Document généré le 14 juil. 2025 17:03

Mesure et évaluation en éducation

Une application de la théorie de la généralisabilité à la planification des enquêtes sur les acquisitions des élèves

Sandra Johnson

Volume 26, numéro 1-2, 2003

Généralisabilité

URI : https://id.erudit.org/iderudit/1088238ar
DOI : https://doi.org/10.7202/1088238ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

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

Découvrir la revue

Citer cet article

Johnson, S. (2003). Une application de la théorie de la généralisabilité à la planification des enquêtes sur les acquisitions des élèves. *Mesure et évaluation en éducation*, *26*(1-2), 37–50. https://doi.org/10.7202/1088238ar

Résumé de l'article

Cet article se centre sur une application de la théorie de la généralisabilité pour évaluer et optimiser des dispositifs d'enquête portant sur les acquisitions des élèves. Il décrit la façon dont l'analyse des réponses à l'une des enquêtes de mathématiques du Programme d'évaluation des acquis en Écosse a pu fonder les décisions prises au sujet de l'organisation des enquêtes futures sur ce sujet.

Tous droits réservés © ADMEE-Canada - Université Laval, 2003

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter en ligne.

https://apropos.erudit.org/fr/usagers/politique-dutilisation/

Cet article est diffusé et préservé par Érudit.

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche.

https://www.erudit.org/fr/

Une application de la théorie de la généralisabilité à la planification des enquêtes sur les acquisitions des élèves

Sandra Johnson[1]

S&R Johnson, Consultants en éducation, France

MOTS CLÉS : Théorie de la généralisabilité, évaluation de systèmes, plan d'enquête, évaluation en mathématiques, mesure absolue, erreur-type de mesure, optimisation

Cet article se centre sur une application de la théorie de la généralisabilité pour évaluer et optimiser des dispositifs d'enquête portant sur les acquisitions des élèves. Il décrit la façon dont l'analyse des réponses à l'une des enquêtes de mathématiques du Programme d'évaluation des acquis en Écosse a pu fonder les décisions prises au sujet de l'organisation des enquêtes futures sur ce sujet.

KEY WORDS : Generalizability theory, system evaluation, survey design, mathematics assessment, absolute measurement, standard error of measurement, optimization

This paper focuses on an application of generalizability theory in evaluating and optimizing the design of pupil attainment surveys. It describes the way in which analysis of response data from one of the mathematics surveys conducted within Scotland's Assessment of Achievement Programme informed decisions about the design of future surveys in this subject.

PALAVRAS CHAVE: teoria da generalizabilidade, avaliação de sistemas, plano de inquérito, avaliação em matemática, medida absoluta, erro-tipo de medida, optimização

Este artigo centra-se na aplicação da teoria da generalizabilidade, para avaliar e optimizar os dispositivos de inquérito que incidem nas aquisições dos alunos. Descreve de que forma a análise das respostas a um dos inquéritos de matemática do Programa de avaliação dos adquiridos, na Escócia, fundamentou as decisões tomadas para a organização de futuros inquéritos sobre este tema.

Note de l'auteure : La recherche présentée dans cet article a été conduite dans le cadre d'un travail de consultant en cours au service du Département de l'Éducation du gouvernement écossais. Rod Johnson a préparé les jeux de données utilisés dans ces analyses, à partir de la banque de données de l'AAP. Jean Cardinet a apporté son concours en appliquant aux données le programme Etudgen et en traduisant le manuscrit à partir de l'anglais. Je les remercie tous les deux pour leur soutien, et aussi pour des commentaires très constructifs sur mon avant-dernier projet d'article.

Introduction

La théorie de la généralisabilité est appelée à jouer un rôle utile dans toute entreprise d'évaluation travaillant par échantillonnage. Elle donne le moyen de quantifier les thèmes qui contribuent à la variance d'erreur, et même de quantifier cette erreur de mesure elle-même (voir Brennan, 1983 ; Brennan, 2001 ; Cardinet & Tourneur, 1985 ; Cronbach, Gleser, Nanda & Rajaratnam, 1972 ; Shavelson & Webb, 1991). Elle permet de plus d'utiliser cette information relative à l'erreur pour établir des stratégies minimisant l'erreur de mesure à l'intérieur d'un cadre de contraintes données.

