Mesure et évaluation en éducation

Cinq dispositifs pour vérifier le progrès

Jean Cardinet

Volume 26, Number 1-2, 2003

Généralisabilité

URI: https://id.erudit.org/iderudit/1088239ar
DOI: https://doi.org/10.7202/1088239ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Cardinet, J. (2003). Cinq dispositifs pour vérifier le progrès. *Mesure et évaluation en éducation*, *26*(1-2), 51–59. https://doi.org/10.7202/1088239ar

Article abstract

The generalizability model offers a number of different strategies for checking an individual's progress. The individual's performance is necessarily observed on two different occasions and then compared. But the resulting judgment may or may not be generalized over the criteria corresponding to the educational objectives, over the indicators associated with each criterion, or over the assessment conditions on the two occasions (these three factors being nested in one another). Five measurement designs are presented and discussed.

Cinq dispositifs pour vérifier le progrès

Jean Cardinet
Ancien responsable du Service de la recherche de l'IRDP, Neuchâtel

MOTS CLÉS : Évaluation individualisée, certification, progrès, généralisabilité, dispositif

Le modèle de la généralisabilité suggère une série de stratégies différentes pour vérifier si une personne progresse. Sa performance doit nécessairement être observée à deux occasions, puis comparée, mais on peut (ou non) généraliser son jugement à l'ensemble des critères qui opérationalisent les objectifs éducatifs, ou des indicateurs utilisés pour chaque critère, ou à toutes les conditions d'observation en ces deux occasions (ces trois facteurs étant emboîtés). Cinq dispositifs de mesure possibles sont présentés et discutés.

KEY WORDS : Individualized evaluation, certification, progress, generalizability, design

The generalizability model offers a number of different strategies for checking an individual's progress. The individual's performance is necessarily observed on two different occasions and then compared. But the resulting judgment may or may not be generalized over the criteria corresponding to the educational objectives, over the indicators associated with each criterion, or over the assessment conditions on the two occasions (these three factors being nested in one another). Five measurement designs are presented and discussed.

PALAVRAS-CHAVE: Avaliação individualizada, certificação, progresso, generalizabilidade, dispositivo

O modelo da generalizabilidade oferece uma série de estratégias diferentes paraverificar se uma pessoa progride. A prestação individual deve, necessariamente, ser observada em dois momentos distintos e depois comparada. Porém, o juízo pode (ou não) generalizar-se ao conjunto dos critérios que operacionalizam os objectivos educativos, aos indicadores utilizados para cada critério, ou a todas as condições de observação nestes dois momentos (estando estes três factores interligados). São apresentados e discutidos cinco dispositivos possíveis de medida.

Cadre conceptuel et problématique

Cette présentation se situe dans le contexte d'un enseignement et d'une évaluation individualisés. Il s'agit de pouvoir certifier qu'un apprenant a fait des progrès par rapport à un certain objectif, sans le comparer à d'autres apprenants (pour éviter les méfaits de la compétition scolaire) et sans prétendre contrôler l'atteinte finale et globale de l'objectif (ce qu'aucun examen, dans les conditions habituelles, ne saurait assurer). La démarche proposée s'appuie sur la théorie de la généralisabilité, mais cherche un modèle théorique de référence qui soit pertinent par rapport à la situation observée. La question traitée est celle-ci : à quelles fluctuations d'échantillonnage va-t-on comparer le progrès, pour s'assurer qu'il n'est pas dû au hasard ?

Méthode utilisée

Un même ensemble de données (huit mesures appareillées, prises avant et après apprentissage) sert à comparer la généralisabilité du progrès dans chacun des cinq dispositifs envisagés : 1) échantillonnage des critères ; 2) des critères et des indicateurs ; 3) des indicateurs seulement ; 4) des indicateurs non appareillés ; 5) des présentations.

Les données

Deux textes ont été évalués à l'aide des huit critères de la grille de Forgette-Giroux et Simon[1] (en annexe). Les résultats recueillis apparaissent à la figure 1. Pour pouvoir effectuer des comparaisons, on admettra que ces valeurs restent valables pour les cinq dispositifs.

