Mesure et évaluation en éducation

Généralisabilité et séquences didactiques : illustration et défense d'un modèle à vocation édumétrique

Daniel Bain

Volume 26, numéro 1-2, 2003

Généralisabilité

URI : https://id.erudit.org/iderudit/1088237ar
DOI : https://doi.org/10.7202/1088237ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Bain, D. (2003). Généralisabilité et séquences didactiques : illustration et défense d'un modèle à vocation édumétrique. *Mesure et évaluation en éducation*, *26*(1-2), 19–36. https://doi.org/10.7202/1088237ar

Résumé de l'article

La généralisabilité est un modèle statistique particulièrement adéquat quand il s'agit de mettre au point des dispositifs visant à évaluer des apprentissages. Il permet de vérifier si l'instrument d'évaluation élaboré possède les qualités métrologiques nécessaires pour situer les résultats des apprenants sur l'échelle des performances attendues, donc par rapport à un barème critériel. Pour l'illustrer, nous prenons l'exemple d'un test de physique appliqué au début et à la fin d'un ensemble de cours en laboratoire sur la notion de chaleur. Nous donnons ainsi un échantillon des possibilités du modèle pour traiter quelques-uns des problèmes qui peuvent se présenter dans le cadre d'une séquence didactique.

Généralisabilité et séquences didactiques : illustration et défense d'un modèle à vocation édumétrique

Daniel Bain
Ex-collaborateur du Service de la recherche en éducation, Genève

Mots clés : Généralisabilité, didactique, édumétrie, dispositifs d'évaluation, progrès

La généralisabilité est un modèle statistique particulièrement adéquat quand il s'agit de mettre au point des dispositifs visant à évaluer des apprentissages. Il permet de vérifier si l'instrument d'évaluation élaboré possède les qualités métrologiques nécessaires pour situer les résultats des apprenants sur l'échelle des performances attendues, donc par rapport à un barème critériel. Pour l'illustrer, nous prenons l'exemple d'un test de physique appliqué au début et à la fin d'un ensemble de cours en laboratoire sur la notion de chaleur. Nous donnons ainsi un échantillon des possibilités du modèle pour traiter quelques-uns des problèmes qui peuvent se présenter dans le cadre d'une séquence didactique.

Key words : Generalizability, didactics, edumetrics, evaluation devices, change measurement

Generalizability is a particularly adequate statistical model when it comes to refining devices for the evaluation of learning. It allows one to verify whether an evaluation instrument possesses the necessary psychometric properties to place learners' results on a given scale, that is to say with respect to a previously defined criterion. To illustrate it, we have chosen the example of a physics test on the subject of heat, given at the beginning and the end of a series of laboratory courses. We present examples of the possibilities of the model for dealing with some of the problems that can arise in a didactical sequence.

Palavras-chave: Generalizabilidade, didáctica, edumetria, dispositivos de avaliação, progresso

A generalizabilidade é um modelo estatístico particularmente adequado quando se trata de preparar dispositivos que visam avaliar as aprendizagens. Ele permite verificar se o instrumento de avaliação elaborado, possui as qualidades metrológicas necessárias para situar os resultados dos aprendentes na escala dos desempenhos esperados, isto é, em relação a critérios previamente definidos. Para ilustrá-lo, tomamos o exemplo de um teste de física aplicado no princípio e no fim de um conjunto de cursos, em laboratório, sobre a noção de calor. Damos, assim, exemplos das possibilidades do modelo para tratar alguns dos problemas que se podem apresentar no quadro de uma sequência didáctica.

Introduction

Pourquoi entreprendre une campagne – voire une croisade – en faveur de la généralisabilité comme nous le faisons depuis plusieurs années dans le cadre de la commission Édumétrie[1]? Nous y voyons au moins deux raisons majeures.

Trop souvent, les dispositifs d'évaluation utilisés en sciences de l'éducation n'ont pas les qualités métrologiques qu'impliquent les décisions qu'elles sont censées légitimer, faute des contrôles adéquats. Une des erreurs les plus courantes est de considérer qu'un instrument mis au point dans une certaine perspective, par exemple pour évaluer les compétences des élèves en lecture, est également fiable dans d'autres utilisations : pour comparer le niveau de différentes classes ou l'efficacité de diverses méthodes d'apprentissages dans ce même domaine, ou encore pour estimer des progrès ou le degré de difficulté de certains types de questions[2].

Trop souvent encore, les techniques utilisées pour mettre au point les instruments de mesure dans le domaine de la formation se réfèrent à une approche psychométrique, inadéquate pour mesurer spécifiquement des apprentissages. Or, le formateur ne s'intéresse pas au premier chef aux différences interélèves se traduisant par un barème normatif (en rangs sur 100, stanines, etc.) ; il privilégie une approche édumétrique[3] et critérielle lui permettant de situer les résultats des apprenants sur une échelle de mesure correspondant à la compétence enseignée (par exemple le degré de maîtrise de la lecture à un stade donné de son apprentissage).

Comme nous le montrerons, le modèle de la généralisabilité[4] fournit un outil statistique conséquent avec ce type d'objectif lorsqu'il s'agit de tester la fiabilité d'un dispositif d'évaluation dans le domaine de la formation. L'objet de cet article, qui s'adresse plus particulièrement aux didacticiens, est de défendre l'intérêt et l'utilité de ce modèle en illustrant par quelques exemples son utilisation dans l'enseignement. Pour ce faire, le cadre d'une séquence didactique nous a semblé particulièrement adéquat : il correspond à une structure souvent proposée actuellement pour l'enseignement dans diverses branches (*cf.* Bain & Schneuwly, 1993 ; Dolz, Noverraz & Schneuwly, 2001) ; il vise à donner par ailleurs leurs cohérence et pertinence pragmatiques aux emplois illustrés.

