Pour mesurer les progrès des élèves en physique

Autor(en): **Cardinet, Jean / Zimmermann, Marie-Louise**

Objekttyp: **Article**

Zeitschrift: **Bildungsforschung und Bildungspraxis : schweizerische Zeitschrift für Erziehungswissenschaft = Éducation et recherche : revue suisse des sciences de l'éducation = Educazione e ricerca : rivista svizzera di scienze dell'educazione**

Band (Jahr): **11 (1989)**

Heft 3

PDF erstellt am: **08.08.2024**

Persistenter Link: https://doi.org/10.5169/seals-786387

Pour mesurer les progrès des élèves en physique

Jean Cardinet/Marie-Louise Zimmermann

L'un des auteurs a développé une méthode d'apprentissage des sciences expérimentales qui s'appuie sur la recherche personnelle des élèves, pour faire évoluer leurs représentations. L'autre auteur a proposé un modèle pour la mesure du progrès individuel, appliquant la théorie de la généralisabilité. Dans cet article, ce dernier modèle statistique est appliqué à des données relatives à l'étude de la chaleur. Il apparaît que les niveaux d'apprentissage peuvent être comparés de façon très fidèle pour chaque élève pris isolément. Ceci démontre la possibilité d'une évaluation individualisée, totalement indépendante des différences entre élèves. Cependant cette conclusion concerne les scores relatifs de progrès, et non les scores absolus qui restent trop affectés par les différences de difficulté des questions.

Origine du travail

La recherche est de plus en plus un travail d'équipe, même dans les sciences de l'éducation. Le premier auteur avait développé un modèle statistique qui semblait utilisable pour la construction de tests de connaissance (Cardinet, 1986 et 1987). Le second auteur avait expérimenté une démarche didactique nouvelle pour l'enseignement de la physique (Zimmermann, en préparation) et avait recueilli des données à cette occasion. Il était tentant d'y appliquer la démarche d'analyse précédente. L'idée fut proposée par Daniel Bain, qui avait traité initialement ces données. Les calculs furent repris par Bernard Muller, sur l'ordinateur du Centre vaudois de recherches pédagogiques. D'autres personnes s'intéressèrent au projet et constituèrent un groupe de travail, qui se réunit une dizaine de fois pour en discuter (Zimmermann et Bain, 1988; Cardinet, 1988). C'est le résultat de tous ces efforts qu'il reste à communiquer, sous une forme accessible aux non-spécialistes, les chercheurs pouvant se reporter aux publications techniques qui viennent d'être citées.

But de l'étude

La question traitée ici concerne l'évaluation du progrès individuel de chaque élève. Peut-on mesurer ce progrès, et surtout, peut-on déterminer la précision de cette estimation, pour pouvoir adapter éventuellement le dispositif de mesure utilisé? Une telle question se pose nécessairement à quiconque introduit une didactique nouvelle, puisqu'il s'agit de contrôler la généralité de l'effet attendu.

Ce problème se distingue d'autres interrogations qui apparaissent souvent dans le même contexte: comment comparer les performances finales des élèves qui ont suivi ce cours, pour distinguer les bons élèves des moins bons? Peut-on être sûr qu'en moyenne les élèves ont progressé? La nouvelle méthode est-elle préférable, en moyenne, à l'ancienne? Cette didactique réussit-elle mieux pour certains objectifs que pour d'autres, ou pour certains types d'élèves?

A chacune de ces questions correspond un traitement statistique particulier, que la théorie de la généralisabilité permettrait de conduire, au même titre que l'étude qui fait l'objet de cet article. Pour simplifier, cependant, un seul point sera abordé ci-dessous: la mesure du progrès pour chaque élève en particulier.

Théories de référence

Du point de vue didactique, ce travail se rattache à l'ensemble des études, inspirées par André Giordan en particulier, qui tentent de partir des représentations scientifiques spontanées des élèves, pour les faire évoluer. Il s'agit, pour l'enseignant, de créer des situations où ces premières représentations soient démenties par les faits, pour que d'autres représentations puissent les remplacer. Mais on sait que les élèves doivent être activement engagés dans la recherche de ces nouvelles explications pour qu'ils «construisent» leur savoir, selon l'expression de Piaget. Ceci implique de proposer des expériences «cruciales», permettant de distinguer entre les hypothèses en compétition, mais de laisser les élèves conduire ces essais eux-mêmes et les interpréter en groupes. Sans cette redécouverte, leurs connaissances restent purement scolaires et non intégrées à leur vision du monde.

La première motivation de la présente recherche est de tenter de confirmer, par rapport à cette théorie didactique, que les représentations des élèves ont effectivement évolué. La seconde motivation est de voir si le modèle statistique utilisé permet de tirer des conclusions utilisables dans une situation scolaire concrète. C'est la théorie de la généralisabilité qui a été appliquée aux données de ce problème. On entend par là une méthode d'estimation de l'importance des fluctuations dues à l'échantillonnage des conditions d'observation. Connaissant l'ampleur des variations de performance que le hasard peut susciter, on peut ensuite y comparer le progrès observé chez tel ou tel élève particulier, pour savoir si le gain obtenu se distingue plus ou moins nettement de cette marge d'imprécision.

