Document generated on 07/13/2025 5:19 a.m.

Mesure et évaluation en éducation

Pour évaluer les qualités docimologiquesdes tests de maîtrise
L'intérêt de recourir à la généralisabilité

Daniel Bain

Volume 33, Number 2, 2010

URI: https://id.erudit.org/iderudit/1024895ar
DOI: https://doi.org/10.7202/1024895ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

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

Explore this journal

Cite this article

Bain, D. (2010). Pour évaluer les qualités docimologiquesdes tests de maîtrise : l'intérêt de recourir à la généralisabilité. *Mesure et évaluation en éducation*, *33*(2), 35–63. https://doi.org/10.7202/1024895ar

Article abstract

Many pedagogical assessments controlling knowledge and competences present themselves as mastery tests. They indeed intend to evaluate achievement stated as objectives to attain by the curriculum, and are often considered as a prerequisite to further learning. Firstly, we demonstrate that the methods frequently used in the elaboration of these kinds of tests are hardly adequate. They strongly encourage eliminating of «too easy» items, although these can be of as great an interest as others to the didactician. We then illustrate the advantages of the generalizability model, using a detailed example (a grammar exam at University level). This model allows us to control the major docimological quality of a mastery test: its capacity to reliably differentiate learners who attain the set cut score from those who do not. It further offers a range of useful information for the development and the interpretation of the test.

Tous droits réservés © ADMEE-Canada - Université Laval, 2010

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

https://apropos.erudit.org/en/users/policy-on-use/

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

https://www.erudit.org/en/

Pour évaluer les qualités docimologiques des tests de maîtrise : l'intérêt de recourir à la généralisabilité

Daniel Bain
Société suisse pour la recherche en éducation

Mots clés : Édumétrie, test de maîtrise, généralisabilité, seuil de réussite, coefficient critérié, analyse d'items

Bien des épreuves pédagogiques, contrôlant périodiquement les connaissances et compétences des apprenants, se présentent comme des tests de maîtrise. Elles visent en effet l'évaluation des acquis fixés comme objectifs par le plan d'études et sont souvent jugés comme des prérequis pour la suite des apprentissages. Nous montrerons en introduction que les approches fréquemment utilisées pour l'élaboration de ce type d'épreuves sont peu adéquates, incitant en particulier à écarter abusivement les items « trop bien réussis », tout aussi intéressants que les autres pour le didacticien. Nous illustrerons ensuite par un exemple détaillé (un examen de grammaire au niveau universitaire) l'intérêt du modèle de la généralisabilité. Il permet en effet de vérifier la qualité docimologique majeure d'un test de maîtrise : sa capacité à distinguer de façon fiable les apprenants qui satisfont ou non au critère de réussite fixé. Il apporte en outre toutes sortes de renseignements utiles pour la mise au point de l'épreuve et pour son interprétation.

Key words: Edumetrics, mastery tests, generalizability, criterion cut-off score, item analysis

Many pedagogical assessments controlling knowledge and competences present themselves as mastery tests. They indeed intend to evaluate achievement stated as objectives to attain by the curriculum, and are often considered as a prerequisite to further learning. Firstly, we demonstrate that the methods frequently used in the elaboration of these kinds of tests are hardly adequate. They strongly encourage eliminating of «too easy» items, although these can be of as great an interest as others to the didactician. We then illustrate the advantages of the generalizability model, using a detailed example (a grammar exam at University level). This model allows us to control the major docimological quality of a mastery test: its capacity to reliably differentiate learners who attain the set cut score from those who do not. It further offers a range of useful information for the development and the interpretation of the test.

Palavras-chave: Edumetria, testes de maestria, generalizabilidade, limiar de sucesso, coeficiente criterial, análise de itens

Muitas provas de avaliação pedagógica que controlam periodicamente os conhecimentos e competências dos alunos apresentam-se como testes de maestria. Estas provas visam, com efeito, a avaliação das aprendizagens definidas como objectivos pelo plano de estudos e muitas vezes consideradas como pré-requisitos para futuras aprendizagens. Em primeiro lugar, demonstraremos que as abordagens frequentemente utilizadas para a elaboração deste tipo de provas são pouco adequadas, incentivando em particular a eliminar abusivamente os itens "mais fáceis", os quais são tanto ou mais interessantes que os outros. De seguida, ilustraremos através de um exemplo detalhado (um exame de gramática no ensino universitário) o interesse do modelo da generalizabilidade. Com efeito, este modelo permite verificar a qualidade docimológica de um teste de maestria: a sua capacidade de distinguir de modo fiável os alunos que satisfazem ou não o critério de sucesso fixado, contribuindo ainda com um conjunto de informações úteis para o desenvolvimento e interpretação do próprio teste.

Note de l'auteur – Toute correspondance peut être adressée comme suit : Daniel Bain, Groupe de travail Édumétrie, Société suisse pour la recherche en éducation, route du Moulin-Roget 49, CH-1237 Avully (Genève, Suisse), téléphone : +41 22 756 34 7, ou par courriel à l'adresse suivante : [daniel.bain@bluewin.ch].

Introduction : un malentendu docimologique

Pour introduire notre propos, qui se situe dans une perspective spécifiquement édumétrique[1] et didactique, nous partirons d'un type d'incidents dont nous avons été témoin plus d'une fois. Une administration scolaire décide d'enquêter sur l'état des connaissances et des compétences d'une cohorte d'élèves. Elle confie cette tâche à un groupe (un consortium) de spécialistes de l'évaluation (docimologues, statisticiens), chargés de construire une épreuve pédagogique dite de référence. Ce consortium fait appel pour cette élaboration à des didacticiens ou à des enseignants qui connaissent bien la matière et le plan d'études concernés ; il leur demande de rédiger un certain nombre de questions représentatives du champ à couvrir. Les items proposés font l'objet d'un essai sur un échantillon de la population visée et sont traités statistiquement.

C'est lors de la sélection des items en vue de la forme définitive du test que s'amorce parfois le malentendu docimologique introductif à notre propos. Les didacticiens ou les enseignants s'entendent dire, en effet, que certaines des questions proposées sont « trop faciles », réussies par la très grande majorité des élèves ; qu'elles sont « peu informatives » ou « peu discriminatives » ; qu'elles n'ont pas les caractéristiques statistiques nécessaires et qu'elles doivent de ce fait être laissées de côté. Pour des pédagogues, une affirmation de ce type est étonnante, voire scandaleuse. Pour eux, effectivement, des questions bien réussies apportent une information importante à plus d'un titre : elles attestent notamment que l'enseignement a été adéquat et efficace, que l'on peut tabler sur certains acquis pour la suite des apprentissages. Leur étonnement est d'autant plus grand si l'enquête évaluative (*survey*) est censée porter sur les « standards de formation minimaux » du plan d'étude commun, lorsqu'il s'agit d'« instruments [qui doivent permettre aux autorités scolaires] de déterminer les compétences de base (standards) que tous les élèves doivent acquérir » (CDIP, 2006, p. 2). En l'occurrence, c'est la définition que donne l'opération HarmoS (Harmonisation de la scolarité obligatoire en Suisse) des épreuves de référence. Elle correspond aussi en France aux contrôles des « compétences de bases » faisant partie de ce que le rapport Thélot (2004)

appelait le « socle commun des indispensables », soit « la maîtrise des connaissances, des compétences et des règles de comportement indispensables pour toute la vie » (p. II).