Les enquêtes sur les acquisitions des élèves offrent une occasion intéressante d'appliquer la théorie de la généralisabilité, étant donné que, à la différence de beaucoup d'autres situations d'évaluation, on n'a pas particulièrement besoin pour une enquête de différencier les élèves, les classes, les écoles, ni aucun autre composant du dispositif d'observation. L'objectif principal de telles enquêtes est en général d'estimer le niveau de capacité, de performance ou de réussite d'une certaine *population* d'élèves, dans un certain domaine du curriculum. Dans ce contexte, les élèves échantillonnés n'ont pas d'autre rôle que de représenter leur population ; on ne s'intéresse pas spécialement à eux en tant qu'individus. Dans ce cas, puisqu'ils sont un échantillon, les élèves (et leurs classes et écoles) contribuent à la variance d'erreur, comme le font les tâches d'évaluation qu'on leur présente (si l'on a l'intention de généraliser au-delà de l'ensemble particulier de questions qu'on a utilisées dans l'enquête). Selon la terminologie de la généralisabilité, les élèves, comme les tâches de l'enquête d'évaluation, constituent des facettes d'instrumentation.

Les principales options stratégiques pour minimiser l'erreur de mesure, dans ce cas, se traduisent en décisions relatives aux nombres de tâches d'évaluation qu'il faudrait utiliser pour bien représenter les différents aspects du savoir auquel on s'intéresse, ainsi qu'au nombre d'élèves qui devraient passer chaque épreuve (voir Johnson, 1989 ; Johnson & Bell, 1985).

Nous exposons ci-dessous comment une analyse de généralisabilité a aidé à fonder des décisions relatives à une nouvelle organisation des enquêtes, rendue nécessaire par une modification de la demande, concernant la façon de présenter les résultats, pour les enquêtes du Programme d'évaluation des acquis, en Écosse (Assessment of Achievement Programme : AAP). Le texte se centrera spécialement sur les enquêtes de l'AAP de 1994 et de 2000 effectuées dans le domaine de la mathématique. (Leurs résultats ont été publiés respectivement

dans Robertson, Meechan, Clarke, Moffat & McCormick, 1995, et AAP, 2001.) Il illustrera comment l'analyse de généralisabilité des données de 1994 a pu être prise en compte pour l'organisation de l'enquête de 2000.

Le dispositif du Programme d'évaluation des acquis (AAP)

Le Programme d'évaluation des acquis (Assessment of Achievement Programme : AAP) en Écosse est un instrument d'évaluation du système scolaire qui fonctionne depuis le milieu des années 1980. Ce programme contrôle les acquisitions des élèves dans les matières principales, à l'aide d'enquêtes par échantillonnage. De ce fait, ses activités recoupent de bien des façons celles des autres programmes nationaux et internationaux d'enquêtes sur le rendement scolaire, comme les programmes IEA (*International Evaluation of Educational Achievement*) et PISA (*Programme for International Student Assessment*). Depuis sa création, l'AAP a suivi les performances des élèves en Langue anglaise, en Mathématiques et en Sciences. En 2002, une quatrième branche du curriculum, les Études sociales, y a été ajoutée. Le programme suit les résultats des élèves pour certains groupes d'âge de l'école primaire et secondaire. Dans ce texte, le groupe d'âge visé est celui des 13-14 ans, c'est-à-dire les élèves effectuant leur deuxième année dans l'enseignement secondaire (degré S2).

