent utiliser des situations de type résolution de problème (proches de situations réelles de tests) pour tenter de modéliser le fonctionnement cognitif.

L'objectif est de différencier les sujets non plus sur leur niveau global de performance mais sur des différences de traitement de l'information (en particulier sur la vitesse de traitement, sur le niveau de précision, sur le choix et/ou le contrôle des processus cognitifs...). Si l'on reprend la distinction proposée par Huteau & Lautrey (1999a, chapitre 8) nous pouvons distinguer trois approches théoriques principales : l'approche corrélationnelle, l'approche composantielle et la modélisation de la dynamique de traitement. Les auteurs, qui présentent les caractéristiques et les limites ² de ces approches, se montrent sceptiques, dans leur conclusion, sur les applications réelles des résultats de ces recherches : « *Le courant qui vient d'être passé en revue n'a donc pas donné naissance à des instruments d'évaluation susceptibles de remplacer les tests d'intelligence classiques* » (Huteau & Lautrey, 1999a, p. 247).

Bien que partageant les interrogations des auteurs, des applications possibles de ces recherches nous paraissent possibles, principalement sur la notion de diagnostic

cognitif (voir sur ce point Richard, 1996) avec, par exemple, l'analyse cognitive de la tâche et l'analyse des stratégies de résolution.

En effet, l'analyse de l'activité cognitive mobilisée dans la résolution des items d'un test d'intelligence va aider à mieux comprendre ce qui est évalué dans chaque épreuve et peut permettre, à l'intérieur d'une même épreuve, d'effectuer une analyse plus « clinique » de la performance d'un sujet en fonction de ses réussites et de ses échecs. Cette analyse des différences de pattern individuels permet alors de compléter et d'enrichir l'information classique, en termes de score global, par des informations plus précises, par exemple, en termes de scores partiels. Nous présenterons en deux exemples de cette démarche réalisée à partir de tests de facteur g: le D70 et le N.N.A.T.

Un autre axe de recherche s'intéresse à l'analyse des différences individuelles dans les procédures de résolution. En effet dans certaines situations, et ici plus précisément dans certains tests, une même tâche peut être résolue de différentes manières et il est alors pertinent de pouvoir distinguer les sujets, non plus seulement sur leur score total au test, mais aussi sur la (ou les) procédure(s) de résolution utilisée(s). Ainsi, de nombreux travaux relevant de la psychologie différentielle se sont orientés, depuis plusieurs années, vers l'analyse des différences inter et intra individuelles dans les stratégies de résolution de problème et illustrent ainsi les propositions théoriques de Reuchlin concernant la vicariance des processus cognitifs (Reuchlin, 1978). Nous illustrerons ces aspects par la présentation des stratégies de résolution dans une épreuve de type cubes de Kohs et d'une procédure informatisée permettant ce type d'analyse.

L'évaluation dynamique

Un autre développement des recherches en psychologie, et plus précisément ici en psychométrie, concerne le déplacement de l'objet même de l'évaluation : de l'observation de la performance du sujet vers l'analyse de son « éducabilité ». En effet, un courant « critique » de la psychométrie classique propose une évaluation dynamique des aptitudes et la mesure d'un potentiel d'apprentissage. Ce potentiel représente « *l'habileté à apprendre et à tirer profit d'une expérience adéquate* » (Ionescu & Jourdan-Ionescu, 1985, p. 920). La principale caractéristique de cette approche est de placer le sujet dans une situation réelle d'apprentissage (un entraînement et/ou des aides sont fournis au sujet pour réaliser les tâches) afin de mieux prendre en compte ses potentialités. Pour ce courant, l'évaluation classique, qualifiée de statique, présenterait un certain nombre de biais, en particulier lorsque le sujet est peu familiarisé avec ce type de situation d'évaluation et/ou qu'il n'appartient pas au même milieu culturel et/ou social que le concepteur de l'épreuve (Loarer, 2001). Dans ces situations, l'évaluation classique risque de sous-estimer les potentialités réelles du sujet.

Sans vouloir faire une présentation exhaustive de cette approche (voir sur ce point Huteau & Lautrey, 1999a, chap. IX) nous en rappellerons les principaux éléments.

L'origine de ce courant de recherche est bien antérieure au développement des recherches en psychologie cognitive car dès les années 30 deux psychologues, Rey en Europe et Vygotsky en U.R.S.S., vont proposer de nouvelles formes d'évaluation.

Rey propose de modifier la situation classique d'évaluation afin de pouvoir observer l'éducabilité du sujet, au lieu de procéder au simple constat de son niveau de performance (Rey, 1934). Dans son article Rey, nous propose à la fois une

problématique, celle de l'éducabilité, et un outil d'évaluation. Dans cette épreuve, qu'il nomme « labyrinthe manuel ³ », le sujet doit soulever des plateaux à l'aide de prises qu'il doit mémoriser. Cet apprentissage est soigneusement observé par le psychologue qui relève différents indicateurs : les erreurs, le temps, le nombre de répétitions nécessaires pour arriver au critère de réussite (fixé à 3 réalisations successives sans erreurs). Il s'agit donc bien ici de faire porter la mesure sur des critères liés à l'apprentissage de la tâche. Mais, Rey n'envisageait pas la possibilité de fournir des aides au sujet.

C'est justement cette dimension que va apporter Vygotsky. En effet, dans les mêmes années, Vygotsky propose de distinguer deux niveaux de développement chez un même enfant : son niveau de développement actuel (que l'on peut évaluer par des tests classiques) et son niveau de développement potentiel (qui correspond à ce qu'il est capable de réaliser avec l'aide d'un adulte). Il formalise la différence entre ces deux niveaux de développement en proposant la notion de « Zone Proximale de Développement » (Z.P.D.). Il illustre l'intérêt de cette notion en prenant l'exemple de deux enfants qui auraient le même âge mental de 7 ans, donc un niveau de développement identique d'après leurs résultats à un test classique, mais ces deux enfants se distingueraient sur leur capacité à tirer profit des aides proposées : *« l'un des deux résout avec un minimum d'aide des tâches pour des enfants de neuf ans, alors que l'autre ne réussit que des tâches adaptées à des enfants de sept ans et demi »* (Vygotsky, in Schneuwly & Bronckart, 1985, p. 108). Ces deux enfants, de niveau équivalent, se différencient donc sur leur capacité de progression, sur l'étendue de leur Z.P.D. ⁴. Et ce type d'information présente un caractère prédictif car : *« Ce que l'enfant est en mesure de faire aujourd'hui à l'aide des adultes, il pourra l'accomplir seul demain »* (Vygotsky, in Bronckart & Schneuwly, 1985, p. 109). C'est bien cette capacité de progression que l'on cherche à évaluer par la procédure d'évaluation dynamique.

Mais ces approches novatrices ne vont bouleverser ni les outils ni les pratiques des psychologues et il faudra attendre plusieurs années pour que ces premiers écrits soient (re)découverts dans les années 80 et suscitent de nombreuses recherches tant en Europe qu'en Amérique du Nord (Loarer & Chartier, 1996).

L'une des spécificités de l'évaluation dynamique est bien de donner une autre position au psychologue : celui-ci va délaisser (en partie) les règles classiques de standardisation de la passation pour intervenir et apporter une aide au sujet. Il peut ainsi, par exemple, dans une phase d'entraînement (phase d'apprentissage), expliquer les différentes règles de résolution des items ou les difficultés spécifiques de l'épreuve ou, au cours même de la passation, apporter des aides en cas d'échec. On distingue ainsi deux procédures d'évaluation dynamique : test-aide-retest ou aide au cours du test. Quelle que soit la procédure utilisée, il s'agit bien ici d'évaluer le Potentiel d'Apprentissage (P.A.) du sujet en le plaçant réellement dans une situation d'apprentissage et en observant dans quelle mesure il est capable de tirer profit des aides proposées. Cette démarche, qui garantit une meilleure familiarisation du sujet avec la tâche, permet également de renforcer sa motivation (et augmente ainsi la validité de la mesure).