Critères	1	2	3	4	5	6	7	8	
Avant	2	2	2	1	3	1	2	3	Moy. = 2
Après	3	4	2	2	4	2	4	3	Moy. = 3
Super-critères	A		B		C		D		

Figure 1. ***Données de base***

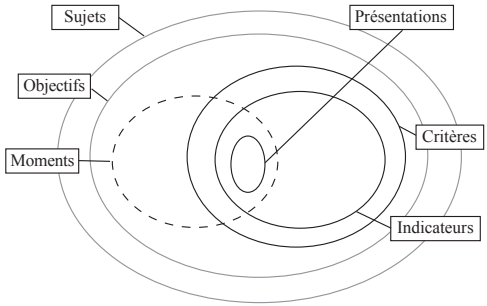

Figure 2. ***Représentation des dispositifs par des diagrammes***

Le progrès correspond à l'effet de la facette Moments, fixée à deux niveaux, Avant et Après, qui est représentée par une ellipse en traitillés épais, vers la gauche du diagramme à la figure 2.

Les fluctuations aléatoires peuvent provenir des facettes en trait plein (Critères, Indicateurs et Présentations), qui sont emboîtées les unes dans les autres[2]. Comme on n'observe qu'un seul sujet et un seul objectif, les facettes correspondantes sont dites cachées, et ici représentées en pointillés.

Principaux résultats

Dispositif 1 : Progrès mesuré par rapport aux fluctuations entre critères

Pour savoir si la moyenne des huit rubriques est plus élevée la seconde fois que la première, on pourrait calculer un « t appareillé » qui donnerait le résultat suivant :

Différence de moyenne : $3 - 2 = 1{,}00 \Rightarrow$ D.L. : $7 \Rightarrow t = 3{,}742 \Rightarrow p = 0{,}0072$

D'après ce résultat, le progrès ne ferait aucun doute, mais on n'aurait pas réfléchi au modèle statistique que l'on emploierait. Or ce dernier suppose que l'on ait tiré au hasard les critères opérationalisant l'objectif, ce qui n'est pas du tout le cas : dans la réalité, les critères sont liés à l'objectif et non aléatoires.

Le premier dispositif envisagé s'appuierait sur les mêmes présupposés (des critères tirés au hasard). C'est pourquoi, même si la généralisabilité était très bonne (0,929), cela n'aurait généralement pas de sens de calculer ce coefficient de fidélité de la mesure du progrès.

Dispositif 2 : Progrès mesuré par rapport aux fluctuations entre critères et entre indicateurs

On pourrait envisager que deux sources de fluctuations aléatoires interviennent à la fois, dues à l'échantillonnage des critères, puis des indicateurs pour chaque critère. On supposerait alors que les mêmes données, fournies par la grille d'évaluation de français, correspondent à quatre critères plus globaux, avec deux indicateurs chacun, comme présentés au tableau 1.

Tableau 1
Critères et indicateurs de la grille d'évaluation de français

Super-critères	*Numéros*	*Indicateurs*
Clarté de la visée	1	Prise en compte du destinataire
	2	Référence au but recherché
Clarté des idées	3	Développement des idées
	4	Organisation des idées
Expression	5	Richesse du vocabulaire
	6	Structure des phrases
Correction	7	Modes et temps des verbes
	8	Orthographe

Le dispositif 2 est un peu plus défendable que le premier. Les quatre «super-critères» sont moins redondants, donc plus indépendants, que les huit premiers. Les deux indicateurs correspondant à chaque super-critère sont assez proches, mais cela correspond bien au modèle qui veut qu'ils soient comme des mesures répétées de la même réalité, représentée par le super-critère.

On n'observe à chaque fois qu'une seule présentation, c'est-à-dire un seul thème à traiter, une seule occasion, un seul contexte, examinateur, correcteur, etc. (Les Présentations restent une facette cachée.) Le coefficient d'importance de l'effet, «Oméga carré relatif», vaut dans ce cas 0,911 et le résultat semblerait fiable si l'échantillonnage aléatoire des critères ne restait pas discutable.

Dispositif 3 : Progrès mesuré par rapport aux fluctuations entre indicateurs (supposés appareillés), les super-critères étant cette fois supposés fixés

La comparaison aux résultats précédents est facile à réaliser. On annonce au logiciel que les mêmes données correspondent à quatre super-critères fixés, avec deux indicateurs tirés aléatoirement pour chacun, parmi une multitude d'autres indicateurs possibles, et identiques avant et après apprentissage.

On continue à n'observer qu'une présentation (thème, occasion, observateur, correcteur) à chaque fois. (Les Présentations sont encore une facette cachée.)

La valeur de «Oméga carré relatif» pour la comparaison des Moments fixés est de 0,906.

L'effet de l'apprentissage reste marqué par rapport aux fluctuations d'échantillonnage des indicateurs et une certification du progrès semble justifiée par la fiabilité de sa mesure.