Méthodes pédagogiques utilisées

Les méthodes APA (apprentissage des sciences expérimentales par l'autonomie) ont été définies de façon précise dans une brochure (Zimmermann et Paillard, 1987). Elles sont maintenant utilisées auprès de 300 élèves de 16 ans. Ces méthodes prennent en compte l'activité propre de l'élève. C'est lui qui a le rôle central, il est le maître d'œuvre de la construction de son savoir. Cette place nouvelle qui lui est dévolue entraîne un changement d'attitude de l'enseignant. Celui-ci quitte l'avant-scène pour un rôle peut-être moins prestigieux, mais pourtant indispensable!

Les phases de l'enseignement sont: investigation, mise en commun, réinvestissement des notions, évaluation.

L'enseignant fixe les thèmes et pose les problèmes sous forme de fiche de travail. Pendant cette phase d'investigation, les élèves ont toute liberté d'expérimentation, de recherche dans les documents. Ce qui est valorisé, ce n'est plus la réponse trouvée, mais toute la démarche mise en œuvre. Non seulement l'autonomie est développée mais, également, la créativité.

Une phase de mise en commun permet d'échanger les réponses, de confronter les arguments et devrait aboutir à la formulation d'une solution acceptée par toute la classe. Lors de cette phase, non seulement les aptitudes à la communication, à la sociabilité, sont entraînées, mais c'est aussi un processus métacognitif qui est mis en œuvre, car les élèves sont obligés de réfléchir à leurs procédures expérimentales.

Une phase de réinvestissement des connaissances permet ensuite aux élèves de vérifier leurs acquisitions, grâce à des tests de connaissances, des tests de réflexion et même des tests pratiques.

Une évaluation formative est réalisée au cours de l'année, des notes de progrès sont données grâce à l'utilisation des pré- et posttests. Dans l'avenir il est prévu de mettre en place une évaluation formatrice, c'est-à-dire faisant appel à une démarche autonome de l'élève (qui s'évalue par rapport à ses objectifs personnels et se construit son propre parcours d'apprentissage).

Objectifs d'apprentissage

Les objectifs généraux du cours de physique étaient les suivants: essayer de faire aimer la physique, développer l'autonomie des élèves, stimuler la réflexion des apprenants, faire acquérir une technique d'expérimentation, éveiller l'esprit critique des élèves et apprendre aux élèves à communiquer.

En ce qui concerne le chapitre particulier de la chaleur, il fallait évidemment faire acquérir quelques notions sur la chaleur, mais ce thème a été limité à trois domaines: thermométrie, différenciation chaleur-température et changement d'état.

Des objectifs spécifiques ont été classés en objectifs conceptuels, savoir raisonner et savoir-faire. Ils ont été spécifiés pour chaque thème (Zimmermann, en cours de réalisation).

Questionnaire

Le questionnaire a été élaboré par un groupe de chercheurs du LDES (Laboratoire de didactique et d'épistémologie des sciences de l'Université de Genève). Il est le fruit d'observations réalisées à partir de deux questionnaires précédents. Pourtant, malgré ces précautions, nous avons constaté que certaines questions étaient ambiguës et que la plus grande partie des erreurs étaient dues au décalage entre les conceptions de l'enseignant et celles de l'apprenant.

Les objectifs du questionnaire étaient, d'une part, d'évaluer le progrès des élèves, et d'autre part de mettre en évidence les erreurs les plus courantes, afin d'essayer de mettre en œuvre des stratégies permettant de faire évoluer les conceptions des élèves.

Ce questionnaire comportait 26 items parmi lesquels 11 items concernaient la thermométrie, 8 items se rapportaient au changement d'état et 7 items avaient pour objet la différence entre sensation de chaleur et température.

Ce questionnaire a été passé avant tout enseignement (prétest) et environ un mois après tout enseignement (posttest); 199 élèves de l'école de Jean Piaget à Genève ont ainsi été testés. La durée du test était de 45 minutes. Le surveillant avait pour consigne de ne répondre à aucune question.

Le dépouillement des tests a nécessité une cotation détaillée permettant de répertorier les réponses justes, mais aussi les réponses non attendues par l'enseignant.

Données recueillies

De l'ensemble des résultats de l'expérience, une partie seulement a été prélevée pour l'étude statistique, satisfaisant divers critères d'homogénéité: il s'agit de cinq classes d'un même degré scolaire (la première année de cette école), qui assurent un complément de culture générale après la fin de la scolarité obligatoire, classes qui avaient bénéficié du curriculum expérimental complet pour le chapitre étudié.

Le programme d'ordinateur disponible ne permet ni données manquantes, ni données supplémentaires. Ceci a obligé à abandonner un certain nombre de questions (quatre qui étaient trop bien réussies par l'ensemble des élèves, et une qui l'était trop mal), pour s'aligner sur le nombre minimum de sept items par thème. De même, dans chaque classe, huit élèves seulement ont été conservés, nombre minimum de dossiers complets (comprenant les résultats à l'épreuve passée avant et après l'étude du chapitre sur la chaleur). L'enseignement se déroulant en laboratoire, chaque classe comportait une dizaine d'élèves: le cas échéant, un ou deux dossiers ont donc été éliminés au hasard.

