Zeitschrift: Schweizer Erziehungs-Rundschau : Organ für das öffentliche und private Bildungswesen der Schweiz = Revue suisse d'éducation : organe de l'enseignement et de l'éducation publics et privés en Suisse

Herausgeber: Verband Schweizerischer Privatschulen

Band: 48 (1975-1976)

Heft: 8

Artikel: Analyse des objectifs et évaluation [suite]

Autor: Chancerel, J.L.

DOI: https://doi.org/10.5169/seals-851948

Schweizer Erziehungs-Rundschau

ST. GALLEN Nr. 8, Seite 185–208 NOVEMBER 1975

Institut de psychologie – Université de Neuchâtel
SEMINAIRE PEDAGOGIQUE DE MONTREUX

Analyse des objectifs et évaluation

J. L. Chancerel

Chapitre IV

DE LA DOCIMOLOGIE CLASSIQUE A L'ADAPTATION DES TESTS AUX FINALITES DE L'EVALUATION

Première partie

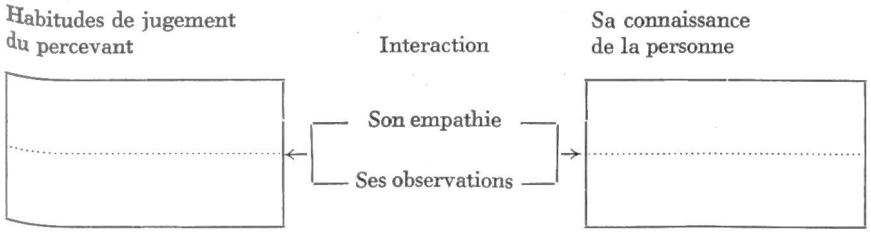
I. *La docimologie* (Principes)

En 1963 paraissait au P. U. F. le livre désormais classique de Henri *Piéron* «Examen et Docimologie». Ce livre présentait un ensemble de résultats de la commission Carnegie sur les notes.

Le terme de docimologie désigne l'étude systématique des examens (gr.: dokimè = épreuve): modes de notation, variabilité des examinateurs, facteurs subjectifs, etc. Depuis une trentaine d'années environ, on se préoccupe du problème des examens et de la notation des épreuves aux Etats-Unis, en Angleterre et en France essentiellement.

L'examen a pour but de *mesurer* une *performance* (généralement un *savoir*) c'est-à-dire d'y faire correspondre une *appréciation*, la plupart du temps chiffrée.

On peut rapprocher l'appréciation conduisant à la notation du schéma de la sensitivité que Smith donne dans la connaissance d'autrui:



Ce qui détermine les prédictions d'un percevant à propos d'une personne.

Cela implique:

- que la performance peut être mesurée.
- que cette mesure est *objective*, c'est-à-dire:
  1. *fidèle* (stable dans le temps: une épreuve notée plusieurs fois par le même correcteur doit aboutir à une même mesure) et stable entre les correcteurs: plusieurs correcteurs doivent apprécier une épreuve de la même manière).
  2. valide: (l'appréciation estime correctement la performance).

Les recherches docimologiques montrent que les examens scolaires traditionnels ne sont ni très fidèles ni très valides.

1. *Fidélité*

a) Lorsque plusieurs examinateurs corrigent les mêmes copies, les notes varient parfois énormément (jusqu'à 13 points sur 20 dans les études sur le baccalauréat français).

Ces écarts sont plus élevés pour certaines matières que pour d'autres. Viennent en tête la dissertation philosophique, la composition française et la version latine. Les écarts sont sensiblement moindres pour les mathématiques et la physique. Quand, après un intervalle de plusieurs mois, voire de plusieurs années, on donne à nouveau à un examinateur des copies qu'il a déjà corrigées, les notes varient aussi. Les écarts sont d'autant plus importants que l'intervalle est long. Il faut considérer cependant que les correcteurs diffèrent passablement les uns des autres: les uns sont «stables», d'autres le sont beaucoup moins.

b) *Les notes n'ont pas la même signification pour tous les correcteurs.* Certains, par exemple, emploient toute l'échelle tandis que d'autres se limitent à une partie seulement. (Supposons que la copie de mathématiques d'un candidat peu doué pour cette discipline soit corrigée par un professeur qui utilise toute l'échelle, il obtiendra peut-être un «3» (sur 20). A supposer que son admission dépende de la moyenne des résultats en mathématiques, en latin et en français et que les copies de latin et de français soient corrigées par deux professeurs qui n'utilisent que les notes centrales, notre candidat obtiendra, s'il a bien réussi, peut-être deux «13». Sa moyenne sera cependant insuffisante).