Cette approche est qualifiée de « dynamique » car elle se centre sur l'évolution du sujet, sur sa modifiabilité, sur ses capacités de progression (en cours d'épreuve ou entre deux passations) et non plus seulement sur son niveau de performance. En d'autres termes, il ne s'agit plus ici de différencier les sujets uniquement sur leur niveau de performance mais également sur leur capacité d'apprentissage.

Ce type d'évaluation, qui serait plus valide (au moins pour un public spécifique : public de bas niveau de qualification, public non francophone...), suscite encore de nombreuses interrogations, tant sur la validité de la mesure que sur sa fidélité (voir sur ces points : Loarer & Chartier, 1996 ainsi que Lautrey, 1994).

Faut-il alors se résigner à ne pas pouvoir utiliser de telles procédures ? Même si, effectivement, les qualités psychométriques des épreuves d'évaluation dynamique ne sont pas satisfaisantes une utilisation de ces procédures — comme compléments à une évaluation classique — nous semble possible. Dans cette optique, nous en présenterons deux exemples, reposant tous les deux sur une épreuve de type cubes de Kohs.

Les apports méthodologiques

Dans les rénovations d'épreuves, les auteurs peuvent être amenés à modifier celles-ci : intégration de nouveaux items ou de nouveaux subtests, modifications de certains items, proposition de nouvelles procédures d'analyse des résultats... Ces modifications doivent se justifier par des éléments théoriques et/ou méthodologiques. Nous illustrerons ces aspects à partir de l'analyse des modifications apparues dans la troisième version de l'échelle de Wechsler pour enfant : le W.I.S.C. III. En effet, nous constatons dans cette version française de 1996 l'apparition, d'une part, d'une nouvelle épreuve (symboles) et, d'autre part, de nouveaux indicateurs (les indices factoriels) qui viennent compléter les indicateurs classiques en termes de Q.I. Ces modifications reposent en fait sur l'apport de techniques d'analyse de données et en particulier sur les techniques d'analyses factorielles.

L'analyse factorielle (ou plutôt les analyses factorielles ⁵) est l'une des principales méthodes statistiques utilisées dans la validation des tests d'intelligence. Elle permet en particulier, dans le cas où une épreuve est constituée de plusieurs échelles, comme dans les épreuves de Wechsler, de valider le rattachement d'un subtest à l'une de ces échelles. En effet, l'addition de scores doit toujours être justifiée pour prendre un sens. Ces éléments de validation figurent dans les manuels des tests et il est souvent très intéressant d'analyser les données proposées afin de mieux comprendre les utilisations et les limites des scores et sous-scores que le praticien est amené à calculer. Nous présenterons en 2.4 une illustration de ces aspects à partir de l'analyse des scores factoriels que le praticien peut calculer dans le W.I.S.C. III.

Quelques applications concrètes des recherches

Après avoir rapidement présenté les axes théoriques dans notre première partie, nous allons maintenant les illustrer en abordant quatre types d'application possible : l'analyse de la tâche (analyse cognitive du test), l'analyse des stratégies (analyse qualitative de la démarche des sujets), l'analyse des capacités d'apprentissage (l'évaluation dynamique) et, pour terminer, l'analyse des apports méthodologiques dans la création ou la rénovation d'épreuves.

L'analyse de la tâche

Nous présenterons ici deux exemples d'épreuves de facteur g, D70 et N.N.A.T., dont l'analyse de la résolution des items permet de fournir d'autres indicateurs que le simple constat du score total.

Le D70

Cette version D70 est la version révisée en 1970 du test original de 1948: le D48. Dans cette épreuve, typique des tests de facteur g, le sujet doit trouver la règle de progression et l'appliquer afin de déterminer les caractéristiques du domino manquant. Ce test collectif, de format classique papier/crayon, est composé de 44 items. Ces items, directement inspirés des items du D48, peuvent se différencier sur leur logique de résolution et le manuel nous indique que, par rapport au D48: « *les lois qui président aux items du D70 ont été variées de façon plus systématique* » (Manuel du D70, 1970, p. 3). Ainsi, dans l'élaboration du D70: « *La première étape a consisté en une exploration systématique des lois qui présidaient à chacun des 44 items du Test D48* » (Manuel du D70, 1970, p. 4). Mais à aucun moment, ni dans le manuel du D48, ni dans celui du D70, ces lois ne sont définies.

Dickes et Martin ont procédé à l'analyse de ces lois de résolution, en se limitant aux items impairs du test, et proposent la distinction de quatre types d'items (Dickes & Martin, 1998) :

- les items spatiaux : qui nécessitent une résolution spatiale (par symétrie, répétition, inversion...),
- les items numériques : il s'agit ici d'appliquer des règles d'incrémentation (par exemple : ajouter 2, retrancher 1...),
- les items mixtes : dans ce cas la résolution d'une des faces est effectuée par une règle spatiale, tandis que l'autre face nécessite l'application d'une règle numérique,
- les items arithmétiques : la solution est trouvée ici par l'application d'une règle arithmétique simple (de type $a + b = c$) entre 3 faces de dominos

Les auteurs nous proposent ainsi une catégorisation des 22 items impairs de l'épreuve et observent alors, d'une part, la forte proportion d'items spatiaux et, d'autre part, des différences dans le niveau de difficulté (les items spatiaux étant les plus faciles, les arithmétiques les plus difficiles).

Il ressort de leur analyse que le test D70 est bien un test de facteur g en raison du caractère composite du score global et du type de raisonnement nécessaire à la résolution des items, principalement dans le cas des items numériques et arithmétiques.

Afin de compléter ces travaux, nous avons réalisé une analyse sur l'ensemble du test ⁶ dont nous présentons les résultats principaux. Nos résultats confirment la prédominance des items spatiaux dans l'épreuve : ils représentent près de la moitié des items du test. Nous pouvons aussi noter la répartition inégale des items selon les autres règles de résolution : il n'y a que 3 items arithmétiques pour 9 mixtes et 12 numériques (voir *tableau 1*).

TABLEAU 1. Répartition des items du D70 (44 items au total)

	ITEMS SPATIAUX	ITEMS NUMÉRIQUES	ITEMS MIXTES	ITEMS ARITHMÉTIQUES
Effectif	20	12	9	3
Pourcentage	45,5	27,3	20,4	6,8

TABLE 1. **DISTRIBUTION OF ITEMS**

Nous retrouvons également des différences en termes de niveau de difficulté avec une réussite plus importante pour les items spatiaux et une difficulté plus importante pour les items arithmétiques (voir *tableau 2*).

TABLEAU 2. Pourcentages de réussite de chaque catégorie d'items sur une population d'élèves de 2' (N = 195)

	ITEMS SPATIAUX	ITEMS MIXTES	ITEMS . NUMÉRIQUES	ITEMS ARITHMÉTIQUES
% de réussite	82	80	52	22

TABLE 2. **PERCENTAGE OF ACHIEVEMENT FOR EACH CATEGORY OF ITEMS AMONG A POPULATION OF FIFTH FORM PUPILS**

Cette catégorisation théorique ⁷ des items, que nous indiquons en annexe (*tableau 4*, annexe 1), permet alors de calculer des sous-scores (un score spatial, un score numérique...) et de repérer ainsi les forces et faiblesses de chaque sujet. Ces indications viennent compléter l'indication classique du score total. Mais, compte tenu du nombre limité d'items dans certaines catégories (en particulier pour les items arithmétiques), ces sous-scores sont à interpréter avec prudence.

On pouvait s'attendre, dans la rénovation récente de ce test (forme D2000), à une application concrète des résultats exposés dans l'article de Dickes et Martin. Les auteurs du manuel nous indiquent bien qu'une catégorisation des items a été effectuée lors de l'élaboration de l'épreuve. Ainsi la forme expérimentale, qui comporte 44 items, se compose de « 21 items numériques, 9 spatiaux, 8 mixtes et 6 numériques ⁸ » (manuel D2000, p. 7) mais les auteurs ne nous indiquent ni le détail de cette catégorisation pour chaque item, ni les utilisations possibles de cette catégorisation en termes d'analyse de pattern. De plus, pour la version définitive (réduite à 40 items), nous n'avons aucune indication sur les items retenus. On peut supposer que la répartition des types d'items est plus équilibrée que dans le D70 (avec ici une plus grande fréquence des items numériques) mais le psychologue intéressé par cet aspect devra effectuer lui-même l'analyse des items s'il souhaite calculer, et interpréter, des sous-scores !