Pour bien des enseignants, ce genre d'épreuve est conçu implicitement comme un *test de maîtrise*, même si le terme n'est généralement pas employé. Dans cette perspective, pour être vraiment utile sur le plan didactique, l'évaluation devrait porter sur un certain *fundamentum*, sur les notions considérées comme des prérequis pour la suite des apprentissages ou pour des activités ultérieures. Elle devrait avoir pour objectifs principaux de distinguer à la fois les élèves qui ont atteint ou non le seuil d'acquisition fixé et les notions (à travers les items) qui peuvent être considérées comme assimilées ou non. La qualité primordiale pour un tel contrôle est une bonne *validité de contenu* : « Un test de connaissance qui ambitionne de faire l'inventaire des acquisitions à la fin d'un cycle d'études, suivant un programme déterminé, doit réellement couvrir les aspects importants de ce programme » (De Landsheere, 1992, p. 323). En outre, les consignes et les modalités de présentation des stimuli et de correction doivent être en adéquation avec l'objet mesuré (Laveault & Grégoire, 2002, chap. 4.2). En pratique, ce type de validité est surtout contrôlé par concordance entre le choix ou la forme des items et les prescriptions du plan d'études. Pour l'enseignant, et plus particulièrement pour le didacticien, s'ajoute l'exigence, primordiale, d'une *validité conceptuelle* (de *construct*) : les questions ou exercices de l'épreuve doivent correspondre – dans leur contenu, dans leur forme et surtout dans les opérations qu'ils mobilisent – à la conception qu'ils se donnent des objets d'enseignement-apprentissage, au modèle théorique qui guide l'enseignement et dont les présupposés sont soumis à diverses vérifications empiriques (De Landsheere, 1992, article *validité de construit* ; Laveault & Grégoire, 2002, chap. 4.4). Par exemple, du point de vue de la validité tant de contenu que de construit, dans une approche de type socio-constructiviste, compléter des phrases à trous par les connecteurs adéquats ne représentera probablement pas valablement la capacité à rédiger un texte argumentatif ; suivre et exécuter un protocole d'expérience ne reflétera que très partiellement la compétence expérimentale visée par la didactique des sciences. Quant au pouvoir discriminatif des items, dans la perspective d'un test de maîtrise, il ne correspond à une qualité attendue que dans la mesure où il contribue à situer les élèves ou les notions par rapport à un *critère de maîtrise* (appelé aussi *seuil de suffisance*). On recourt alors à un *indice de*

discrimination au seuil de maîtrise (Pini, 2009). Et dans ce cas, la forme de la distribution des résultats escomptée est une courbe en J (asymétrique à droite) supposant une majorité de questions réussies par une majorité d'apprenants.

En revanche, pour les consortiums chargés d'élaborer les enquêtes évaluatives ou les épreuves de référence, la perspective est souvent *de facto* bien différente, même si leurs objectifs semblent convergents avec ceux que nous venons d'évoquer et si, pour eux aussi, les items doivent être représentatifs des objectifs et du contenu du plan d'études. Reprenons, à titre d'illustration, l'opération HarmoS déjà mentionnée ci-dessus. Là où les enseignants entendent *standards (minimaux) de formation*, les spécialistes de l'évaluation, conformément aux instructions reçues, construisent leurs tests à partir de « *standards de performance* fondés […] sur un cadre de référence incluant des *niveaux de compétence* » (CDIP, 2007, p. 4, art. 7.2a). On peut détecter facilement en arrière-plan une référence à des modèles de compétence et à des niveaux analogues à ceux définis par l'enquête PISA (*Programme for International Student Assessment*) en littératie, en mathématiques ou en sciences (OCDE, 2007). Tout le dispositif d'évaluation mis en place (sur le plan fédéral ou cantonal) caractérise les épreuves élaborées comme des *tests de niveaux*. Il s'agit de différencier des niveaux de compétences allant des « attentes fondamentales » à la « maîtrise de problèmes complexes »[2]. En outre, les « instruments de développement et d'assurance qualité » élaborés dans un but de coordination et de monitorage doivent permettre de situer – donc de différencier – les performances à l'échelon des systèmes scolaires cantonaux ou des établissements, selon les cas[3]. Le traitement des données et l'étalonnage de ce type de tests font conséquemment appel à la *théorie des réponses aux items* (TRI). C'est souvent en référence à ce type de modèles statistiques (et plus précisément aux courbes caractéristiques et courbes d'information d'items) que certaines questions sont considérées comme « peu informatives » parce trop bien réussies et médiocrement discriminatives, même si certaines d'entre elles sont conservées comme représentatives des premiers niveaux de compétences. En trop grand nombre, des items « trop faciles » affaiblissent la validité de l'épreuve compte tenu de ses objectifs effectifs.

Or, le malentendu docimologique que nous voulons mettre ici en évidence tient précisément à une question de *validité*, entendue au sens général que Bertrand et Blais (2004) donnent à ce concept : « De notre point de vue, ce sont les interprétations des scores qui doivent être considérées comme valides ou non, pas le test en lui-même » (pp. 238-239). L'important est donc ce que

l'on fait ou veut faire des résultats de l'évaluation, ainsi que les modalités de l'évaluation, en accord avec l'objectif fixé. Or, contrairement à une opinion trop souvent répandue, et opérationnalisée dans des pratiques sur le terrain, une épreuve n'est pratiquement jamais « bonne à tout faire » (Bain,Weiss & Agudelo, 2008; Cardinet, 1977). Par exemple, apte à tester les performances des élèves en mathématiques, telle épreuve se révélera d'une médiocre fiabilité pour évaluer des différences de réussite selon les établissements et les classes ou pour contrôler la progression des apprentissages. Par ailleurs, le modèle statistique utilisé pour évaluer la qualité docimologique de l'épreuve doit être conséquent avec l'objectif de l'évaluation, principe qui nous amène au sujet de notre contribution : l'évaluation de la qualité des tests de maîtrise.

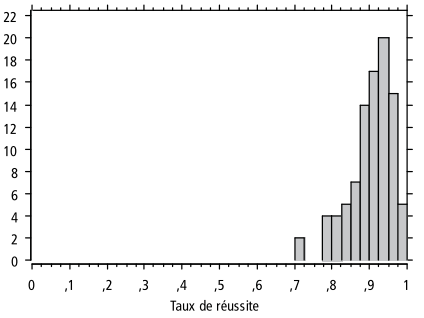
Test de maîtrise : définition et illustration

Nous reprenons et précisons ici la définition d'un test de maîtrise, déjà esquissée plus haut, en nous référant à la description qu'en font Cardinet et Tourneur dans leur ouvrage de 1985 *Assurer la mesure* (p. 252) :

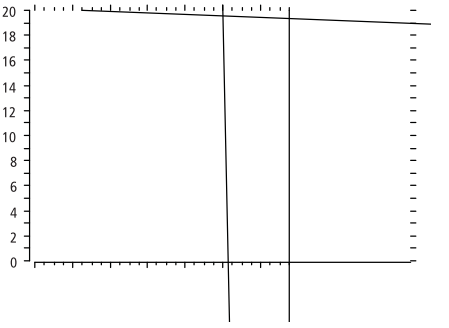
« Un test de maîtrise possède les propriétés suivantes :

1. Au contraire des tests classiques, la performance d'un étudiant n'est pas située par rapport à une performance moyenne d'un groupe qui sert de référence ; elle est comparée à un seuil absolu de réussite dans un univers de tâches ;
2. L'univers doit être suffisamment bien défini pour qu'on puisse en extraire un échantillon aléatoire de tâches ou d'items, et surtout donner une définition précise à la performance qui est observée. Le modèle statistique que nous utilisons[4] (l'analyse de variance) ne suppose pas l'homogénéité à l'intérieur de l'univers, mais seulement l'échantillonnage aléatoire des items ;
3. L'intérêt de l'examinateur étant de savoir si le score univers de l'étudiant (le score que ce dernier obtiendraiul

Examen 2005 (Année 1)



Examen 2008 (Année 2)



évidente, il est très important de vérifier que les fluctuations dans la difficulté des questions échantillonnées ne créent pas d'injustice d'un examen à l'autre. Dit autrement, d'une année à l'autre les examens devraient être en principe des versions parallèles du même test et de difficulté comparable.