Dans chaque enquête sur une branche scolaire, une série de tâches sont choisies pour l'évaluation, représentant divers aspects d'un cadre de référence ayant fait l'objet d'un accord préalable. Les enquêtes s'appuient surtout sur des tests papier-crayon, mais d'autres modes d'évaluation s'y ajoutent, comportant des tâches pratiques en mathématiques et en sciences, l'écoute de bandes sonores dans l'évaluation de la langue maternelle, des interviews individuels et des discussions de groupes pour la langue maternelle ou les démarches d'investigation, et depuis tout récemment, des travaux présentés sur ordinateur pour toutes les branches. Un échantillonnage matriciel est employé pour répartir les tâches entre les élèves, de façon que chaque élève ne passe qu'une partie de l'ensemble des tâches d'évaluation. En général, les épreuves papier-crayon sont présentées aux élèves sous forme de cahiers de tests.

Un des aspects les plus intéressants de ce programme est la façon dont il a évolué pour répondre à de nouvelles demandes. Dans les premières années du projet, le rapport sur les performances des élèves présentait les résultats tâche par tâche (une « tâche » pouvant d'ailleurs comprendre plusieurs items de test). Cette façon de faire conduisait inévitablement à une situation où on ne

pouvait parler de l'évolution des performances dans le temps qu'en s'appuyant sur des tâches qui restaient les mêmes, et étaient réutilisées dans les enquêtes successives (constituant donc des tâches « communes »).

En 1991, cependant, furent publiées les *Directives nationales 5-14 pour l'Écosse*, concernant la langue maternelle et les mathématiques (SOED, 1991a, b). Ces directives, et celles qui les suivirent pour les mathématiques (SOEID, 1999), pour l'étude de l'environnement (SOED, 1993, révisées dans Scottish Executive, 2000) et pour d'autres branches scolaires encore, allaient avoir un impact important, non seulement sur l'enseignement et l'évaluation dans les écoles, mais aussi sur le programme AAP, et en particulier sur la façon dont ce dernier devait faire rapport sur les acquisitions des élèves.

Les Directives nationales concernent le curriculum des écoles et son évaluation, pour les élèves âgés de 5 à 14 ans. Elles catégorisent les sujets d'étude en « Thématiques », subdivisées en « Thèmes ». Elles définissent le résultat à atteindre pour un certain thème en fonction d'objectifs d'acquisition et de niveaux de progression (du niveau A au niveau F). Chaque niveau est censé couvrir à peu près deux ans d'éducation. Selon les Directives (voir, par exemple, SOED, 1991b, p. 10), le niveau E devrait être atteint par la plupart des élèves en S2.

Pour donner un exemple, Types et Grandeur de Nombres (qui fera l'objet de l'étude de généralisabilité au point suivant) est l'un des thèmes mathématiques à l'intérieur de la thématique des « Nombres et Mesures ». Le tableau 1 présente quelques-uns des objectifs d'acquisition de ce thème.

Tableau 1
Quelques objectifs d'acquisition du thème Types et Grandeur de Nombres

Niveau A	Travailler avec les nombres entiers de 0 à 20 (compter, ordonner, lire/écrire des phrases, entrer dans la calculatrice)
Niveau C	Travailler avec des nombres entiers jusqu'à 10 000 (compter, ordonner, lire/écrire)
Niveau E	Travailler avec toutes les fractions d'usage courant et utiliser l'équivalence entre elles et les décimales (dans des applications)

Après la publication des Directives nationales, l'AAP a senti de plus en plus nécessaire de formuler les résultats de ses enquêtes sur les performances scolaires en fonction des niveaux décrits verbalement dans ce nouveau cadre de référence. Ce changement, depuis une présentation assez détaillée des résultats moyens pour des tâches isolées jusqu'à la formulation des résultats

atteints par les élèves par rapport à des thèmes et à des niveaux, obligeait à revoir entièrement la catégorisation des tâches utilisées jusque-là pour l'évaluation, et aussi à repenser toute l'organisation des enquêtes sur les performances scolaires. En mathématiques, les décisions relatives à cette nouvelle organisation furent prises sur la base d'une étude de généralisabilité effectuée à partir du fichier qu'on avait conservé des réponses des élèves (Johnson, 1997). La section ci-dessous décrit cette analyse.

