HAL
open science

L'examen ou l'évaluation psychométrique : où en sont les praticiens ?

Stéphane Vautier, Nadine Mazet

To cite this version:

Stéphane Vautier, Nadine Mazet. L'examen ou l'évaluation psychométrique : où en sont les praticiens ?. 2019. halshs-02318261

HAL Id: halshs-02318261

https://shs.hal.science/halshs-02318261

Preprint submitted on 16 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

L'EXAMEN OU L'EVALUATION PSYCHOMETRIQUE : OU EN SONT LES PRATICIENS ?

Stéphane Vautier et Nadine Mazet
Université de Toulouse[1]

[1] Université de Toulouse Jean Jaurès, 5 allées Antonio Machado, 31058 Toulouse cedex 9. Courriel : vautier@univ-tlse2.fr

Résumé

Objectif : proposer un état des connaissances de l’examen psychométrique

Résultat : il est nécessaire de poser le problème de la valeur du savoir qu’il produit.

Argument :

1. L’examen psychométrique vise au moins un savoir
2. Or les scores ne sont pas des mesures au sens classique du terme
3. L’interprétation psychométrique, c’est-à-dire probabiliste, des scores, est aussi dépourvue de signification psychologique
4. Les scores décrivent (mal) un comportement ponctuel et par défaut transitoire de la personne
5. Le savoir visé par l’examen psychométrique ne porte pas sur le comportement mais sur une partie de l’identité ou du moi de la personne
6. Le savoir résultant de l’examen psychométrique est évaluatif, c'est-à-dire attributif
7. Son caractère attributif est douteux, scientifiquement tant qu’on ne sait pas contrôler expérimentalement la variation de la grandeur évaluée, politiquement, tant qu’on ignore son utilité.
8. Conclusion : s’il revendique une responsabilité professionnelle, le praticien doit répondre de ce qu’il sait à propos de la vérité scientifique de son savoir, le point est épistémologique et, aussi, de la fonction psychologique ou sociale de ce savoir, le point est politique.

Abstract

Goal: to provide a state of knowledge about psychometric assessment

Result: the value of the knowledge resulting from psychometric assessment is an issue.

Argument:

1. Psychometric assessment aims at some knowledge
2. But scores are not measurements in the classical sense of the term
3. Moreover, the psychometric, i.e., probabilistic, interpretation of scores, has no more psychological meaning
4. Scores depict (badly) the intermittent and by default transient behavior of people
5. Psychometric assessment does not aim at behavioral knowledge but at knowledge on part of identity or ego/self

6. The knowledge resulting from psychometric assessment is evaluative, that is, attributive
7. Its attributive character is doubtful, scientifically as long as one ignores how to control experimentally the variation of the assessed quantity, politically as long as one ignores its usefulness.
8. Conclusion: If the practitioner wants to assert her/his professional responsibility, she/he has to be responsible for what she/he knows about the scientific truth of her/his knowledge, the point is epistemological, and also, about the psychological or social function of this knowledge, the point is political.

L’examen ou l’évaluation psychométrique : où en sont les praticiens ?

L’état des connaissances proposé est un résumé de la situation (description en compréhension) et non pas un inventaire (description en extension). L’article développe chaque point de l’argument présenté en résumé de l’article. Les termes « examen » et « évaluation » psychométrique sont considérés comme synonymes.

1. L’examen psychométrique vise au moins un savoir

S’il ne visait aucun savoir, dans quel but l’examen serait-il entrepris ? Si la réponse à la question était "aucun but", l’examen ne serait pas entrepris. Or l’examen psychométrique est une pratique répandue, par exemple dans le cadre du bilan de compétences, du dépistage des troubles neuropsychologiques ou encore du recrutement. Donc l'examen psychométrique vise au moins un savoir. De quel savoir s'agit-il ? Les tests étant utilisés pour fournir des scores, quelles informations ces scores délivrent-ils ?

2. Or les scores ne sont pas des mesures au sens classique du terme

Si les scores étaient des mesures au sens classique du terme, c’est-à-dire au sens du mesurage de grandeurs théoriques permis par un appareil de mesure, appareil dont le fonctionnement suppose (condition nécessaire et non suffisante) que la variation du mesurant dépende de manière monotone de celle du mesuré, les réponses au test seraient comparables, ce que réfute régulièrement l’expérience. Pour plus de détails, voir Lacot, Afzali et Vautier (2016) et Vautier (2015).