Pour des raisons pédagogiques également, l'épreuve devrait en outre permettre de distinguer de façon fiable les domaines (parties du test) et notions (questions) plus ou moins bien réussies dans l'ensemble. Les résultats des analyses guideront ainsi les remédiations à apporter au cours dans une perspective de régulation proactive (Allal, 2007), visant à éviter certains échecs aux cohortes suivantes.

Le modèle de la généralisabilité a été décrit et illustré en détail dans un numéro spécial de *Mesure et évaluation en éducation* (Cardinet, 2003b) et d'autres renseignements sont disponibles sur le site Edumétrie [http://www.irdp.ch/edumetrie/generalisabilite.htm]. Nous ne donnerons donc ici que l'information nécessaire pour comprendre les analyses présentées. Le lecteur pressé – ou peu intéressé par les aspects « techniques » de ces analyses – peut parcourir rapidement les pages suivantes pour se centrer sur les conclusions que nous tirons des données statistiques.

Le dispositif d'évaluation

Les objectifs d'évaluation que nous venons de fixer supposent que nous ne nous contentions pas d'un simple plan Étudiants x Questions. Nous devons construire un dispositif d'évaluation précisant les moyens et conditions de la mesure. C'est en effet un avantage méthodologique de la généralisabilité que de contraindre le chercheur à expliciter ce que d'autres techniques statistiques (par exemple l'alpha de Cronbach[7]) laissent souvent dans l'implicite.

Tableau 1
Plans d'observation et d'estimation

Facette	*Étiquette*	*Niveaux*	*Univers*
Années de passation de l'examen	A	2	10
Étudiants dans chaque examen	E:A	94	INF
Parties (exercices) du test	P	6	6
Questions	Q:AP	4	INF

Les *plans d'observation et d'estimation*, première étape de l'analyse[8] (tableau 1) définissent :

- les différentes *facettes* (facteurs) à prendre en considération parce qu'influant sur les résultats qui nous intéressent ; elles sont désignées dans le programme EduG par leur initiale ;
- le nombre de *niveaux* (modalités) que ces facettes comportent ;
- la dimension des *univers* dont ces niveaux sont extraits ; cette dernière information caractérise le mode d'échantillonnage des facettes, déclarées selon les cas comme fixées (si le nombre de niveaux de l'univers égale le nombre de niveaux observés), aléatoires finies (si le nombre de niveaux univers est supérieur au nombre de niveaux observés dans l'échantillon aléatoire tout en étant fini), ou aléatoires infinies (si le nombre de niveaux univers est considéré si grand qu'il n'est pas possible d'en trouver le nombre exact) ;
- les relations entre les facettes : croisées ou nichées (incluses : symbole de la relation de nichage = le deux-points, soit « : »).

Nous prendrons en compte dans notre étude les facettes suivantes (énumérées dans l'ordre du plan d'observation et d'enregistrement des données) :

- Les années de passation de l'examen (A). Les deux épreuves ont été sélectionnées aléatoirement parmi les examens passés sur une période d'une décennie (facette aléatoire finie) ;
- Les étudiants (E), en l'occurrence leurs différents niveaux de compétence en grammaire influant sur les réponses et le score total. Ces étudiants sont évidemment différents d'une année (d'un test) à l'autre ; cette facette E est donc nichée dans la facette année de passation (E:A). Ces étudiants sont 94 à avoir passé le premier test (2005) et 97 le second (2008) ; on a écarté aléatoirement trois étudiants du second corpus pour avoir un plan équilibré. Les 94 candidats ayant passé chacun des deux examens sont considérés comme choisis au hasard parmi le grand nombre d'étudiants ayant passé ce type d'épreuve ces dernières années (univers considéré comme pratiquement infini = INF ; facette aléatoire infinie) ;
- Les six parties (exercices) de chaque test (P) ; celles-ci couvrent l'ensemble des six domaines grammaticaux que l'on souhaite contrôler. Chaque examen comporte les mêmes six exercices. Cette facette est donc fixée et croisée avec les facettes A et E:A ;

- Les quatre groupes de questions nichées dans les différentes parties de l'examen. Elles diffèrent naturellement d'une partie (d'un exercice) à l'autre; elles sont également différentes, pour la même partie, d'un test à l'autre. Dit autrement, elles sont spécifiques à la fois à chaque partie (exercice) et à chaque test ; cette facette est donc étiquetée Q:AP (= nichée dans l'intersection A x P). Elles sont considérées comme sélectionnées aléatoirement parmi le grand nombre de questions que l'on peut générer à partir du contenu de chaque chapitre du cours (facette aléatoire infinie). Chaque question (ou groupe de questions) est cotée de 0 à 1 en fonction de la proportion de réponses justes données aux sous-questions (par exemple, 3 réponses justes sur 4 → 0,75).

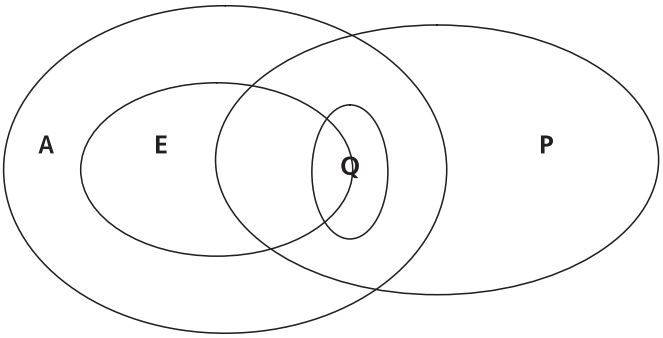


Figure 2. ***Diagramme d'Euler-Venn illustrant les relations entre facettes dans le plan d'observation***

La figure 2 illustre par un diagramme d'Euler-Venn les relations entre les différentes facettes du plan d'observation et la complexité du dispositif d'évaluation.

En outre, une analyse de variance (ANOVA) sur ces données, première étape d'une étude de généralisabilité, fournit une première estimation des composantes de variances (effets principaux et interactions) intervenant dans le dispositif (tableau 2). Chaque analyse de généralisabilité donnera un certain statut à ces différentes composantes, selon qu'elles interviennent pour caractériser l'objet ou les objets de mesure ou au contraire les erreurs qui affectent cette mesure. Mais on peut déjà noter à partir de ces données (*cf.* tableau 2, colonne 8) :

- la composante nulle attachée à l'effet principal A : toutes choses égales par ailleurs, les épreuves se présentent comme de difficulté pratiquement équivalente d'une année à l'autre ;
- la faible importance relative de la variance liée aux compétences des étudiants (E:A ; 7,3 % de la variance totale), constat conséquent avec l'objectif pédagogique d'un cours visant en principe à homogénéiser (vers le haut) le niveau des apprenants et avec une épreuve de maîtrise qui se veut non sélective ;
- la variance d'une certaine ampleur caractérisant les différences entre parties (P ; 24,2 %) et questions de l'épreuve (Q:AP ; 18,0 %) : certaines parties ou questions se révèlent moins bien réussies que d'autres ;
- le fait que la variance la plus importante est celle de l'interaction Étudiants x Questions (EQ:AP ; 40,0 %) : plusieurs étudiants répondent mieux ou moins bien qu'attendu aux diverses questions compte tenu de leur niveau de capacité et du degré de difficulté des exercices ; noter que cette composante intègre également toutes les sources non identifiées de variance.