Analyse de généralisabilité des données de l'enquête de 1994

L'enquête sur les mathématiques de 1994 était la quatrième que l'AAP ait conduite dans ce domaine (Robertson et al., 1995). Pour préparer cette enquête, les items de test déjà existants et ceux qui avaient été rédigés récemment furent catégorisés par thème et par niveau[2]. L'intention était d'essayer de présenter les résultats sous la forme de taux de facilité moyens pour chaque niveau dans chaque thème (les items étant tous notés de façon dichotomique).

Comme mentionné plus haut, le thème Types et Grandeur de Nombres est le sujet de l'étude de généralisabilité décrite ici. Les items de test suivants, de niveau D, peuvent servir à illustrer son contenu (voir Robertson et al. 1995, pp. 64, 66, 69, pour une présentation des items sous la forme que recevaient les élèves).

- Écris ce nombre en chiffres : cent vingt mille.
- Range les nombres ci-dessous par ordre, en commençant par le plus petit :

  3,7 5 0,4 1,3 0,86
- Quel nombre est un de moins que 30 000 ?
- Dans un test, un garçon a obtenu 32 sur 50. Écris son résultat sous forme de pourcentage.

Les items relatifs aux Types et Grandeur de Nombres qui ont été administrés dans l'enquête visant le degré secondaire S2 ont été répartis dans 15 cahiers de tests. Chaque cahier contenait de 35 à 40 items correspondant à divers thèmes. Il était prévu que les élèves prennent environ une heure pour y répondre. (Chaque élève passait deux cahiers, lors de sessions séparées, dans les deux heures de temps scolaire accordées pour l'enquête.) Malheureusement, le nombre de questions relatives aux Types et Grandeur de Nombres variait selon les cahiers de 0 à 11 pour l'enquête au degré secondaire. Les niveaux étaient aussi représentés de façon inégale entre les cahiers. Ces variations réduisaient la

possibilité d'effectuer une analyse englobant la majorité des items, car il aurait fallu pouvoir traiter des plans très déséquilibrés, ce qui pose de nombreux problèmes (voir Bell, 1985, et Verhelst, 2000, pour des commentaires sur la difficulté accrue dans ce cas d'estimer les composantes de variance, et sur le manque de logiciels conviviaux pour l'analyse de données non équilibrées).

Pour obtenir un jeu de données équilibré en vue d'une étude de généralisabilité, on choisit trois cahiers qui contenaient chacun au moins huit items relatifs aux Types et Grandeur de Nombres. On prit les huit items sur ce sujet du premier cahier et on sélectionna les huit premiers items sur le même sujet dans les deux autres cahiers. On avait ainsi huit items par cahier et 24 items en tout. (La moitié d'entre eux environ étaient de niveau D et le reste était un mélange de niveau C et de niveau E.) Étant donné que le plus petit nombre de garçons ou de filles pour chacun des cahiers était de 227, quelques élèves furent exclus au hasard de l'ensemble des données, de façon à conserver finalement 227 garçons et 227 filles pour chacun des trois cahiers.

Comme il n'y avait pas de recouvrement entre les trois cahiers (ils étaient passés par des élèves différents et contenaient des items différents), on peut considérer que les items sont nichés dans les cahiers, de même que les élèves. Mais les élèves sont aussi nichés dans les deux sexes, masculin et féminin. Pour chaque cahier, les élèves sont croisés avec les items, dans le sens que chaque élève passe tous les items du cahier. Cahiers et sexes sont aussi croisés, puisque garçons et filles ont passé chaque cahier. Sexe est aussi croisé avec items, étant donné qu'à la fois les garçons et les filles ont essayé chaque item. Cette situation est symbolisée dans la figure 1, qui illustre en même temps la partition correspondante de la variance totale des réponses aux items.