Modèle de l'analyse de la variance

La théorie de la généralisabilité repose sur un modèle statistique connu, l'analyse de la variance, dont le principe est simple: il postule que chaque résultat

observé (le score obtenu par chaque élève à chaque question) est le résultat de l'addition d'un certain nombre d'effets ou sources de variation, que des comparaisons systématiques permettent de mettre en évidence.

Ces sources sont, par exemple, les élèves, avec leurs compétences variables, ou les questions, qui peuvent être aussi de difficultés différentes. A ces effets principaux, s'ajoutent les interactions entre facteurs. On entend par là, par exemple, le fait que certaines questions sont plus familières (et donc faciles) pour certains élèves, alors que d'autres questions conviennent mieux à d'autres. Comme l'addition de l'effet de la compétence générale de l'élève et de la facilité habituelle de la question ne suffit pas à rendre compte du score de chaque élève à chaque question, la différence est attribuée à cet effet d'interaction «élève-question».

Ainsi, si l'on connaissait les valeurs vraies des différents effets pour les facteurs principaux et toutes leurs interactions (deux à deux, trois à trois, etc.), on pourrait reconstituer les scores observés par simple addition de ces composants de scores.

On peut démontrer qu'une analyse semblable peut s'effectuer au niveau de la variance totale, qui peut être décomposée en «composantes de variance», correspondant à chacun des facteurs principaux et de leurs interactions. Cette analyse permet de savoir quelles influences sont les plus marquantes dans la détermination des résultats.

Plan d'observation et plan d'estimation

Pour pouvoir calculer ces effets, il faut recueillir des observations selon un plan équilibré, permettant de comparer les divers niveaux d'un même facteur «toutes choses égales par ailleurs».

On appelle plan d'observation ce mode d'organisation systématique des données. Un tel plan précise d'abord le nombre des facteurs étudiés (appelés facettes). Il décrit ensuite les relations existant entre ces facettes, qui peuvent être soit croisées (lorsque tous les niveaux d'une facette sont observés pour chacun des niveaux de l'autre, comme pour un produit cartésien), soit emboîtées (comme le sont les élèves dans des classes parallèles, par exemple). Il définit enfin les nombres de niveaux échantillonnés pour chaque facette.

L'idée d'échantillonnage est essentielle en analyse de la variance, car tout l'intérêt des statistiques est de pouvoir contrôler les effets imprévisibles liés au tirage au hasard d'un petit nombre d'observations, pour étudier des ensembles beaucoup plus grands, voire infinis.

Le plan d'estimation traduit les décisions de l'expérimentateur concernant l'échantillonnage des facettes du plan d'observation. Si le chercheur peut observer tous les niveaux possibles, il n'a plus de fluctuations à craindre: la facette correspondante est dite fixée. Mais s'il est obligé de recourir à un échantillonnage au hasard, la facette correspondante est dite aléatoire et introduit nécessairement de l'incertitude dans les résultats.

Il est possible d'estimer quelle part de la variation observée pour un facteur (ou une interaction de facteurs) est liée, en moyenne, aux fluctuations causées

par ces échantillonnages aléatoires. En soustrayant ces influences perturbatrices, on peut obtenir une estimation purifiée de l'importance de chaque source de variation. C'est là que s'arrête d'ordinaire une analyse de la variance et c'est là que commence une étude de généralisabilité.

Application aux données de l'étude

Le plan d'observation comporte, dans le cas présent, les cinq facettes suivantes: Classes, Elèves, Thèmes, Items et Phases (c'est-à-dire les pré- et posttests, avant et après apprentissage). A ces effets principaux, s'ajoutent les interactions de ces facteurs pris deux à deux, et trois à trois (par exemple l'interaction Elèves X Phases X Thèmes, qui traduit les progrès variables des élèves selon les thèmes).

Les relations entre ces facettes sont explicitées sur la figure 1, qui correspond à la traduction graphique du plan d'estimation (diagramme de Cronbach). La disposition des ellipses montre que les Elèves sont emboîtés dans les Classes et les Items emboîtés dans les Thèmes. Comme chaque élève répond à chaque item, les facettes E et C sont croisées avec les facettes I et T.

Enfin, l'ensemble des mesures est effectué deux fois, de sorte que la facette P est croisée avec toutes les autres.

Le plan d'observation peut donc être symbolisé ainsi:

$$(E:C) \times (I:T) \times P$$

Le plan d'estimation est aussi explicité sur cette figure, en tenant compte de la convention suivante: les facettes fixées (sans échantillonnage) sont représentées en pointillés, les facettes aléatoires en traits pleins. Ici seules les facettes Items, Elèves et Classes sont échantillonnées au hasard. Les trois thèmes étudiés sont considérés comme les seuls à prendre en compte dans l'évaluation des connaissances relatives à la chaleur. Les moments d'examen, avant et après apprentissage, ne sont pas non plus remplaçables par d'autres. Les facettes Thèmes et Phases sont donc fixées.

Résultats des calculs

Les résultats de cette analyse de la variance apparaissent au tableau I. Chacune des composantes correspond à une plage du diagramme de Cronbach, donné à la figure 1.