Tableau 2
Analyse de variance

				Composantes				
Source	*SC*	*dl*	*CM*	*Aléatoires*	*Mixtes*	*Corrigées*	*%*	*Er. St.*
A	0,18678	1	0,18678	-0,00002	-0,00024	-0,00024	0,0	0,00008
E:A	15,38565	186	0,08272	0,00214	0,00281	0,00281	7,3	0,00036
P	43,21956	5	8,64391	0,01125	0,01112	0,00926	24,2	0,00615
Q:AP	23,86761	36	0,66299	0,00689	0,00689	0,00689	18,0	0,00162
AP	0,92505	5	0,18501	-0,00131	-0,00131	-0,00131	0,0	0,00048
EP:A	29,12165	930	0,03131	0,00400	0,00400	0,00400	10,4	0,00037
EQ:AP	51,32796	3348	0,01533	0,01533	0,01533	0,01533	40,0	0,00037
Total	164,03428	4511					100 %	

Fiabilité de l'évaluation des compétences grammaticales des étudiants

Ce qui nous intéresse au premier chef est l'évaluation des compétences grammaticales des étudiants par rapport au seuil de réussite (75 % de rendement) fixé par l'enseignant. Mais auparavant, pour calculer le coefficient critérié Phi(lambda) qui teste cette fiabilité, nous allons passer par une étape intermédiaire : il s'agit de contrôler comment le dispositif mesure – différencie – les compétences des étudiants sans prendre en compte à ce stade le seuil de maîtrise en question.

Plan de mesure et analyse de généralisabilité

Chaque étude de généralisabilité implique que l'on définisse un *plan de mesure* spécifiant le rôle des différentes facettes dans la mesure, soit :

- ce qui est l'*objet de mesure*, ce que l'on cherche à différencier (*face de différenciation* D) ; dans la présente étude, qui a pour but d'estimer la fiabilité des scores des étudiants, les facettes de différenciation sont E ainsi que A, dans laquelle elle E est nichée[9] ;
- les *instruments, moyens ou conditions de cette mesure* (*face d'instrumentation* I), en l'occurrence les exercices ou parties (P) de l'examen et les questions (Q) incluses dans l'intersection des facettes Années et Parties (AP).

Une barre de fraction sépare les faces D et I du plan de mesure ; celui-ci s'exprime donc dans ce cas par la formule AE/PQ : les compétences des étudiants qui ont passé les examens lors des deux années échantillonnées sont évaluées au moyen des questions contenues dans les différentes parties de l'épreuve. Soulignons encore le fait que, par ce type de plan de mesure, nous vérifions la capacité du dispositif à différencier l'ensemble des étudiants (n = 188) ayant passé les examens lors des deux années échantillonnées (*cf*. AE sur la face de différenciation).

Tableau 3

Analyse de généralisabilité pour le plan de mesure AE/PQ

Sources de var.	*Variance de différ.*	*Sources de var.*	*Variance. d'err.rel*	*% rel.*	Variance d'err.abs.	% abs.
A	(0,00000)		.....		.....	
E:A	0,00281		.....		.....	
	.....	P	.....		*(0,00000)*	*0,0*
	.....	Q:AP	0,00029	31,0	*0,00029*	*31,0*
	.....	AP	(0,00000)	0,0	*(0,00000)*	*0,0*
	.....	EP:A	(0,00000)	0,0	*(0,00000)*	*0,0*
	.....	EQ:AP	0,00064	69,0	*0,00064*	*69,0*
Total des variances	0,00281		0,00093	100 %	*0,00093*	*100 %*
Écarts-types	0,05299		Erreur type relative : 0,03043		*Erreur type absolue : 0,03043*	
Coef_G relatif	0,75					
Coef_G absolu	*0,75*					

Moyenne générale pour les niveaux traités : 0,91048
Variance d'échantillonnage de la moyenne générale pour les niveaux traités : 0,00016
Erreur-type sur la moyenne générale : 0,01272

En fonction des données des plans d'observation et d'estimation, le modèle de la généralisabilité calcule les composantes de variance attribuables aux facettes figurant sur la face de différenciation du plan de mesure et à leurs interactions (colonnes 1 et 2 du tableau 3) d'une part ; les composantes de variance attribuables aux facettes faisant partie de la face d'instrumentation, à leurs interactions entre elles ainsi qu'avec les facettes de différenciation (colonnes 3 à 7 du tableau 3), d'autre part. La variance d'erreur, quant à elle, est calculée à partir des composantes de variance associées aux facettes d'instrumentation aléatoires[10]. Elle diffère dans sa composition selon le type de mesure visée :

- mesure relative (tableau 3, colonnes 4 et 5) : quand on cherche à classer les objets ou les individus sur une échelle relative, par rapport à un barème normatif, comme le font les scores en rangs sur 100 ou en stanines en psychométrie ;

- mesure absolue (tableau 3, colonnes 6 et 7) : quand on veut situer des valeurs par rapport à une échelle dont les échelons sont définis *a priori* ; c'est le cas lorsque l'évaluation est critériée, notamment quand on fixe (comme pour l'examen qui nous intéresse) différents seuils pour caractériser divers degrés de réussite (notes de A à F, en l'occurrence).

Interprétation des résultats de l'analyse de généralisabilité

Sans entrer dans un commentaire détaillé du tableau 3, nous nous intéresserons à trois données (en italique) apportées par l'analyse en nous centrant sur la *mesure absolue* : le coefficient de généralisabilité absolue (dernière ligne du tableau) indiquant la fiabilité de cette mesure ; les composantes d'erreur influant sur cette fiabilité (colonnes 6 et 7) ; l'estimation de l'erreur de mesure fournie par l'écart-type du total des erreurs absolues (antépénultième ligne, colonne 6).

Les deux coefficients de généralisabilité (Coef_G) caractérisant les deux types de mesure, relative et absolue, correspondent au rapport entre la variance de différenciation et la somme de la variance de différenciation plus la variance d'erreur considérée (relative ou absolue). Variant de 0 à 1, ce coefficient indique une *fiabilité* ou *fidélité* du dispositif d'évaluation insuffisante lorsqu'il tend vers 0 et une fiabilité suffisante ou bonne lorsqu'il se rapproche de 1. On admet généralement que la fiabilité est satisfaisante lorsque la valeur de Coef_G est supérieure ou égale à 0,80, c'est à dire lorsque la variance de différenciation représente au moins 80 % de la variance totale.

On constate que le *coefficient de généralisabilité absolue* (0,75) est légèrement en dessous de cette valeur de référence de 0,80. On peut l'expliquer d'abord par la relative homogénéité des résultats des étudiants, comme le montrent les distributions des scores (en % de réussite) aux deux examens de la figure 1 ci-dessus. Cette homogénéité a pour conséquence évidente de limiter la variance de différenciation. On notera que celle-ci dépend uniquement des différences entre étudiants (E:A), la variance de A étant nulle, comme nous l'avons déjà signalé. On peut donc légitimement comparer et mesurer les scores des 2 x 94 étudiants. La variance de différenciation se révèle ainsi relativement restreinte par rapport aux deux sources d'erreur absolue, qui introduisent un « flou » dans l'évaluation : une certaine hétérogénéité dans la difficulté des six parties (exercices) de l'épreuve (P : 31 % de la variance totale d'erreur) et surtout l'interaction Étudiants x Questions (EQ:AP ; 69 %), difficile à contrôler.