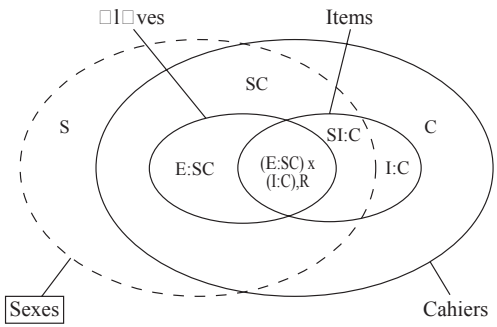

Figure 1 ***Illustration du plan d'observation (E:SC) x (I:C) et représentation de la décomposition de la variance totale («r» représente la variance résiduelle, liée à des sources non déterminées).***

Les élèves ont été tirés au hasard dans leur groupe d'âge, et bien que l'effectif de ce groupe soit fini, il est suffisamment grand pour qu'on puisse considérer qu'on se situe, du point de vue du modèle, dans le cas aléatoire infini. Les items aussi sont considérés comme ayant été tirés aléatoirement à partir d'une grande banque d'items – l'univers infini des items relatifs aux Types et Grandeur de Nombres – même s'il n'y a pas eu, en réalité, de véritable tirage au hasard dans ce cas. Les cahiers aussi sont considérés comme des échantillons de l'univers infini de tous les cahiers possibles, c'est-à-dire de tous les sous-ensembles d'items possibles. Le sexe est une facette fixée (représentée en pointillés sur la figure 1). Ces détails sont rappelés au tableau 2.

Tableau 2
Dispositif d'évaluation : plans d'observation et d'estimation

Facettes	*Niveaux*	*Univers*	*Nom*	*Réduction*
C	3	INF	Cahiers	Aucune
I:C	8	INF	Items	Aucune
S	2	2	Sexe	Aucune
E:SC	227	INF	Élèves	Aucune

Le but de la mesure est d'estimer le niveau de réussite de la population d'élèves de S2 pour le thème Types et Grandeur de Nombres sous la forme d'une moyenne générale X_{ESIC} (les lettres majuscules indiquant qu'on calcule la moyenne sur les niveaux échantillonnés pour chaque facette), moyenne qui fut en pratique présentée comme l'estimation de la facilité moyenne des items.

En réalité, c'est l'erreur type de la mesure absolue qui nous intéresse (*absolute standard error of measurement* : SEM). Nous n'avons pas ici à différencier des élèves, ni des items, ni les niveaux d'aucune autre facette. À l'exception du sexe, qui est l'unique facette fixée, toutes les facettes aléatoires et leurs interactions contribuent, dans notre cas, à la variance d'erreur pour la mesure absolue (SEM). On peut dire que toutes sont par conséquent des facettes d'instrumentation. (Pour le détail des calculs des paramètres de généralisabilité, on se reportera à Brennan, 2001 ; Cardinet & Allal, 1983 ; Cardinet & Tourneur, 1985 ; Cardinet, Tourneur & Allal, 1981).

L'estimation de la variance d'erreur absolue est donnée dans ce cas par la formule suivante :

$$\hat{\sigma}^2(X_{ESIC}|S)=\hat{\sigma}^2(c|S)/n_c+\hat{\sigma}^2(i:c|S)/n_i n_c+\hat{\sigma}^2(e:sc|S)/n_e n_s n_c+\hat{\sigma}^2(ei:sc,r|S)/ n_e n_i n_s n_c$$

Les quatre composantes de variance qui contribuent à la variance d'échantillonnage (la facette S, étant fixée, n'y contribue pas, de même que ses interactions) sont chacune divisées par les effectifs des échantillons de niveaux tirés pour les facettes étudiées. La barre verticale indique simplement la présence de la facette S fixée dans le modèle. La lettre «r» ajoutée à la dernière composante indique que l'interaction (élèves x items) est confondue avec la variance d'erreur résiduelle provenant de toutes les sources d'influence non déterminées (école, classe, ou classification de l'item dans un niveau, dans un format de présentation, etc.). La racine carrée de $\hat{\sigma}^2(X_{ESIC}|S)$ fournit l'estimation de l'erreur type sur la mesure absolue (Brennan, 2001, p. 11), et à partir de là, donne la base pour calculer un intervalle de confiance autour de l'estimation de la moyenne générale.