c) Les correcteurs sont sensibles à l'*effet de halo:* pour autant que le correcteur connaisse l'élève, l'attitude en classe de ce dernier, sa plus ou moins grande application, les notes déjà obtenues, etc. peuvent influencer la notation. De plus, on a tendance à surestimer une bonne copie si on la corrige après une copie médiocre et à sous-estimer une copie moyenne (ou médiocre) après avoir noté une bonne épreuve. En outre, il semble que l'on soit plus sévère dans l'appréciation des pre-

mières copies d'une série que dans celle des dernières (référence à un *modèle* implicite ou explicite tout d'abord, d'où sévérité relative; puis adaptation des notes à une certaine moyenne des performances).

d) La fatigue du correcteur exerce aussi un effet néfaste sur l'objectivité.

2. *Validité*

a) La corrélation entre les *notes d'écrit* et les *notes d'oral* (baccalauréat, licence) est très faible (en moyenne: environ 0.20). Cela revient à dire qu'en mettant au hasard les notes d'oral, on aboutirait à peu près aux mêmes résultats.

b) Si l'on examine les pourcentages d'admission (baccalauréat français), on constate qu'ils varient non seulement d'une année à l'autre (jusqu'à 20 % ici aussi). Ces écarts ne s'expliquent pas uniquement par des différences d'aptitudes ou de préparation des étudiants. Il s'agit bien d'inégalités dans les exigences des jurys.

3. On constate aussi que les *moyennes générales* diffèrent selon les matières.

En résumé, la réussite ou l'échec d'un candidat tient en partie au hasard. La subjectivité des examinateurs joue un rôle notable, d'autant plus important que l'examen (ou une partie seulement) est oral, qu'il est court, que la notation est confiée à un seul individu, que celui-ci est «instable». Elle dépend en outre de la matière, de l'époque et de la circonscription scolaire.

2. *Conséquences pédagogiques d'une évaluation par classement*

Jean Cardinet, dans un exposé au 3e Congrès de l'A.I.P.E.L.F. à Bruxelles (Avril 1972) (nous nous référerons dorénavant à ce texte) écrivait:

«Le succès effectif de ces méthodes pour assurer une meilleure prédiction de la réussite scolaire, et donc une sélection plus efficace, a longtemps pour justifier la méthodologie des tests inventée par Galton. Ses dangers commencent seulement à être pris en considération».

On peut voir plusieurs conséquences de ce type d'évaluation:

- l'accent n'est plus mis sur l'apprentissage scolaire mais sur la compétition.
- le classement stable étant le critère de fidélité on retire les questions peu discriminatives qui ont leur rôle dans l'évaluation.
- la sélection des plus aptes (basée sur la stabilité) assure «la reproduction par l'école de la hiérarchie sociale actuelle qui en est la conséquence».

(La deuxième partie, la troisième partie, la quatrième partie de l'exposé seront constituées par des extraits de la conférence de Jean Cardinet au 3e Congrès de l'Association Internationale de Pédagogie expérimentale de Langue française).

Deuxième partie

Les finalités pratiques de l'évaluation

1. Les trois facettes d'une classification générale

A l'occasion de cours de formation sur les problèmes de l'évaluation, donnés à des enseignants secondaires, un grand nombre d'avis ont été rassemblés concernant les utilisations possibles de la note scolaire et des autres appréciations pédagogiques. Il était possible de classer ces finalités de multiples façons. Nous avons choisi trois principes de classement qui paraîtront sans doute arbitraires, mais qui avaient l'avantage de suggérer des modes d'évaluation différenciés. Les trois facettes de la classification proposée correspondent aux questions ci-dessous:

a) L'évaluation porte-t-elle sur un individu ou sur un groupe?
b) L'évaluation porte-t-elle sur le passé, le présent ou l'avenir?
c) Quelle phase du processus pédagogique est-elle concernée?

