La psychotechnique des aptitudes : pour différencier une sociotechnique de l'évaluation sans mesurage et une psychologie balbutiante de la compréhension de la performance

Stéphane Vautier

La psychotechnique des aptitudes
Pour différencier une sociotechnique de l'évaluation sans mesurage et une psychologie balbutiante de la compréhension de la performance

Stéphane Vautier
Université de Toulouse[1]

Résumé

Une conception répandue consiste à considérer que des tests psychotechniques validés permettent de mesurer des aptitudes intellectuelles à partir du scorage des performances observées. Cet article (i) développe une conception falsifiable du mesurage ordinal, (ii) montre que les performances observées falsifient vraisemblablement cette conception et (iii) analyse comment la modélisation psychométrique satisfait l'impératif comparatif qui sous-tend l'évaluation des aptitudes. Mais l'efficacité évaluative s'établit au détriment de la connaissance scientifique des déterminants de la performance. La pratique de l'examen psychologique est ensuite analysée comme une sociotechnique de l'évaluation sans mesurage.

Mots-clés : tests psychologiques, mesurage, psychométrie

Abstract

A widespread view consists in considering that validated psychotechnical tests enable one to measure intellectual abilities with the help of the scoring of observed performances. This paper (i) elaborates a falsifiable conception of ordinal measurement, (ii) shows that it is likely that the observed performances falsify it, and (iii) analyzes how psychometric modeling fulfils the comparative imperative that underpins the assessment of abilities. But the evaluative efficacy builds up to the detriment of scientific knowledge of the performance's determinants. The practice of psychological assessment is then thought of as a sociotechnics of assessment without measurement.

Key-words: psychological testing, measurement, psychometrics

[1] Universtité de Toulouse-Le Mirial. Pavillon de la Recherche, Octogone. 5 allées Antonio Machado, 31058 Toulouse cedex 9. Courriel : vautier@univ-tlse2.fr. Je remercie Philippe Chartier, Jean-Philippe Gaudron et Valérie Tartas pour leurs commentaires pendant l'élaboration du présent article.

> La crainte qu'éprouve le fils authentique de la civilisation moderne à l'idée de s'éloigner des faits qui sont déjà schématiquement préformés par les conventions dominantes de la science, du commerce et de la politique, est la même que la crainte qu'inspire la déviation sociale.
> Max Horkheimer et Theodore W. Adorno, *La dialectique de la raison*.

1. Introduction

Tout praticien des tests d'aptitude sait que les scores ou les notes qu'il attribue aux personnes qu'il teste ne mesurent pas de grandeur analogue à la longueur ou la température d'un corps. Il sait aussi l'importance du label « test validé » dans sa pratique : on n'entreprendrait pas sans risque professionnel une évaluation des aptitudes avec des tests qui ne sont pas validés. Ce label, qui est en fait un label d'utilisabilité, est parfois considéré dans la communauté des utilisateurs comme un label de scientificité (e.g., Gaillard, Colasse, Guihard, & Michel, 2011, p. 155). Je voudrais montrer que cette dernière opinion est discutable et que, par conséquent, la finalité du scorage psychotechnique doit être assumée dans une perspective franchement normative et donc sociotechnique. Un tel éclairage entraîne des conséquences politiques puisqu'il s'agit de dé-naturaliser l'objet de l'évaluation psychotechnique en lui reconnaissant le caractère d'un fait social par opposition à un fait brut (cf. Searle, 1995).

La position que je souhaite mettre à l'épreuve est la suivante : l'évaluation d'un niveau de performance ou d'aptitude[2] à l'aide d'un score qu'on rapporte à une norme statistique ne constitue pas une opération de mesurage, ce qui implique que l'utilisation du terme de *mesure* est scientifiquement trompeuse. Si la communauté des utilisateurs de tests souhaite assumer sa responsabilité scientifique, elle doit alors « faire le ménage » dans ses modes d'expression pour clarifier son domaine de compétences tant vis-à-vis de ses membres que des membres de la société civile au sens large, en évacuant de sa terminologie les termes évoquant des grandeurs mesurables et en assumant l'évaluation comme un processus qui assigne à l'individu une ou des propriétés qu'il ne possède pas intrinsèquement.

Le terme d'évaluation possède une ambiguïté descriptive et appréciative redoutable. Si on dit qu'on évalue la longueur d'un objet, on entend qu'on cherche à connaître le nombre réel par quoi il faut multiplier une unité de mesure de référence pour décrire correctement cet objet du point de vue de sa longueur. La connaissance de ce nombre, nécessairement approchée, n'implique aucun

[2] Bien que performance et aptitude (ou compétence) ne soient pas des concepts équivalents, le praticien vise l'aptitude en regardant la performance. Pour ne pas alourdir inutilement l'analyse, je négligerai la distinction sauf cas particulier.

jugement de valeur. Le jugement de valeur qui s'appose au résultat du mesurage est éventuellement possible si un contexte d'utilisation de l'objet détermine sa valeur d'usage (il ne doit pas être *trop* long par exemple). C'est pourquoi je préfère le verbe « mesurer » au verbe « évaluer » lorsqu'il s'agit de déterminer un *rapport* numérique (pour une introduction à la notion de mesurage, voir par exemple Michell, 2003a).

Si on dit qu'on évalue l'intelligence d'un enfant, alors il s'agit non pas de mesurer une grandeur, mais bien d'assigner une place à l'enfant dans une échelle de scores ; les tests d'aptitude sont faits pour cela. En soi, un score n'est pas un jugement de valeur, mais ce n'est pas non plus le résultat d'un mesurage : le score n'est jamais compris comme l'encadrement d'un nombre réel qui indiquerait par quoi il faut multiplier une unité de référence pour obtenir une grandeur caractéristique de l'enfant, ou, plus justement, de sa performance. Le score, interprété comme une propriété de l'enfant par une sorte de jeu verbal (le score « traduit » la performance, laquelle est le « produit » de l'intelligence de l'enfant ; donc le score « décrit » l'intelligence de l'enfant), constitue la condition de possibilité pour que la formulation d'un jugement de valeur sur l'enfant, en fonction du contexte dans lequel il s'agit de l'*insérer*, acquière une factualité suffisante, quoique trompeuse du point de vue scientifique.

« Insérer une personne dans un contexte social », en l'occurrence l'enfant évalué, implique une construction de significations à propos de la personne qui devient objet d'attention, objet à spécifier, à positionner et vis-à-vis duquel se positionner, objet à insérer dans un réseau d'enjeux relationnels et/ou institutionnels, le plus souvent implicites. Par exemple, Binet et Simon (1907) proposent leur « échelle métrique de l'intelligence » pour « une situation où des doutes planent sur les causes du retard scolaire » (p. 92) et où l'enjeu consiste à « envoyer l'élève à la classe de perfectionnement » ou bien à le renvoyer « à l'école ordinaire ». La performance de l'enfant *doit* alors être la variable d'une fonction compatible avec l'évaluation, c'est-à-dire d'une fonction dont les valeurs sont compatibles avec les notions pratiques de « pas assez », « trop », « suffisamment » : la performance doit être *suffisamment élevée* pour un renvoi à l'école ordinaire, ou bien *assez basse* pour une orientation en classe de perfectionnement. Comme la description de la performance n'est pas un nombre ordinal (on verra que c'est généralement un *m*-uplet), le scorage prépare son évaluation en transformant sa description en scalaire, ou encore, en nombre, toujours lisible comme degré dans un ordre simple[3], auquel il suffit d'adjoindre des seuils ajustables à la situation.

Ainsi peut-on exhumer de la pratique du scorage psychotechnique ce que j'appellerai l'*impératif de comparabilité*. La psychotechnique répond à une demande sociale de comparabilité. Non pas qu'il s'agisse de comparer à tout va ; ce qui importe, c'est de *pouvoir* comparer *si* le besoin s'en fait sentir. L'intérêt social de la psychotechnique comme technicité qui s'exerce sur autrui dépend de sa capacité à satisfaire l'impératif de comparabilité. Mais, malgré ce qu'affirment

[3] Un ordre simple est un ensemble dont les éléments pris par paires peuvent toujours être ordonnés l'un par rapport à l'autre (plus, moins, ou aussi que).