Pourquoi ne pas avoir profité de cette rénovation pour proposer au psychologue les outils (catégorisation de chaque item, procédure de calcul des sous-scores, étalonnages spécifiques) permettant d'effectuer une analyse plus précise des résultats du sujet qui compléterait l'information classique donnée par le score global ?

Une application très concrète des résultats de recherches était donc possible ici et, sans remettre en cause l'intérêt du D2000, on ne peut que regretter le caractère trop classique de cette rénovation.

Néanmoins, à partir de la catégorisation que nous avons effectuée (voir annexe 1), nous proposons au psychologue les premiers éléments d'une telle approche ⁹ sur l'ancienne version de l'épreuve : le D70. En utilisant une telle procédure, le psychologue peut différencier les sujets sur leur score global mais aussi procéder à l'analyse des patrons de réussite : il peut être intéressant de savoir qu'un sujet, par exemple, a échoué massivement aux items numériques mais a réussi tous les items spatiaux alors qu'un autre, avec le même score global, peut avoir un pattern de réussite différent. Cette approche, que l'on pourrait qualifier de clinique, est particulièrement pertinente dans le cadre d'un diagnostic cognitif (voir Richard, 1996).

Le N.N.A.T.

Dans la même logique d'analyse, nous présentons maintenant une autre épreuve de facteur g : le test N.N.A.T. (Naglieri Non Verbal Ability Test). Il s'agit d'une adaptation récente du test des matrices analogiques de Naglieri de 1985 (manuel N.N.A.T., 1998). Ce test collectif, de format papier/crayon, se compose de sept formes (A à G) adaptées à différents niveaux scolaires (de la maternelle au Bac). Le sujet doit résoudre des problèmes du type Matrices de Raven : choisir une matrice qui correspond à la suite logique d'une série de matrices proposées ou qui correspond à une partie d'une figure globale. Il s'agit donc, là aussi, d'une tâche typique d'une épreuve de facteur g mais avec la spécificité suivante : l'épreuve permet, en plus du score global, le calcul de sous-scores en fonction des types d'items. Les auteurs distinguent ainsi quatre types d'items :

- des items de complètement de pattern (P.C.) dans lesquels le sujet doit retrouver la partie manquante de la figure globale,
- des items de raisonnement analogique (R.A.) dans lesquels le sujet doit reconnaître les relations logiques entre les différentes formes géométriques,
- des items de raisonnement en série (S.R.) où le sujet doit comprendre la logique de la série,
- des items de représentation spatiale (S.V.) qui nécessitent de combiner deux dessins.

La répartition de ces items varie en fonction de la forme de l'épreuve et, par exemple, pour la forme E (niveau C.M.2/6e) qui comprend 38 items, nous pouvons observer une sur-représentation des items S.V. qui forment la moitié des items du test (voir *tableau 3*).

TABLEAU 3. Répartition des items de la forme E du N.N.A.T.

	P.C.	R.A.	S.R.	S.V.
Nombre d'items	5	6	8	19
Pourcentage	13	16	21	50

TABLE 3. *Distribution of items*

Les auteurs indiquent qu'il existe des différences de difficulté entre ces types d'items : les items P.C. étant les plus faciles, les items S.V. les plus difficiles.

Comme dans la procédure que nous avons proposée pour le D70, le praticien peut donc calculer, classiquement, le score global à l'épreuve mais aussi, et c'est l'une des originalités de cette épreuve, des sous-scores selon le type d'item (P.C., R.A., S.R. et S.V.). La démarche d'analyse est ici facilitée, car la catégorisation de chaque item figure dans le manuel, ainsi que des recommandations concernant l'utilisation de ces sous-scores. Par contre, les étalonnages par sous-scores ne sont pas fournis et le manuel nous incite à une certaine prudence dans l'utilisation de cette procédure : « *Les sous-scores ne devront être interprétés que de façon qualitative, compte tenu du faible nombre d'items qui les composent* » (manuel N.N.A.T., 1999, p. 12).

Cette épreuve peut donc être considérée comme une illustration d'une possible analyse clinique ¹⁰ d'un test cognitif. Nous pouvons néanmoins formuler quelques remarques sur la validité de ces sous-scores :

- pourquoi ne pas avoir conçu cette épreuve de manière à disposer d'un nombre suffisant d'items de chaque catégorie afin d'améliorer ainsi la validité des sous-scores ?
- les définitions des types d'items ne nous semblent pas assez claires (en particulier la distinction entre R.A. et S.R.),
- les justifications des rattachements des items nous semblent parfois discutables, en particulier lorsque l'analyse de la résolution de l'item ne correspond pas complètement à la catégorisation proposée (exemple pour la forme E, le rattachement de l'item 14 au type P.C. peut se discuter) ou lorsque l'item présente une saturation importante sur une autre catégorie que sa catégorie d'appartenance (manuel N.N.A.T., 1999, p. 41 et Annexe 2 du manuel).

L'ensemble de ces éléments, même s'ils ne remettent pas en cause la pertinence de ce type d'analyse, doit nous inciter à la prudence dans l'interprétation de ces sous-scores.

Les deux exemples d'analyse récente de test de facteur g que nous avons présentés ici, à partir du D70 et du N.N.A.T., illustrent bien quel peut être l'intérêt de cette démarche : apporter des informations complémentaires au score global classique. Mais cette démarche d'analyse nécessite des informations spécifiques : une catégorisation valide des items, des étalonnages spécifiques et des grilles d'interprétation des sous-scores.

L'analyse des stratégies

Dans certaines situations, et plus précisément ici dans certains tests, des sujets différents peuvent résoudre une même tâche par différentes procédures et il est alors pertinent de caractériser les sujets sur leur niveau de performance mais également sur la (ou les) procédures utilisées dans le test : « *Une même performance peut être obtenue par des voies très différentes et souvent il est plus informatif de repérer la voie suivie que de mesurer la performance* » (Huteau, 1994, p. 24).

Ainsi, dès 1974, Hunt distingue deux possibilités pour résoudre certains items des Matrices de Raven : par une approche globale ou par une approche analytique (Hunt, 1974). Il s'agit bien ici d'une approche cognitive et différentielle qui vise à identifier les différentes stratégies de résolution d'un même item ¹¹. Une grande partie des recherches actuelles en psychologie différentielle est centrée sur l'analyse de la démarche des sujets dans des situations de type résolution de problèmes, situations

proches de celles d'items de tests. Mais là encore, le passage des recherches à la pratique ne s'effectue pas sans difficultés et les adaptations de procédures expérimentales sont assez rares.

Nous présentons l'une des seules applications actuellement disponibles en France : le logiciel Samuel. Cette épreuve est une adaptation de l'épreuve créée par Kohs dans les années 1920, dans laquelle le sujet doit reproduire différents modèles à l'aide de cubes colorés. Cette épreuve classique a fait l'objet de plusieurs adaptations (elle représente, par exemple, l'un des subtests des échelles de Wechsler) mais les différentes versions évaluent toujours le niveau de performance (reproduction correcte dans un minimum de temps, avec des bonifications éventuelles si le sujet est rapide) et ne prennent pas en compte la démarche suivie par le sujet. Pourtant, Wechsler, dès les années 40, dans son adaptation de l'épreuve pour la première version de son échelle d'intelligence, distinguait déjà deux types de résolution selon que le sujet s'attache à la figure globale ou qu'il l'analyse en cubes séparés (Grégoire, 2000, P. 196). Ces premières observations ont été prolongées ensuite par plusieurs recherches qui font bien apparaître des différences inter individuelles dans les stratégies de résolution. On peut ainsi distinguer trois procédures de résolution dans ce type d'épreuve : une stratégie globale, une stratégie analytique et une stratégie synthétique (Rozencwajg & Huteau, 1996). Chaque stratégie se caractérise sur différents traits, essentiellement sur le type de représentation mentale de la situation (décomposition du modèle en cubes élémentaires pour la stratégie analytique ; décomposition en groupe de cubes, selon les formes géométriques, pour la stratégie synthétique ; difficulté dans la décomposition du modèle pour la stratégie globale), sur le type de construction (aléatoire pour la stratégie globale ; par colonne ou par ligne pour la stratégie analytique ; par formes géométriques pour la stratégie synthétique) et sur le niveau d'efficacité (un temps de résolution plus long et plus d'erreurs pour la stratégie globale).