Cependant, cette confiance plus grande qu'on peut avoir en la pertinence et l'applicabilité du modèle est payée par un univers de généralisation plus étroit. L'appréciation du progrès est limitée aux quatre super-critères fixés avec lesquels on a travaillé.

Dispositif 4 : Progrès mesuré par rapport aux fluctuations entre indicateurs (cette fois non appareillés), les super-critères restant supposés appareillés et fixés

Les données à recueillir pour utiliser ce dispositif 4 se présentent de la même façon que pour les dispositifs 1 à 3, à la figure 1. Il n'est besoin encore que de huit observations avant, et huit après apprentissage. Deux observations sont recueillies chaque fois pour chaque super-critère, mais elles se rapportent à deux indicateurs choisis au hasard au moment considéré. (L'indicateur X n'est donc pas le même Avant et Après.) Les facettes Indicateurs et Présentations sont ainsi confondues, et emboîtées dans l'interaction Moments X Critères.

La valeur calculée pour le coefficient «Oméga carré relatif» pour la comparaison des Moments fixés baisse considérablement et se limite à 0,684.

Les résultats sont donc cette fois insuffisamment fiables. Ceci provient du fait que, les indicateurs n'étant plus appareillés, on ne peut plus mesurer le progrès par une comparaison par paire qui annulerait l'effet de leur difficulté propre. Les différences de difficulté des indicateurs font partie des fluctuations d'échantillonnage, de même que les différences entre présentations (thèmes, etc.).

On peut examiner alors quelle aide apporteraient les démarches d'optimisation. Le logiciel permet d'estimer quelle serait la fidélité du dispositif si l'on échantillonnait davantage d'indicateurs. On voit au tableau 2 qu'à condition de recueillir quatre indicateurs par critère, on obtiendrait une fidélité suffisante de la mesure, même sans que les indicateurs soient appareillés.

Tableau 2
Optimisation du dispositif 4

Nombre d'indicateurs par critère	2	3	4	5	6	7
Nombre total d'observations	16	24	32	40	48	56
Oméga carré	0,684	0,765	0,813	0,844	0,867	0,883

Ce résultat est important, parce qu'il indique un type de dispositif où l'on a beaucoup plus de chances de maîtriser effectivement les fluctuations aléatoires, parce que le modèle statistique de référence est défendable. On peut penser qu'il correspond à la réalité.

Dispositif 5 : Progrès mesuré par rapport aux fluctuations entre présentations, les critères et indicateurs étant appareillés et fixés

On objectera peut-être au dispositif précédent qu'il est difficile de trouver quatre indicateurs par critère : d'où un cinquième dispositif envisageable, échantillonnant les Présentations.

Pour estimer la variance entre Présentations, il faut qu'il y ait deux observations pour chaque cellule, au moins. Ceci oblige à doubler le nombre d'observations.

On a alors des données du genre de celles de la figure 3, où les moyennes par cellule correspondent aux valeurs utilisées précédemment :

	Quatre super-critères							
	Clarté de la visée		Clarté des idées		Expression		Correction	
Avant	1	3	2	1	2	1	1	2
	3	1	2	1	4	1	3	4
Après	2	3	1	3	5	3	3	2
	4	5	3	1	3	1	5	4
Indicateurs	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**

Figure 3. ***Données comportant deux observations par cellule***

La valeur du coefficient « Oméga carré relatif » pour la comparaison des Moments fixés est alors de 0,662.

Les résultats sont du même genre que ceux obtenus avec le dispositif précédent parce que, si Critères et Indicateurs sont fixés et ainsi appareillés entre les moments Avant et Après, les Présentations, au contraire, ne peuvent pas être appareillées. Chaque observation suppose une nouvelle formulation de la question, un nouveau contexte, etc.

Au vu de l'instabilité du sujet considéré, on doit conclure que le progrès n'est pas suffisamment assuré, malgré que l'on ait recueilli quatre observations par critère, un nombre qui semblait devoir être suffisant d'après le dispositif précédent. Mais on échantillonnait des indicateurs dans ce dernier cas, et ce sont des présentations cette fois-ci. Or la variance entre présentations est plus du double de la variance entre indicateurs. (Il s'agit là d'une circonstance particulière, propre aux données fictives que nous avons choisies, pour conserver les moyennes antérieures.)

Une étude d'optimisation effectuée avec ce dispositif 5 montre qu'il faudrait échantillonner chaque indicateur à quatre reprises (et non plus deux) avant et après l'apprentissage, pour assurer statistiquement la mise en évidence du progrès de cet individu spécialement instable. (Oméga carré relatif atteindrait 0,797 avec quatre observations par cellule.)