L'*erreur type absolue* nous donne une estimation de la précision de la mesure des compétences : elle est d'environ 3 % de l'échelle des scores (0,03043). On peut, dans un premier temps, la considérer comme relativement satisfaisante, l'ambition d'une épreuve pédagogique étant rarement d'être beaucoup plus précise quant à l'estimation des compétences des apprenants. Toutefois, la qualité de cette précision doit être reconsidérée du fait que l'enseignant attribue des notes, de A à F. Or, les évaluations de A à E se logent dans l'espace relativement restreint de l'échelle allant de 100 % à 75 % de réussite, chaque note couvrant environ 5 % de cette échelle. L'*intervalle de confiance* autour de chaque score étant d'environ 6 % (1,96 x 0,03043), on doit considérer la note comme une estimation approximative du degré d'excellence en grammaire ; on ne peut affirmer avec une grande assurance que tel score (93 %, par exemple) vaut la note B plutôt que A ou C. Ce même intervalle de confiance appliqué au seuil lui-même doit également inciter l'enseignant à faire preuve de circonspection – voire de tolérance – dans l'examen des cas d'étudiants dont le score se situe juste en dessous du *seuil de suffisance* fixé (*cf.* figure 1, examen de 2005, le cas de deux étudiants qui selon notre correction obtiendraient un score de 72 %).

Tentative d'optimisation du dispositif pour améliorer la généralisabilité

Conçu notamment pour la mise au point de dispositifs d'évaluation, le modèle de la généralisabilité ouvre la possibilité, dans une étape dite d'*optimisation*, d'examiner quelles seraient les conséquences – et notamment les améliorations – si l'on modifiait à l'avenir (donc virtuellement) les plans d'observation ou d'estimation. Le module d'*optimisation* d'EduG permet de simuler de tels changements. Ceux-ci sont guidés par les résultats de l'analyse de généralisabilité préalable (tableau 3). Elle montre (colonnes 6 et 7) dans le cas présent que la source de variance d'erreur absolue sur laquelle il paraît possible d'agir[11] est la facette Question (Q:AP ; 31 % du total de la variance d'erreur absolue) : la difficulté variable des questions à l'intérieur des deux examens et des six parties (*cf.* figure 4) introduit un certain flou dans l'estimation des scores des étudiants. On peut donc envisager, théoriquement, d'augmenter le nombre de ces questions (donc la quantité d'information fournie par cette facette) pour vérifier l'effet de cette modification sur la fiabilité (Coef_G) et la précision (erreur type) de la mesure absolue.

L'analyse d'optimisation (dont nous ne présenterons pas ici le détail) montre que pour franchir le seuil de 0,80 en ce qui concerne le coefficient absolu (qui passerait à 0,82), il faudrait six questions par partie, ce qui allongerait l'épreuve de moitié (36 questions vs 24) ; l'erreur type serait encore de 2,5 % (0,02484). Le jeu n'en vaudrait pas la chandelle et, compte tenu du temps généralement à disposition pour cet examen, une telle solution ne serait pas applicable. Nous allons voir par ailleurs à l'instant qu'une telle optimisation ne s'impose pas si l'objectif majeur de l'examen est celui d'un test de maîtrise : contrôler quels sont les étudiants qui satisfont – ou ne satisfont pas – aux critères minimaux de réussite fixés par l'enseignant.

La fiabilité du dispositif par rapport à un seuil de maîtrise : le coefficient critérié Phi(lambda)

Dans cette phase de l'analyse, on réintroduit l'information constituée par le seuil de maîtrise défini par l'enseignant (75 % → 0,75) et par rapport auquel il veut situer les résultats de ses étudiants[12]. Sans entrer dans le détail de la formule de calcul du coefficient critérié Phi(lambda)[13], signalons que celle-ci prend en considération la variance d'échantillonnage de la moyenne de l'échantillon observé et la distance entre cette moyenne et le seuil fixé.

Tableau 4
Coefficient critérié Phi(lambda)

Plan de mesure : AE/PQ
Seuil = lambda = 0,75
Phi(lambda) = 0,96858

Pour le plan de mesure considéré (AE/PQ), ce coefficient peut être estimé excellent (0,97), et l'évaluation des compétences grammaticales très fiables si l'on prend comme référence le critère de réussite de 75 % (tableau 4)[14].

Ce constat est important pour notre propos : il correspond en effet à l'objectif effectif de l'épreuve, considérée comme un test de maîtrise au sens de la définition donnée plus haut. Un tel résultat valide une évaluation visant une certification de type dichotomique : réussite-échec (*pass-fail*). Cet examen pourrait être au contraire jugé de qualité insuffisante si le but qui lui était attribué était de distinguer différents niveaux de compétences, comme c'est le cas pour les tests PISA.

Signalons qu'à titre de vérification, nous avons refait les analyses ci-dessus pour chaque examen séparément[15]. Dans les deux analyses, on obtient sans surprise (l'année de passation ne jouant pratiquement aucun rôle sur la réussite des étudiants) des valeurs quasi identiques à celles mentionnées ci-dessus pour les coefficients de généralisabilité et pour les coefficients critériés.

Information complémentaire apportée par d'autres analyses de généralisabilité sur le dispositif d'évaluation

La généralisabilité permet d'explorer la fiabilité du dispositif d'évaluation sous d'autres angles édumétriques, notamment en prenant comme objets de mesure successivement les années de passation de l'examen (facette A), les différentes parties de l'épreuve (P) et les questions à l'intérieur de ces six parties (exercices). Nous nous limiterons, dans le cadre de cet article, à mentionner rapidement les principaux résultats des trois analyses de généralisabilité en question.

Différenciation des années d'examens (A)

Peut-on distinguer de façon fiable des degrés de réussite entre les examens selon l'année de passation ou au contraire peut-on considérer ces deux épreuves comme deux versions parallèles de même difficulté ? Pour répondre à cette interrogation, il n'est pas besoin de réaliser l'analyse de généralisabilité sur le plan de mesure A/EPQ : on en connaît d'avance le résultat. La variance de A étant nulle dans l'ANOVA (tableau 2), la variance de différenciation, au numérateur des deux Coef_G, sera également nulle. Les différentes versions du même examen peuvent donc être considérées, toutes choses égales par ailleurs, comme de difficulté identique : les deux distributions (figure 1) sont pratiquement superposables et leurs taux de réussite moyens très proches : 90 % et 92 %. Ceci garantit une certaine équité dans l'évaluation d'une année à l'autre. On notera simplement en passant qu'une telle vérification est très rarement faite pour la plupart des examens supposés équivalents d'année en année.

Différenciation des parties de l'examen

Dans ce cas également, l'ANOVA de départ reste la même (tableau 2), le plan de mesure de l'analyse de généralisabilité étant alors P/AEQ : les différences de réussite aux six parties (exercices) de l'épreuve (P) sont mesurées au moyen des questions qu'elles contiennent (Q) et des réponses des étudiants (E) des deux années échantillonnées (A).

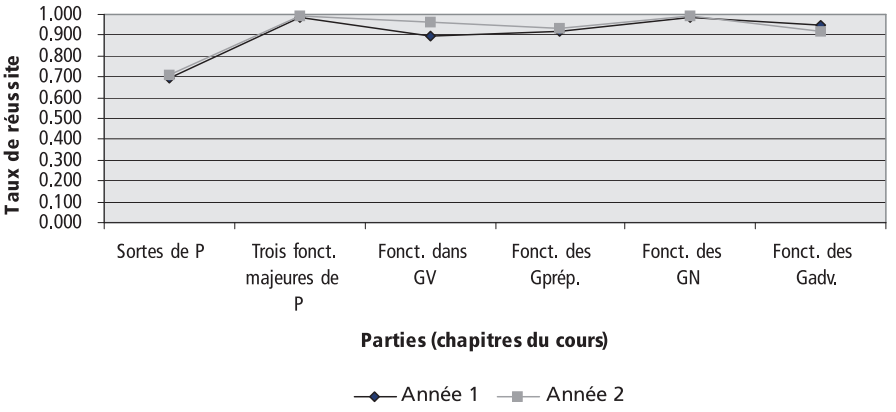


Figure 3. ***Profils des moyennes aux six parties pour chacun des deux examens***

Nous ne donnerons pas ici le détail de cette analyse. Disons simplement que dans ce cas le coefficient de généralisabilité absolue, relativement élevé (0,91), atteste que l'on peut mesurer avec une bonne fiabilité les différences de réussite selon les parties des examens (*cf.* figure 3). Le modèle et le logiciel permettent dans ce cas également de calculer un coefficient critérié. Nous fixons ce seuil à 0,90, donc à une réussite des exercices en moyenne par 90 % des étudiants, ce qui correspond à ce qu'on peut considérer comme une bonne assimilation des chapitres du cours par une grande majorité des apprenants. On obtient ainsi un coefficient Phi(lambda) de 0,91[16]. On peut donc estimer avec une bonne assurance qu'avec 70 % de réussite, le premier exercice (sur les *sortes de phrases*) se situe nettement au-dessous du seuil défini. Il ne serait alors pas inutile d'explorer plus avant les raisons des échecs à cet exercice, notamment en considérant les caractéristiques (contenu, forme, type d'opérations grammaticales exigées) des questions posées dans cette partie de l'examen.