Il existe une théorie générale et falsifiable du mesurage psychologique au sens classique du terme "mesurer" et cette théorie permet de savoir que l'hypothèse, conforme à l'état des connaissances empiriques, est qu'on ne sait mesurer aucune grandeur psychologique. Appeler des scores des mesures et les tests qui les permettent des instruments de mesure, ne suffit pas pour qu’on sache « prendre les mesures psychologiques » des gens.

3. L’interprétation psychométrique, c’est-à-dire probabiliste, des scores, est aussi dépourvue de signification psychologique

La théorie classique des tests, qui demeure le cadre de référence probabiliste de la grande majorité des tests dits validés, définit tout score comme résultat d’une expérience

aléatoire et postule l'existence d'une espérance mathématique, nommée "score vrai", par définition "mesurée" par ce score avec une "erreur" dite de mesure : score = score vrai + erreur de mesure[2]. Un jeu de mots suffit pour faire du score auquel on a soustrait son erreur un score sans erreur donc vrai. Mais contrairement à la mesure au sens classique, la théorie classique des tests ne fournit pas de lien logique entre le score vrai et la quantité de grandeur psychologique auquel l’utilisateur du test s’intéresse, comme par exemple une quantité d’intelligence générale (facteur g) ou d’anxiété. C'est que, pour des raisons d'impossibilité empirique, personne n'est utilisable pour permettre l'étude de la variation expérimentale de ses scores en contrôlant constante sa quantité d'intelligence par exemple. Le score vrai de la théorie classique des tests est un concept statistique *et* non-expérimental, c'est-à-dire vide de signification scientifique au sens d'une science du comportement des gens testés.

Par exemple, considérons un score, baptisé « score d’anxiété » parce qu’il est calculé à partir des réponses observées avec un questionnaire, lui-même baptisé « questionnaire d’anxiété » parce que les items décrivent un état anxieux. Le score vrai mesuré par ce score n’est pas conçu comme le mesurant d’une quantité d’anxiété, au sens où la variation du score vrai dépendrait de la variation de la quantité d’anxiété. Une telle conception est possible mais elle est par définition non-testable, donc non-corroborable (par l'expérience), puisque le score vrai n’est pas une quantité empirique (seuls les scores observés sont des données empiriques).

De plus, c'est pour nous une remarque nouvelle dans le contexte de la théorie classique des tests, quels que soient deux scores obtenus avec un même test, il existe une mesure de possibilité telle que la quantité de possibilité que le score vrai mesuré par le premier score soit supérieur au score vrai mesuré par le second score d’une part, et la quantité de possibilité opposée que le score vrai mesuré par le premier score soit inférieur au score vrai mesuré par le second score d’autre part, soient égales (0,5/1). La modélisation probabiliste des scores par la théorie classique des tests garantit que la comparaison de deux scores ne permet pas de comparer les scores vrais respectifs. Cette indétermination tient au fait qu’ayant accepté l’existence d’une loi de probabilité gouvernant l’expérience aléatoire dont le résultat est un

[2] Le score est axiomatiquement posé comme résultat d’une expérience aléatoire, laquelle est modélisée par une variable aléatoire, qui suppose qu’on accepte l’existence d’une loi de probabilité déterminant la probabilité de chaque score de l’échelle de scores utilisée.

score particulier, on ne sait pas identifier cette loi de probabilité[3]. Ainsi par exemple, la possibilité que le score vrai d'un QI de 130 points soit *supérieur* au score vrai d'un QI de 80 points est la même que la possibilité que le score vrai mesuré par 130 soit *inférieur* au score vrai mesuré par 80 (elle mesure 0,5/1).

Il résulte de 2 et 3 que si les scores sont entendus comme des mesures, ces mesures n’ont pas la signification (scientifique) d’une relation corroborée entre une grandeur empirique et une grandeur psychologique, telle qu’on saurait par exemple prendre les mesures d’anxiété et d’intelligence de Paul et Julie.