Le tableau 3 présente les résultats de l'analyse de la variance pour ce modèle. Il montre que ni le sexe, ni les cahiers ne semblent avoir d'influence, ni aucune interaction impliquant ces deux facteurs. Les trois principales sources de variance sont les élèves, les items, et l'interaction entre les élèves et les items. La proportion de la variance des scores (1 ou 0) attribuable aux différences de facilité des items est plus grande que la part attribuable aux différences entre les moyennes des élèves. Comme on pouvait s'y attendre, l'interaction (élèves x items) est élevée : elle représente 64 % du total des composantes de variance, résultat similaire à ce qui avait été observé dans une enquête parallèle en Science (voir Johnson, 1989, p. 93).

Tableau 3

Analyse de variance pour le modèle mixte (E:SC) x (I:C) pour le thème Types et Grandeur de Nombres pour le degré S2 en 1994 (trois cahiers, les huit premiers items sur le sujet dans chacun, et 454 élèves par cahier)

Source de variation	*Somme des carrés*	*d. l.*	*Carré moyen*	*Estimation Composante de variance*
Cahiers (C)	43,6338	2	21,8169	-0,0014
Sexes (S)	0,1619	1	0,1619	-0,0001
Sexes x Cahiers (SC)	1,7583	2	0,8791	0,0002
Items dans les Cahiers (I:C)	548,9477	21	26,1404	0,0572
(Sexes x Items) dans Cahiers (SI:C)	4,6746	21	0,2226	0,0003
Élèves dans (Sexes x Cahiers) (E:SC)	551,6520	1356	0,4068	0,0313
(Élèves dans Sexes et Cahiers) x (Items dans Cahiers) (E:SC) x (I:C) plus l'erreur résiduelle	1482,1277	9492	0,1561	0,1561

***** Pour des commentaires sur les estimations négatives, voir Brennan, 2001, p. 84-85.

En remplaçant dans la formule donnée plus haut les symboles des composantes de variance et des tailles d'échantillons pour chaque facette par les estimations appropriées, on trouve que la variance d'erreur absolue pour la moyenne générale des items sur le sujet Types et Grandeur de Nombres est de 0,0025. La racine carrée de cette variance d'erreur, 0,05, est l'erreur type de mesure (SEM). On obtient alors un intervalle de confiance à 95% pour la moyenne générale égal à ± 0,098, ou ± 9,8 points pour cent autour de la facilité moyenne des items.

La question est maintenant de savoir si cette erreur peut être réduite, et si oui, comment le faire en restant à l'intérieur des contraintes données.

Le temps à disposition des enquêteurs pour faire passer les tests papier-crayon est limité à deux heures pour chaque élève au niveau S2, et cette attribution ne peut pas être augmentée de façon substantielle. Il ne serait pas non plus acceptable d'augmenter beaucoup la taille de l'échantillon total des élèves touchés par l'enquête, qui représente environ 5% de la population des élèves de l'âge visé. De plus, il ne serait pas souhaitable d'augmenter le nombre d'items portant sur les Types et Grandeur de Nombres, si on le faisait aux dépens d'autres sujets également importants. Par conséquent, toute augmentation dans les enquêtes futures du nombre d'items portant sur les Types et Grandeur de Nombres ne pourra se faire qu'en augmentant le nombre de cahiers utilisés et en diminuant le nombre d'élèves passant chaque cahier.

Le tableau 4 présente les résultats de trois études d'optimisation (D-studies) pour différents plans d'échantillonnage des élèves et diverses stratégies d'administration des items. (La variance d'erreur est calculée en utilisant la formule donnée plus haut, en remplaçant les effectifs d'échantillons utilisés dans l'étude de généralisabilité par les effectifs qu'on envisage.)