On y serait encouragé en outre par une autre information issue de l'analyse. Le parallélisme des deux courbes de la figure 3 illustre le fait que l'interaction entre A et P est nulle ; le degré de réussite des six parties de l'épreuve est pratiquement parallèle dans les deux examens échantillonnés. On est ainsi assuré que les échecs plus nombreux à la première partie de l'épreuve ne sont pas le fait d'un simple « incident de parcours » lié à l'échantillonnage des questions. Par ailleurs, l'erreur type est relativement faible (0,03030, soit environ 3 %), attestant une bonne précision de l'évaluation des moyennes de parties.

Différenciation des questions de l'examen

La réussite aux différentes questions à l'intérieur des six parties de chaque examen peut aider à cerner les notions qui font problème du point de vue de l'enseignement-apprentissage, à condition toutefois qu'on puisse distinguer de façon fiable des degrés de réussite nettement différenciés.

Tableau 5
Analyse de généralisabilité pour le plan de mesure QAP/E

Sources de var.	*Variance de différ.*	*Sources de var.*	*Variance. d'err.rel*	*% rel.*	Variance d'err.abs.	% abs.
A	(0,00000)		.....		.....	
	.....	E:A	0,00003	12,7	*0,00003*	*12,7*
P	0,00926		.....		.....	
Q:AP	0,00689		.....		.....	
AP	(0,00000)		.....		.....	
	.....	EP:A	0,00004	18,1	*0,00004*	*18,1*
	.....	EQ:AP	0,00016	69,3	*0,00016*	*69,3*
Total des variances	0,01615		0,00024	100 %	*0,00024*	*100 %*
Écarts-types	0,12710		Erreur type relative : 0,01535		*Erreur type absolue : 0,01535*	
Coef_G relatif	0,99					
Coef_G absolu	*0,99*					

Dans ce cas, le plan de mesure de l'analyse de généralisabilité (tableau 5) est QAP/E. Sur la face de différenciation (à gauche de la /), cette formule reflète le fait que les questions sont nichées dans l'intersection Années x Parties (*cf.* figure 2) ; sur la face d'instrumentation (à droite de la /) le fait que les étudiants sont les « instruments » permettant d'évaluer le degré de difficulté des questions. L'étude portera donc sur l'ensemble des 48 questions – toutes différentes – posées dans les deux examens (*cf.* figure 4).

L'analyse de généralisabilité apporte plusieurs renseignements dignes d'intérêt. Compte tenu de l'un des objectifs possible de l'examen (contrôler la maîtrise des diverses notions enseignées dans le cours), on s'intéressera de nouveau à la mesure absolue. On constate tout d'abord (tableau 5) que l'évaluation (degrés de réussite ou de difficulté) des questions est très fiable : le Coef_G absolu est de 0,99 ; le total de la variance d'erreur absolue (0,00024) est très faible comparé au total de la variance de différenciation (0,01615)[17].

On relève ensuite que ce dernier total est l'addition de quatre composantes de variances de différenciation émanant des trois facettes qui figurent sur la face de différenciation : A, P et Q:AP ainsi que de l'interaction AP (tableau 5, colonnes 1 et 2). Deux composantes, A et AP sont nulles, comme le montre déjà l'ANOVA (tableau 2). Ceci indique que l'année de passation n'intervient pas dans la différenciation des questions (composante A) : le degré global de difficulté des deux épreuves est pratiquement le même ; qu'il y a très peu de différence dans les profils des questions d'une année à l'autre (interaction AP), ce que confirme très clairement le graphique de la figure 4.

En conséquence, les différences entre les questions n'ont que deux sources : P et Q:AP, d'importance inégale (respectivement 57 % et 43 % du total de la variance de différenciation). La difficulté (ou le taux de réussite) des questions tient donc un peu plus à leur rapport avec tel chapitre du cours, contrôlé par telle partie de l'épreuve, qu'à la notion vérifiée spécifiquement par la question.

Comme on l'a déjà constaté ci-dessus, et comme le confirme le graphique de la figure 4, les questions relatives aux *sortes de phrases* (P1) ont été en général moins bien assimilées que celles relatives aux autres chapitres de la grammaire. Par ailleurs, un type de questions a plus particulièrement échoué : celles qui concernent les *phrases non standards*[18] (Q4 dans P1 pour les deux examens).

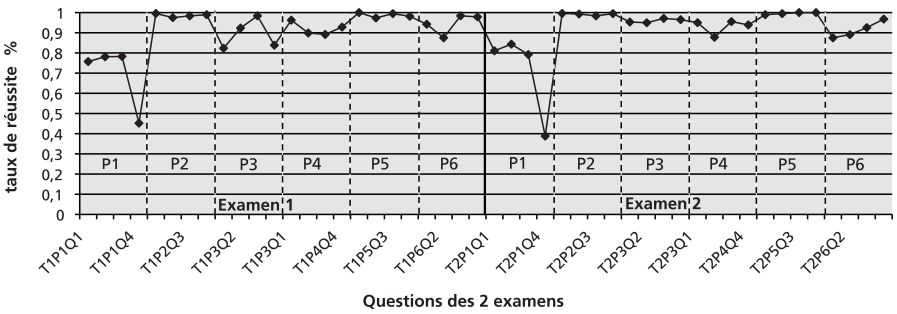


Figure 4. ***Profils des questions pour les deux examens***

Une troisième information encourage enfin à distinguer différents degrés de réussite parmi les questions ; il s'agit de l'*erreur type* de mesure (absolue) : elle n'est que de 1,5 % (0,01535). En outre, dans ce cas également, il est possible de calculer un coefficient critérié pour vérifier par exemple s'il est possible de différencier de façon fiable les items qui se situent en-dessus et en-dessous d'un *seuil de suffisance* que nous fixerons de nouveau à 0,90 (90 % de réussite). Le coefficient Phi(lambda) de 0,99 ainsi obtenu confirme la pertinence d'une analyse plus approfondie des questions les plus échouées par les étudiants (notamment celles proches de ou inférieures à 80 % de réussite).

Discussion et conclusion

Comme on l'aura saisi en lisant notre introduction, notre propos ici n'est pas simplement de « faire l'article » pour un modèle trop souvent méconnu : la généralisabilité. Il s'agit surtout de plaider pour une bonne adéquation entre, d'une part, les techniques statistiques utilisées pour évaluer la qualité des tests et, d'autre part, les objectifs pédagogiques des instruments de mesure. La question primordiale en évaluation devrait être en effet : évaluer pour quoi faire ? Et immédiatement : quels sont les moyens adéquats pour le faire et vérifier que cette évaluation est fiable ?