Les recherches dans ce domaine sont le plus souvent centrées sur les différences inter individuelles, et catégorisent alors les sujets sur leur stratégie dominante (Rozencwajg & Huteau, 1996) mais, en fonction des caractéristiques des modèles présentés, on peut également observer une variabilité intra-individuelle ¹² importante dans ce type d'épreuve (Chartier, 1999). Ainsi, la présence de formes géométriques dans les modèles (un triangle composé de 2 cubes, un losange, une bande...) va ainsi inciter, plus ou moins fortement, les sujets à utiliser ces repères dans leur construction, donc à mettre en oeuvre une stratégie synthétique. Le diagnostic peut alors porter sur la stratégie préférentielle du sujet mais également sur sa flexibilité (capacité à utiliser plusieurs stratégies).

Face à la difficulté de saisir en temps réel la stratégie des sujets, les chercheurs ont utilisé des procédures vidéos pour le recueil des données (analyse de l'ordre de placement des cubes, de la qualité de l'anticipation, du contrôle visuel, du temps de réalisation, des erreurs...), procédures peu compatibles avec l'activité de praticien (en raison de la lourdeur du dispositif et du temps important nécessaire aux analyses).

Pour permettre un recueil et une analyse des données plus rapides, une version informatisée a été élaborée et le logiciel Samuel est maintenant diffusé (Rozencwajg *et al.*, 1999; Rozencwajg & Corroyer, à paraître). Cette épreuve permet d'établir un diagnostic des stratégies utilisées par les sujets dans la résolution des modèles. À partir d'une passation informatisée, le recueil (et l'analyse) de plusieurs indices est automatisé : chaque sujet peut être caractérisé par son niveau de performance mais

également, et c'est là la spécificité de cette épreuve, par le type de stratégie utilisée. Il s'agit réellement d'un test diagnostique qui nous donne différentes informations sur la ou les démarche(s) du sujet dans la réalisation de la tâche. Une caractérisation en termes de stratégie dominante est obtenue en fin de passation, mais une analyse plus fine concernant les différents indicateurs (ordre de construction, temps de réalisation, temps d'observation du modèle, qualité de la réalisation, correction des erreurs...) et/ou concernant l'analyse de la variabilité intra-individuelle (changement de stratégie en cours d'épreuve) est aussi possible.

Pour analyser ces données, des étalonnages sont fournis pour des sujets âgés de 9 à 25 ans (5 groupes d'âge). Ainsi, par exemple, l'utilisation de la stratégie globale tend à diminuer avec l'âge (74 % chez les enfants de 9 ans et 13 % chez les adultes) au profit de l'utilisation des stratégies plus élaborées (analytique ou synthétique). Ces étalonnages concernent également les sept indicateurs observés. Le manuel contient également un guide d'interprétation ainsi que des exemples d'analyse de protocoles.

L'utilisation possible de cette épreuve dans le domaine scolaire est illustrée dans un article récent qui analyse les liens entre type de stratégie au test Samuel et adaptation scolaire (Rozencwajg & Francequin, 1999). Ainsi ce type d'épreuve permet de recueillir des informations pertinentes et originales qui peuvent, par exemple, éclairer des difficultés d'apprentissage ou établir des liens plus étroits entre évaluation psychologique et remédiation.

Même si cette situation informatisée modifie fortement la situation du test d'origine ¹³ (les cubes sont ainsi des carrés, impossibilité ici de tourner le cube dans la main pour trouver la bonne face mais obligation de sélectionner la bonne orientation de la face du carré...) et qu'elle présente certaines limites (le diagnostic ne repose pour l'instant que sur 4 items ¹⁴), ce test est un bon exemple d'une application directe des apports théoriques à une pratique d'évaluation.

L'analyse des capacités d'apprentissage (l'évaluation dynamique)

Comme nous l'avons noté dans la première partie, l'évaluation dynamique apporte des informations supplémentaires sur les sujets par rapport à la simple mesure de leur niveau de performance (Büchel, 1995). Même si cette procédure présente encore des imperfections elle semble plus valide qu'une évaluation classique pour des populations spécifiques : faible niveau scolaire, familiarisation insuffisante avec des situations de type évaluation (Loarer, 1996 et 2001). Les recherches utilisent le plus souvent des tests collectifs de type facteur g (Matrice de Raven...) ou des épreuves individuelles adaptées de tests classiques (par exemple : l'épreuve des cubes de Kohs) accompagnées de procédures particulières de passation (formalisation des aides, calcul de scores...). Mais aucune de ces épreuves n'a fait l'objet d'une adaptation en France, sans doute en raison de problèmes théoriques et méthodologiques (validité, fidélité, différences inter individuelles non prises en compte dans les apprentissages...). Malgré ce manque d'outils, l'utilisation d'une telle démarche d'évaluation nous semble possible, même s'il est nécessaire de rester prudent dans l'interprétation des résultats (Guthke, 1990). Nous présentons ici deux procédures d'évaluation dynamique : l'une relativement connue des chercheurs du domaine (mais peut-être moins connue des praticiens), qui repose sur des aides au cours de l'épreuve, l'autre qui semble peu connue mais qui pourtant est, à notre connaissance ¹⁵, l'une des seules épreuves de ce type éditée en France : le test des carrés. Ces deux épreuves présentent des points communs, en particulier, elles

reposent sur une tâche de type cubes de Kohs et nécessitent une passation individuelle, mais l'une relève de la procédure aide au cours du test, alors que l'autre relève de la procédure test-aide-retest.

La procédure d'aide au cours du test

Galifret-Granjon et Santucci vont présenter dans les années 50 une procédure directement inspirée de celle de Goldstein. Il s'agit de modifier l'épreuve classique de Kohs : « *La transformation essentielle que Goldstein apporte à l'épreuve réside dans l'aide concrète qui est donnée au sujet* » (Galifret-Granjon & Santucci, 1958, p. 264). Après chaque échec, des aides progressives sont ainsi proposées à l'enfant jusqu'à l'obtention d'une reproduction correcte de la figure modèle. Trois niveaux d'aide sont envisagés :

- aide 1: modèle agrandi (même échelle que les cubes réels),
- aide 2: modèle agrandi avec les limites des cubes apparentes,
- aide 3 : on présente à l'enfant le modèle tout construit.

Dès que l'enfant réussit à reproduire un modèle, on revient à l'item échoué afin d'évaluer le transfert dans l'apprentissage. Les auteurs proposent pour chaque item une notation de 0 à 10 en fonction du niveau de réussite ainsi qu'un étalonnage pour des enfants de 7 à 13 ans. Une attention particulière est également portée sur les erreurs dans la construction.

Cette procédure sera reprise et développée quelques années plus tard par Ionescu et son équipe (Ionescu *et al.*, 1983, 1985 et 1987) avec les modifications suivantes :

- un niveau d'aide supplémentaire (avant l'aide 3) : une aide verbale à partir du modèle de l'aide 2;
- une cotation différente de chaque réussite, avec pondération en fonction de la difficulté du modèle (7 point à 17,2 points) ;
- l'élaboration de plusieurs indicateurs de la performance : note spontanée ¹⁶ (somme des points des items résolus sans aide), note aide, note de transfert, note globale... (des indications pour le calcul de ces différents indicateurs figurent en annexe 2).

Sur une population d'adultes présentant un « *déficit mental léger ou moyen* » (Ionescu *et al.*, 1990, p. 160) les auteurs comparent les résultats d'une évaluation classique (les matrices de Raven) aux résultats obtenus par cette évaluation dynamique, principalement sur leur validité prédictive concernant l'adaptation professionnelle. Les résultats montrent que l'évaluation dynamique (avec comme indicateur la note globale) a une validité prédictive plus élevée que l'évaluation classique. Ce type de résultat, qui complète les premiers travaux de Budoff et qui a été souvent observé (Huteau & Lautrey, 1999a), confirme la pertinence de ce mode d'évaluation pour ce public spécifique.