La dernière colonne du tableau 1 montre que la plus grande partie de la variance relève des trois sources suivantes: les variations de difficulté des items (I:T), les réactions différentes des élèves à ces items (EI:TC) et leurs apprentissages différents (EPI:TC).

La variance entre classes pour ces items apparaît au contraire comme négligeable, y compris dans ses interactions avec les autres facteurs. Les élèves peuvent donc être mélangés en un seul ensemble dans la suite des calculs. Un

diagramme de Cronbach simplifié apparaît à la figure 2, sans la facette Classes.

La facette Thèmes n'a pas non plus été dessinée sur ce second graphique, parce qu'elle n'intervient pas dans les calculs. Son échantillonnage, étant exhaustif, ne peut causer de fluctuations imprévues. Il n'est donc pas source d'erreur.

Les variations entre Phases (Avant-Après) et entre Elèves sont de même ordre de grandeur (5 et 8 % de la variance totale). Les autres composantes de variance sont négligeables, même l'interaction Elèves X Phases, ce qui tend à prouver que les élèves réagissent de façon assez homogène à l'enseignement qu'ils reçoivent.

Plan de mesure

La seconde phase de l'analyse statistique implique l'estimation de la variance «vraie» entre objets d'étude et de la variance d'erreur, due à l'échantillonnage des conditions d'observation.

L'intérêt de la théorie de la généralisabilité est de montrer que n'importe quelle facette peut être considérée comme objet d'étude ou bien comme condition d'observation. Selon le point de vue de l'expérimentateur, la même source de variation peut ainsi contribuer à la variance vraie ou à la variance d'erreur.

Tableau 1: Analyse de la variance

Sources de variation		Composantes de variance	Part du total
Symbole	Signification: variance entre:		
C	Classes	0.0015	0.58
E:C	Elèves dans les Classes	0.0194	7.68
P	Phases	0.0137	5.41
CP	Classes X Phases	- 0.0005	0.00
EP:C	Elèves X Phases dans les Classes	0.0045	1.78
T	Thèmes	0.0132	5.21
CT	Classes X Thèmes	- 0.0002	0.00
ET:C	Elèves X Thèmes dans les Classes	0.0044	1.76
PT	Phases X Thèmes	0.0009	0.36
CPT	Classes X Phases X Thèmes	0.0007	0.30
EPT:C	Elèves X Phases X Thèmes dans Cl.	0.0034	1.34
I:T	Items dans les Thèmes	0.0572	22.58
CI:T	Classes X Items dans les Thèmes	0.0058	2.31
EI:TC	Elèves X Items dans Thèmes et Cl.	0.0755	29.81
PI:T	Phases X Items dans les Thèmes	- 0.0003	0.00
CPI:T	Classes X Phases X Items dans Thèmes	0.0007	0.29
EPI:TC	Elèves X Phases X Items dans Th. et Cl.	0.0522	20.62

(Les valeurs négatives traduisent de petites inexactitudes dans l'estimation des composantes, résultant d'effets aléatoires. Elles sont traitées comme égales à zéro.)