Or, on a souvent l'impression sur le terrain que le moyen prend le pas sur le but ; qu'il n'y a de bonne évaluation que si l'épreuve est discriminative et peut être soumise pour analyse aux modèles exigeant une variance appréciable des scores aux items ; que l'échantillonnage des questions doit se soumettre aux contraintes de ces modèles, qu'il faut donc écarter les items trop bien réussis ; que même s'il s'agit de tester des *standards minimaux de formation*, il est nécessaire docimologiquement de pouvoir distinguer des niveaux de performances, voire d'excellence.

Ces représentations de l'évaluation et de ses contraintes sont particulièrement inadéquates quand il s'agit de tests de maîtrise ou plus généralement de toute épreuve visant à contrôler l'assimilation des compétences de base exigibles à la fin d'un cycle de formation. À titre d'exemple, appliquer le modèle de la TRI à des données comme celles des nos deux examens suscite plusieurs problèmes[19]. Tout d'abord, d'un point de vue théorique, il est difficile de postuler, pour les contrôles de maîtrise qui nous occupent, « l'hypothèse de la normalité de la distribution du trait latent » (Bertrand & Blais, 2004, p. 205).

D'autre part, selon les auteurs cités à l'instant, « il existe des problèmes avec tous les modèles [de la TRI] lorsque des sujets ont répondu correctement ou incorrectement à tous les items » (p. 231). Il est alors nécessaire d'utiliser des méthodes bayesiennes pour l'estimation du paramètre thêta, ce qui oblige à faire des hypothèses (*a priori* ou *a posteriori*) sur la distribution de thêta. On constate dans la pratique que les constructeurs de tests utilisant la TRI préfèrent éliminer ce type d'items « trop bien réussis », ou « faiblement discriminants », qualifiés de « peu informatifs », sans toujours remettre en question leur pertinence par rapport au champ évalué et à l'objectif de l'épreuve. Par ailleurs, les estimations des paramètres des items et celle de l'habileté thêta exigent en principe (en particulier pour les modèles à deux ou trois paramètres) des effectifs de sujets (étudiants) importants pour garantir une bonne fiabilité des évaluations. Dans un cas comme celui d'un examen universitaire, il est rare que l'on puisse compter sur de tels effectifs[20].

De plus, sur le plan pédagogique, il n'est pas innocent de transformer un test de maîtrise en test de niveau sous prétexte que (comme dans PISA notamment) tel de ces niveaux opérationnalise le standard de performance minimal exigible des apprenants à ce stade de leurs études et que les autres niveaux apportent des données intéressantes sur les compétences des étudiants au-delà de ce seuil plancher. Il y a 30 ans déjà, Cardinet (1977) signalait dans une telle façon de faire une « confusion des perspectives » : on tend à transformer un test d'acquisition en test d'aptitude, à passer d'un objectif sommatif à un but pronostique, sans que les critères de validité soient ajustés et vérifiés en conséquence. On met les élèves implicitement en compétition dans une optique souvent sélective qui se camoufle en contrôle des acquis. Sur le plan psychologique, l'épreuve de contrôle étant un aspect important – parce que particulièrement visible – du *contrat didactique* (Brousseau, 1984) inhérent à toute situation d'enseignement, ce changement de perspective peut être considéré comme une espèce de tromperie : le contenu et la forme de l'épreuve ne correspondent plus aux objectifs annoncés pour le cours.

Certes, le modèle de la généralisabilité reposant sur l'ANOVA suppose une certaine variance dans les scores observés et ses résultats doivent être considérés avec précaution lorsque les présupposés de l'analyse de variance ne sont que partiellement satisfaits. Néanmoins, comme concluent les auteurs d'un ouvrage sur la généralisabilité paru récemment (Cardinet, Johnson & Pini, 2009), le modèle de l'ANOVA est relativement robuste et surtout « on doit bien utiliser ces modèles imparfaits, tant qu'on n'en a pas de meilleurs à

disposition pratiquement ». Disons aussi – un peu ironiquement – que le risque est faible qu'un examen soit tellement bien réussi que les résultats soient tout à fait homogènes (dispersion des scores quasi nulle), et ce, pour deux raisons au moins. Les enseignants sont souvent très optimistes en ce qui concerne l'efficacité de leur enseignement ou les compétences acquises par leurs élèves, et certaines des questions posées sont plus difficiles qu'ils ne le pensent. D'autre part, les étudiants sont parfois tout aussi optimistes quant à leur assimilation du cours et certains sous-estiment l'importance des révisions nécessaires. Dans ces conditions, comme le montrent les distributions de la figure 1 ci-dessus, il subsiste une dispersion appréciable des scores des étudiants, et on constate par ailleurs une différence non négligeable entre les réussites aux différentes parties ou questions de l'épreuve. En outre, le coefficient critérié Phi(lambda), qui nous intéresse plus particulièrement ici, prend également en compte une autre variance : celle liée à la distance par rapport au seuil de réussite fixé.

En conclusion, nous résumerons et soulignerons ainsi les avantages de la généralisabilité dans le cas d'un test de maîtrise. En tant que modèle d'échantillonnage (*vs* d'étalonnage dans le cas de la TRI), il est conséquent avec une épreuve pédagogique qui se veut représentative des conduites et contenus visés par le cours. Il fournit un moyen adéquat d'évaluer la fiabilité d'un test critérié. Sur le plan didactique, le modèle présente docimologiquement deux avantages majeurs. En premier lieu, comme le souligne Johnson (2008), il représente un outil puissant pour identifier et estimer les diverses sources d'erreurs qui affectent un dispositif d'évaluation, selon l'objectif visé. Il permet en outre de prédire, lors d'une phase dite d'optimisation, l'effet des modifications de ce dispositif, qu'il s'agisse de rendre celui-ci moins onéreux ou plus fiable et plus précis (diminution ou augmentation du nombre d'items ou de sujets à tester, selon le cas). D'autre part, contrairement à d'autres modèles, la généralisabilité ne se limite pas à analyser le plan de mesure : Sujets/Items. Par sa flexibilité et par sa polyvalence, elle permet d'explorer toutes les facettes du dispositif de mesure. Comme nous l'avons vu dans l'exemple traité ici, il est possible grâce à ce modèle de tester l'équivalence de deux formes d'examen, les différences de réussite aux différentes parties ou questions de l'épreuve. Dans d'autres cas encore, cette approche statistique aide à vérifier la fiabilité d'évaluations portant sur des progrès (différence prétest – posttest) ou sur l'effet de

différentes méthodes d'enseignement. Sur le plan didactique, que nous avons privilégié ici, c'est finalement un moyen précieux pour faire une « analyse spectrale » d'un enseignement sous l'angle de ses résultats et des facteurs qui influent sur son fonctionnement.