Cette procédure d'aide en cours de test est souvent utilisée dans les recherches (c'est d'ailleurs l'exemple classique utilisé par différents auteurs pour illustrer la procédure d'aide au cours du test) mais elle n'a toujours pas donné naissance à une épreuve éditée, directement utilisable. Sans revenir sur les raisons de ce constat (des interrogations demeurent sur la fidélité et la validité de la mesure), une utilisation de cette démarche d'évaluation nous semble pourtant possible si elle se limite à fournir des informations *complémentaires* à une évaluation classique des capacités du sujet. Mais en l'absence d'épreuve disponible, l'utilisateur doit créer son matériel à partir des indications qu'il trouvera dans les différents articles. Avec un tel dispositif, le

psychologue peut recueillir des informations spécifiques, peut-être plus qualitatives que réellement quantitatives, sur les capacités d'un sujet à tirer profit d'aides proposées par un adulte.

Le test des carrés

À notre connaissance, ce test est la seule épreuve d'évaluation dynamique diffusée actuellement en France. Mais, paradoxalement, elle semble beaucoup moins utilisée par les chercheurs que la procédure précédente ¹⁷.

Cette épreuve de 1985 est directement inspirée de la procédure expérimentale décrite en 1982 par les mêmes auteurs (Mary & Mariel, 1982). Elle est présentée comme étant un test « d'éducabilité » destiné à des élèves d'école maternelle (4 ans à 6 ans). C'est une épreuve individuelle, inspirée (elle aussi) de l'épreuve des cubes de Kohs, mais ici le sujet doit manipuler des palets (carrés) de différentes couleurs et non pas des cubes. Le matériel comporte un carnet « test » (composé de 21 modèles de niveau croissant de difficulté) et un carnet « apprentissage » (les mêmes modèles mais avec des couleurs différentes). Le sujet doit reproduire les figures modèles à l'aide des palets. La procédure, de type test-aide-retest, comporte les étapes suivantes :

- passation classique des items du carnet test (arrêt après 3 échecs consécutifs) ;
- reprise alors de l'item précédent le premier item échoué, mais avec le carnet d'apprentissage et avec 3 niveaux d'aide : aide verbale ; démonstration sur la moitié de la figure ; démonstration complète avec l'ensemble des palets. Dès que le sujet réussit, on lui présente le modèle originel échoué (carnet test) ;
- application de la procédure d'aide à l'ensemble des items échoués dans la première phase ;
- arrêt après 3 échecs consécutifs.

À l'issue de la passation, le sujet est caractérisé par deux notes :

- une note test : note de réussite spontanée (première passation des items du carnet test) ;
- une note retest : note test à laquelle se rajoutent les items réussis grâce aux aides (seconde passation des items).

Les auteurs fournissent deux étalonnages séparés (test et retest) pour chaque classe d'âge (4 ans à 6 ans). Dans leur échantillon d'étalonnage, ils notent de grandes différences inter individuelles de progression entre test et retest (donc des différences de sensibilité à l'apprentissage) : « *l'aide n'a donc pas les mêmes effets sur tous les enfants et certains en tirent un plus grand profit* » (Mary & Mariel, 1985, p. 8). On retrouve ici les propositions de Vygotsky concernant les différences inter individuelles dans cette « Zone Proximale de Développement ». Et cette phase d'aide semble avoir encore des effets trois mois plus tard avec le maintien des scores d'aide, alors même que la passation se déroulait cette fois sans aide.

Bien que l'on puisse situer clairement les auteurs dans le courant de l'évaluation dynamique : « *on ne se contente plus d'évaluer un niveau obtenu par l'enfant, on essaie en outre d'agir sur ce dernier, de le modifier.* » (Mary & Mariel, 1982, pp. 33-34), les auteurs ne font référence à aucune recherche dans ce domaine ¹⁸ et n'utilisent, ni en 1982 ni en 1985, les notions et les concepts clefs de ce courant (potentiel d'apprentissage, Zone Proximale de Développement...) alors même qu'ils proposent une évaluation de cette Zone Proximale de Développement ! Mais la mesure quantitative ne semble pas essentielle aux auteurs. Selon eux : « *l'important n'est-il pas tellement le score obtenu, mais plutôt l'observation de l'enfant dans sa démarche, ses tentatives pour surmonter la difficulté* » (Mary & Mariel, 1985, p. 7).

Nous retrouvons ici nos propres réflexions concernant l'intérêt et les limites de l'utilisation de telles démarches. Et cette épreuve, ou tout du moins la démarche proposée de type test-aide-retest, est particulièrement pertinente puisqu'elle nous donne ici un véritable score « spontané »¹⁹ qui correspond au niveau de performance classique du sujet.

Ce courant de l'évaluation dynamique présente donc un intérêt certain pour les psychologues, malgré les problèmes théoriques et méthodologiques que nous avons évoqués. Face à ces difficultés, on peut remarquer une orientation récente de certains travaux vers l'élaboration d'épreuves dynamiques de connaissance scolaire (Bonthoux *et al.*, 2000). Toutefois, ces problèmes ne sont plus réellement insurmontables si le psychologue utilise ces procédures comme *compléments* d'une évaluation : elles apportent, au moins à un niveau qualitatif, des informations complémentaires aux évaluations classiques.

Les apports méthodologiques dans la création ou la rénovation d'épreuves

Nous illustrerons cette partie par l'analyse des changements opérés dans la dernière version de l'échelle de Wechsler pour enfant : le W.I.S.C. III.

La rénovation régulière des tests est indispensable, principalement afin d'obtenir des normes récentes pour situer les performances du sujet (effet Flynn²⁰) et, secondairement, pour actualiser le matériel. Les auteurs peuvent également profiter de la rénovation pour intégrer les résultats les plus récents des recherches portant sur leur épreuve. Ainsi pour le W.I.S.C. III, l'un des objectifs de la rénovation vise au « *renforcement de la structure factorielle* » (Wechsler, 1996, p. 11).

Nous développerons ici ce dernier aspect et plus précisément l'apport méthodologique, mais aussi théorique, qui sont à la base des scores factoriels proposés dans cette épreuve. En effet, le praticien qui est amené à calculer ces trois nouveaux indices peut s'interroger sur leur validité et sur leur utilité.

Ces scores factoriels sont directement issus de résultats de recherches. Ainsi pour expliquer les scores au W.I.S.C. R, forme précédente du W.I.S.C. Kaufman proposait un regroupement non pas en deux sous-ensembles (les 2 échelles classiques : performance et verbale) mais en trois sous-ensembles, correspondant aux trois facteurs repérés par les analyses factorielles (Grégoire, 2000). Cette analyse en trois sous-scores serait alors théoriquement plus valide que l'analyse classique en deux échelles qui ne reposait, à l'origine, que sur une conception pragmatique de l'intelligence. Ces trois facteurs sont interprétés en termes de Compréhension Verbale (C.V. : la facette verbale de l'intelligence), Organisation Perceptive (O.P. : la facette visuo-spatiale) et Attention Concentration (A.C. : indépendance à la distraction, capacité à maintenir son attention). Dans l'élaboration du W.I.S.C. III, les auteurs ont souhaité renforcer le 3e facteur (qui ne regroupait que deux subtests du W.I.S.C. R: Arithmétique et Mémoire des chiffres) en créant un nouveau subtest (*Symbole*) afin d'améliorer la structure factorielle de l'épreuve. Par cette modification les éditeurs américains ont bien cherché à intégrer les apports des recherches récentes afin d'obtenir trois mesures valides et robustes. Mais sur l'échantillon américain de l'expérimentation du W.I.S.C. III, les auteurs obtiennent non pas trois facteurs mais quatre ! En effet, en plus des 3 facteurs attendus (C.V., O.P. et A.C.) ils obtiennent un 4e facteur (qui regroupe *Code* et *Symboles*) qu'ils interprètent

comme un facteur de Vitesse de traitement (V.T.). Ils proposent donc le calcul de quatre scores factoriels (Grégoire, 2000, p. 112). Lors de l'adaptation française, quelques années plus tard, les auteurs s'attendaient à retrouver ces quatre facteurs, mais les résultats obtenus sont différents : la solution avec quatre facteurs correspond mal aux données françaises. Plus précisément, la stabilité des quatre facteurs ne se retrouve pas à toutes les classes d'âge (Grégoire, 2000, p. 114). C'est pour cette raison que la version française du W.I.S.C. III ne comporte pas le facteur *Attention/Concentration* de la version américaine et que les trois scores factoriels n'utilisent pas tous les subtests (*Arithmétique* et *Mémoire des chiffres* ne sont rattachés à aucun indice factoriel dans la version française).