4. Les scores décrivent (mal) un comportement ponctuel et par défaut transitoire de la personne

Un test psychométrique propose un langage de description standardisé d’une réponse m-variée, où m est le nombre d’items du test. En termes éthologiques, cette réponse est un comportement. Le scorage consiste en une application d’agrégation qui, à chaque m-uplet possible, associe un et un seul score dans l’échelle de scores du test. Le cardinal de l’ensemble des m-uplets possibles est généralement d’un ordre de grandeur très supérieur à celui du cardinal de l’échelle des scores possibles. Les scores non-extrêmes ont donc plusieurs antécédents et ne permettent donc pas d’identifier la réponse au test.

Par exemple, si on considère une échelle de stress composée de 10 items dont la réponse est codée dans une échelle de Likert en quatre points (de 1 à 4 points), il existe 4^{10} 10-uplets possibles – l’ordre de grandeur est de un million – ayant chacun une et une seule image dans l’échelle de scores S = (10, 11, …, 40), dont le cardinal vaut 31. La « compression » (Coombs, 1964) effectuée par le scorage ne respecte pas la structure logique des réponses (comportements) observables – cette structure est le produit cartésien des 10 échelles de Likert. Le remplacement de ce produit cartésien par une échelle de scores procède d’une convention qui n’a pas de raison scientifique si on s’intéresse au comportement : comment un score (non-extrême) pourrait-il décrire ou encore représenter 10 réponses sans les défigurer ?

[3] Il est sûr qu’un modèle gaussien est inapproprié parce que l’intervalle de l’erreur de mesure est fini (Vautier, Veldhuis, Lacot, & Matton, 2012), ce qui invalide les calculs probabilistes discutés par Vautier et Gaudron (2002).

5. Le savoir visé par l’examen psychométrique ne porte pas sur le comportement mais sur l’identité ou le moi de la personne.

Éliminons la visée comportementale avant d’en venir à la notion psychologique d’identité. Qu’entendre par « comportement » ? S’agit-il du comportement au test ou bien du comportement prédit par le comportement ou le score au test ? Lorsque les tests sont présentés comme des techniques prédictives, les prédictions portent rarement sur des comportements et la validité prédictive est assez faible pour invalider qu’une prédiction soit utile au cas par cas. Donc si le savoir visé par l’évaluation psychométrique porte sur le comportement, c’est sur le comportement au test (par exemple en vue de déduire un diagnostic relatif aux processus qui sous-tendent le comportement au test).

Si le savoir visé par l’évaluation psychométrique portait sur le comportement au test, il porterait sur les réponses aux items du test (comportement) et non pas sur le score, qui ne décrit pas exactement ces réponses (cf. 4). De plus, la variabilité intra-individuelle relative aux comportements possibles au test n’est généralement pas l’objet de l’examen, puisque son observation nécessite un design longitudinal tandis que l’examen psychométrique est une approche interindividuelle (design transversal où l’on compare le score à une norme statistique). Une approche intra-individuelle supposerait aussi de prendre en compte la familiarisation progressive du sujet avec le test (ou, en cas de formes dites parallèles, du contenu du test), ce qui supposerait d’admettre un éventuel processus d’apprentissage (mémorisation) des items et des réponses, ce qui complexifie encore plus, si c’est possible, la théorisation des processus psychologiques (et non pas stochastiques dans le cas des modèles statistiques) de réponse.

Venons-en maintenant à la notion psychologique d’identité. Si le savoir visé par l’évaluation psychométrique ne porte pas sur le comportement de la personne, sur quoi porte-t-il ? Il porte sur la personne tout court, c’est-à-dire sur son identité ou encore son moi.

6. Le savoir résultant de l’examen psychométrique est évaluatif, c'est-à-dire attributif

La subjectivité humaine est caractérisée par sa réflexivité langagière, c’est-à-dire par le fait qu’un sujet s’appelle lui-même « je » et appelle un autre sujet « il » ou « elle ». Comment faire la différence entre « je » et « il » ou « elle » ? Il n’y a pas de différenciation possible des sujets sans identification des êtres des sujets. Or l’identification suppose un espace d’identifiants dans lequel les êtres (ou objets) puissent être localisés ou inscrits pour

pouvoir être reconnus (identifiés). De même que le patronyme est un identifiant *attribué*, et non pas révélé comme une propriété inhérente à l’être du sujet, le profilage psychologique qui résulte de l’évaluation psychométrique est une *attribution* (et non une *mesure*, cf. 2 et 3). C’est une désignation (elle s'opère de l'extérieur de je/moi et m'engage dans un échange à la Mauss parce que je/moi *dois* mon pouvoir d'être reconnu à mon désignateur).