Séquence didactique et problèmes d'évaluation

Dispositifs didactiques et d'évaluation

Pour cette illustration utilisée à titre exemplatif, nous reprendrons, en les aménageant, des données récoltées par Marie-Louise Zimmermann dans la présentation de sa thèse sur l'enseignement de la notion de chaleur dans le cadre d'une recherche-innovation (1990)[5]. Étant donné l'objectif de cet article, nous nous contenterons d'une présentation très sommaire de cette séquence. Ce cours de physique en laboratoire a été donné à l'École de culture générale Jean Piaget à des élèves de 15-19 ans, de niveau scolaire plutôt faible, dont certains avaient déjà suivi quelques leçons sur le concept de chaleur au Cycle d'orientation. Selon les principes mêmes régissant la mise en œuvre d'une séquence didactique, l'enseignante a fait passer aux classes visées un prétest, qui lui a permis d'analyser les connaissances et les conceptions des élèves avant le cours pour les prendre en compte lors de la séquence. À la fin de la séquence, un posttest lui a fourni un bilan de l'opération lui permettant d'estimer les progrès de ses élèves et d'ajuster ultérieurement son enseignement. L'instrument (test) utilisé lors de chacune des deux évaluations encadrant les leçons en laboratoire comportait trois sous-échelles, correspondant aux thèmes suivants :

- T1 : *Thermométrie* ; exemple d'items : lire la température sur des images de thermomètres (nécessite d'interpréter correctement les graduations de l'instrument) ;
- T2 : *Différence entre sensation de chaleur et température* ; exemple : « Dans une pièce, vous touchez la poignée de la porte, un tapis, un journal. […] Classez les corps du plus chaud au plus froid. Est-ce que le classement par température est le même ? Si oui, pourquoi ? Sinon, lequel est-il ? » ;
- T3 : *Changement d'état* ; exemple : « Adèle met un thermomètre dans un creuset contenant du zinc liquide, elle relève la température toutes les minutes. Elle obtient : 500°, 480°, 460°, 420°, 420°, 420°, 420°, 420°… Pourquoi le thermomètre indique-t-il plusieurs fois 420 ? »

Les items se présentent sous la forme de QCM ou de questions ouvertes. Les réponses ont été cotées en juste (1) ou faux (0).

Pour les besoins de notre analyse, nous avons retenu sept items par thème et nous avons considéré les résultats de 40 élèves répartis dans cinq classes. Dans la perspective et la terminologie de l'analyse de la généralisabilité, le *dispositif d'évaluation* présente donc les cinq *facettes* (facteurs) suivantes : les Classes (C), les Élèves dans les Classes (E:C), les Phases de la séquence (P : début et fin), les Thèmes des sous-échelles (T) et les Items à l'intérieur des Thèmes (I:T) définis dans le plan d'observation du tableau 1. Pour chaque facette on précise (deuxième colonne) le nombre de *niveaux* (modalités) qu'elle comporte et l'*univers* dans lequel ces niveaux ont été échantillonnés. Univers considéré comme très grand, pratiquement infini, pour les Classes, les Élèves et les Items (facettes aléatoires infinies) ; univers limité aux deux Phases et aux trois Thèmes observés dans le cas des deux autres facettes (facettes fixées ; *cf.* plan d'estimation du tableau 1).

Tableau 1
Dispositif d'évaluation : plans d'observation et d'estimation

Facettes	*Niveaux*	*Univers*	*Nom*	*Réduction*
C	5	INF	Classes	
E:C	8	INF	Élèves	
P	2	2	Phases: prétest et posttest	Sélection niveau 1 ou 2*
T	3	3	Thèmes des sous-échelles	
I:T	7	INF	Items dans chaque thème	

***** Cette sélection permet d'analyser spécifiquement et successivement le prétest et le posttest.

Dispositif didactique et questions d'évaluation

Dans une telle séquence didactique et avec l'information disponible, on peut se demander quelle est la fiabilité (généralisabilité) du dispositif d'évaluation adopté (test et facettes observées) pour diverses utilisations didactiques et quelles seraient, le cas échéant, les améliorations à lui apporter. Dans cette étude de généralisabilité, nous nous situons fictivement d'un point de vue docimologique dans une étape expérimentale préalable, où il s'agirait de tester le dispositif expérimenté par l'enseignante pour savoir s'il vaut la peine d'y recourir dans d'autres séquences semblables, à quelles conditions, avec quelles restrictions ou moyennant quelles améliorations.

Les questions suivantes correspondent, pour la généralisabilité, à différents *plans de mesure* (*cf*. tableau en annexe) testant différents problèmes d'évaluation.

1. Au *prétest*, peut-on de façon fiable

1.1 évaluer le niveau de connaissances préalables des *élèves* pour les regrouper dans des cours à niveaux ou pour différencier l'enseignement en conséquence ?

1.2 différencier les *classes* selon le niveau de connaissances préalables pour en tenir compte dans l'enseignement ?

1.3 différencier le niveau de connaissances préalables correspondant aux trois *thèmes* pour adapter en conséquence l'enseignement de ces trois sous-chapitres ?

1.4 relever, en différenciant le niveau de réussite aux *items*, les notions ou compétences plus ou moins maîtrisées par les élèves pour les aborder de façon adaptée dans la séquence ?

2. En comparant les résultats au *posttest et* au *prétest,* peut-on de façon fiable mesurer des *progrès* ?

3. Au *posttest*, peut-on de façon fiable

3.1 repérer les *élèves,* qui à la fin de la séquence, ont des résultats particulièrement faibles et ont encore besoin d'un appui ?

3.2 repérer les *classes* qui ont un niveau de maîtrise insuffisant pour analyser ce qui s'est passé et organiser éventuellement un complément de formation ?

3.3 différencier le degré de réussite selon les *thèmes*, identifier celui ou ceux qui «ont mal passé», pour envisager les domaines de révisions qui s'imposent ou pour adapter ultérieurement l'enseignement dans ces domaines ?