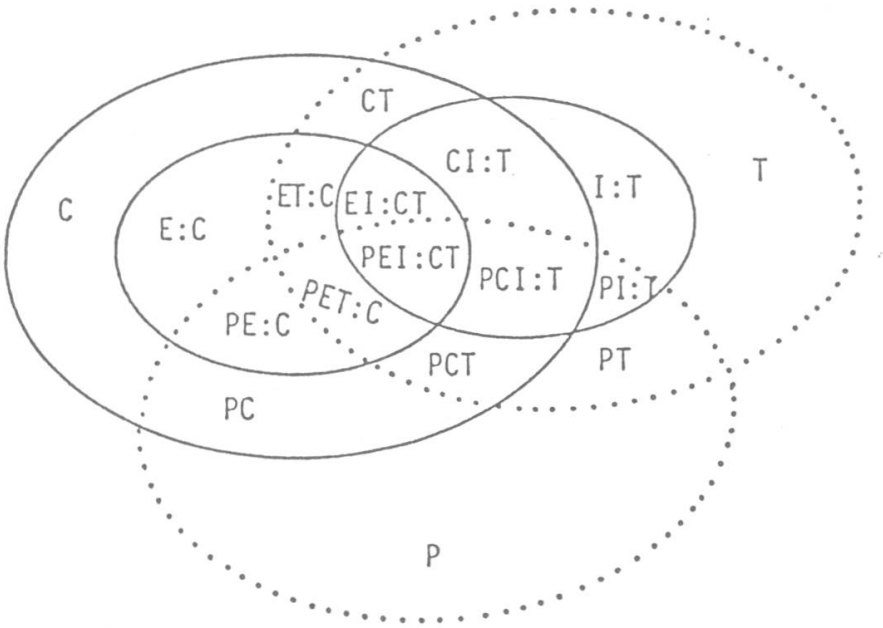

Figure 1: Diagramme de Cronbach complet

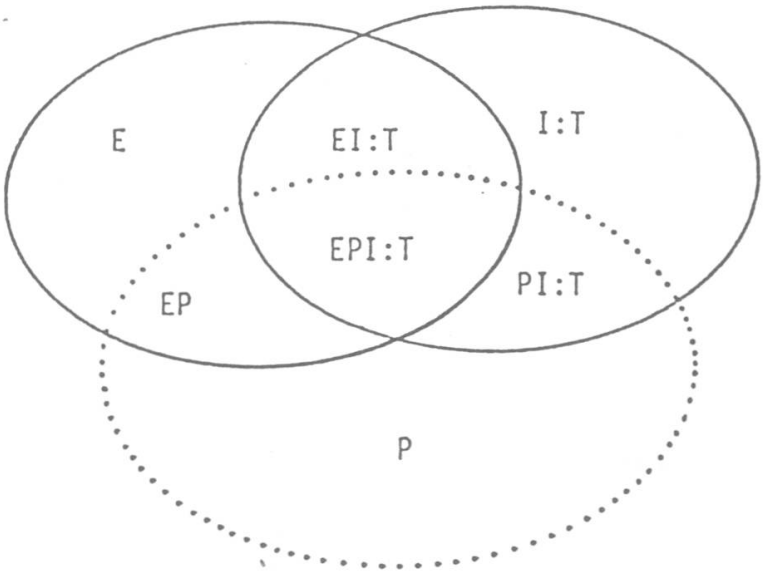

Figure 2: Diagramme de Cronbach simplifié

Par exemple, si l'on veut différencier les élèves en vue de leur proposer une orientation, on utilise comme moyen (ou instrument) de mesure un ensemble de questions. Mais si l'on s'intéresse à comparer la difficulté de diverses formes de questionnement, on recourt à un ensemble d'élèves pour mesurer cette difficulté. Dans le premier cas, la source d'erreur réside dans l'échantillonnage des questions ; dans le second cas, dans l'échantillonnage des élèves.

Ces deux exemples montrent bien pourquoi l'utilisateur doit préciser, au début d'une étude de généralisabilité, quel est l'objet de son étude et de sa mesure (face de différenciation) et quels moyens ou instruments il veut prendre pour cela (face d'instrumentation).

Dans le cas présent, l'objet d'étude est le progrès de chaque élève. Un bon test de progrès doit différencier les phases de l'apprentissage, avant et après enseignement. La facette Phases est donc facette de différenciation et la facette Items facette d'instrumentation.

La signification de la facette Items mérite d'être explicitée. C'est sur elle que porte la généralisation, c'est-à-dire que l'on cherche (en se fondant sur la représentativité assurée par l'échantillonnage aléatoire), à généraliser le constat obtenu à partir de ce groupe particulier de questions à tout l'univers de questions correspondant au même objectif. La démarche logique est la même que celle qui conduit à estimer le résultat d'une élection à partir d'un sondage d'opinion effectué sur un échantillon aléatoire de répondants.

On notera que le dispositif étudié ne comporte qu'une seule facette de généralisation : celle des Items. Il serait possible d'élargir le plan d'observation et de faire corriger chaque copie par plusieurs correcteurs, par exemple. Dans ce cas, la généralisation pourrait porter sur les deux facettes I et C (l'ensemble des items, notés par l'ensemble des correcteurs). La qualité de la mesure est toujours fonction du type de généralisation que l'on souhaite pouvoir faire.

La facette Elèves, tout en faisant aussi partie de la face de différenciation, joue ici un rôle particulier, dans la mesure où l'on s'intéresse à mesurer le progrès de chaque élève pris pour lui-même, indépendamment des résultats des autres. C'est une facette dite «découpée». La variance entre élèves est donc laissée de côté. Les interactions de cette facette avec d'autres sources de variation sont confondues avec ces dernières, comme on le verra plus loin.

Le plan de mesure, finalement, situe les facettes Phases et Elèves sur la face de différenciation et la facette Items sur la face d'instrumentation.

Etude de généralisabilité

Puisque ce sont les deux niveaux de performance (initial et final) de chaque élève qui constituent les objets d'étude à différencier, on pourrait calculer pour chaque sujet la composante de variance correspondant à son progrès. L'estimation qu'on en ferait serait cependant très instable. Si le but est de vérifier l'adéquation du dispositif d'évaluation utilisé à la population examinée, c'est plutôt la valeur discriminative moyenne qu'il faut chercher à apprécier.

Dans ce but, il faut d'abord estimer la variance systématique qu'introduisent les apprentissages individuels, en moyenne. Si le progrès était le même pour

tous les élèves, ce serait la variance entre Phases: $\sigma^2(p)$; mais en fait, il faut y ajouter l'effet des variations des élèves autour de ce progrès moyen: $\sigma^2(ep)$. La variance systématique intra-individuelle liée à l'apprentissage est donc en moyenne:

$$\sigma^2(\tau) = \sigma^2\ (p) + \sigma^2(ep) = 0{,}0182$$

Pour calculer la variance d'erreur, il faut savoir si l'on veut utiliser les mêmes questions, ou non, pour les deux phases. Dans le premier cas, le progrès est mesuré de façon «relative»: l'estimation du gain n'est pas affectée par le choix de questions faciles ou difficiles, puisque la difficulté des questions reste la même. Il ne reste plus que deux sources d'erreur possibles: 1) le fait que certains items soient plus sensibles que d'autres à l'apprentissage, effet dont la variance est donnée par $\sigma^2(pi{:}t)$; 2) la réaction différente des élèves à ces items en début et fin d'apprentissage, mesurée par $\sigma^2(epi{:}t)$. La variance d'erreur «relative» est ainsi:

$$\sigma^2(\delta) = [\sigma^2(pi{:}t) + \sigma^2(epi{:}t)] \ / \ [n(t) \cdot n(i)]$$