Huteau et Lautrey (1999, p. 76), lorsqu'ils écrivent que la mesure de l'efficience intellectuelle – via l'observation de performances à des items de tests – est *fondée* au niveau ordinal, on ne peut pas laisser croire que les connaissances de la science psychologique permettent de fonder une telle technicité. C'est la raison pour laquelle je propose de classer la psychotechnique dans la catégorie des sociotechniques, parce que c'est une technique (ou une ingénierie) sociale et qu'elle ne constitue l'exploitation d'aucune loi psychologique connue ni, a fortiori, d'aucun principe de mesurage.

Dans le cadre de cet article, je me bornerai à définir, dans une première partie, ce que serait un *mesurage ordinal* en prenant l'exemple d'un test bien connu, et à montrer comment le discours psychométrique entérine le fait qu'on ne sache mesurer ordinalement aucune grandeur théorique avec des réponses (ou des performances) à des items de test. Puis j'analyserai les pratiques linguistiques en cours dans la littérature psychotechnique pour montrer comment l'emploi des mots masque ce fait, en prenant comme exemple le manuel d'un test bien connu. Cette analyse sera complétée d'une petite mise en scène qui vise à rendre sensibles les tensions logiques et éthiques que doit affronter le psychologue clinicien lorsqu'il sert la démarche évaluative.

2. Mesurer une grandeur avec le test Cubes du WISC-IV

La stratégie que je vais appliquer consiste à utiliser un exemple concret pour développer l'argument suivant. Soient des conditions suffisantes pour le mesurage ordinal d'une grandeur théorique dans une certaine population d'unités d'observation. Ces conditions forment une hypothèse théorique qui est fausse. Par conséquent, l'argument selon lequel on dispose d'une hypothèse de laquelle *déduire* qu'on sait mesurer une grandeur de manière ordinale dans cette population est logiquement valide mais il est logiquement non valable parce que sa prémisse est fausse. Il résulte d'une telle analyse qu'en l'absence d'hypothèse alternative, on ne sait pas justifier qu'on sache mesurer la grandeur théorique avec la performance dans cette population. Cette grandeur n'est tout simplement pas un concept scientifique.

Le WISC-IV est une batterie de tests utilisée par les psychologues cliniciens dans le cadre de l'examen psychologique de l'enfant et de l'adolescent (voir aussi Chartier & Loarer, 2008; Grégoire, 2009; Jumel & Savournin, 2013). Elle comprend 15 tests et permet de calculer, en fonction des réponses observées, des scores, appelés notes ou indices, de Compréhension Verbale, de Raisonnement Perceptif, de Mémoire de Travail, de Vitesse de Traitement, ainsi qu'une note Totale (Wechsler, 2005a). Je me concentre ici sur la question technique du mesurage d'une grandeur théorique par la performance observée au test Cubes.

2.1. La description de la performance au test Cubes

La description de la performance au test Cubes mobilise un langage dont la syntaxe et le lexique sont codifiés de la manière suivante. Tout d'abord, comme le test comprend 14 tâches (ou items), la performance au test est un 14-uplet. Cette notion est fondamentale pour la compréhension de ce qui suit, c'est pourquoi il

convient de s'y attarder quelque peu en partant d'un exemple didactique. Supposons pour simplifier que le test ne comprenne que trois tâches, toujours administrées dans le même ordre. La performance au test est alors décrite sous la forme d'un triplet (un 3-uplet), par exemple le triplet (1, 1, 0), qu'on peut aussi abréger par « 110 ». Le premier « 1 » indique le résultat issu de l'observation de l'enfant lorsqu'il est confronté à la première tâche ; le deuxième « 1 » indique le résultat issu de l'observation de l'enfant lorsqu'il est confronté à la seconde tâche : le « 0 » indique le résultat issu de l'observation de l'enfant lorsqu'il est confronté à la troisième tâche. La syntaxe de la description de la performance à ce petit test prend la forme « 1 *puis* 1 *puis* 0 ». La description de la performance aux 14 tâches du test est un 14-uplet. Cette description est multivariée, plus précisément 14-variée[4].

Après la syntaxe, penchons-nous sur le lexique de la description. Le résultat de l'observation de l'enfant face à une tâche s'exprime à l'aide d'un lexique spécifique. Par exemple, si les chiffres « 0 » et « 1 » signifient que la tâche est respectivement échouée ou réussie, la description « 110 » indique deux réussites successives puis un échec. Le test Cubes comprend trois lexiques. Le premier vaut pour la description du résultat obtenu à chacune des trois premières tâches et comprend trois modalités descriptives : « 0 » signifie l'échec, « 1 » signifie la réussite partielle et « 2 » la réussite totale de la tâche. Le second lexique vaut pour la description du résultat à chacune des tâches n° 4 à 8 et comprend deux modalités : « 0 » signifie l'échec et « 4 » signifie la réussite. Enfin, le troisième lexique vaut pour la description du résultat à chacune des six dernières tâches et comprend cinq modalités : « 0 » signifie l'échec et les chiffres « 4 », « 5 », « 6 » et « 7 » signifient des degrés croissants de réussite. Il n'est pas nécessaire pour le propos de préciser comment le psychologue utilise ces lexiques ; supposons seulement que les psychologues qui pourraient effectuer la description d'une certaine performance (qu'on aurait filmée par exemple) soient interchangeables – le test est réputé « cotation-objectif ».

2.2. Le principe de mesurage : éléments théoriques

Ce qui précède définit le cadre descriptif des phénomènes empiriques qu'on peut décrire « cotation-objectivement » avec le test. Voyons maintenant comment on peut imaginer un principe général permettant de relier la grandeur théorique visée par le test à l'ensemble des performances observables. Par « observables », il faut entendre « qui peuvent être observées lorsqu'on procède à une observation », par opposition à l'énumération de toutes les possibilités logiques générées par le langage descriptif de la performance, qui ne dépend d'aucune observation, et qui constitue l'ensemble de réponses *logiquement* possibles[5], par opposition à l'ensemble des réponses *empiriquement* possibles

[4] En réalité, la passation du test obéit à ce qu'on appelle une règle de départ et une règle d'arrêt, ce qui signifie que dans certaines conditions, la performance n'est pas un 14-uplet, auquel cas le test n'est pas complètement standardisé. Il n'est pas nécessaire de tenir compte de cette particularité ici.

[5] Le référentiel du test (Vautier, 2011; 2013).

(i.e., celles qui s'observent en fait). Il faut imaginer un principe pour chaque tâche, avant de résoudre le problème à l'échelle de la description 14-variée. Comme il s'agit de démontrer que la construction théorique requise pour fonder l'idée que la performance au test mesure une grandeur psychologique est fausse, il ne sera pas nécessaire de développer toute la démarche. Il suffit d'en développer une partie et de montrer que cette partie est fausse pour que la théorie complète, qui contient la théorie partielle, soit fausse.

Considérons une tâche dont le résultat est décrit selon le lexique {0, 4}, qui est le plus parcimonieux des lexiques du test Cubes. On suppose une grandeur psychologique que le résultat à la tâche n° 4 permet de mesurer. Cette grandeur possède par hypothèse une origine naturelle, qu'on peut noter O, en posant qu'elle désigne l'absence de quantité – on admet qu'une quantité négative d'aptitude n'existe pas. On peut aussi considérer que la grandeur possède un maximum qu'on notera « max ».

Ainsi, le problème consiste à définir une relation du segment [O, max] dans la paire {0, 4}. À tout point de la grandeur, on veut faire correspondre une valeur observable. On veut aussi que tout point de la grandeur ne corresponde qu'à une valeur, sinon cette relation ne pourrait pas être utilisée comme un principe de mesurage. On veut donc une *application* de [O, max] dans {0, 4}. Enfin, comme la valeur descriptive « 4 » indique par définition un niveau théorique supérieur à celui qu'indique la valeur descriptive « 0 » dans {0, 4}, cette application doit être croissante.