Les auteurs du manuel proposent donc au praticien de calculer, en plus des scores classiques (Q.I.V., Q.I.P. et Q.I.T.) ces trois scores factoriels (C.V., O.P. et V.T.) qui se présentent dans la même métrique que les Q.I. (moyenne de 100 et écart type de 15). Quel est alors l'intérêt de ces indices factoriels dans la pratique ?

Ces indices seraient plus valides que les indications en termes de Q.I. car leur organisation repose sur le résultat d'analyses factorielles et non plus sur l'approche pragmatique de Wechsler, à la base de la structure de l'épreuve. Ainsi, pour Grégoire, : « *les indices Compréhension Verbale et Organisation Perceptive sont des mesures plus homogènes que la Q.I. Verbal et le Q.I. de performance. Ils n'incluent en effet que des épreuves fortement saturées par le facteur en question* » (Grégoire, 2000, p. 114). L'indice C.V. serait une mesure plus fine de l'intelligence verbale (par rapport au Q.I.V.) et l'indice O.P. une mesure plus précise de l'intelligence visuo-spatiale (par rapport au Q.I.P.). Mais, à l'intérieur même de chaque indice, les subtests peuvent se différencier selon leur degré de représentativité du facteur considéré. Ainsi, par exemple pour le facteur O.P., l'épreuve de *Cubes*, avec une saturation de .70, peut être considérée comme une évaluation plus fiable de ce facteur que le subtest *Arrangement d'image* pour lequel la saturation n'est que de .45 (Wechsler, 1996, p. 204). Pour atteindre le but fixé dans l'élaboration du W.I.S.C. HI (obtenir des mesures plus fiables par les indices factoriels) on pouvait s'attendre à une pondération des subtests, selon l'importance de leur saturation, dans le calcul des indices factoriels correspondants. Ainsi, pour reprendre l'exemple du facteur O.P., il faudrait accorder plus de poids à la note obtenue au subtest *Cubes* et moins de poids à celui d'*Arrangement d'images*. Cette possibilité, envisagée par les auteurs, n'a pas été retenue : « *une procédure plus rigoureuse, mais plus complexe, consisterait à pondérer les notes standard des subtests en fonction du degré de saturation factorielle de ces derniers* » (Wechsler, 1996, p. 208). Une telle procédure, qui consisterait simplement à affecter un coefficient à chaque subtest, ne nous semble pourtant pas si complexe à mettre en oeuvre et améliorerait la validité de ces indices factoriels. La modalité de calcul retenue aboutit ici, selon nous, à une diminution de l'intérêt théorique de ces indices. De plus, les deux indices principaux (C.V. et O.P.) sont très proches des indices classiques Q.I.V. et Q.I.P. (la seule différence se limite dans les deux cas à un seul subtest) et l'on peut se demander alors ce qu'ils apportent de plus que les indicateurs classiques. Le premier réflexe est de se tourner vers le manuel. Mais celui-ci comporte très peu d'indications spécifiques sur l'utilisation de ces nouveaux indices. C'est dans un ouvrage plus récent que l'on peut trouver quelques éléments de réponse : « *l'usage des Indices n'a de sens que lorsque le Q.I.V. et/ou le Q.I.P. sont déprimés du seul fait d'une note anormalement basse en Arithmétique et/ou en Code* » (Grégoire, 2000, p. 153). De telles informations mériteraient d'être présentes dans le manuel du W.I.S.C. III afin d'être accessibles à tous les utilisateurs de cette épreuve.

Pour le troisième indice factoriel V.T. (Vitesse de Traitement) son utilisation est également limitée mais pour d'autres raisons. D'une part, il ne comprend que deux subtests (la mesure est moins fiable) et, d'autre part, l'interprétation de ce facteur en termes de vitesse de traitement est discutable. En effet, les deux subtests concernés (Code et Symboles) sont bien 2 épreuves chronométrées, mais elles mesurent également d'autres aspects du fonctionnement cognitif (aspects grapho-moteurs...). De plus, d'autres subtests évaluent également la vitesse de traitement car ils donnent lieu à des bonifications si le sujet est rapide. L'utilisation de cet indice doit donc être effectuée avec prudence (voir sur ces points Grégoire, 2000, pp. 114-115 et p. 153).

Cette présentation de certains apports de la recherche dans la rénovation d'épreuves à partir de l'exemple du W.I.S.C. III, visait à rendre plus visibles les liens entre recherche et pratique. Ainsi, des résultats de recherches peuvent avoir pour conséquence de modifier une épreuve classique afin d'améliorer sa validité. Mais nous avons vu ici, à partir du W.I.S.C. III, d'une part, que l'ajout d'un subtest n'a pas obligatoirement les effets escomptés et, d'autre part, que la logique de calcul des indices factoriels n'a pas été suivie jusqu'à son terme. Au bout du compte, les « nouveaux indicateurs » (les indices factoriels) n'apportent pas réellement une information différente (ou suffisamment fiable) de celle apportée par les indicateurs classiques de type Q.I. Cet exemple nous permet également d'illustrer l'importance pour le praticien de conserver un regard critique sur les tests qu'il utilise et, ici plus précisément, sur les scores qu'il est amené à calculer et à interpréter.

Si les traitements statistiques (en particulier, ici, les analyses factorielles exploratoires et confirmatoires) permettent bien de valider a posteriori des épreuves anciennes, elles contribuent également à assurer la validité théorique de nouvelles épreuves en vérifiant, par exemple, les regroupements théoriques proposés par les chercheurs (regroupement effectué, dans le meilleur des cas, à partir d'un cadre théorique de référence). Une rapide illustration nous en est donnée dans l'élaboration du K. ABC. L'objectif de cette épreuve étant d'évaluer trois aspects de « l'intelligence » (deux types de traitement de l'information — processus séquentiels et processus simultanés — et une échelle de connaissance) les auteurs ont élaboré des subtests pour chacune des trois échelles. Sur les 50 subtests de la version expérimentale ils ne retiennent que ceux présentant une saturation élevée dans l'une de ces échelles. Ils s'aperçoivent ainsi que le subtest *série de photos* qui avait été conçu, *a priori*, comme une mesure des processus séquentiels était saturé plus fortement dans le facteur qui correspond aux processus simultanés. Les auteurs ont donc modifié la structure de leur épreuve en déplaçant ce subtest vers l'échelle de processus simultanés (Kaufman *et al.*, 1993, p. 55). Ils ont amélioré ainsi la validité de leur épreuve.

Conclusion

Dans cet article, nous avons montré, à partir d'exemples précis sur des tests publiés, qu'il existait des applications concrètes de certains résultats de recherches. Nous avons également souligné, à partir de l'exemple du W.I.S.C. III et du K. ABC, que de tels liens entre recherche et pratique se retrouvent dans les modifications apportées lors de la rénovation d'épreuves classiques ou dans l'expérimentation de nouveaux tests.

Nous avons présenté plusieurs démarches de lecture « clinique » de tests d'intelligence (analyse des sous-scores du D70 et du N.N.A.T. ; analyse des stratégies dans une épreuve

de type cubes de Kohs avec le logiciel Samuel) ainsi que deux procédures d'évaluation du potentiel d'apprentissage. Ces différentes utilisations d'épreuves classiques permettent d'obtenir des indications complémentaires sur le fonctionnement cognitif du sujet qui viennent enrichir l'indication classique sur le niveau de performance du sujet. Nous sommes bien ici dans une logique de diagnostic cognitif (Richard, 1996) et/ou dans celle d'aide à la connaissance de soi (Chartier, 2001). Ces procédures doivent permettre également de resserrer les liens entre évaluation et remédiation : mieux connaître les caractéristiques du sujet pour lui apporter des réponses adaptées.