2. Classement des finalités selon leur visée individuelle ou collective

La différence entre les finalités à visée individuelle ou collective est assez évidente pour que quelques exemples suffisent à la concrétiser. D'un côté il s'agit de prendre des décisions concernant des élèves considérés isolément; on peut vouloir s'assurer des connaissances qu'ils ont acquises en mathématiques; on peut souhaiter définir les exercices les plus utiles à leur proposer; on peut rechercher la section où ils auraient le plus de chance de réussir ou de s'épanouir. De l'autre côté, on parlera de visée collective lorsqu'il s'agit par exemple de s'informer sur les caractéristiques d'un groupe, d'apprécier le degré d'assimilation d'un programme, de découvrir les difficultés particulières de certaines catégories d'élèves, de s'assurer des effets de nouvelles méthodes d'enseignement, etc.

Les problèmes à résoudre dans l'un ou l'autre cas sont radicalement différents. Considérons d'abord les tests à visée individuelle. Le problème est alors d'obtenir pour l'individu considéré une information fidèle, sûre, répétable. Il faut que l'on puisse, à partir des réponses données par l'élève aux quelques questions que nous lui avons posées, estimer quelle aurait été sa réponse aux autres questions du programme. Autrement dit, en utilisant le vocabulaire des statistiques, il faut que l'échantillon observé nous renseigne sur la population sous-jacente. Il faut donc que les questions choisies soient représentatives de l'ensemble des questions possibles. Ceci oblige à contrôler la validité conceptuelle, ou de construction, des épreuves que l'on établit.

Pour pouvoir, à partir d'un échantillon, tirer des conclusions sur la population d'ensemble, il ne suffit pas cependant que l'échantillon soit représentatif. Il faut encore qu'il soit suffisamment grand pour ne pas risquer d'être trop influencé par des cas extrêmes. La dimension de l'échantillon nécessaire dépend naturellement de la variabilité des mesures et de la précision que l'on recherche. Les procédures classiques de construction des tests peuvent être considérées comme une façon d'optimiser l'échantillonnage des questions, de façon à pouvoir généraliser valablement des résultats obtenus dans un test à l'ensemble des questions concernant ce domaine.

Tableau I *Caractéritiques de l'évaluation selon sa visée*

	Individuelle	*Collective*
Exemples:	Attribution de diplômes Choix de curriculum Conseil d'orientation	Assimilation d'un programme Effet d'une méthode Prévision d'effectif
Généralisation à:	Population de questions définies	Population d'élèves définie
Implique de connaître:	Fluctuations dues à l'échantillonnage des questions	Fluctuations dues à l'échantillonnage des élèves
Pour optimiser:	Choix des questions (analyse d'items) Nombre de questions	Choix des élèves (plans (plans incomplets équilibrés) Nombre d'élèves

Examinons maintenant de façon parallèle les tests à visée collective. Il s'agit dans ce cas d'obtenir des informations sur une certaine population d'élèves: les connaissances acquises, les difficultés qui subsistent, les motivations sur lesquelles l'enseignement peut s'appuyer dans ce groupe particulier. La source d'erreur la plus évidente est alors le choix d'un échantillon d'élèves non représentatif de la population visée. Les résultats ne seront pas les mêmes, en effet, si nous nous adressons à des élèves d'un quartier à niveau socio-culturel élevé, à des enfants traumatisés par un échec récent, à des élèves sélectionnés du fait de la non-promotion de leurs camarades, etc. La valeur de nos conclusions dépendra essentiellement de la représentativité de l'échantillon que nous aurons choisi par rapport à la population que nous voulions étudier.

Là encore, la précision des mesures obtenues pourra être déterminée à l'aide des lois habituelles de la statistique. Plus l'échantillon sera grand et la variabilité d'élève à élève faible, plus notre estimation sera précise. Le choix des questions, par contre, ne pose plus de problème. Une question unique peut très bien donner lieu à une enquête. Même si l'épreuve proposée comporte une série de problèmes, les résultats seront beaucoup plus faciles à interpréter question par question que si l'on part de la note totale de chaque élève. En résumé, les tests à visée collective doivent se préoccuper de l'échantillonnage des élèves, mais non de celui des questions, alors que c'est l'inverse pour les tests à visée individuelle.

Cette opposition se marque encore plus nettement lorsque l'on cherche à optimiser la prise d'information, c'est-à-dire à obtenir le maximum de renseignements pour le coût minimum. Si l'évaluation a une visée individuelle, on emploiera les méthodes de «tests sur mesure», c'est-à-dire que l'on choisira des questions de difficulté moyenne pour l'élève considéré. Si la visée est collective, on utilisera plutôt la technique de «l'éventail», consistant à donner dans chaque classe une série de formes différentes de la même épreuve, pour couvrir de façon plus exhaustive le domaine sur lequel porte l'enquête.