3.4 relever à travers les *items* les notions ou compétences mal maîtrisées par une majorité d'élèves pour en tirer diverses conséquences didactiques ?

On constate la grande diversité des exploitations possibles du dispositif d'évaluation. Un logiciel (*Etudgen* pour Macintosh ou *EduG* pour PC) permet de traiter facilement les données selon les divers plans évoqués ci-dessus. Nous laisserons ici de côté les problèmes d'utilisation du modèle et des logiciels en renvoyant pour cela le lecteur au «mode d'emploi» que nous avons élaboré (Bain & Pini, 1996). Nous nous concentrerons sur les analyses suggérées ci-dessus et dont les résultats sont résumés dans le tableau annexé, auquel nous invitons le lecteur à se référer tout au long du chapitre suivant.

Radiographie d'une séquence didactique

Apports de la généralisabilité pour l'analyse d'un dispositif d'évaluation

Avant de commenter les résultats et pour faciliter la lecture du tableau en annexe, présentons sommairement l'information statistique et docimologique que nous fournit l'étude de généralisabilité pour analyser notre *dispositif d'évaluation*. Précisons d'abord que nous entendons par là non seulement l'instrument (épreuve, test, échelle, questionnaire, grille d'évaluation) mais également les conditions et le contexte de cette évaluation, définis ici par les plans d'observation et d'estimation du tableau 1.

Dans une première étape, l'analyse de généralisabilité

- fournit deux indices globaux de la fiabilité du dispositif sous la forme de *coefficients de généralisabilité relative et absolue* (ρ^2 *rel.* et *abs.*)*;* le premier est utilisé quand on vise seulement à sérier, hiérarchiser les résultats ; le second, quand on cherche en plus à les situer sur l'échelle (de performance, d'attitude, etc.) du test considéré ; chacun des deux coefficients est considéré comme satisfaisant s'il est égal ou supérieur à 0,80 ;
- signale les facettes ou combinaisons de facettes (interactions) qui contribuent le plus à l'*erreur de mesure*, favorisant ainsi le diagnostic des facteurs ou biais affectant l'évaluation et donnant des pistes en ce qui concerne les améliorations envisageables ou non ;
- calcule l'*écart type de l'erreur de mesure* permettant d'estimer un *intervalle de confiance* ou *intervalle d'incertitude* à prendre en compte quand on fixe des seuils sur l'échelle de l'instrument (par exemple 50 % de réussite).

Dans une seconde étape, visant à étudier les possibilités d'amélioration du dispositif, l'analyse

- aide à relever dans certaines facettes les *niveaux* (par exemple certains items) qui contribuent à diminuer la fiabilité du dispositif et dont la suppression améliorerait par conséquent la généralisabilité ;
- fournit une estimation de la généralisabilité du dispositif si on lui apporte certaines modifications (par exemple en augmentant le nombre d'items).

L'évaluation des connaissances initiales : analyses du prétest

Le principe même de la séquence didactique telle que nous l'entendons (Bain & Schneuwly, 1993) implique qu'on s'intéresse aux savoirs et aux représentations déjà là au départ. L'expérience montre en effet que, dans plusieurs branches, sur beaucoup de chapitres du programme (en particulier en première secondaire), une partie des élèves possèdent avant même tout enseignement formel des connaissances ou des compétences non négligeables. C'est *a fortiori* le cas quand, comme dans l'exemple étudié ici, le chapitre a été abordé antérieurement pour une partie des élèves. Il est donc légitime de se demander si les données apportées par le prétest ne pourraient pas aider l'enseignant à différencier son enseignement. La réponse est négative ou réservée en ce qui concerne les trois premiers objectifs d'évaluation définis ci-dessus (1.1, 1.2 et 1.3) : le dispositif d'évaluation n'a pas la fiabilité nécessaire pour distinguer des différences de réussite exploitables pédagogiquement entre les élèves, les classes ou les thèmes (ρ^2 rel. ou abs. < 0,80 ; *cf.* tableau annexé, §1.1, 1.2 et 1.3). C'est le cas en particulier si l'on cherche à situer les résultats sur l'échelle des performances ou de rendement (de 0 % à 100 % de réussite) propre au test, ce qui implique que l'on considère le ρ^2 absolu.

Aurait-on pu s'attendre à un tel résultat ? On nous objecte souvent que cela n'a pas de sens d'imposer aux élèves un contrôle préalable sur des notions qui n'ont pas encore été enseignées ou révisées, que dans ce cas les réponses des élèves sont nécessairement très aléatoires (*cf.* en particulier les réponses au hasard dans les QCM). Cette objection n'est que partiellement pertinente. Ce type de flou dans les réponses explique probablement une partie de l'interaction Élèves x Items (EI:CT), source d'erreur majeure qui entache l'évaluation du niveau des *élèves* au départ de la séquence (ρ^2 abs. = 0,58). En conséquence, si l'on cherchait comme prévu à fixer un seuil pour constituer deux groupes de niveau ou des degrés contrastés de maîtrise, par exemple à 40 % de réussite, la marge d'incertitude (*intervalle de confiance* autour du seuil) serait telle (± 19 %) que l'entreprise n'aurait pratiquement pas de sens : entre 20 % et 60 % de réussite, il faudrait recourir à d'autres données pour diminuer l'aléa de la décision de placement. Si l'on voulait améliorer la généralisabilité du dispositif dans cette perspective – et diminuer l'intervalle d'incertitude –, l'analyse dite d'*optimisation* montre qu'il faudrait pratiquement tripler la longueur du test, ce qui serait probablement irréalisable sur le plan pratique.