Même si les épreuves et/ou les procédures présentées doivent encore être, pour certaines, améliorées et validées (fidélité de la mesure, validité des indices, étalonnages spécifiques...), elles témoignent de l'apport des recherches actuelles aux pratiques d'évaluation.

BIBLIOGRAPHY

- Baccini, P. (2000). *Étude de la nature des règles de résolution des items du D70, test d'intelligence générale*. Mémoire de D.E.C.O.P. (non publié). Paris : C.N.A.M./I.N.E.T.O.P.
- Bonthoux, F., Cobessi, N., & Touron, S. (2000). L'évaluation des compétences scolaires par le potentiel d'apprentissage : étude au C.E.1. *Psychologie Française*, 45, 3, 261-268.
- Büchel, F.-P. (1995). L'évaluation du potentiel d'apprentissage In J. Lautrey (Éd.), *Universel et Différentiel en Psychologie*. Paris : P.U.F.
- Michel, F.-P., De Ribaupierre, A., & Scharnhorst, U. (1990). Le diagnostic du potentiel d'apprentissage par L.P.A.D. : une étude de la fidélité. *European Journal of Psychology of Education*, 5, 135-158.
- Büchel, F. P., & Paour, J. L. (Eds.), (1990). Assessment of learning and potential development : Theory and practices. *European Journal of Psychology of Education*, 5, 2, special issue. Chartier, P. (1999). Étude de la variabilité intra et inter-individuelle dans la résolution d'une épreuve du type cubes de Kohs. In M. Huteau & J. Lautrey (Éds), *Approches différentielles en Psychologie*. Rennes : P.U.R.
- Chartier, P. (2001). Quel usage des tests d'efficience dans le cadre de l'éducation à l'orientation ? *L'Orientaion Scolaire et Professionnelle*, 30, 240-245.
- Chartier, P., & Vrignaud, P. (en préparation). L'analyse cognitive d'un test de facteur g: le D70.
- Dickes, P. (1996). L'analyse factorielle linéaire et ses deux logiques d'application. *Psychologie Française*, 41, 1, 9-22.
- Dickes, P., & Martin, R. (1998). Les composantes de l'intelligence générale du D70. *Psychologie et Psychométrie*, 19, 1, 27-51.

- Flieller, A. (2001). Problèmes et stratégies dans l'explication de l'effet de Flynn. In M. Huteau (Éd.), *Les figures de l'intelligence*. Paris : E.A.P. Galifret-Granjon, N., & Santucci, H. (1958). Test adapté de Kohs-Goldstein. In R. Zazzo (Éd.), *Manuel pour l'examen psychologique de l'enfant*. Neuchâtel : Delachaux et Niestlé.
- Gardner, H. (1997). La conception standard de l'intelligence est fausse. *Sciences Humaines*, 69, 34-37.
- Gaudron, J.-P. (1999). La psychométrie assistée par ordinateur : problématiques en question et perspectives de recherches, *L'Orientation Scolaire et Professionnelle*, 28, 1, 31-62.
- Grégoire, J. (2000). *L'évaluation clinique de l'intelligence de l'enfant. Théorie et pratique du W.L.S.C.-HI. Sprimont : MARDAGA.*
- Guthke, J. (1990). Les tests d'apprentissage comme alternative ou complément aux tests d'intelligence : un bilan de leur évolution. *European Journal of Psychology of Education*, 5, 117-133.
- Hunt, E. (1974). Quote the Raven ? Nevermore ! In J. Gregg (Ed.), *Knowledge and Cognition*, 129-157. Hillsdale NJ: Erlbaum.
- Huteau, M. (1990). La variété des formes et des composantes de l'intelligence. In M. Reuchlin, J. Lautrey, C. Marendaz, & T. Ohlmann (Éds.), *Cognition : l'individuel et l'universel*. Paris : P.U.F.
- Huteau, M. (1994). L'évaluation psychologique des personnes : problèmes et enjeux actuels. In M. Huteau (Éd.), *Les techniques psychologiques d'évaluation des personnes*. Issy-les-Moulineaux : E.A.P.
- Huteau, M. (1995). Les tests d'intelligence et la psychologie cognitive. In J. Lautrey (Éd.), *Universel et différentiel en psychologie*. Paris : P.U.F.
- Huteau, M. (Éd.). (2001). *Les figures de l'intelligence*. Paris : E.A.P.
- Huteau, M., & Lautrey, J. (1978). L'utilisation des tests d'intelligence et de la psychologie cognitive dans l'éducation et l'orientation. *L'Orientation Scolaire et Professionnelle*, 7, 2, 99-174.
- Huteau, M., & Lautrey, J. (1997). *Les tests d'intelligence*. Paris : La découverte.
- Huteau, M., & Lautrey, J. (1999 a). *Évaluer l'intelligence. Psychométrie cognitive*. Paris : P.U.F.
- Huteau, M., & Lautrey, J. (1999 b). *Approches différentielles en Psychologie*. Rennes : P.U.R.
- Ionescu, S., Jourdan-Ionescu, C., & Toselli-Toschi, M.-R. (1983). Nouvelles utilisations dans l'utilisation d'un test d'intelligence (cubes de Kohs). *Enfance*, 4, 363-380.
- Ionescu, S. et Jourdan-Ionescu, C. (1985). L'évaluation du potentiel d'apprentissage, 1: Utilisation du test des cubes. *Bulletin de psychologie*, 38, 919-927.
- Ionescu, S., Jourdan-Ionescu, C., & Alain, M. (1987). L'évaluation du potentiel d'apprentissage, 2: Une nouvelle méthode de quantification. *Bulletin de psychologie*, 40, 481-487.
- Ionescu, S., Jourdan-Ionescu, C., & Fortin, P. (1990). Potentiel d'apprentissage et adaptation professionnelle des déficients mentaux. *European Journal of Psychology of Education*, 2, 159-166.
- Kaufman, A.-S., & Kaufman, N.-L. (1993). *K.A.B.C. Batterie pour l'examen psychologique de l'enfant*. Paris : E.C.P.A.
- Lautrey, J. (1994). L'évaluation du potentiel d'apprentissage : état de la question. In M. Huteau

- (Ed.), *Les techniques psychologiques d'évaluation des personnes*. Issy-les-Moulineaux E.A.P. Lautrey, J. (1996). La recherche des « particules élémentaires » de l'intelligence : une impasse ? *Psychologie Française*, 41, 1, 23-33.
- Lautrey, J. (2001). L'évaluation de l'intelligence : état actuel et tentatives de renouvellement. In M. Huteau, (Ed.), *Les figures de l'intelligence*. Paris : E.A.P.
- Lautrey, J., & Huteau, M. (1990). L'évaluation du développement et des compétences cognitives chez l'enfant. Quoi de neuf ? *Revue de Psychologie Appliquée*, 40, 4, 437-464.
- Loarer, E. (1996). Le pronostic de l'adaptation scolaire des jeunes étrangers nouvellement arrivés en France par la mesure du potentiel d'apprentissage. *Migrants-Formation*, 104, 98-110.
- Loarer, E. (2001). L'évaluation dynamique des aptitudes comme moyen de limiter les biais culturels dans les tests d'intelligence. In M. Huteau (Ed.), *Les figures de l'intelligence*. Paris : E.A.P.
- Loarer, E., Chartier, D., Huteau, M., & Lautrey, J. (1995). *Peut-on éduquer l'intelligence ?* Bern : Peter Lang.
- Loarer, E. et Chartier, D. (1996). L'évaluation dynamique des aptitudes : révolution ou gadget ? *Psychologie Française*, 41, 1, 35-46.
- Mary, C., & Mariel, R. (1982). Le développement intellectuel chez les enfants de 4 à 6 ans : le test des carrés. *Psychologie scolaire*, 40, 33-51.
- Mary, C., & Mariel, R. (1985). *Le test des carrés*. Issy-les-Moulineaux : E.A.P. Rémy, L., & Gilles, P.-Y. (1999). Stratégies de résolution spatiale et numérique du D70. In
- M. Huteau & J. Lautrey (Éds), *Approches différentielles en Psychologie*. Rennes : P.U.R. Reuchlin, M. (1978). Processus vicariants et différences individuelles. *Journal de Psychologie Normale et Pathologique*, 2, 133-145.
- Reuchlin, M., & Bacher, F. (1989). *Les différences individuelles dans le développement cognitif de l'enfant*. Paris : P.U.F.
- Rey, A. (1934). D'un procédé pour évaluer l'éducabilité. *Archives de Psychologie*, 24, 297-337. Rey, A. (1991). *Le test des plateaux*. Issy-les-Moulineaux : E.A.P.
- Richard, J.-F. (1996). Les différentes approches de l'analyse des compétences cognitives. *Psychologie Française*, 41, 1, 3-8.
- Rozencwajg, P., & Huteau, M. (1996). Les stratégies globale, analytique et synthétique dans les cubes de Kohs. *Psychologie Française*, 41, 1, 57-64.
- Rozencwajg, P., & Corroyer, D. (à paraître). Strategy development in a block design task. *Intelligence*.
- Rozencwajg, P., Corroyer, D., & Altman, P. (1999). *Samuel : Diagnostic du fonctionnement cognitif (manuel)*. Cergy : Delta Expert.
- Rozencwajg, P., & Francequin, G. (1999). Contributions de l'analyse des stratégies de résolution de problème à l'examen psychologique. *L'Orientation Scolaire et Professionnelle*, 28, 1, 63-82.
- Schneuwly, B., & Bronckart, J.-P. (1985). *Vygotsky aujourd'hui*. Neuchâtel : Delachaux et Niestlé.
- Test N.N.A.T. (1998). *Manuel d'application*. Paris : E.C.P.A.
- Test D70 (1970). *Manuel d'application*. Paris : E.C.P.A.
- Test D2000 (2000). *Manuel d'application*. Paris : E.C.P.A.