La seule solution possible est une fonction par palier. Ainsi, dire que le lexique de la tâche n° 4 *mesure* la grandeur revient à invoquer une fonction à deux paliers, les deux paliers étant séparés par un seuil dans [O, max], dont on ignore la valeur. Lorsque, par une « expérience de pensée », on fait varier la grandeur de O jusqu'au seuil, on pose qu'on observe le résultat « 0 » ; quand la grandeur dépasse le seuil et varie jusqu'à son maximum, on pose qu'on observe le résultat « 4 ». Cette construction théorique n'est pas falsifiable, puisqu'on ne connaît pas la valeur de la grandeur et qu'on peut toujours observer soit « 0 », soit « 4 ». Mais elle fournit un cadre logique pour relier intelligiblement le lexique descriptif de la réponse à la tâche et la grandeur que la réponse est supposée mesurer.

On applique la même démarche à la tâche n° 5, en inventant un autre seuil. La question qui se pose maintenant est de savoir comment ordonner les deux seuils sur le segment [O, max], étant donné qu'on suppose que les deux tâches mesurent la même quantité théorique. Notons A et B les deux seuils respectifs. Le langage de la grandeur implique que soit A < B, soit A = B, soit B < A. Comme la tâche n° 4 est supposée plus facile que la tâche n° 5, A se trouve avant B. En effet, l'ordre de difficulté des deux tâches implique la possibilité qu'un enfant possède une quantité théorique telle qu'elle lui permet de réussir la tâche n° 4 mais pas la tâche n° 5. Dans ce cas, cette quantité théorique est supérieure à A et inférieure à B. Donc A est inférieur à B. En d'autres termes, la performance (4, 0) signifie que la quantité théorique de l'enfant se trouve après A – d'où le « 4 » de (4, 0) – et avant B – d'où le « 0 » de (4, 0).

Une conséquence capitale découle de ce qui précède. Cet enfant ne peut théoriquement pas exhiber la performance (0, 4), puisque s'il réussit la tâche n° 5,

c'est que sa quantité théorique est supérieure au seuil B, et donc qu'elle est aussi supérieure au seuil A[6]. D'après la fonction par palier de la tâche n° 4, on devrait observer une réussite et non pas un échec à la tâche n° 4.

2.3. La falsifiabilité du principe de mesurage et ses conséquences techniques

Nous disposons d'un cadre logique pour relier la grandeur théorique et la performance observable avec deux items, et ce cadre d'interprétation possède un falsificateur, qui est l'observation (0, 4). Donc la théorie est falsifiable, ou encore testable (Popper, 1973). Si on admet que la quantité théorique peut varier lorsque l'enfant passe d'un item à l'autre, la théorie n'est plus falsifiable mais tautologique. Supposons qu'on considère maintenant que la théorie s'applique à tout enfant satisfaisant un certain nombre de conditions (conditions initiales). On peut alors énoncer la loi suivante : quel que soit un enfant dans ces conditions, il ne peut pas produire la performance (0, 4) puisque la performance mesure sa quantité théorique selon la fonction de mesurage que nous venons d'élaborer. Autrement dit, nous venons de dire que la probabilité d'observer l'événement (0, 4) dans ces conditions est nulle (pour une élaboration de la notion de loi en psychologie, voir Vautier, 2011; 2013; Vautier, Lacot, & Veldhuis, in press).

Supposons que des observations, nombreuses, corroborent cette prédiction – aucun « 04 » n'a été observé. Alors le langage de la grandeur théorique est une commodité linguistique pour énoncer cette loi empirique[7] de manière concise (pour une analyse de la fonction descriptive de la théorie en physique, voir Duhem, 2007). Nous ne savons pas si la grandeur existe en tant que telle, nous savons seulement que la fonction de mesurage dont elle constitue le domaine de définition est un modèle commode et prometteur. Supposons maintenant que quelques observations falsifiantes aient été rapportées dans la littérature scientifique[8]. Un nouveau problème scientifique se pose : de quoi d'autre que la quantité théorique dépendent de telles observations ? Quelles que soient les solutions envisageables, l'existence du problème crée un impératif technique : ce qui est observé est une anomalie au regard de la théorie (ou fonction) de mesurage ; un argument qui interprète ce qui est observé en termes de niveau de la grandeur théorique n'est pas *valide*. L'exploitation de la technique de mesurage doit prendre en compte le fait que parfois, les données sont aberrantes. Il ne s'agit pas d'une erreur de mesure au sens d'un manque de précision conduisant à la nécessité d'utiliser un encadrement de la valeur théorique plutôt qu'une valeur ponctuelle, mais d'une aberration théorique qui nécessite une élucidation parce qu'elle signale qu'on ne comprend pas ce qui se passe. Dès lors, une précaution élémentaire consiste à ne pas qualifier ces observations comme des données

[6] Ici, on a besoin de postuler que la quantité que l'on veut mesurer varie de manière négligeable entre le moment où l'enfant traite l'item n° 4 et le moment où il traite l'item n° 5.

[7] Cette loi structurale est aussi connue sous le nom d'échelle de Guttman (1944).

[8] Ce qui, soit dit en passant, est quasi-impossible étant donné que les politiques éditoriales des revues d'évaluation quantitative en psychologie décrètent que ce type d'étude manque de portée.

exploitables et l'utilisateur doit affirmer clairement qu'il ne peut rien conclure de ses observations parce qu'elles sont théoriquement inintelligibles – la performance ne dépend pas que de la quantité théorique, donc la théorie ne « marche pas ». Supposons enfin que de nombreuses observations falsifiantes aient été rapportées. Alors l'intérêt scientifique de la théorie de mesurage est négatif : on a appris qu'une telle construction théorique est fausse, ce qui constitue une authentique connaissance scientifique – une connaissance en creux.

2.4. Incertitudes à propos de l'incertitude

L'incertitude est une notion vague tant qu'on ne précise pas sur quoi elle porte. Lorsqu'on dispose d'un modèle de mesurage ordinal corroboré qui est fondé sur des observations multivariées, on dispose d'une théorie générale dont la vérité est incertaine. La généralité de la théorie est limitée à la population des êtres qu'on peut évaluer. Par exemple, la proposition « quels que soient les enfants qui rempliraient certaines conditions (l'âge, et d'autres attributs descriptifs liés à la manière dont se déroule la passation du test), la performance observée serait 00, 40 ou 44 » est une proposition générale. Comme la proposition est contrefactuelle, le nombre d'unités d'observation est infini et on ne peut donc pas vérifier la proposition unité par unité. On sait au mieux qu'un certain nombre de tests (au sens poppérien du terme) corroborent cette proposition. Face à l'incertitude irréductible de cette proposition, on se contente de considérer qu'elle est vraie jusqu'à preuve du contraire.

Supposons qu'on décide de croire en une telle loi parce qu'elle a toujours été corroborée. Une autre incertitude s'y attache, qui prend la forme d'une indétermination intrinsèque. La gradation des performances 00, 40 et 44 définit trois segments sur [O, max], dont on connaît l'ordre mais pas l'étendue, ce qui implique qu'on est rigoureusement incapable d'assigner une valeur numérique à la performance. En effet, on ne sait pas définir ce que signifie expérimentalement, ou encore empiriquement, la proposition théorique « 1 + 2 = 3 » par exemple, parce qu'on ne dispose d'aucune unité de mesure dotée d'une signification expérimentale (ou empirique, ou encore opératoire). C'est pourquoi la psychotechnique peut viser l'objectif de découvrir un mesurage ordinal, c'est-à-dire une loi d'ordre simple sur le domaine des performances observables – ce qui correspond à l'affirmation de l'impossibilité d'au moins un type de performance logiquement possible (Vautier, 2011; 2013). Si on additionne les valeurs ordinales des données dans les couples 00, 40 et 44, on obtient respectivement 0, 4 et 8, mais il est évident que la proposition « 0 + 4 = 4 », par exemple, est scientifiquement absurde bien que mathématiquement vraie. L'addition n'a pas de sens psychologique. Ces « scores » signifient seulement que la grandeur détectée par l'observation « 0 » (i.e., 00) est plus petite que la grandeur détectée par l'observation « 4 » (i.e., 04), qui est elle-même plus petite que la grandeur détectée par l'observation « 8 » (i.e., 44). Supposons enfin qu'on augmente le nombre d'items jusqu'à un nombre *m* et qu'il soit possible d'identifier une fonction par palier à *m'* seuils (le nombre de seuils dépendant du nombre de valeurs descriptives associées à ces items). On aura affiné le grain de l'échelle

ordinale, mais la mesure demeurera ordinale, c'est-à-dire que l'addition des scores demeurera une absurdité psychologique (ou scientifique).