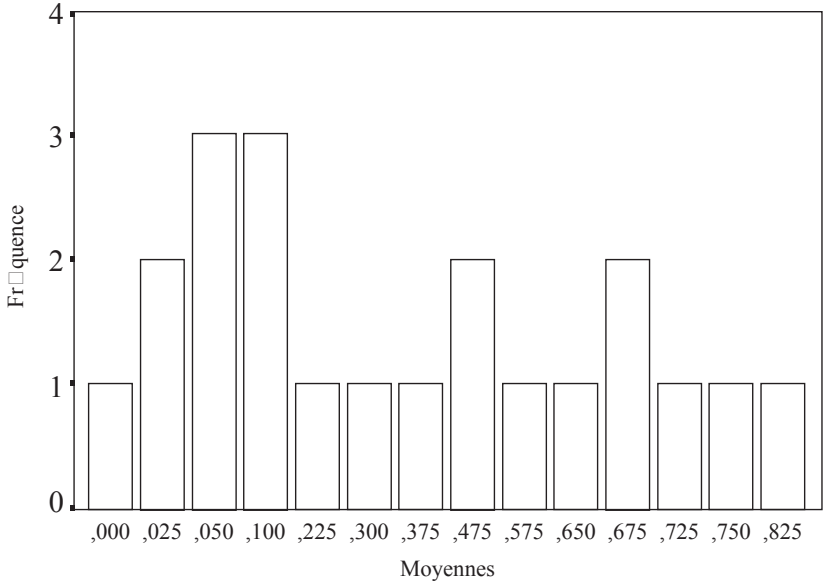

Figure 1. ***Prétest : distribution des moyennes de réussite par item***

Mais dans les trois cas cités (évaluation des élèves, des classes et des thèmes), on relève une autre source d'erreur qui pourrait tenir à la situation particulière où le test contraste fortement (*cf. infra*) les notions maîtrisées par une majorité d'élèves des autres notions non encore abordées ou participant moins d'une culture scientifique commune. La composante d'erreur principale tient en effet à la grande diversité de réussite des différents *items* (figure 1), qui affecte spécifiquement la généralisabilité absolue. C'est ainsi que, s'ajoutant à l'hétérogénéité des élèves dans chaque groupe, la forte variabilité de réussite des items empêche d'estimer avec précision le niveau moyen des *classes* (annexe §1.2 : ρ^2 abs. : 0,322), malgré des différences non négligeables entre ces dernières (de 43 % à 27 %). Et dans ce cas, les tentatives d'optimisation aboutissent à des solutions pratiquement irréalistes : il faudrait plusieurs milliers d'items ! On peut aussi voir dans ces résultats la conséquence d'une stratégie d'organisation des classes qui a veillé à composer des groupes à peu près de même niveau, sans prendre en compte spécifiquement la formation reçue ou non antérieurement en physique.

A fortiori, la même source d'erreur I:T affecte fortement la mesure des différences de difficulté des *thèmes* (annexe §1.3). La grande hétérogénéité des taux de réussite des items à l'intérieur de chacun d'entre eux ne permet

pas de cibler avec précision et fiabilité les moyennes de ces sous-chapitres (ρ^2 rel. et abs. = env. 0,70). En d'autres termes, la variance intrathèmes est nettement plus importante que la variance interthèmes. De ce fait, on ne peut pas affirmer sans autres précautions que, dans leur ensemble, les notions de *thermométrie* (moyenne 56%) sont mieux connues au départ que celles relatives à la *différence entre sensation de chaleur et température* (31%), l'intervalle d'incertitude (27%) étant plus grand que la différence entre les deux moyennes[6]. Pour assurer la mesure (ρ^2 rel. et abs. $\geq$ 0,80), il faudrait doubler la longueur du test (14 items par thème), réduisant ainsi l'intervalle d'incertitude à 20%.

Comme nous l'annoncions plus haut, la quatrième analyse sur le prétest (1.4) montre qu'il est effectivement possible de différencier avec une bonne assurance différents degrés de maîtrise (ou de difficulté potentielle) des notions contrôlées à travers les *items* considérés individuellement (ρ^2 abs. : 0,939). La figure 1 illustre le fait que les taux de réussite varient de 0% à 83%.

Les questions qui sont déjà assimilées par les deux tiers ou plus des élèves concernent essentiellement la *thermométrie*; celles qui sont maîtrisées par moins d'un tiers d'entre eux portent surtout sur la *différence entre sensation de chaleur et température*. En se basant sur une analyse plus clinique du contenu des questions, l'enseignant peut donc, dans une certaine mesure, anticiper les principaux obstacles qu'il va rencontrer (pour la différence entre sensation de chaleur et température, par exemple), la stratégie à mettre en œuvre en fonction des différentes notions à enseigner, ainsi que le temps à leur consacrer (il pourrait notamment passer plus rapidement sur certains points du sous-chapitre thermométrie).

La mesure des progrès : analyse des différences entre les deux moments d'évaluation

Peut-on évaluer avec quelque sécurité les *progrès* accomplis lors de la séquence au moyen du dispositif adopté? La généralisabilité est un des rares modèles qui permettent de le vérifier en prenant comme facette de différenciation (objet à mesurer) les *phases*, soit les deux moments de l'évaluation[7]. Les coefficients relatif et absolu (0,98 et 0,78) attestent que l'on peut estimer avec une bonne fiabilité le progrès moyen des élèves (en moyenne 24%) et le situer sur l'échelle du test, même si le second coefficient est légèrement inférieur à 0,80. Il suffirait d'ailleurs d'un item de plus pour que ce seuil soit atteint (*cf.* plan d'optimisation pour huit items par thème); la principale source d'erreur absolue est en effet, dans ce cas également, la diversité des items

(I:T). La faiblesse des interactions entre P et les autres facettes (dans l'analyse de variance précédant celle de généralisabilité) montre en outre que les progressions entre les deux phases sont relativement parallèles quand on compare entre les deux moments les classes, les thèmes ou les items.