Toussaint, I. (2000). *Analyse cognitive d'un test d'intelligence générale : le D70*. Mémoire D.E.C.O.P. (non publié). Paris : C.N.A.M./I.N.E.T.O.P.

Vrignaud, P. (1996). Les tests au XXI^e siècle. Que peut-on attendre des évolutions méthodologiques et technologiques dans le domaine de l'évaluation psychologique des personnes ?, *Pratiques Psychologiques*, 4, 5-27.

Wechsler, D. (1996). *Manuel du W.L.S.C. III*. Paris : E.C.P.A.

NOTES

1. Deux dimensions importantes de l'intelligence dans nos sociétés, mais dont la présentation et l'analyse dépassent le cadre de cet article. On peut consulter sur ces points Gardner, 1997.
2. Certaines d'entre elles semblent même être dans une impasse (Lautrey, 1996).
3. Cette épreuve expérimentale n'a été éditée sous forme de test qu'en 1991 sous le nom de « test des plateaux » (Rey, 1991).
4. Ces premières « intuitions » de Vygotsky ont ensuite été confirmées par plusieurs expériences : ainsi Budoff distingue bien, dans un échantillon d'élèves relevant de l'enseignement spécialisé, deux types de sujets de niveau comparable (si l'on s'en tient à une procédure classique d'évaluation) : les gagnants et les non gagnants. Seuls les gagnants tirent profit des aides proposées et Budoff suggère alors de ne considérer comme déficients que les « non gagnants » (Ionescu & Jourdan-Ionescu, 1985). Cette distinction entre des sujets qui apparemment sont de même niveau n'est possible que par une évaluation dynamique.
5. Pour une présentation des analyses factorielles exploratoires et confirmatoires voir Reuchlin & Bacher (1989) et Dickes (1996).
6. Ce travail a été mené en collaboration avec P. Baccini et I. Toussaint dans la cadre de leur mémoire D.E.C.O.P. (Baccini, 2000; Toussaint, 2000).
7. Notre catégorisation est en cours de validation (Chartier & Vrignaud, en préparation).
8. Je reprends ici les termes exacts du manuel, mais je suppose qu'il s'agit d'une erreur de celui-ci et que les auteurs distinguent ici les items arithmétiques...
9. Pour compléter notre travail, il reste encore à élaborer des étalonnages par sous scores.
10. Pris ici dans le sens : analyse des réussites et des échecs, analyse du pattern.
11. Contrairement aux travaux présentés en 2.1 dans lesquels le type d'item contraignait fortement la démarche du sujet (par exemple un item *spatial* du D70 nécessite obligatoirement une procédure *spatiale* pour être résolu), les travaux ici se réfèrent à des situations moins contraignantes qui laissent au sujet la possibilité d'utiliser différentes procédures de résolution (sur l'analyse des stratégies au test D70 voir Rémy & Gilles, 1999).
12. Utilisation par un même sujet de plusieurs stratégies au cours d'une même épreuve.
13. Sur la problématique plus générale de l'informatisation des tests voir Gaudron, 1999.
14. Le test comporte 10 items mais le diagnostic des stratégies ne repose que sur les items les plus complexes comportant 9 « cubes ».
15. Mis à part l'épreuve de Rey, déjà présentée, et l'épreuve de Feurstein (le L.P.A.D.) qui est plus une étape préliminaire à des activités de remédiation (P.E.1.). Pour une présentation de cette épreuve, on peut consulter Loarer *et al.*, 1995, ainsi que Büchel & Paour, 1990.
16. Qui n'est pas réellement une note spontanée car elle prend en compte les effets des aides précédentes.
17. Ainsi nous n'avons trouvé aucun article récent sur cette épreuve.
18. Excepté la thèse de 1966 de Hurtig.
19. Ce qui n'est pas le cas, comme nous l'avions fait remarquer, dans la démarche précédente de type aide au cours du test.

20. À travers l'analyse de plusieurs recherches, Flynn a observé une évolution progressive du niveau moyen de performance dans les tests d'intelligence (de l'ordre de 1/3 de point de Q.I. par an). Des étalonnages récents sont donc nécessaires pour situer les résultats d'un sujet dans une population de référence qui lui correspond. Pour une présentation récente de l'effet Flynn, voir Flieller, 2001.

ABSTRACTS

The synthesis articles about research as far as intelligence is concerned most of the time come to the conclusion that this research has but a low impact on the current test practices.

The aim of this paper is to present different tests (and/or analysis proceedings) that can really be applied today. We'll illustrate these uses according to four main ones: task analysis (with D70 and N.N.A.T.), the analysis of resolution strategies (introducing the Samuel software), the assessment of learning potential (two methods will be put forward) and the methodological contribution to the renovation tests (analysis of W.I.S.C. 111's factorial indices). The limits to the use of those new tests, and/or to those new approaches to former ones, will also be dealt with.

Les articles de synthèse concernant les recherches dans le domaine de l'intelligence concluent, le plus souvent, aux faibles répercussions de ces recherches sur les pratiques actuelles d'évaluation. L'objectif de cet article est de présenter des tests (et/ou des procédures d'analyse) effectivement applicables aujourd'hui. Nous illustrerons ces utilisations selon quatre axes : l'analyse de la tâche (à partir des tests D70 et N.N.A.T.), l'analyse des stratégies de résolution (présentation du logiciel Samuel), l'évaluation du Potentiel d'Apprentissage (deux méthodes seront proposées) et les apports méthodologiques dans la rénovation de tests (analyse des indices factoriels du W.I.S.C. III). Les limites d'utilisation de ces nouvelles épreuves, et/ou de ces nouvelles lectures de tests plus anciens, seront également abordées.

INDEX

Keywords: Competencies assessment, career guidance, evaluation of career counseling practices, career transitions, self-esteem, career maturity

Mots-clés: Bilan de compétences, orientation professionnelle, évaluation de pratique de conseil, transitions professionnelles, estime de soi, maturité vocationnelle

AUTHOR

PHILIPPE CHARTIER

est Conseiller d'Orientation-Psychologue à l'I.N.E.T.O.P. (Service de Recherche). Ses travaux portent sur les problématiques d'évaluation (évaluation psychologique et évaluation scolaire)

Contact : I.N.E.T.O.P., 41, rue Gay-Lussac, 75005 Paris. Courriel : chartier@cnam.fr.

Remerciements : L'auteur remercie Serge Blanchard et Pierre Vrignaud pour leur lecture critique d'une première version de cet article.