Les estimations du tableau 1 donnent $\sigma^2(pi{:}t)$ comme nulle, mais $\sigma^2(epi{:}t)$ comme égale à 0,0522. Il en résulte une variance d'erreur relative pour les gains égale à 0,0025, qui paraît suffisamment faible par rapport à la variance «vraie» de 0,0182 pour que les mesures soient bonnes. Le calcul d'un coefficient de fidélité permet de s'en assurer.

On définit habituellement la fidélité comme la proportion de variance vraie dans la variance observée. Cette dernière serait, si l'on conserve les mêmes items:

$$\sigma^2(X) = \sigma^2(\tau) + \sigma^2(\delta) = 0{,}0182 + 0{,}0025 = 0{,}0207$$

La fidélité de la mesure relative des effets d'apprentissage est donc:

$$\rho^2(\delta) = \sigma^2(\tau)/\sigma^2(X) = 0{,}0182 \ / \ 0{,}0207 = 0{,}879$$

ce qui est tout à fait satisfaisant. On voit donc que l'épreuve utilisée permet de déterminer si le gain d'un élève particulier est important ou non, sans que l'échantillonnage des items utilisés puisse perturber sensiblement cette estimation.

On peut se poser une seconde question cependant, impliquant la mesure absolue des niveaux de départ et d'arrivée. Pour déterminer l'apprentissage réalisé, il n'est pas nécessaire de poser les mêmes questions avant et après apprentissage. Au contraire, il serait préférable de tirer aléatoirement des questions pour constituer deux épreuves différentes.

Ceci aurait l'avantage d'éviter l'effet de la mémorisation des questions initiales. Par contre, la difficulté variable des questions ainsi tirées au hasard viendrait grandement perturber l'estimation des deux niveaux considérés. La variance d'erreur comprendrait, en plus des sources déjà prises en compte ci-dessus, la variance des difficultés des items $\sigma^2(i{:}t)$ et la variance due à la connaissance différente que les élèves peuvent avoir de ces items $\sigma^2(ei{:}t)$. Il s'agit des deux sources de variation les plus importantes d'après le tableau 1. L'estimation de la variance d'erreur «absolue» est alors:

$$\sigma^2(\Delta) = [\sigma^2(pi:t) + \sigma^2(epi:t) + \sigma^2(i:t) + \sigma^2(ei:t)] / [n(t) \cdot n(i)]$$

c'est-à-dire 0,0088. La variance observée de l'effet d'apprentissage individuel est alors estimée à:

$$\sigma^2(X) = \sigma^2(\tau) + \sigma^2(\Delta) = 0{,}0182 + 0{,}0088 = 0{,}0270$$

et la fidélité tombe à 0,0182 / 0,0270 = 0,674.

Cette valeur est habituellement considérée comme insuffisante pour permettre une mesure assurée. Il faudrait donc modifier le dispositif, si l'on voulait poser des questions différentes avant et après l'étude du chapitre sur la chaleur.

Calcul de la marge d'erreur absolue

Dans l'optique d'une évaluation par objectifs, un enseignant peut se poser un autre type de question: tel élève a-t-il atteint une maîtrise du domaine suffisante pour qu'il puisse arrêter son étude et entreprendre un autre apprentissage?

Dans l'école où a été effectué ce travail, le seuil habituellement admis est de 70% de bonnes réponses au test final.

L'estimation du niveau de l'élève est nécessairement affectée par l'échantillonnage des items, et en particulier par la difficulté des questions posées. La variance d'échantillonnage affectant le calcul du taux de réussite après apprentissage vient d'être calculée: c'est la variance d'erreur «absolue», c'est-à-dire $\sigma^2(\Delta) = 0{,}0088$.

L'écart-type des fluctuations d'échantillonnage est la racine carrée de cette valeur (0,0938). On sait que très peu d'observations se situent à plus de deux écarts-types (ici 0,1876) de la moyenne. Très peu d'élèves dont le niveau réel se situe à 70% (c'est-à-dire maîtrisant 70% de tous les items possibles concernant les trois thèmes) auront donc des résultats:

- soit inférieurs à (0,70 – 0,1876), c'est-à-dire 0,51 (51%)
- soit supérieurs à (0,70 + 0,1876), c'est-à-dire 0,89 (89%).

Cette marge d'erreur (70% ± 19%) peut paraître importante. L'incertitude de la mesure provient de ce que la difficulté des questions est mal contrôlée. On pourrait en tout cas certifier la réussite des élèves obtenant 89% de réussite et donner un appui à ceux qui n'atteignent pas 51% de bonnes réponses. Dans les cas intermédiaires, on pourrait recourir à d'autres informations pour trancher, par exemple aux observations faites pendant les exercices pratiques.