L'*analyse de facette* appliquée à la facette Classes[8] montre qu'en écartant un des cinq groupes, on passerait tout juste le seuil de $\rho^2 = 0{,}80$. Mais cette amélioration serait minime et il serait difficile de dire en quoi et pourquoi cette classe fonctionne autrement que les autres.

Le lecteur didacticien sera peut-être déçu du progrès constaté en moyenne. Nous lui conseillons de réaliser une telle opération dans son propre enseignement, avant de porter un jugement sur ce résultat : nous sommes souvent très optimistes sur le rendement de nos cours faute d'avoir mesuré les connaissances préalables des apprenants.

L'évaluation des connaissances finales : analyses du posttest

Le posttest livre des renseignements sur une situation que l'on ne peut vraiment qualifier de finale sur le plan didactique : des interventions sont encore possibles dans la suite de l'enseignement et des leçons à tirer pour une autre année.

Si l'on analyse les scores finaux des *élèves* (dispositif 3.1), on constate que la séquence a eu notamment deux résultats. Elle a augmenté plutôt les différences entre les apprenants : l'écart type de la distribution des scores moyens passe de 14 % au prétest à 21 % au posttest (différence cependant statistiquement non significative). Cette tendance va à l'encontre des objectifs d'une pédagogie de maîtrise (Bloom, 1972 ; Hubermann, 1988) ; elle est cependant très couramment observée : elle tient à la difficulté de gérer pédagogiquement les différences d'origines diverses entre les élèves (Perrenoud, 1991). Ce premier effet a aussi pour corollaire de contribuer à améliorer la généralisabilité du dispositif pour l'évaluation du niveau des élèves en augmentant la variance de différenciation (au numérateur de la formule de la généralisabilité ; les variances d'erreurs, au dénominateur, n'ont pas augmenté corrélativement).

Il résulte de ce dernier fait que, malgré une amélioration de la généralisabilité, les intervalles d'incertitude (relatif et absolu) sont restés pratiquement identiques à ceux observés au prétest : ± 16 % et ± 19 % sur l'échelle du test. Ces intervalles dépendent en effet uniquement des variances d'erreurs relative ou absolue, qui restent importantes, principalement à cause des fluctuations aléatoires dues à l'interaction Elèves x Items (EI:CT) et aux items (I:T). De ce

fait, il serait pratiquement difficile de fixer un seuil de réussite (par exemple 50% de rendement au test) par rapport auquel on déciderait d'un appui. On prendrait alors quelque risque d'attribuer inutilement une telle remédiation (par exemple entre 30% et 50%) ou d'écarter à tort des élèves apparemment au-dessus du seuil fixé (entre 50% et 70%). Pour diminuer l'intervalle d'incertitude de moitié (de 19% à 9,5%), des essais d'*optimisation* aboutissent à la conclusion qu'il faudrait pratiquement quadrupler la longueur du test !

L'expérience montre qu'une telle marge d'incertitude se présente assez souvent et que nous avons tendance à surestimer la précision des résultats que nous obtenons avec des instruments de ce type, assez représentatifs des épreuves fabriquées habituellement par les enseignants. Si d'autres données ne sont pas prises en compte (*cf.* les stratégies séquentielles proposées par Cronbach et Gleser, 1965), nos décisions fondées sur des scores ou des notes risquent bien d'être affectées d'un arbitraire certain.

La généralisabilité fournit en outre une réponse à une autre question intéressante si l'on cherche à évaluer la *réussite globale* de la séquence : le dispositif permet-il d'estimer où se situe la moyenne de la population d'élèves par rapport à un certain standard de performance ? En d'autres termes, compte tenu des erreurs d'échantillonnage, peut-on mesurer avec fiabilité la distance entre la réussite moyenne observée au posttest (58%) et un seuil déterminé, par exemple 67% (l'élève moyen réussit les deux tiers des questions) ? Le calcul d'un coefficient critérié (phi(lambda) de Brennan et Kane, à évaluer comme un coefficient de généralisabilité absolue) montre que c'est possible ($\rho^2 = 0{,}82$). En se référant à ce seuil, on conclurait que la cible fixée arbitrairement à la séquence n'a pas été atteinte.

La mesure des différences entre *classes* (3.2) se révèle clairement non généralisable à ce stade de la séquence et découragerait toute tentative d'interventions différenciées en fonction de ce mode de regroupement des élèves. En effet, une fois prise en compte la variance des autres composantes intervenant dans le dispositif, la variance du facteur Classes, et par conséquent la variance de différenciation dans le plan de mesure C/EP_2TI, est nulle. L'importante dispersion des résultats dans chacun des cinq groupes (écart type d'environ 20%) confirme que cela n'aurait pas de sens de considérer telle classe comme *globalement* plus faible que les autres.

Il est de même difficile de distinguer de façon fiable la réussite finale pour les trois *thèmes* (3.3 : ρ^2 abs. : 0,376), même si les taux de réussite sont nettement plus bas (43%) pour les items relatifs à la différence entre sensation de

chaleur et température que pour ceux appartenant aux deux autres sous-chapitres (70 % et 61 %). L'intervalle d'incertitude (pour la comparaison entre deux modalités de ce facteur) est trop grand (28 %) et la source principale d'erreur est de nouveau la grande variabilité de réussite des items à l'intérieur de chaque thème. Une optimisation du dispositif est, dans ce cas aussi, pratiquement impossible.

On en est donc réduit à considérer finalement les résultats aux *items* individuels (3.4). Comme dans le prétest, il apparaît en effet possible de distinguer de façon fiable différents degrés de réussite aux questions opérationalisant diverses notions du chapitre testé (marge de réussite : de 20 % à 98 % ; coefficients de généralisabilité rel. et abs. > 0,90 ; intervalles d'incertitude pour la comparaison entre items respectivement de 21 % et 22 %). Les progrès étant relativement parallèles d'un item à l'autre, les notions distinguées comme les moins connues ou les plus difficiles au départ se révèlent finalement aussi les moins bien maîtrisées. On retrouve ainsi une situation analogue à celle décrite ci-dessus pour le prétest (figure 1) avec un décalage sur la droite de 5 % à 35 % selon les items. Les obstacles rencontrés – notamment les freins épistémologiques relativement bien connus (Bachelard, 1986 ; Zimmermann, 1990) – opposent une résistance que nos moyens didactiques ont de la peine à surmonter dans les conditions habituelles de l'enseignement.