NOTES

1. Pour une discussion sur la perspective édumétrique (*vs* psychométrique), *cf.* dans les bulletins de l'ADMEE-Europe les articles de Demeuse (2002), Cardinet (2003a) et Mokonzi (2003).
2. Cet aspect des épreuves PISA devrait être développé à partir de 2012 : « L'informatisation de l'administration des épreuves devrait, à plus long terme, permettre d'améliorer l'alignement des tests sur les niveaux de compétence des élèves » (OCDE, 2007, p. 16) par l'introduction d'une forme de test adaptatif (Bertrand & Blais, 2004, chap. 9).
3. « La CDIP [Conférence suisse des directeurs cantonaux de l'instruction publique] et les régions linguistiques se concertent au cas par cas pour développer des tests de référence sur la base des standards de formation » (CDIP, 2007).
4. Dans les analyses de généralisabilité.
5. Nous remercions le professeur J.-P. Bronckart (Faculté de psychologie et des sciences de l'éducation, Université de Genève) d'avoir mis à notre disposition les données nécessaires à cet exemple. Nous avons cependant réaménagé celles-ci en modifiant quelque peu les modalités de correction et de cotation pour les besoins de nos analyses et pour satisfaire aux contraintes du modèle de la généralisabilité. Nos résultats ne correspondent donc que partiellement aux données originales, mais la corrélation entre les scores des deux corrections est très élevée : r_{BP} respectivement 0,98 et 0,96 pour les deux examens analysés.
6. Entre parenthèses, les contenus effectivement testés dans l'un ou l'autre examen.
7. *Cf.* Cardinet, 2007.
8. Les analyses qui suivent ont été réalisées au moyen du logiciel EduG 6.0 f, téléchargeable sur le site du groupe Edumétrie : [http://www.irdp.ch/edumetrie/logiciels.htm].
9. Sur la face de différenciation, une facette nichée (E) est nécessairement accompagnée de sa facette nichante (A) ; on cherche en effet à différencier les scores de tous les étudiants (= 2 x 94).
10. Les facettes d'instrumentation fixées (ici la facette P) et leurs interactions avec les facettes de différenciation (AP et EP:A) ne sont pas prises en compte pour la détermination de la variance d'erreur puisqu'elles ne créent pas de fluctuations d'échantillonnage. Pour le détail des calculs, cf. par exemple Mokonzi, 2003, dans le numéro de Mesure et évaluation en éducation coordonné par Cardinet.
11. Il est en effet pratiquement impossible d'agir directement sur la principale source d'erreur relative et absolue : l'interaction entre les facettes Étudiants et Questions (EQ:AP ; 69 %), qui intègre également toutes les sources non identifiées de variance.
12. Ce seuil a été fixé par l'enseignant en fonction de son expérience. Une discussion sur la fixation d'un tel score et sa pertinence est hors de notre propos, d'autant plus que nous ne disposons par de critère extérieur pour estimer sa validité.
13. *Cf.* Bertrand & Blais, 2004, pp. 95-96.
14. Le coefficient critérié prenant en compte la distance entre la moyenne de l'échantillon (ici 0,91) et la valeur du critère (0,75) sera d'autant plus élevé que cette distance est importante (ici 0,91 – 0,75 = 0,16).

15. Ce que permet de faire facilement la routine de *réduction du plan* du logiciel EduG.
16. Dans ce cas comme dans le suivant (différenciation des questions), le critère de réussite choisi (0,90) étant très proche de la moyenne de l'échantillon (0,92), le calcul du coefficient critérié aboutit à un résultat identique à celui du coefficient de généralisabilité absolu (Phi(lambda) bridé ; *cf.* l'*Aide* du logiciel EduG sous *Coefficient critérié*).
17. Le Coef_G absolu est donc calculé de la façon suivante, comme indiqué ci-dessus : 0,01615/(0,00024 + 0,01615).
18. Phrases telles que « Pierre cherche désespérément **où loger** » ou « **Plus que les autres**, les enfants doivent être écoutés » (exemples donnés par le cours). Noter que ce type de phrases ne fait pas partie du programme de structuration à l'école primaire.
19. Précisons que nous ne mettons pas en cause d'autres utilisations de ce modèle.
20. La généralisabilité, elle, prend en compte le paramètre *n* des effectifs (du nombre de modalités ou de *niveaux*) dans ses calculs des erreurs de mesure et des Coef_G.

RÉFÉRENCES

Allal, L. (2007). Régulations des apprentissages : orientations conceptuelles pour la recherche et la pratique en éducation. In L. Allal & L. Mottier Lopez (éds), *Régulation des apprentissages en situation scolaire et en formation*. Bruxelles : De Boeck.

Bain, D., Weiss, L., & Agudelo, W. (2008). Radiographie d'une épreuve commune de mathématiques au moyen du modèle de la généralisabilité. In L. Mottier Lopez, Y.-E. Dizerens, G. Marcoux & A. Perréard Vité (éds), *Entre la régulation des apprentissages et le pilotage des systèmes : évaluations en tension. Actes du 20e colloque de l'ADMEE-Europe, Université de Genève (Genève, 9-11 janvier 2008)*. Genève : Université, Sciences de l'éducation. Accès : [https://plone.unige.ch/sites/admee08/symposiums/j-s8/j-s8-1].

Bertrand, R., & Blais, J.-G. (2004). *Modèles de mesure : l'apport de la théorie des réponses aux items.* Sainte-Foy : Presses de l'Université du Québec.

Bronckart, J.-P. (2004). *Didactique de la grammaire*. Genève : Département de l'instruction publique, Enseignement primaire, Secteur Langues et Cultures – Français.

Brousseau, G. (1984). Le rôle central du contrat didactique dans l'analyse et la construction des situations d'enseignement et d'apprentissage des mathématiques. In *Actes de la 3e École d'été de didactique des mathématiques.* Grenoble : Université, IMAG.

Cardinet, J. (1977). *Objectifs pédagogiques et fonctions de l'évaluation*. Neuchâtel : Institut romand de recherches et de documentation pédagogiques.

Cardinet, J. (2003a). Pourquoi faut-il parler d'édumétrie ? *Bulletin de l'ADMEE-Europe no 2002/3 et 2003/1*, 5-7.

Cardinet, J. (2003b). Numéro spécial sur la généralisabilité. *Mesure et évaluation en Éducation*, *26*(1-2) et *26*(3).

Cardinet, J. (2007). Résumé de l'article de Lee J. Cronbach : « Ce que je pense maintenant du coefficient alpha et de ses suites ». *Bulletin de l'ADMEE-Europe no 2007/1*, 4-7.

Cardinet, J., Johnson, S., & Pini, G. (2009). *Applying Generalizability Theory using EduG.* New York: Routledge/Taylor & Francis (Quantitative Methodology Series).

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

CDIP (2006). *Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Brève information (février 2006).* Berne : Conférence suisse des directeurs cantonaux de l'instruction publique.

CDIP (2007). *Accord intercantonal sur l'harmonisation de la scolarité obligatoire du 14 juin 2007*. Berne : Conférence suisse des directeurs cantonaux de l'instruction publique.

De Landsheere, G. (éd.) (1992). *Dictionnaire de l'évaluation et de la recherche en éducation*. Paris : Presses Universitaires de France.

Demeuse, M. (2002). Psychométrie et édumétrie. *Bulletin de l'ADMEE, 2*, 3-4.

DIP (2007). *Plan d'études de l'enseignement primaire 1E - 6 P.* Genève : Département de l'instruction publique, Service de l'enseignement.

Johnson, S. (2008). The versatility of G-theory for exploring and controlling measurement error. *Mesure et évaluation en éducation*, *31*(2), 55-73.

Laveault, D., & Grégoire, J. (2002). *Introduction aux théories des tests en sciences humaines.* Bruxelles : De Boeck.

Mokonzi, G. B. (2003). L'édumétrie : domaine de mesure au service de l'apprentissage ? *Bulletin de l'ADMEE-Europe n° 2002/3 et 2003/1*, 7-8.

Mons, N., & Pons, X. (2006). *Les standards en éducation dans le monde francophone. Une analyse comparative.* Neuchâtel : Institut de recherche et de documentation pédagogique.

OCDE (2007). *PISA – Programme international pour le suivi des acquis des élèves*. Paris: OCDE.

Pini, G. (2009, 28 juin). Édumétrie : lexique. [Page Web]. Accès : [http://www.irdp.ch/edumetrie/lexique.htm].

Thélot, Cl. (2004). *Pour la réussite de tous les élèves. Rapport de la Commission du débat national sur l'avenir de l'École présidée par Claude Thélot.* Paris : La documentation française.

Date de réception : 4 décembre 2009

Date de réception de la version finale : 4 février 2010

Date d'acceptation : 1er mars 2010