Les psychométriciens savent bien que la fonction par palier qui est nécessaire pour fonder l'idée d'un mesurage ordinal d'une grandeur à l'aide d'une performance multivariée est fausse (e.g., Bertrand, El Ahmadi, & Heuchenne, 2008; Borsboom, 2008). Il serait tout de même surprenant qu'un phénomène aussi complexe qu'une performance à un test d'aptitude obéisse à un principe aussi simple, qui revient à expliquer la performance à l'aide d'une seule « variable latente » (ou théorique). Le fait que le modèle soit faux nous apprend (i) que pour expliquer ces phénomènes, une théorie plus riche est nécessaire et (ii) qu'il est impossible de déduire de l'observation des performances quoique ce soit en terme de niveau de la grandeur.

Mais, au lieu de prendre acte de ces connaissances pour clamer que le jugement évaluatif ne peut être fondé sur nos connaissances scientifiques faute de mesurage, et, éventuellement, pour encourager un programme de recherche ciblé sur les processus de réponse à des items de tests, les psychométriciens ont conservé l'impératif d'une interprétation unidimensionnelle et quantitative de la performance. Pour ce faire, ils ont modifié le modèle théorique que je viens de développer en introduisant la notion de probabilité d'observer telle réponse conditionnellement à telle valeur numérique de la grandeur, laquelle est définie sur une échelle d'intervalle grâce notamment au postulat de l'existence d'une fonction caractéristique de l'item (cf. Fischer, 1995). Cette transition est explicitée par Bertrand et al. (2008), ce qui permet d'examiner comment ils la justifient (pour une analyse des conséquences catastrophiques pour la falsifiabilité de la théorie à l'échelle individuelle, voir Vautier, Veldhuis, Lacot, & Matton, 2012).

Suivons les auteurs pas à pas.

> « Si la modélisation de la réussite de sujets à des items veut être réaliste, la théorie précédente est trop abrupte et doit être assouplie. Il est exceptionnel qu'une échelle de Guttman, à cause de sa rigidité, s'applique parfaitement aux données expérimentales » (p. 31).

La signification de l'adjectif « réaliste » est ici non pas descriptive, mais opérative. Il faut opérer malgré l'ignorance dans laquelle les descriptions dont nous disposons nous plongent. Le réalisme invoqué est en fait un appel à la soumission à la demande sociale d'un savoir fondateur de l'évaluation. Du point de vue scientifique, l'existence d'anomalies théoriques est reconnue, mais pas leur fréquence, ni leur caractère falsifiant, ni l'invalidation de l'inférence théorique à partir du modèle et des données. Les auteurs poursuivent en adoptant la position suivante :

> « Il est naturel d'interpréter les écarts au modèle en concédant un caractère aléatoire à la relation empirique 'réussir' de S [ensemble des sujets, unités d'observation] vers I [ensemble des items]. Ce caractère aléatoire est dû aux autres variables – non explicitement prises en compte comme la compétence des sujets et la difficulté des items – qui peuvent influencer la réussite ou l'échec ; on imagine aisément qu'elles sont

> complexes et nombreuses : humeur du sujet, environnement physique et social, mode de présentation de l'item, etc. […] » (p. 31).

Les auteurs reconnaissent explicitement que les performances dépendent d'une multitude de causes inconnues tant d'un point de vue théorique que pratique. Mais je ne vois pas en quoi cette ignorance implique que la performance observée doive être conçue comme le résultat d'une expérience aléatoire (pour une introduction à la notion d'expérience aléatoire, voir Falmagne, 2003). Il me semble que les auteurs confondent les probabilités subjectives, qui servent à jauger la confiance qu'on a en certaines propositions, et les probabilités objectives, qui supposent le postulat d'une indétermination intrinsèque des phénomènes (pour une introduction aux problèmes d'interprétation de la notion de probabilité, voir Hacking, 2002). Avant d'en tirer les conséquences, poursuivons encore avec eux.

> « On conçoit donc que dans la situation où un sujet s est confronté à un item i, la réussite de i par s, au lieu d'être toujours réalisée quand $\gamma(s) \geq \delta(i)$, jamais quand $\gamma(s) < \delta(i)$ [$\gamma(s)$ et $\delta(i)$ désignent respectivement la position de s et de i sur le domaine de la grandeur], est gouvernée par une tendance floue : la réussite (et son contraire l'échec) a une certaine probabilité de survenir. Désormais ce n'est plus le fait de réussir, mais les chances de réussir qui seront fonctions de $\gamma(s)$ et $\delta(i)$. La probabilité que s réussisse i, notée $\pi(s,i)$, dépend de la compétence de s comme de la difficulté de i. » (p. 31).

Les anomalies théoriques sont maintenant éliminées par un récit probabiliste qui invente une « tendance floue » à produire telle ou telle performance. Le « flou » s'exprime par des probabilités et ce qu'il y a de permanent dans la « tendance » est sous-tendu par la grandeur nommée « compétence ».

Les psychométriciens ont poussé l'art de la rhétorique jusqu'à appeler les modèles psychométriques des « modèles de mesure ». L'intuition quantitative est sauvegardée mais c'est au prix d'un renoncement à la connaissance théorique. La préférence pour la grandeur comme cadre conceptuel assimilateur dépasse l'intérêt pour la compréhension proprement scientifique de la performance, laquelle n'a finalement qu'un rôle auxiliaire. On pourra désormais *estimer* une valeur numérique, ce qui n'est pas *mesurer*, quand bien même on admet volontiers que cette valeur ne permet pas de comprendre comment la performance a été produite puisque, étant donnée n'importe quelle valeur de la grandeur, toute performance peut être observée – avec une probabilité plus ou moins importante, mais jamais égale à 0 ni à 1. Autrement dit, on accepte l'opacité de la performance et on assigne des valeurs numériques à des performances dans une ignorance que pour ma part je ressens comme vertigineuse[9]. En dépit de

[9] En particulier, l'estimation de la valeur numérique de la grandeur à un instant *t* pour une personne donnée est logiquement différente de la tendance centrale de la grandeur pour cette personne, si tant est qu'une telle notion admette une interprétation psychologique. Rien n'exclut que la tendance centrale soit 'significativement' différente de l'estimation issue d'une performance ponctuelle.

« l'évidence », la psychométrie moderne a sauvé l'entreprise comparative du fait qu'on ne sache mesurer de manière ordinale aucune grandeur psychologique, en substituant l'estimation statistique au mesurage expérimental.

3. Les praticiens des tests peuvent-ils revendiquer une fonction d'évaluation *et* une responsabilité scientifique ?

Dans la partie précédente, j'ai montré (i) comment une fonction par palier reliant la grandeur théorique à une performance multivariée permet de comparer des performances distinctes, et (ii) pourquoi un tel modèle est certainement faux, ce qui implique que l'interprétation ordinale de la performance observée n'est pas valable du point de vue logique faute de modèle alternatif valable. De plus, j'ai montré que le recours aux probabilités pour sauvegarder l'intuition quantitative via la modélisation psychométrique consiste à voiler notre ignorance des déterminants de la performance pour satisfaire un impératif non scientifique qui va être discuté ici (dans un contexte plus large, voir aussi Pestre, 2013, chapitre 3). Ce type d'analyse suggère que l'évaluation psychotechnique des aptitudes constitue un métier extraordinairement ingrat, parce que le praticien doit spécifier comment il articule le besoin social d'assimiler les personnes à des organismes dotés de diverses formes de capacités intellectuelles, conçues comme des grandeurs empiriquement indéterminées mais essentielles pour l'évaluation des individus – « les construits » –, et les questions, toujours ouvertes, (i) de ce qui, dans des conditions particulières (décrites au mieux grossièrement), détermine les performances aux items des tests d'aptitude[10], et (ii) de ce que ces performances déterminent à leur tour.