Pour réduire cette marge, il faudrait soit accepter un risque plus grand dans les décisions prises, soit améliorer l'épreuve.

Possibilités d'adaptation de l'épreuve

Pour améliorer la mesure, une première possibilité est de contrôler si certains groupes d'items ne perturbent pas l'estimation de la réussite moyenne des

élèves. Si c'était le cas, l'univers des conditions d'observation admissibles pourrait être réduit pour les exclure. En contrôlant mieux la difficulté des questions, on obtiendrait en effet une fidélité meilleure.

L'étude a été effectuée ici pour chacun des trois thèmes. Il est apparu que la fidélité baissait si l'un d'entre eux était écarté: ils sont donc bien tous à conserver.

Une seconde possibilité est d'allonger l'épreuve. En doublant le nombre d'items, par exemple, on diminuerait de moitié la variance d'échantillonnage. La marge d'erreur se ramènerait à 70% ± 13%.

Il n'est pas sûr que le temps nécessaire pour cet examen supplémentaire vaille le gain de précision qu'on peut en attendre.

Conclusions sur le dispositif d'enseignement et d'évaluation choisi

La démarche utilisée pour contrôler la mesure des gains individuels pourrait être adaptée très simplement au contrôle de l'effet du traitement pédagogique. Mais comme l'efficacité de ce dernier ne fait aucun doute et que cet article se centre davantage sur les problèmes d'évaluation, c'est l'idée de mesurer les gains individuels qui sera surtout discutée ici.

Il vaut la peine de noter d'abord que la variance entre élèves n'intervient plus nulle part dans les formules proposées plus haut. Ceci correspond bien à l'intention de la majorité des maîtres, qui souhaiteraient individualiser leur enseignement et éviter de mettre beaucoup de leurs élèves en échec, par le simple fait de les classer.

Mais il fallait s'assurer que l'abandon des comparaisons entre élèves ne conduise pas à l'abandon de toute évaluation. La théorie des tests classique suppose en effet toujours que ce sont les différences entre élèves qui font l'objet de la mesure. Le but de cette étude était de vérifier si une solution de rechange était disponible et praticable.

A partir des résultats du tableau 1, on peut calculer que la fidélité de l'épreuve utilisée dans la perspective compétitive classique serait de 0,853. Or les calculs présentés plus haut montrent que l'on obtient une fidélité *supérieure* dans l'estimation des gains individuels, (c'est-à- dire la comparaison entre les deux phases de l'apprentissage). Ceci provient du fait que la source d'interaction EI:T, qui est très importante, affecte la comparaison des élèves, mais non l'estimation de leur gain individuel.

Ainsi une justification technique, fondée sur cet exemple concret, vient soutenir l'argumentation philosophique: mettre l'élève en compétition avec lui-même plutôt qu'avec les autres peut permettre des mesures plus précises, ou si l'on préfère un allégement de la longueur des épreuves, compensant le temps perdu par la répétition des observations.

L'intérêt socio-affectif d'une centration sur les gains individuels est finalement aussi à faire valoir. L'élève peut ainsi observer ses progrès réels, alors que dans le système de notation habituel, ses notes baissent d'année en année, au fur et à mesure que ses apprentissages progressent (parce qu'elles donnent lieu évidemment à un plus grand nombre de possibilités d'erreurs). Ici, au contrai-

re, les gains possibles sont d'autant plus grands que l'élève est plus faible au départ. Cette forme d'évaluation devrait renforcer la motivation des élèves en difficulté.

Implications de la théorie statistique utilisée

Plusieurs fois dans ce texte, une allusion a été faite à d'autres dispositifs de mesure, répondant à d'autres questions. C'est là en effet le leitmotiv de la théorie de la généralisabilité utilisée dans toutes ces analyses: il importe de clarifier quels sont les objets à différencier et les conditions d'observation échantillonnées pour leur mesure. Il n'existe pas de note vraie, ni d'erreur en soi, mais seulement des observations dont les variations peuvent tantôt être occasions d'études intéressantes, tantôt être considérées comme sources de bruit à minimiser, selon la perspective d'utilisation dans laquelle on se place.

Une conséquence demeurée souvent inaperçue est que l'erreur de mesure, dans l'évaluation des élèves n'est pas où les enseignants la cherchent le plus souvent. Les différences entre correcteurs, aussi graves qu'elles apparaissent dans les études de docimologie, ne constituent en effet qu'une source de variation minime par rapport à l'incertitude concernant l'objet même de la mesure. Les objectifs pédagogiques varient largement d'un enseignant à l'autre, et les méthodes ou techniques d'évaluation également. De plus, la finalité de la mesure peut différer totalement de cas en cas, et l'on est logiquement amené à construire des épreuves différentes (si l'on vise à déterminer un niveau moyen dans une branche, ou au contraire à détecter l'effet d'un apprentissage récent, par exemple). Il est clair que des instruments de mesure à visées divergentes et à contenus distincts, appliqués dans des conditions différentes, ne peuvent pas donner les mêmes résultats.