La distinction entre visée individuelle et collective a des répercussions pratiques évidentes. Combien de fois a-t-on vu des psychologues construire l'histogramme de la note totale à un test de connaissances pour décrire les résultats de la population examinée? En fait, l'information transmise était nulle, car cette procédure n'est utile que dans une perspective individuelle, pour situer le résultat d'un sujet particulier dans la distribution de tous les résultats observés. Inversement, combien d'enseignants ne croient-ils pas avoir «validé» leur test parce qu'ils ont étudié le pourcentage de réussite à chaque question? Qu'un problème soit réussi ou non par un grand nombre d'élèves n'a d'intérêt que pour celui qui veut connaître les résultats d'une population donnée. Au niveau individuel, on se soucie plutôt de la possibilité de généraliser les observations du test à une autre population de comportements; on se préoccupe donc plutôt des corrélations entre les questions. On voit ainsi que les procédés d'analyse d'items se répartissent également en deux classes, correspondant aux visées individuelle et collective.

La confusion de ces deux visées conduit à des paradoxes qu'il est intéressant de souligner. Supposons que l'on veuille étudier les résultats du nouveau programme de mathématique moderne. Il serait absurde, lors de la mise au point de l'épreuve de connaissances, d'éliminer les questions trop «faciles» ou «difficiles», sous prétexte qu'elles sont peu discriminatives. On n'obtiendrait que des résultats moyens, du fait même du rejet préalable des questions extrêmes. La procédure traditionnelle d'analyse d'items adapte le test à une évaluation de nature individuelle et non plus collective.

Un second exemple mettra en lumière l'incompatibilité fréquente de ces deux visées. Dans un des systèmes scolaires de Suisse romande, on a introduit l'emploi systématique des épreuves communes. Plusieurs fois chaque année, l'ensemble des élèves répondent à des questions représentatives des connaissances qu'ils auraient dû assimiler à cette date. En principe, la visée de ces épreuves est collective. Les enseignants doivent pouvoir se rendre compte des difficultés qui subsistent chez leurs élèves, et surtout du retard que certaines classes peuvent prendre par rapport à d'autres. Le but final est d'assurer à tous un enseignement de qualité comparable. On emploie malheureusement les mêmes épreuves pour situer les élèves les uns par rapport aux autres et leur donner des notes qui servent à leur orientation. Pourquoi mettre de mauvaises notes et imposer la fréquentation d'une section moins intellectuelle à des élèves qui ont été surtout retardés par un enseignement défectueux? Inversement, si

l'on admet que le retard des élèves provient d'un manque d'aptitude réel de leur part, pourquoi cette compétition des enseignants? Les compromis que l'on essaie d'établir entre la visée individuelle et la visée collective ne font qu'ajouter à la confusion. Ainsi, on évite de poser des questions sur un domaine qui n'a pas encore été étudié dans une classe, pour ne pas pénaliser excessivement les élèves de cette classe. Ce faisant, on renonce à faire apparaître le retard que ces élèves ont pris par rapport au programme, alors que c'était l'objectif de départ de l'épreuve commune de déceler ces inégalités. Le malaise des enseignants et des élèves provient, à notre avis, de ce que ces deux types de finalités n'ont pas été suffisamment distingués. (à suivre)

Gehirn, Fernsehen und die Aggressivität des Menschen

Pierrette Posmowski interviewte David Hamburg

Was macht die Menschen aggressiv? Wissenschaftler vieler Fachrichtungen, von der Biochemie bis zur Soziologie, suchen eine Antwort auf diese Frage in der Hoffnung, daß es uns eines Tages möglich sein wird, Konflikte zwischen Einzelnen oder Gruppen zu vermeiden oder wenigstens merklich zu verringern. In dem folgenden Interview spricht Dr. David Hamburg, Leiter des Departements für Psychiatrie am Medizinischen Zentrum der Stanford University (Kalifornien), mit Pierrette Posmowski über einige physiologische und soziale Faktoren, die bei der menschlichen Aggressivität eine Rolle spielen.