L'impératif qui motive la méthodologie du scorage des performances intellectuelles est le même que celui qui institue la notation scolaire, ou qui conduit Duhem (2007, partie 2, chapitre 1), dans une analyse lumineuse de la quantité et de la qualité, à tenter, sans y parvenir, de justifier le mesurage de la qualité « être un bon géomètre ». Cet impératif est formulé de façon concise par Perron dans une des discussions de la Conférence de consensus sur l'examen psychologique de l'enfant et de l'adolescent (Voyazopoulos, Vannetzel, & Eynard, 2011) : « [les scores] sont des jugements comparatifs de valeur qui se répercutent au niveau sociologique général, dans une société qui a besoin de hiérarchiser, à l'école ou dans l'entreprise, au niveau de la micro-sociologie et au niveau des jugements de valeur que l'individu porte sur lui-même » (p. 234). Michell (2003b) propose une analyse historique de ce qu'il appelle l'impératif quantitatif, mais sans développer la fonction sociale de l'évaluation, laquelle nécessite seulement la projection des performances à évaluer sur une échelle

---

Mais cette incertitude est en quelque sorte renvoyée dans le « pré-conscient épistémologique » du chercheur puisque pour pouvoir l'analyser empiriquement, il faudrait savoir mesurer la grandeur (Vautier et al., 2012).

[10] « […] il est peu réaliste de penser qu'on est aujourd'hui capable d'expliciter les mécanismes psychologiques susceptibles de générer les réponses aux items d'un test ou d'un questionnaire » (Juhel, Gilles, Bouvard, Boy, Fouques, Guimard et al., 2011, pp. 186-187).

ordinale (Vautier et al., 2012). Coombs (1964, chapitre 13) formule clairement la nécessité sociale de compresser les observations sur une « ligne de décision ». Cette nécessité est comparative. Si deux performances sont incomparables, alors deux personnes représentées par ces performances sont aussi incomparables, ce qui est rédhibitoire du point de vue de la satisfaction de la demande sociale.

La dynamique de l'examen psychologique des aptitudes repose sur deux aspirations téléologiquement distinctes, évaluer vs. comprendre, potentiellement incompatibles. Comme les performances observables ne sont pas simplement ordonnées, les méthodologies déployées sont incompatibles dès lors que la première opère un forçage descriptif par le scorage. On ne peut alors pas évaluer et comprendre la performance dans le même mouvement. Pour l'évaluer, il faut la scorer, c'est-à-dire la faire littéralement disparaître sous le nombre, lequel tirera sa signification d'une référentialisation que Danziger (1987; 1990) qualifie de galtonienne – l'étalonnage du score, ou encore le rapport à une distribution de référence. Tandis que pour comprendre la performance, il faut, adoptant une posture expérimentale, en découvrir les tenants – variables indépendantes –, ce qui suppose de s'appuyer pleinement sur le langage descriptif qui permet d'identifier les changements intervenant au niveau de la variable dépendante. Du point de vue temporel, le calcul du score et son interprétation normative (ou, de manière synonyme, évaluative) prennent un instant, tandis que l'investigation de ce qui et de ce que détermine la performance est un effort de pensée, qui conduit peut-être à quelques spéculations ou hypothèses dans le cadre même de l'examen, ou davantage si on dispose de connaissances générales pertinentes – sinon, pas plus, parce que les personnes qui viennent se faire évaluer n'entrent pas *de ce fait* dans un programme de recherche particulier.

Le fait que la communauté des praticiens de l'évaluation psychotechnique cherche une légitimation de cette pratique dans la doctrine de la validation des tests (e.g., Juhel et al., 2011), me paraît constituer un obstacle épistémologique (Bachelard, 1983) majeur au progrès de la connaissance scientifique de ce qui détermine la réponse aux items de tests. La doctrine de la validation des tests est organisée pour ne pas accuser réception de l'argument présenté dans la première partie de cet article, parce qu'elle tient la mesurabilité des grandeurs psychologiques comme un postulat fondateur, quitte, renonçant à l'exigence méthodologique de la valeur logique (soundness) des conclusions qu'elle tirerait des observations individuelles, à aménager la signification des notions de mesurage[11], de validité[12] et de généralité[13] pour les usagers d'un îlot sociotechnique, spécialisés dans l'évaluation des aptitudes.

---

[11] Mesurer c'est attribuer un nombre.

[12] Un test valide est un test qui mesure bien ce qu'il est censé mesurer ; un argument valide est un argument approximativement vrai ; pour des discussions 'orthodoxes' de la validité en psychologie, voir par exemple Cizek (2012), Kane (2006) ou Newton (2012) et pour des discussions critiques, voir Borsboom, Cramer, Kievit, Scholten et Franic (2009) et Michell (2009; 2013).

[13] Dans la psychologie dite, abusivement, « nomothétique », le général n'est plus ce qui s'applique à toute unité d'une classe de référence, mais ce qui particularise

Ce qui est validable dans le cadre de cette doctrine, ce sont des propositions qui portent sur les conséquences de l'application du postulat de mesurabilité des grandeurs psychologiques à l'échelle de populations statistiques. L'immense mobilisation académique qu'entraîne la validation des tests aboutit à proposer des tâches typiques (parfois confidentielles comme dans le cas du WISC-IV), des règles de scorage, des espaces dimensionnels (ou factoriels) associés à des termes parlants mais vagues (les construits – e.g., le Raisonnement Perceptif), et des relations entre variables statistiques. Ce discours est capable d'assimiler toute unité d'observation – toute personne à telle date – comme point parmi une infinité de points possibles, mais la trajectoire individuelle de ces points est reconnue comme totalement indéterminée, tout en étant solennellement estimée. D'où, dans le contexte de l'examen psychologique qui est une situation individualisée, l'impression que « quelque chose cloche ».

La lecture du manuel d'interprétation du WISC-IV (Wechsler, 2005b) permet sans surprise de cueillir quelques spécimens du brouillage practico-conceptuel qui résulte de la subordination réciproque des projets sociotechniques (observer pour évaluer) et scientifique (observer pour comprendre) qui est opérée par la psychotechnique contemporaine, dont le dernier sort perdant parce qu'il devient impossible de renoncer à la mythologie des grandeurs psychologiques que suggère le langage profane. Je me bornerai à rapporter quatre erreurs conceptuelles dignes d'un enseignement d'épistémologie de la psychologie pour des étudiants de licence.

3.1. Quand le score vaut pour le fonctionnement cognitif

On lit dans le manuel que « L'usage des notes standard normalisées en fonction des âges permet au praticien de comparer le fonctionnement cognitif de chaque enfant avec celui des enfants du même âge » (Wechsler, 2005b, p. 85). On peut toujours comparer un nombre à un nombre moyen. Par exemple, une note qui se trouve à 1,2 écart type de la moyenne de référence est supérieure à cette moyenne. Découle-t-il de ce fait algébrique que l'enfant à qui on vient d'attribuer cette note a un fonctionnement cognitif supérieur à celui de l'enfant moyen auquel il est comparé ? Non parce qu'il est évident que le fonctionnement cognitif n'est pas une grandeur mais un processus, lequel, par ailleurs, échappe largement à la psychologie scientifique contemporaine parce qu'elle s'intéresse à un sujet fictif.

On ne peut être d'accord avec l'affirmation du manuel sur une base strictement logique. Mais on peut être d'accord avec le manuel si on prétend aboutir à une évaluation du fonctionnement cognitif de l'enfant, ce qui est autre chose. Dans ce cas, le pouvoir du verbe, qui permet qu'un nombre vaille pour un processus pourvu qu'on se focalise sur la *valeur* de ce processus, joue à plein, et nous avons quitté le registre de la pensée scientifique parce qu'il n'est plus question de décrire mais d'évaluer.

3.2. Quand le score vaut pour la performance

---

la classe de référence (Danziger, 1987; 1990; Lamiell, 1998; Salvatore & Valsiner, 2010; Vautier, 2011; 2013).