Au moins est-on sûr, avec ce dernier dispositif, de comparer le progrès réalisé à une estimation raisonnable de l'erreur d'échantillonnage sur la mesure.

Il faudrait encore s'assurer que les quatre observations se font dans des contextes aussi variés que possible.

Conclusions

Le but de cette présentation était de mettre en garde contre une utilisation non réfléchie des outils statistiques. Ce n'est pas parce qu'on peut calculer un test de «t» que son résultat a un sens. De même, ce n'est pas parce qu'on peut calculer un coefficient de généralisabilité qu'on peut l'interpréter directement. Il faut que le modèle utilisé (ici le choix des facettes échantillonnées) puisse raisonnablement correspondre à la réalité observée.

Les objectifs éducatifs ne peuvent pas être choisis au hasard, et la plupart du temps les critères et les indicateurs non plus. Ils font corps avec l'objectif. Ce sont les façons de les observer : contextes d'examen, juges, moments de l'observation, forme du questionnement, etc., qui peuvent varier. Ce sont donc ces facettes qu'il faut échantillonner.

Finalement, il faut bien comprendre ce que signifie le seuil choisi, c'est-à-dire un coefficient de généralisabilité supérieur à 0,80. On ne fait que comparer le progrès observé aux erreurs causées par les fluctuations d'échantillonnage, pour s'assurer qu'il est nettement plus grand. Toute interprétation ou conclusion dépend donc de ce choix relativement arbitraire de l'étalon d'erreur. C'est une faiblesse certaine, mais on ne peut pas faire mieux.

Si on est conscient de cet arbitraire, mais qu'on peut défendre son choix, on peut par contre présenter sa décision de certification (ou de non-certification) avec assurance, parce qu'on s'appuie sur un modèle éprouvé et contrôlé de la réalité. C'est alors qu'on peut vraiment prétendre avoir une démarche scientifique.

Annexe

Exemple d'un instrument d'évaluation utilisable pour contrôler le progrès en français, pour des textes de type argumentatif, explicatif ou informatif

(adaptation de la grille proposée par Forgette-Giroux et Simon, p. 35)

	Critères	*Niveau 1*	*Niveau 2*	*Niveau 3*	*Niveau 4*
1	Prise en compte du destinataire	pas d'indication de prise en compte de l'autre	peu de souci apparent du destinataire	destinataire sollicité parfois	attention au destinataire maintenue
2	Référence au but recherché	l'intention d'écriture n'apparaît pas	intention d'écriture inégalement respectée	adaptation de l'écrit au but poursuivi	texte bien organisé en vue de l'effet recherché
3	Développement des idées	certains points essentiels manquent	développement des idées insuffisant	nombre d'idées suffisant	beaucoup d'idées et de développements
4	Organisation des idées	manque d'ordre dans les idées	organisation des idées peu apparente	marqueurs soulignant la progression	planification souple et subtile
5	Richesse du vocabulaire	répétitions, erreurs de sens	vocabulaire pauvre, mots passe-partout	vocabulaire approprié	texte riche et complexe
6	Structure des phrases	phrases incomplètes ou incohérentes	phrases minimales, sans variété dans la forme	subordonnées, compléments circonstanciels	style varié adapté à l'effet recherché
7	Modes et temps des verbes	emploi incorrect des modes ou des temps	mélange des temps faisant difficulté	modes et temps employés correctement	utilisation de toutes les possibilités de la langue
8	Orthographe	fautes gênant la compréhension du texte	fautes évitables avec outils de référence	orthographe généralement correcte	pas de faute, même sur des points délicats

NOTES

1. Adaptation de la grille proposée par R. Forgette-Giroux et M. Simon (1997), *Une nouvelle façon d'évaluer le progrès de l'élève : le dossier d'apprentissage, p. 35*. Communication présentée à la 19e Session d'études de l'ADMEE-Canada : *Construire une culture de l'évaluation*, Hull.
2. Dans le dispositif 4, les Indicateurs sont confondus avec les Présentations.

RÉFÉRENCES

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.

Cardinet, J. (1994). Control of the value of an intra-subject measurement design. In D. Laveault, B.D. Zumbo, M.E. Gessaroli & M.W. Boss (Eds), *Modern Theories of Measurement : Problems and Issues* (pp. 181-212). Ottawa : Faculté d'éducation, Université d'Ottawa.