Conclusion : de l'intérêt de la généralisabilité en didactique

Si la généralisabilité peut être utilisée avec profit par tous les évaluateurs en sciences de l'éducation, ce modèle nous semble particulièrement adéquat pour traiter les problèmes docimologiques que se posent les didacticiens. Nous espérons en avoir déjà fait la démonstration par l'exemple exposé ci-dessus. En conclusion, nous voudrions souligner les points sur lesquels cette approche de la mesure présente à nos yeux un intérêt majeur.

C'est avant tout sur ses vertus méthodologiques que nous voudrions insister. Le modèle incite le chercheur ou le praticien à prendre en compte, dans la phase d'élaboration du *plan de mesure,* les différents facteurs (facettes) du *contexte d'enseignement-apprentissage* susceptibles d'influencer l'évaluation, et souvent de jouer également un rôle dans la séquence didactique : les caractéristiques individuelles des apprenants, mais aussi leur appartenance à divers

groupes tels que sexe, catégorie sociale, classe, établissement ou filière ; ou encore la phase de l'enseignement dans laquelle on se situe, ou la méthode pratiquée. On constate parfois que les résultats individuels des élèves dépendent pour une part non négligeable de l'environnement dans lequel ils travaillent. Dans l'exemple ci-dessus, disposant de l'information Classe, nous avons pu vérifier que cette facette avait pratiquement un poids faible dans la variance de différenciation des élèves au prétest et une valeur pratiquement nulle au posttest (dispositifs 1.1 et 3.1) ; et que dans l'estimation des progrès (2.) cette facette ne contribuait que peu aux erreurs relative et absolue. Dans le cas de cette séquence, le groupe dans lequel les apprenants étaient intégrés n'influait donc pas ou guère sur l'évaluation de leurs performances. Trop souvent ces facteurs contextuels sont ignorés et on ne prend *de facto* en considération dans le dispositif d'évaluation que les caractéristiques individuelles, appauvrissant ou biaisant ainsi les analyses.

En ce qui concerne les *caractéristiques de l'instrument*, le modèle permet également de contrôler l'influence du regroupement des items selon différents critères : objectifs, chapitres ou thèmes ; niveaux de complexité *a priori* des notions, des problèmes ou des textes ; modalités de formulation des questions ou des réponses. L'adjonction d'une facette Correcteur peut aider, le cas échéant, à vérifier l'importance d'une certaine subjectivité dans l'appréciation des réponses (en introduisant une correction multiple).

Sur le plan docimologique, notre exemple montre, avec bien d'autres (*cf.* Bain, à paraître), qu'il n'existe pas d'instruments « bons à tout faire ». Le test présenté ci-dessus n'a pas les qualités métrologiques nécessaires pour mesurer des différences entre Classes ou entre Thèmes. Or, on constate sur le terrain la tendance de certaines administrations scolaires (*cf.* par exemple SKBF, 2002) à utiliser des épreuves standardisées évaluant les *élèves* pour contrôler le niveau des *classes,* sans toujours vérifier préalablement la fiabilité d'une telle différenciation. Dans ce cas, convoquer les maîtres des classes les plus faibles pour leur demander de s'expliquer sur leurs résultats risque bien de tenir d'un certain arbitraire, quand on connaît la marge d'erreur qui affecte ce type de mesure.

Dans le domaine de la recherche en didactique, un plan comme celui que nous avons présenté ci-dessus pourrait être complété de façon intéressante en ajoutant notamment les facettes Enseignants et Méthodes (facettes cachées dans notre exemple), ce qui supposerait naturellement qu'on engage dans

l'expérience d'autres enseignants pratiquant d'autres approches pédagogiques. On éviterait de cette façon de généraliser sans vérification nos constats à d'autres situations didactiques.

D'un point de vue méthodologique et docimologique, le modèle de la généralisabilité, de par sa nature même, contraint à définir le *mode d'échantillonnage* des niveaux (modalités) des facettes considérées, donc le type d'erreurs qu'on cherche ou non à contrôler et par conséquent les généralisations qu'on s'autorise ou s'interdit. En fixant la facette Thèmes dans le dispositif ci-dessus, on tient compte du fait que la sélection de ces trois objectifs à l'intérieur du chapitre Chaleur n'est pas aléatoire. Du même coup, on neutralise les erreurs d'échantillonnage dues à ce facteur, mais on renonce à généraliser à d'autres thèmes (à la calorimétrie, par exemple).

Une autre vertu cardinale du modèle est de permettre, en recourant au coefficient absolu, de vérifier la fiabilité du dispositif d'évaluation quand on cherche à situer les résultats par rapport à certains points de *l'échelle utilisée par l'instrument*. C'est souvent le cas en didactique, où l'on s'efforce de donner aux contrôles un caractère représentatif des compétences visées par la séquence et où l'on recourt à des *barèmes critériels*. En effet, on veut généralement situer les résultats par rapport à un *seuil*, souvent exprimé en pourcentage de rendement de l'épreuve. On a vu que le coefficient qui correspond à cet objectif d'évaluation, le ρ^2 absolu, est relativement exigeant, considérant dans les sources d'erreurs non seulement les composantes d'interaction entre la ou les facettes de différenciation et les autres facettes (notamment l'interaction Élèves x Items) mais aussi les effets principaux. Selon les cas, la grande diversité des élèves ou des items échantillonnés crée un certain flou par rapport à ce que l'on cherche à cibler par l'évaluation (par exemple, le niveau des Classes ou la difficulté des Thèmes). Ce flou – et l'intervalle d'incertitude qui le traduit – doivent nous inciter à la prudence dans nos conclusions, et encore plus dans nos décisions.