Dans la même page du manuel, on lit « Une note standard représente la performance d'un enfant à un subtest comparativement à la performance de ses pairs du même âge » (Wechsler, 2005b, p. 85). Le même processus de projection de significations sur le nombre est proposé, puisque cette fois-ci le nombre représente une performance. Nous avons vu que la performance se décrit comme un vecteur (un *m*-uplet), ce qui implique qu'un nombre représente une performance seulement si les performances qu'on peut observer sont simplement ordonnées. Si, comme c'est probablement le cas, les performances qu'on peut observer ne sont pas simplement ordonnées, une note standard représente une performance seulement s'il existe une communauté d'utilisateurs qui veut bien se donner une telle convention, et alors ce n'est plus la performance qui est représentée mais ce qu'elle vaut. La note ne représente que la place qu'on attribue à l'enfant via sa performance lorsqu'elle est cryptée dans une échelle de valeurs.

3.3. Le score est l'apparence trompeuse d'une vérité cachée

Les promoteurs du WISC-IV écrivent encore que « La note vraie est le reflet de la véritable aptitude du sujet, combinée avec un certain degré d'erreur de mesure » (Wechsler, 2005b, p. 87), en référence à la théorie classique des tests qui postule que le score observé est la somme d'un score vrai et d'une erreur de mesure. Ici, on est abasourdi par l'illogisme et le flou du discours. Supposons qu'on admette l'existence de la « véritable aptitude du sujet », et qu'on admette aussi que la note vraie, c'est-à-dire la moyenne de tous les scores qu'aurait pu avoir le sujet, soit identique à – et non pas « reflète » – sa véritable aptitude (laquelle, donc, serait une quantité objective). Alors ce qui est « combiné avec un certain degré d'erreur de mesure » n'est pas la note vraie mais le score observé.

Vautier, Lacot et Veldhuis (in press) ont analysé le rôle interprétatif de la notion d'erreur de mesure dans le paradigme néo-galtonien. Dans un modèle psychométrique, l'erreur de mesure garantit qu'on ne pourra jamais ne serait-ce qu'encadrer la valeur de la grandeur théorique qu'on projette sur le score. Ainsi, l'interprétation psychométrique protège de toute falsification le postulat de l'existence de la grandeur. Ce postulat est intimement lié à la nécessité de l'évaluation, puisqu'il garantit la comparabilité au prix d'une incertitude irréductible due à l'erreur de mesure. Le désir qui projette une grandeur là où on ne sait pas la mesurer est conforté dans sa toute puissance par l'acceptation de la fatalité de l'erreur de mesure inhérente au scorage de la grandeur.

3.4. Le complément d'objet direct du verbe « mesurer » doit être une grandeur

Les incohérences conceptuelles se superposant les unes aux autres, l'utilisateur des tests est progressivement acculturé à un discours qui emploie les verbes « mesurer » et « évaluer » de façon interchangeable. Ainsi, le verbe « évaluer » admet des compléments d'objet direct (COD) qui ne sont pas nécessairement des grandeurs, tandis que le verbe « mesurer » n'admet pour COD que des grandeurs.

Ce détournement de sens est pratique courante dans le manuel du WISC-IV. Par exemple, « L'Indice de Compréhension Verbale du WISC-IV est une mesure de la formation de concepts verbaux, du raisonnement verbal et des

connaissances acquises dans le propre environnement du sujet » (Wechsler, 2005b, p. 89). La formation de concepts verbaux etc. ne dénote pas une grandeur. Dans la phrase « L'ICV actuel peut être considéré comme une mesure affinée et plus pure du raisonnement verbal et de la conceptualisation que […] » (Wechsler, 2005b, p. 89), le raisonnement verbal n'est pas une grandeur, pas plus que le raisonnement perceptif et fluide qu'on trouve dans la citation : « L'Indice de Raisonnement Perceptif est une mesure du raisonnement perceptif et fluide, du traitement spatial et de l'intégration visuomotrice » (Wechsler, 2005b, p. 89)[14].

3.5. Pourquoi ne pas assumer que l'évaluation psychologique soit une sociotechnique ?

Une attitude qui me semble propice pour que les praticiens des tests ne soient pas empêtrés dans le « désir psychométrique » consiste à considérer que la légitimation de l'évaluation psychotechnique relève *aussi* d'une science des contraintes sociales. Par exemple, c'est une contrainte sociale : on n'entre pas dans une formation d'élève pilote de ligne à l'Ecole Nationale de l'Aviation Civile si on occupe une position jugée rédhibitoire dans un certain espace évaluatif. Une telle anthropologisation de l'évaluation psychotechnique, qui se présenterait alors comme une sociotechnique sans mesurage, permettrait au chercheur en psychologie d'étudier la performance aux tests, héritage précieux pour l'observation des performances humaines s'il en est, avec la liberté idéologique d'en interroger tant l'indétermination que la détermination, ce qui supprimerait l'utilité du refoulement des faits falsifiants évoqués plus haut au sein même de la communauté concernée. Elle permettrait au praticien de tourner son attention vers les formes de manifestation du pouvoir social qui s'exercent sur l'individu qu'il inscrit dans un espace de valeurs dont il a l'expertise, et sur lui-même lorsqu'il se fait l'agent de l'évaluation (cf. Canguilhem, 1958).

Dans le cadre de la pratique clinique, je voudrais m'aventurer par le jeu de la simulation théâtrale dans la problématique qui consiste, pour le praticien, à ne pas tricher avec l'enfant qui lui est amené en ce qui concerne ce dont celui-ci est l'enjeu. Je vais mettre en scène un psychologue qui essaie de dire la vérité en répondant à un enfant curieux. En guise de précaution, j'ajoute que ce dialogue n'a pas de vocation normative ni réaliste. Pour les besoins de l'analyse, je laisse au psychologue la possibilité de consulter la documentation technique du test en présence de l'enfant.

Le test Cubes a été administré à un enfant de huit ans et un mois et la performance observée est la suivante : (2, 2, 2, 0, 4, 4, 0, 0, 0, 0, 0, 0, 0, 0).

---

[14] On peut aussi ajouter les deux exemples suivants. (1) « L'indice de Mémoire de Travail procure une mesure de la capacité de la mémoire de travail de l'enfant » (Wechsler, 2005b, p. 90). (2) « L'IVT [indice de vitesse de traitement] fournit une mesure de l'aptitude de l'enfant à inspecter rapidement et correctement des informations visuelles simples, à les traiter de manière séquentielle et à les discriminer […] Cette note composite est également une mesure de mémoire visuelle à court terme, d'attention et de coordination visuomotrice » (Wechsler, 2005b, p. 90).

L’enfant :

- C’est bien ?

Le psychologue :

- Ta note à ce test est de 14 points, ce qui fait une note standard de 6 points, d’après ce tableau [il montre la table d’étalonnage qui se trouve p. 210 dans le Manuel d’administration et de cotation (Wechsler, 2005a)].
- Mais est-ce que c’est bien ?
- La plupart des enfants de ton âge réussissent mieux.
- Alors c’est pas bien.
- Ta note est moins bonne que la note qui sert de point de repère.
- C’est parce que je suis pas assez intelligent ?
- Qu’est-ce que ça veut dire, être assez ou pas assez intelligent ? Tout dépend de ce que tu veux faire.
- Je ne veux pas aller à l’école.
- Ce que je peux te dire, c’est que les gens qui ont fabriqué le test l’on fait passer à des enfants qui avaient à peu près le même âge que toi, qu’ils ont calculé la moyenne de leurs notes, et que cette note moyenne est plus grande que ta note.
- Alors c’est normal.
- Qu’est-ce qui est normal ?
- Que ma note ne soit pas la moyenne, parce que tous les enfants ne peuvent pas avoir la même note.
- Oui. Ta note est plus petite.
- C’est quoi la moyenne ?
- C’est 10.
- C’est qui ces enfants ?
- Je ne sais pas exactement. Les gens qui ont fait le test indiquent, dans ce document [Manuel d’interprétation, p. 23 (Wechsler, 2005b)], qu’ils ont fait passer le test à 23 garçons et 23 filles de ton âge.
- C’était quand ?
- En 2004.
- Il y a longtemps. Peut-être que je suis aussi intelligent que 46 enfants de mon âge maintenant.
- Les gens qui ont fait ce test pensent sans doute que si on faisait passer le test à 46 enfants aujourd’hui, la note moyenne ne changerait pas beaucoup.
- Et toi, qu’est-ce que tu en penses ?
- Je n’en sais rien. Il y a beaucoup de choses qu’on ignore. Par exemple, on ignore ce qu’est l’intelligence et on ne sait certainement pas la mesurer.
- Oui mais tes tests, ils permettent de savoir que je suis moins intelligent que les autres.
- Ils permettent de savoir que ta note est plus petite que la note moyenne d’un certain groupe d’enfants. On utilise ce groupe d’enfants comme une image de l’enfant typique de ton âge.