L’évaluation psychologique attribue des « propriétés psychologiques » (dans la rhétorique naturalisante de la psychologie de la personnalité), c’est-à-dire des identifiants. Elle nourrit la « construction de l’identité psychologique ». Comme l’identification est une nécessité subjective (sinon « qui suis-je » a pour réponse « rien »), l’évaluation est fonctionnelle ou encore adaptative parce qu'elle est identifiante. Notre propos n’est pas d’analyser en détail les mécanismes de la véridiction, terme que nous empruntons à Lordon (2019), ou encore le pouvoir performatif, des descriptions psychologiques, étant entendu que le point de vue scientifique est en échec (cf. 2 à 4). La vérité, ou plus pragmatiquement la crédibilité d’une description est ce qui résulte de la véridiction[4].

Il suffit de remarquer que, puisque l’identification est une nécessité subjective et sociale (pléonasme), puisque l’évaluation est une forme d’identification, l’évaluation est inévitable parce qu'elle est constitutive du champ social. L’évalué est automatiquement engagé dans la véridiction de son évaluation et de son évaluateur, à la mesure, si on nous permet d'abuser à notre tour du mot "mesure", de son besoin de reconnaissance (sauf cas limite, lorsque les conditions sont réunies pour une insurrection contre le « système » d’identification, ce qui suppose probablement qu’un espace identificatoire alternatif soit disponible) : dans le régime de l’intersubjectivité, n’importe quel identifiant vaut mieux que rien. L’enthousiasme dont peuvent témoigner les personnes qu’on interroge sur les résultats de leur bilan de compétences prouve qu'une évaluation peut s’accompagner d’affects positifs.

Un verdict comme « Vous êtes 20,33 % plus stressé qu’une personne normale »[5] suscite rarement une demande d’éclaircissement. Peu importe que le pourcentage décrive un

[4] Le fait que le comportement au test ait son origine dans la personne testée constitue une condition nécessaire au pouvoir identificatoire de l’interprétation psychologique qui en est faite : les signes qui me concernent ne viennent pas de n’importe qui mais de moi-même, donc ils parlent de moi.
[5] Voir par exemple le site https://www.idrlabs.com/fr/anxiete-stress-depression/test.php

intervalle de scores sur une distribution de référence, sans lien logique avec la quantité q de stress de la personne évaluée, ni avec la quantité q_{ref} de stress de « une personne normale ». En admettant le stress comme une grandeur, comment pourrait-on savoir, en effet, que 23,33 = 100 $(q – q_{ref})/q_{ref}$, si on ne sait mesurer ni q ni q_{ref} ?

7. Son caractère attributif est douteux,

7.1. Scientifiquement tant qu’on ne sait pas contrôler expérimentalement la variation de la grandeur évaluée,

La vérité des descriptions psychologiques ne relève pas de la vérité scientifique, laquelle est une vérité logique et suppose une ontologie, au sens informatique du terme, cadre conceptuel qui reste largement à clarifier et à expliciter (par exemple les notions de grandeur mesurable, de dépendance causale du mesurant par rapport au mesuré, d’erreur, de précision). Un examen attentif conduit à rejeter que les tests d'évaluation psychologique soient des instruments de mesure si on entend que mesurer suppose la maîtrise expérimentale d'un principe causal où le mesurant est la "variable dépendante" et le mesuré la "variable indépendante".

7.2. Politiquement, tant qu’on ignore son utilité

Remarquons d’abord que connaître l'utilité d'une évaluation est une condition nécessaire mais non suffisante pour attester sa valeur épistémologique, parce qu’alors il faut disposer d’une épistémologie utilitariste qui doit définir ce qu’on appelle un bienfait ou encore un bien (au moins psychologique ou social).