Posmowski: Die Unesco-Tagung in Brüssel, an der Sie teilgenommen haben, beschäftigte sich mit den Problemen menschlicher Aggressivität. Ich glaube, es ist wichtig, genau zu wissen, was mit diesem Wort gemeint ist. Wenn wir von einem aggressiven Führer sprechen, meinen wir nicht unbedingt, daß er voller Haß, gewalttätig oder destruktiv ist, sondern vielmehr, daß er über eine gute Position Initiative, Ausdauer und Vitalität verfügt. Steuern nun die gleichen Gehirnpartien beide Arten von Verhalten, das gewalttätige und das selbstbewußte?

Hamburg: Wir wissen es nicht genau. Die Forschung auf diesem Gebiet der Verhaltensneurophysiologie ist erst neueren Datums. Ausgehend von den beschränkten Ergebnissen, die bis jetzt vorliegen, kann ich jedoch einige Mutmaßungen wagen. Wahrscheinlich besteht eine gewisse Ueberschneidung bei den Leitungsströmen des Gehirns, die die beiden weitgespannten Verhaltensarten vermitteln. Beide erfordern Perioden intensiver Aktivität, Perioden der Anspannung und die Fähigkeit, Anstrengungen zu ertragen. Daraus ist zu folgern, daß sehr viele Leitungen daran beteiligt sind, diese Aktivität zu bewirken, ferner Veränderungen der Hormone und des Stoffwechsels des Körpers, um die für solche Anstrengungen nötige Energie zu mobilisieren, und auch Veränderungen in den Herz- und Blutgefäßen, um das Blut in die großen Muskeln zu transportieren und für eine gute Gehirndurchblutung zu sorgen, damit eine intensive oder anhaltende Tätigkeit möglich wird. Dies sind grundlegende physiologische Voraussetzungen.

Einige Leitungsbereiche scheinen indessen wirklich für bestimmte drohende und aggressive Verhaltensweisen mehr oder weniger zuständig zu sein, d. h. für den mehr gewalttätigen Sektor des gesamten Spektrums aggressiven Verhaltens. Diese besonderen Leitungen liegen in den im evolutionären Sinne älteren Teilen des Gehirns, wie Hypothalamus, Mittelhirn und limbischer Cortex. Und sie sind wahrscheinlich zu einem guten Teil unabhängig von den Leitungen, die für eine allgemeinere Aktivität beansprucht werden.

Posmowski: Eines der fundamentalsten Probleme bei der Untersuchung menschlicher Aggressivität ist das Erlernen furchtsamer, verächtlicher und feindlicher Verhaltensweisen gegenüber Menschen, die zu anderen Gruppen gehören. In welchem Alter beginnen Kinder sich solche Haltungen anzueignen?

Hamburg: Hierüber ist sehr wenig bekannt, obwohl jüngere Untersuchungen darauf hindeuten, daß das Kind in der Zeit zwischen der Mitte des ersten und der Mitte des zweiten Lebensjahres unbekannten Personen und Orten gegenüber empfindlich wird. Diese emotionale Reaktion, die meiner Meinung nach eine Art verhaltensmäßiges Erbe von unserer Entwicklung als Primaten her ist, scheint in das Nervensystem des Kindes irgendwie eingebaut zu sein.

Sie hat wahrscheinlich keine allzugroße praktische Bedeutung, wenn die Mutter oder andere Menschen in der Umgebung des Kindes diese Zeit nicht dazu nutzen, dem Kind ein Gefühl von Furcht oder Gefahr im Zusammenhang mit allem Fremden oder einer bestimmten Gruppe von Fremden zu vermitteln – vielleicht denen, die eine dunkle Haut oder andere sichtbare Merkmale haben. Umgekehrt ist diese frühe Periode der Empfindlichkeit von besonderem Interesse, weil sie die erste Gelegenheit für eine Mutter, eine Familie oder eine Gruppe sein könnte, eine positive Haltung des Kindes gegenüber Fremden herauszubilden.

Der Anblick von Aggression hat bleibende Wirkung

Posmowski: Aus dem Gesagten ergibt sich, daß kleine Kinder sowohl für freundliches wie für aggressives Verhalten empfänglich sind. Da nun in den ersten Lebensjahren Beobachtung und Nachahmung die hauptsächlichsten Lernarten sind, ist es nicht wahrscheinlich, daß sie oft aggressive Neigungen entwikkeln?

Hamburg: Professor Albert Bandura von der Stanford University hat auf diesem Gebiet bahnbrechende Forschung geleistet, die kürzlich durch die Arbeit in anderen Forschungsstätten bestätigt worden ist.