Car le modèle se présente aussi comme une aide précieuse à la décision, en particulier ses algorithmes d'*analyse de facettes*, aidant à relever les sources d'erreurs (par exemple certains items ou certaines classes) et sa routine d'*optimisation*, testant différents plans d'observation et d'estimation pour améliorer la qualité ou l'efficience du dispositif. Il indique les pistes à explorer et signale sans équivoque certaines impasses, par exemple celles dans lesquelles on

s'engagerait si, dans la séquence sur le concept de Chaleur, on envisageait d'aménager le dispositif pour différencier les classes ou les thèmes au posttest en augmentant le nombre de questions.

Par ses exigences, le modèle ne satisfera pas le constructeur de tests qui ne vise qu'à estampiller son instrument au moyen d'un indice simple comme un coefficient de fidélité. Il apportera en revanche beaucoup de renseignements précieux à celui qui veut mieux saisir le fonctionnement de l'évaluation dans son contexte et profiter de l'éclairage que celle-ci peut lui apporter sur les apprentissages qu'il cherche à mesurer et à améliorer.

Annexe

Présentation sommaire des résultats des analyses de généralisabilité sur les données de la séquence didactique

Objectif d'évaluation et plan de mesure	*Coefficients de généralisabilité*	*Information complémentaire et remarques*
1. Prétest		***Évaluer les connaissances initiales***
1.1 Évaluer le niveau de connaissances préalables des *Élèves* pour les regrouper en cours à niveaux ou différencier l'enseignement. EC/P_1TI	ρ^2 rel. : 0,667 ρ^2 abs. : 0,581	Moyenne générale des scores au prétest : 34 % de réussite avec un écart type de 14 %. Intervalles d'incertitude par rapport à un seuil donné : rel. = ± 16 %; abs. = ± 19 %. Sources d'erreurs absolues les plus importantes : EI:CT = 66 % et I:T = 31 % de la variance d'erreur totale. Optimisation : ρ^2 rel. = 0,80 pour I = 14; ρ^2 abs. = 0,80 pour I = 20.
1.2 Différencier les *Classes* selon le taux moyen de réussite pour y adapter l'enseignement C/EP_1TI	ρ^2 rel. : 0,513 ρ^2 abs. : 0,322	Moyennes des 5 classes : 0,40, 0,43, 0,31, 0,30, 0,27. Principales sources d'erreur abs. : I:T = 55 % et E:C = 25 %. Optimisation pratiquement impossible : impliquerait un nombre irréaliste d'items.
1.3 Différencier le degré de connaissances préalables selon les *thèmes* pour en tenir compte dans l'enseignement $T/CE\ P_1I$	ρ^2 rel. : 0,707 ρ^2 abs. : 0,690	Moyennes de réussite pour chaque thème = resp. 56 %, 16 %, 31 %. Intervalles d'incertitude pour une comparaison entre thèmes : rel. = 27 %; abs. = 28 %. Principale source d'erreur relative et absolue : I:T = resp. 92 % et 85 % de la variance d'erreur totale. Optimisation pour I = 14 : ρ^2 rel. = 0,82 et ρ^2 abs. = 0,80.

1.4 Identifier à travers les *items* le degré de maîtrise ou de difficulté des notions à enseigner IT/ CEP_1	ρ^2 rel. : 0,947 ρ^2 abs. : 0,939	Marge de réussite aux 21 items : de 0 % à 83 % (*cf.* figure 1). Intervalle d'incertitude pour une comparaison entre items. rel. = 19 %; abs. = 20 %.
2. Posttest – pretest		***Evaluer les progrès***
Mesurer le *progrès moyen* des élèves. P/CETI	ρ^2 rel. : 0,99 ρ^2 abs. : 0,78	Différence entre posttest et prétest : 58 % - 34 % = 24 %; écart type : 16 %. Intervalles d'incertitude par rapport à un seuil donné : rel. = ± 3 %; abs. = ± 12 %. Principale source d'erreur absolue : I:T = 71 %. Optimisation : ρ^2 abs. = 0.80 pour I = 8.
3. Posttest		***Évaluer les connaissances finales***
3.1 Évaluer le niveau de connaissances final des *Élèves* pour distinguer ceux qui ont besoin d'un appui. EC/P_2TI	ρ^2 rel. : 0,853 ρ^2 abs. : 0,805	Moyenne générale des scores au posttest : 58 % de réussite avec un écart type de 21 %. Intervalles d'incertitude par rapport à un seuil donné : rel. = ± 16 %; abs. = ± 19 %. Sources d'erreurs absolues les plus importantes : EI:CT = 68 %, I:T = 29 % de la variance d'erreur totale. Coefficient critérié comparant la moyenne générale à un seuil de 67 % de réussite = 0,82.
3.2 Repérer les *classes* qui ont un niveau de maîtrise insuffisant pour prendre les mesures adéquates. C/EP_2TI	ρ^2 rel. : 0,000 ρ^2 abs. : 0,000	Moyennes des classes resp. : 68 %, 63 %, 56 %, 50 %, 53 %. Variance de différenciation C = 0,0, d'où impossibilité d'optimiser le dispositif par rapport à cet objectif de mesure.
3.3.Relever le ou les *thèmes* moins bien réussis et qui nécessiteraient une révision. T/CE P_2I	ρ^2 rel. : 0,399 ρ^2 abs. : 0,376	Moyennes de réussite pour chaque thème = resp. 70 %, 43 %, 0,61 %. Intervalles d'incertitude pour une comparaison entre thèmes : rel. = 27 %; abs. = 28 %. Principale source d'erreur relative et absolue : I:T = resp. 84 % et 76 % des variances d'erreur totales relative et absolue. Optimisation : pratiquement impossible.
3.4 Repérer à travers les *items* les notions non ou mal maîtrisées pour les réviser. IT/CEP_2	ρ^2 rel. : 0,918 ρ^2 abs. : 0,905	Marge de réussite aux 21 items : de 20 % à 98 %. Intervalle d'incertitude pour une comparaison entre items : rel. = 21 %; abs. = 22 %.