- J’aime pas passer pour un imbécile et ma note dit que je suis un imbécile à ce test.
- Dans la vie, nous sommes tous, à un moment ou à un autre, comparés aux autres. Ta note sert à te comparer à d’autres enfants lorsque tu essaies de résoudre les problèmes du test.
- Ah, alors c’est comme sur un podium.
- Oui, tu n’es pas parmi les premiers.
- Comment tu le sais ?
- Parce que j’utilise un modèle qui me dit comment les enfants sont répartis sur le podium.
- Mais ce modèle, il est vrai ?
- Non. Mais puisque je veux te comparer aux autres, je l’utilise. Les psychologues qui font passer ce test font comme ça, nous utilisons le même modèle comme ça on parle de la même chose.
- Même s’il est faux ?
- Comme beaucoup de monde considère qu’il n’est pas trop faux, il devient vrai entre nous.
- Mais toi tu sais qu’il est faux.
- Oui.
- Alors ce n’est pas grave.
- Qu’est-ce qui n’est pas grave ?
- Que cette note me fasse passer pour un nul à ce test.
- Je vais quand même être obligé de communiquer ta note aux personnes qui m’ont demandé de te faire passer le test.

La mise en nombre de la performance au test – le 14-uplet – n’est pas permise par le fait qu’elle serait surdéterminée par une loi d’ordre simple (une loi de structure). Elle constitue cependant la condition de possibilité d’une comparaison sociale de l’enfant lorsqu’il se comporte dans une certaine scène sociale – le test. De ce point de vue, on peut saluer la clarté avec laquelle Reuchlin (1969) définit la finalité des tests : « ils fournissent les moyens d'exprimer ces observations [les réponses] sous une forme telle que soient possibles la comparaison [des] individus entre eux et la comparaison de chacun avec les "normes" (descriptives) de la population à laquelle ils appartiennent » (p. 22).

Si on veut éviter de contredire Huteau et Lautrey (1999) lorsqu’ils affirment que la mesure de l’efficience intellectuelle est fondée au niveau ordinal, on doit ajouter que ce fondement ne réside pas dans une loi expérimentale mais dans un impératif social : il faut comparer et pour comparer il faut ordonner, d’où les conventions nécessaires – le barème de notation et la normalisation des notes dites brutes.

La conséquence qui découle de cette analyse est que si la demande sociale d’examen psychologique n’exige pas une comparaison sociale, le psychologue n’a pas besoin d’utiliser de notes standard, lesquelles sont la seule justification des notes brutes (qui ne sont pas ontologiquement brutes mais bien socialement construites pour satisfaire l'impératif de comparabilité sociale, cf. Searle, 1995). Je

poursuis maintenant le dialogue entre le psychologue et l'enfant pour illustrer en quoi la demande sociale est une demande de comparaison sociale.

L'enfant :

- Pourquoi tu vas dire mes notes ?
- Parce que les gens qui se préoccupent de ton avenir ont besoin de cette information.
- Alors tant mieux s'ils croient que je suis un imbécile, j'irai pas à l'école.
- Ils ont besoin de décider dans quelle école tu vas aller.
- Ils vont me mettre dans une école pour imbéciles ?
- Je ne sais pas. Ils essaient de trouver quelle est la meilleure solution pour toi parce que ta maîtresse pense que tu risques d'avoir des difficultés si tu restes dans ton école.
- J'aime pas cette école.
- Penses-tu que tu serais capable d'avoir de bonnes notes si tu restais dans cette école ?

J'évite délibérément de poser la question de ce qui conduit l'enfant à ne pas aimer son école, pour introduire la question de l'interprétation non pas évaluative au sens de la comparaison sociale, mais diagnostique et pronostique, au sens de la mobilisation des connaissances scientifiques pertinentes, de la performance au test. Poursuivons le dialogue en l'infléchissant vers l'indétermination de la performance.

L'enfant :

- Je sais pas.
- Si tu as de bonnes performances aux tests, cela veut dire que tu es capable d'avoir de bonnes notes à l'école et on te fait confiance.
- Pourquoi ?
- Parce que, grosso modo, c'est la même intelligence qui te permet de faire les exercices aux tests et de faire les exercices à l'école.
- Oui mais t'as bien vu que je suis nul à ce test.
- Si tes performances aux tests ne sont pas bonnes, ça ne veut pas forcément dire que tu n'es pas capable d'avoir de bonnes notes à l'école.
- J'aimerais bien mais j'y arrive pas.

Qu'est-ce qui bloque la performance de *cet* enfant ? Quelles contraintes opèrent dans son fonctionnement psychologique ? Ces contraintes permettent-elles de prévoir un échec scolaire dans des conditions de scolarisation dont on ignore les détails ? Répondre à de telles questions en s'appuyant sur des connaissances scientifiques générales, au sens nomothétique du terme (Lamiell, 1998), paraît aujourd'hui prématuré.

Si on se réfère à la modélisation psychométrique, la réponse observée à un item est le résultat d'un processus aléatoire. Par exemple, l'échec enregistré à l'item n° 6 aurait pu, étant donné le score latent qu'on prête à l'enfant tout en

l'ignorant, être une réussite. Autrement dit, on ignore absolument la trame causale de l'échec à cet item à cette date. Le psychologue pourrait suggérer à l'enfant de recommencer la sixième tâche afin que tous deux puissent tenter de comprendre ce qui n'a pas fonctionné.

Quoiqu'il en soit, affirmer que l'enfant est incapable de réussir cet item ne serait pas une proposition valide faute d'une prémisse de laquelle la déduire. Une prémisse convenable serait l'absence d'une condition nécessaire à la réussite de l'item, mais, à ma connaissance, aucun manuel de test ne propose une liste de conditions nécessaires à la réussite des items. La performance au test n'a donc aucune signification diagnostique ni pronostique valable pour *cet* enfant (pour une analyse de l'invalidité d'un argument individuel fondé sur une prémisse statistique, voir aussi Lamiell, 2006; Vautier, 2012).

Dans une perspective proprement idiographique, le psychologue tient une position de détective. Il doit identifier les significations que les tâches proposées revêtent pour l'enfant. En particulier, il paraît essentiel d'identifier l'intérêt de l'enfant pour les performances tant scolaires que psychotechniques, même si ce type d'interprétation pose de sérieux problèmes méthodologiques. Un enfant dont on peut dire qu'il ne s'engage pas dans la tâche est un enfant qui refuse de montrer ce qu'il sait faire comme ce qu'il ne sait pas faire, auquel cas le psychologue devrait, me semble-t-il, rapporter qu'il n'a pas pu « observer l'intelligence » de cet enfant parce qu'il ne peut pas assurer que l'enfant s'est « approprié la situation d'examen »[15]. On voit bien que la problématique de l'évaluation des aptitudes possède une dimension signifiante complexe. La notation d'une performance qui n'est pas naturellement ordinale constitue une forme d'autorité sociale vis-à-vis de la personne évaluée. Le clinicien doit alors évaluer de quelle liberté il jouit en tant qu'agent de l'évaluation vis-à-vis de l'impératif évaluatif. La personne testée exhibe une performance dont la signification émerge dans un champ signifiant plus ou moins explicite pour la personne testée. On peut supposer que la détermination signifiante de la performance, comme support d'une comparaison sociale ou bien produit de l'activité cognitive développée par un organisme « motivé » par la réussite de la tâche, conditionne ce qui est effectivement exhibé par la personne testée.