Le problème de l’utilité réside moins dans l’immensité de l’ensemble des buts utiles, qui doivent être définis dans un référentiel descriptif hypercomplexe – ce qui pose un problème technique ou mathématique –, que dans ce que nous devons entendre par « on », dans « tant qu’on ignore son utilité ». La réponse immédiate serait : « on » est le peuple. Le problème est de savoir ce que « peuple » signifie dans l’épistémè contemporaine pour reprendre le terme de Foucault. S’agit-il de la totalité des humains ? Ou bien, dans l’hypothèse où les humains seraient organisés en classes sociales, quitte à fabriquer un concept de classe mouvante, nuages de points se déplaçant dans un espace multidimensionnel et évolutif : « on » est une certaine partie du peuple ?

A l’heure actuelle, il semble suffisant que quelques-uns seulement n’ignorent pas son utilité pour que l’évaluation psychométrique se développe dans la société. C’est devoir faire confiance en la justesse de jugement des gens qui pratiquent cette utilité, parce qu’il faudrait aussi qu’ils sachent que cette utilité n’est pas un mal psychologique ou social, ce qui demeure un problème ouvert tant qu’on ne dispose pas d’une théorie du bien.

8. Conclusion : s’il revendique une responsabilité professionnelle, le praticien de l’examen psychologique doit répondre de ce qu’il sait à propos de la vérité scientifique de son savoir, le point est épistémologique et, aussi, de la fonction psychologique ou sociale de ce savoir, le point est politique

L’argument développé ici vise non pas à évaluer les praticiens de l'examen psychométrique, mais à encourager un débat critique, comme un exercice de vitalisation déontologique d’une communauté professionnelle ou académique menacée d’agentisme au sens de Milgram, non pas par goût, mais par la conséquence de l’évolution de ce qu’on nomme en sciences sociales le marché de l’emploi, à savoir alors par la conséquence de sa précarisation et donc par la nécessité de répondre à la demande sociale dont la fiabilité politique est, comme toute institution au sens de Lordon, affectée par l’anarchie des valeurs en régime néocapitaliste qui, cela est corroboré par les récentes et récurrentes crises économiques, ignore en tant que régime ce qu’est le bien de chaque humain pris un par un et encore davantage le bien des futurs humains (Bouleau, 2018; Pelletier, 2019). Il existe plusieurs façons de défendre l’utilité de tel type d’examen psychométrique, ou plutôt d’évaluation psychologique : à prouver.

Il existe des cas où l’interprétation psychologique naturaliste d’un score est possible. Par exemple, un score de QI très élevé permet d’attester que le répondant est capable[6] de produire les bonnes réponses aux sous-tests de telle « échelle d’intelligence », sans qu’il soit nécessaire que cette capacité soit scientifiquement élucidée. Le présent article tente d’attirer l’attention du lecteur sur la différence du sens causal des relations reliant mesurant et mesuré, qui interdit de confondre « mesurer » et « évaluer » : tandis que le mesurage fait dépendre le mesurant du mesuré, parce qu’il exploite une loi expérimentale, l’évaluation inverse la

[6] En négligeant les bonnes réponses fortuites, ce qui est possible si les formats de réponse sont ouverts et la cotation des réponses objective – pas de variation de la cotation en fonction du juge.

causalité, parce qu’elle exploite le pouvoir social de véridiction de ses descriptions : le niveau scolaire (mesuré) par exemple, dépend de la mention au baccalauréat (mesurant), comme le niveau d’intelligence (mesuré) dépend du QI (mesurant).

Références

Bouleau, N. (2018). *Le mensonge de la finance : Les mathématiques, le signal-prix et la planète*. Paris: Les Editions de l’Atelier.

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

Lacot, E., Afzali, M. H., & Vautier, S. (2016). Test validation without measurement : Disentangling scientific explanation of item responses and justification of focused assessment policies based on test data. *European Journal of Psychological Assessment*, *32*, 204□214.

Lordon, F. (2019). *La condition anachirque : Affects et institution de la valeur*. Paris: Seuil.

Pelletier, W. (Éd.). (2019). *Manuel indocile des sciences sociales*. Paris: La Découverte.

Vautier, S. (2015). La psychotechnique des aptitudes : Pour différencier une sociotechnique de l’évaluation sans mesurage et une psychologie balbutiante de la compréhension de la performance. *Pratiques Psychologiques*, *21*, 1□18.

Vautier, S., & Gaudron, J. P. (2002). Intégrer l’erreur de mesure dans l’interprétation quantitative des scores individuels. *Pratiques Psychologiques*, *2*, 97□108.

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.