Beaucoup d'erreurs de mesure sont ainsi inévitables, tant que les enseignants n'ont pas clarifié ce qu'ils voulaient mesurer. L'avantage de la démarche illustrée dans ce texte est au contraire d'obliger à préciser le but que l'on poursuit, de rendre manifeste le résultat atteint, et de faire réfléchir au choix que l'on préfère.

Nécessité de poursuivre le débat

En conclusion, on ne procédera pas de la même façon et on n'obtiendra pas la même information: (1) si l'on veut comparer les connaissances d'un élève à celles de ses camarades d'étude; (2) si l'on veut le situer dans la population des élèves de son âge; (3) si l'on veut établir un bilan du degré de maîtrise qu'il a acquis pour un objectif pédagogique étroitement défini; (4) si l'on cherche à estimer le niveau de connaissance que le même élève a atteint dans un certain domaine du savoir (en laissant varier alors la forme du questionnement); ou encore (5) si l'on cherche à mesurer son progrès individuel par rapport à son niveau de départ, comme l'idée en a été défendue ci-dessus. Il serait même possible d'affiner plus encore les types de mesure à considérer, en diversifiant

davantage les utilisations envisagées (diagnostic de difficultés temporaires, contrôle des préacquis, etc.).

Le choix entre ces différents modèles d'évaluation est lourd de conséquences pédagogiques, psychologiques et sociales. Il ne pourra être fait lucidement que si chercheurs et praticiens commencent à confronter leurs représentations et leurs expériences de l'appréciation du travail des élèves.

L'idée qu'un modèle quantitatif de mesure se révèle finalement inadéquat pour accompagner et guider une démarche éducative globale n'est nullement à exclure. Cette hypothèse semble même constituer une des issues probables du débat proposé. Mais l'accepter serait invalider en même temps le système des notes et des moyennes sur lequel s'appuie encore tout le système scolaire.

C'est dire que le questionnement devrait être réciproque, entre théorie et pratique, pour rechercher ensemble des solutions valables dans ce champ si confus de l'évaluation.

Références

Cardinet, J. L'apport de la théorie de la généralisabilité à l'évaluation sommative individualisée. In: Cardinet, J. *Evaluation scolaire et mesure,* Bruxelles: De Boeck, 1986, p. 119–141.

Cardinet, J. *La construction de tests d'apprentissage selon la théorie de la généralisabilité.* Neuchâtel: Institut romand de recherches et de documentation pédagogiques, 1987. Coll. Recherches 87.107.

Cardinet, J. La mesure d'un apprentissage en physique. Compte rendu des séances du Groupe de travail «Edumétrie». Neuchâtel: Institut romand de recherches et de documentation pédagogiques, 1988. Coll. Recherches 88.111, Cahiers du GCR nº 16.

Zimmermann, M.-L. Contribution à l'étude des conceptions d'élèves et de leurs utilisations dans un processus d'apprentissage. Thèse de doctorat (en préparation). Genève: Université de Genève, Faculté de psychologie et des sciences de l'éducation.

Zimmermann, M.-L. et Bain, D. Que vaut notre évaluation des connaissances? Les avatars d'un questionnaire de physique. In: Actes des Rencontres de Chamonix nº 10: *Communication, éducation et culture scientifique et industrielle: innovations et recherches.* Paris: André Giordan et Jean-Louis Martinand, 1988 p. 503-508.

Zimmermann, M.-L. et Paillard, B. *Apprentissage des sciences expérimentales par l'autonomie (A.P.A.).* Genève: Laboratoire de didactique et d'épistémologie des sciences de l'Université de Genève, 1987.

Zur Messung der Fortschritte der Schüler in Physik

Zusammenfassung

Der eine der beiden Autoren hat eine Lernmethode für wissenschaftliche Fächer entwickelt; sie beruht auf den persönlichen Bemühungen der Schüler, die für Fortschritte in ihren Arbeiten notwendig sind. Der andere Autor schlägt ein Modell zur Messung des individuellen Fortschritts vor und wendet dafür die Theorie der Verallgemeinerbarkeit an. Dieses statistische Modell wurde im

vorliegenden Artikel auf die Angaben bei der Untersuchung des Kapitels Wärme angewendet. Es geht daraus hervor, dass die Lernstufen für jeden einzelnen Schüler sehr genau verglichen werden können. Dies beweist die Möglichkeit einer individuellen Bewertung, vollkommen unabhängig von den Unterschieden zwischen den Schülern. Jedoch betrifft dieses Ergebnis die rein relativen Leistungen und nicht die absoluten, die stark von den unterschiedlichen Schwierigkeitsgraden der Fragen abhängig bleiben.

To measure pupils' progress in physics

Summary

One of the authors developed a method for learning experimental sciences, which rests on the personal research endeavour of each pupil, as a way to modify his representations. The other author proposed a statistical model for measuring individual progress, applying generalizability theory. In this paper, this last model is applied to observations concerning the study of heat. The results show that initial and terminal learning levels can be compared very reliably, for each pupil taken singly. This proves that an individualized evaluation can be carried out, totally independently of between-pupils comparisons. However, this conclusion applies to relative scores of progress and not to absolute scores, which remain too much affected by differences in question difficulty.