4. Conclusion

Les tests d'aptitude ne sont pas des instruments de mesure parce que si c'était vrai, on saurait quels principes ou lois de mesurage opèrent chez les personnes soumises aux items de tests. En confondant le scorage et le mesurage,

---

[15] « R6 : L'enfant doit exprimer son accord et s'approprier la situation d'examen » (Voyazopoulos et al., 2011, p. 44). Il semble que ce type de critère nécessite un important travail théorique pour qu'il devienne opérationnel dans une perspective de description « cotation-objective ». Comment fait-on pour décider que l'enfant simulé dans le dialogue « s'approprie la situation d'examen » ? Un psychologue désireux de satisfaire des demandeurs peu sensibles aux subtilités de l'activité mentale, sera plus enclin à juger que l'enfant s'est approprié la situation d'examen qu'un psychologue qui travaille pour des demandeurs moins directifs.

les psychologues qui défendent la pratique du testage (de l'anglais testing) psychologique ne réussiront pas à faire admettre que leurs pratiques sont scientifiquement fondées s'ils s'adressent à des scientifiques, pas plus qu'à dissiper les doutes du profane qui, tout en faisant confiance au professionnalisme des psychologues, ne les crédite pas, à juste titre, d'une science du mesurage. C'est pourquoi je pense que la communauté des testeurs n'a pas grand-chose à perdre à reconnaître que l'évaluation psychotechnique soit une sociotechnique, c'est-à-dire un art de préparer des jugements évaluatifs selon les contextes dans lesquels on fait appel à leurs compétences, et n'a aucun devoir de défendre une conception exceptionnelle de ce en quoi consiste le mesurage (Michell, 1997; 2000).

En revanche, une telle lucidité dans la communauté des psychologues de tous bords libèrerait la psychologie scientifique de l'impératif de fonder sinon tous du moins une grande partie de ces programmes de recherche sur la mythologie des grandeurs psychologiques. Si la psychométrie statisticienne a besoin d'une telle mythologie pour développer ses modèles à variables latentes, la recherche scientifique en psychologie n'a pas nécessairement besoin des modèles psychométriques (Vautier et al., 2012). Elle a essentiellement besoin du droit à rendre compte de l'immensité de notre ignorance de ce qui cause les comportements qu'on sait observer (Vautier, 2012), laquelle ne peut qu'éclairer la spécificité de la tâche qui consiste à évaluer autrui dans un contexte social quelconque.

Références

Bachelard, G. (1983). *La formation de l'esprit scientifique* (12e ed.). Paris: Vrin.

Bertrand, D., El Ahmadi, A., & Heuchenne, C. (2008). D'une échelle ordinale de Guttman à une échelle de rapports de Rasch. *Mathématiques et Sciences Humaines, 4*, 25-46.

Binet, A., & Simon, T. (1907). Le développement de l'intelligence chez les enfants. *L'Année Psychologique, 14*, 1-94.

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives, 6*, 25-53.

Borsboom, D., Cramer, A., Kievit, R. A., Scholten, Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 135-170). Charlotte, NC: Information Age Publishing.

Canguilhem, G. (1958). Qu'est-ce que la psychologie ? *Revue de Métaphysique et de Morale, 1*, 12-25.

Chartier, D., & Loarer, E. (2008). *Evaluer l'intelligence logique : approche cognitive et dynamique*. Paris: Dunod.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification of test use. *Psychological Methods, 17*, 31-43.

Coombs, C. H. (1964). *A theory of data*. New York: Wiley.

Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Krüger, G. Gigerenzer, & M. S.

Morgan (Eds.), *The probabilistic revolution, Vol. 2: Ideas in the sciences* (pp. 35-47). Cambridge: MIT Press.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research*. New York: Cambridge University Press.

Duhem, P. (2007). *La théorie physique, son objet, sa structure*. Paris: Vrin.

Falmagne, J. C. (2003). *Lectures in elementary probability theory and stochastic processes*. Boston: McGraw Hill.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15-38). New York: Springer-Verlag.

Gaillard, F., Colasse, M., Guihard, C., & Michel, R. (2011). Pertinence et nécessité de l'examen psychologique de l'enfant et de l'adolscent. In R. Voyazopoulos, L. Vannetzel, & L.-A. Eynard (Eds.), *L'examen psychologique de l'enfant et utilisation des mesures : conférence de consensus* (pp. 125-178). Paris: Dunod.

Grégoire, J. (2009). *L'examen clinique de l'intelligence de l'enfant : fondements et pratique du WISC-IV* (2e ed.). Collines de Wavre: Pierre Mardaga Editeur.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139-150.

Hacking, I. (2002). *L'émergence de la probabilité*. Paris: Seuil.

Huteau, M., & Lautrey, J. (1999). *Evaluer l'intelligence. Psychométrie cognitive*. Paris: Presses Universitaires de France.

Juhel, J., Gilles, P.-Y., Bouvard, M., Boy, T., Fouques, D., Guimard, P. et al. (2011). Validité des modèles et des outils de l'examen psychologique. In R. Voyazopoulos, L. Vannetzel, & L.-A. Eynard (Eds.), *L'examen psychologique de l'enfant et utilisation des mesures : conférence de consensus* (. Paris: Dunod.

Jumel, B., & Savournin, F. (2013). *L'aide-mémoire du WISC-IV* (2e ed.). Paris: Dunod.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Washington, DC: American Council on Education/Praeger.

Lamiell, J. T. (1998). 'Nomothetic' and 'idiographic': Contrasting Windelband's understanding with contemporary usage. *Theory & Psychology, 8*, 23-38.

Lamiell, J. T. (2006). La psychologie contemporaine des "traits" dans le cadre de la recherche néogaltonienne : comment elle est censée fonctionner et pourquoi en réalité elle ne fonctionne pas. *Psychologie Française, 51*, 337-335.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology, 88*, 355-383.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10*, 639-667.

Michell, J. (2003a). Measurement: A beginner's guide. *Journal of Applied Measurement, 4*, 298-308.

Michell, J. (2003b). The quantitative imperative: Positivism, naïve realism, and the place of qualitative methods in psychology. *Theory & Psychology, 13*, 5-31.

Michell, J. (2009). Invalidity in validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 111-133). Charlotte, NC: Information Age Publishing.

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology, 31*, 13-21.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*, 1-29.

Pestre, D. (2013). *À contre-science : politiques et savoirs des sociétés contemporaines*. Paris: Seuil.

Popper, K. R. (1973). *La logique de la découverte scientifique*. Paris: Payot.

Reuchlin, M. (1969). *Les méthodes en psychologie*. Paris: Presses Universitaires de France.

Salvatore, S., & Valsiner, J. (2010). Between the general and the unique: Overcoming the nomothetic versus idiographic opposition. *Theory & Psychology, 20*, 817-833.

Searle, J. R. (1995). *The construction of social reality*. New York: The Free Press.

Vautier, S. (2011). The operationalisation of general hypotheses versus the discovery of empirical laws in Psychology. *Philosophia Scientiae, 15*, 105-122.

Vautier, S. (2012). Propos sur la responsabilité scientifique du psychologue. Essai d'épistémologie appliquée. *Pratiques Psychologiques, 18*, 373-383.

Vautier, S. (2013). How to state general qualitative facts in pychology? *Quality and Quantity, 47*, 49-56.

Vautier, S., Lacot, E., & Veldhuis, M. (in press). Puzzle-solving in psychology: The neo-Galtonian vs. nomothetic research focuses. *New Ideas in Psychology*.

Vautier, S., Veldhuis, M., Lacot, E., & Matton, N. (2012). The ambiguous utility of psychometrics for the interpretative founding of socially relevant avatars. *Theory & Psychology, 22*, 810-822.

Voyazopoulos, R., Vannetzel, L., & Eynard, L.-A. (2011). *L'examen psychologique de l'enfant et l'utilisation des mesures : conférence de consensus*. Paris: Dunod.

Wechsler, D. (2005a). *WISC-IV - Manuel d'administration et de cotation*. Paris: Les Editions du Centre de Psychologie Appliquée.

Wechsler, D. (2005b). *WISC-IV : manuel d'interprétation*. Paris: Les Editions du Centre de Psychologie Appliquée.

Conflit d’intérêt : néant.