NOTES

1. Cette commission est un des groupes de travail de la Société suisse pour la recherche en éducation. Y participent actuellement, outre le soussigné, madame D. Hexel, messieurs J. Cardinet, Fr. Ducrey et G. Pini.
2. *Cf.* aussi à ce sujet Bain, à paraître.
3. Rappelons que l'édumétrie, selon la définition de V. de Landsheere (1988, p. 59) est un « mot créé par Carver (1974) sur le modèle de psychométrie, pour désigner l'étude quantitative des variables relatives aux apprentissages suscités par l'éducation : influence d'une action pédagogique, performance effective par rapport à une performance attendue, épreuves centrées sur les objectifs… ». (Cité par Demeuse, 2002, p. 3.)
4. Pour la présentation de la généralisabilité, voir ici même l'article de Mokonzi ; pour une présentation statistique, se référer à Cardinet et Tourneur, 1985 ; pour un « mode d'emploi » du modèle, consulter Bain et Pini, 1996.
5. Ces données avaient été traitées à l'époque par la commission Édumétrie. Soulignons le fait qu'elles ne correspondent qu'à une petite partie des données récoltées par la doctorante, qu'elles ont été aménagées pour le traitement par le modèle de la généralisabilité et qu'elles ne rendent pas justice à l'ensemble du travail pédagogique réalisé par l'enseignante à l'occasion de cette séquence.
6. L'intervalle d'incertitude donne un point de référence commode, mais approximatif, parce qu'il néglige le risque d'erreur lié aux comparaisons multiples.
7. Pour une évaluation individualisée des progrès, *cf.* l'article de Cardinet dans cette revue.
8. Dans le plan d'observation adopté, l'analyse de facettes ne peut pas être appliquée aux facettes Élèves et Items, les deux logiciels disponibles ne permettant pas de traiter des facettes nichées (incluses dans d'autres comme I:T et E:C). Par ailleurs, des analyses réalisées sur un plan d'observation différent, laissant de côté la facette Thèmes, ne signalent pas d'items dont la suppression améliorerait substantiellement la généralisabilité.

RÉFÉRENCES

Bachelard, G. (1986). *La formation de l'esprit scientifique* (13e éd.). Paris : Vrin.

Bain, D. (à paraître). Qualité de la formation et qualité des dispositifs d'évaluation : défense et illustration du modèle de la généralisabilité. *Actes du 15e colloque international de l'ADMEE-Europe et congrès annuel de la SSRE : La qualité dans la formation et l'enseignement, comment la définir, comment l'évaluer ?* Lausanne : Institut suisse de pédagogie pour la formation professionnelle (ISPFP).

Bain, D. & Pini, G. (1996). *Pour évaluer vos évaluations. La généralisabilité : mode d'emploi.* Genève : Centre de recherches psychopédagogiques, Direction générale du Cycle d'orientation.

Bain, D. & Schneuwly, B. (1993). Pour une évaluation formative intégrée dans la pédagogie du français : de la nécessité et de l'utilité des modèles de référence. In L. Allal, D. Bain & Ph. Perrenoud (dir.), *Évaluation formative et didactique du français* (pp. 51-79). Neuchâtel et Paris : Delachaux et Niestlé.

Bloom, B.S. (1972). *Apprendre pour maîtriser*. Lausanne : Payot (traduction G. Lorenz).

Cardinet, J. (à paraître dans ce numéro). Cinq dispositifs pour vérifier le progrès. *Mesure et évaluation en éducation.*

Cardinet, J. & Tourneur, Y. (1985). *Assurer la mesure*. Berne : Peter Lang.

Cronbach, L.J. & Gleser, G.C. (1965). *Psychological tests and personnel decisions*. Urbana : University of Illinois Press.

de Landsheere, V. (1988). *Faire réussir, faire échouer. La compétence minimale et son évaluation*. Paris : Presses universitaires de France.

Demeuse, M. (2002). Édumétrie et psychométrie. *Bulletin de l'ADMEE-Europe*, n° 2002/2, 34.

Dolz, J., Noverraz, M. & Schneuwly, B. (2001). *S'exprimer en français. Séquences didactiques pour l'oral et pour l'écrit. Notes méthodologiques, vol. IV*. Bruxelles : De Boeck, Corome.

Hubermann, A.M. (dir.) (1988). *Assurer la réussite des apprentissages scolaires ? Les propositions de la pédagogie de maîtrise*. Neuchâtel et Paris : Delachaux et Niestlé.

Mokonzi, G.B. (à paraître dans ce numéro). Le modèle de la généralisabilité : une théorie de la mesure en éducation. *Mesure et évaluation en éducation.*

Perrenoud, Ph. (1991). Des différences culturelles aux inégalités scolaires : l'évaluation et la norme dans un enseignement indifférencié. In L. Allal, J. Cardinet & Ph. Perrenoud (éds), *L'évaluation formative dans un enseignement différencié* (6e éd., pp. 25-65). Berne : Peter Lang.

SKBF (2002). BL : Mini-PISA an Progymnasien. *Newsletter, juillet 2002* (d'après la Basler Zeitung, 22.3.2002). Aarau : Schweizerische Koordinationsstelle für Bildungsforschung.

Zimmermann-Asta Riccardo, M.-L. (1990). *Concept de chaleur. Contributions à l'étude des conceptions d'élèves et de leurs utilisations dans un processus d'apprentissage.* Thèse n° 172, Genève, Faculté de psychologie et des sciences de l'éducation.