Tableau 4

Résultats de l'étude d'optimisation pour Types et Grandeur de Nombres au degré deuxième secondaire

	Étude G	*Plan I*	*Plan II*	*Plan III*
Nombre de cahiers	3	15	15	20
Nombre d'items par cahier	8	2	4	4
Nombre d'élèves par cahier	454	250	250	200
Nombre total d'items	24	30	60	80
Nombre total d'élèves	1362	3750	3750	4000
Variance d'erreur absolue	0,0025	0,0019	0,0010	0,0007
Erreur type de mesure	0,050	0,044	0,031	0,027
Intervalle de confiance à 95% en points pour cent	± 9,8	± 8,6	± 6,1	± 5,3

Le Plan I montre ce qu'on pourrait attendre si le nombre total d'items était augmenté de 24 (nombre utilisé dans l'étude de généralisabilité) à 30, les items étant alors distribués entre 15 cahiers au lieu de trois, avec moins d'items dans chacun. En même temps, le nombre total d'élèves serait multiplié presque par trois, passant de 1 362 à 3 750, mais le nombre d'élèves par cahier serait presque divisé par deux, passant de 454 à 250. L'étude d'optimisation prédit que la variance d'erreur baisserait de 0,0025 à 0,0019.

Le Plan II double le nombre total d'items administrés, par rapport au Plan I, allant de 30 à 60. Il place deux fois plus d'items dans chaque cahier, mais il garde constant le nombre de cahiers et le nombre d'élèves. Comme on pouvait s'y attendre, la variance d'erreur décroît à nouveau jusqu'à 0,0010.

Le Plan III augmente encore le nombre total d'items, qui se monte à 80 maintenant. Pour contenir ce nombre supérieur, il faut augmenter le nombre de cahiers, de 15 à 20, tandis que le nombre d'items sur le sujet des Types et Grandeur de Nombres dans chaque cahier reste inchangé à quatre. On augmente le nombre total d'élèves testés, qui monte de 3 750 à 4 000, quoique le nombre d'élèves par cahier se réduise à 200. Le résultat est que la variance d'erreur est baissée encore davantage, jusqu'à 0,0007.

Le Plan III aurait pu être réalisé dans l'enquête suivante, si l'on considérait uniquement le sujet des Types et Grandeur de Nombres, en testant à peu près le même nombre d'élèves qu'en 1994. Mais cette taille d'échantillon n'aurait pas permis de consacrer une place égale, pour ce qui est du nombre d'items, aux autres thèmes de mathématiques. C'est pourquoi le Plan II a semblé le dispositif le plus approprié pour les futures enquêtes, étant donné les contraintes pratiques limitant de telles études.

Sur la base de cette analyse et d'autres encore, l'évaluation effectuée après l'enquête de 1994 incluait parmi ses recommandations les propositions suivantes. Le nombre d'items utilisés par thème et par niveau dans les futures enquêtes devrait être accru (le sujet des Types et Grandeur de Nombres était le thème qui avait été le plus représenté dans l'enquête de 1994). Le nombre d'élèves par cahier, et par conséquent par item, devrait être réduit de moitié. Chaque thème de mathématique devrait être testé par le même nombre d'items. La distribution des items dans les cahiers devrait accorder une place équilibrée à chaque thème de mathématique. Ces recommandations furent mises en pratique dans l'enquête de 2000.

Le plan d'enquête de 2000

Pour l'enquête sur les mathématiques de l'an 2000, on décida que tout thème qui pouvait être testé par une procédure papier-crayon serait représenté à chaque niveau par 30 items, et que chaque item serait passé par environ 250 élèves. Étant donné le nombre de thèmes (11 au degré S2), cet effectif était le maximum qu'on pouvait introduire, compte tenu de la taille de l'échantillon total d'élèves et du temps de test qui étaient acceptables en pratique. Pour le degré deuxième secondaire, il y eut 30 items de niveau D et 30 de niveau E représentant tous les thèmes, dont Types et Grandeur de Nombres. (Dans quelques cas, mais pas pour ce dernier thème, on introduisit aussi des items de niveau F.) Les items furent distribués entre les cahiers de façon que tout élève passe juste deux items de chaque thème à chaque niveau, parmi les quelque 50 items au total qu'il pouvait passer en deux heures.

Cette fois, tous les items relatifs aux Types et Grandeur de Nombres furent inclus dans une étude de généralisabilité, reprenant le dispositif de la figure 1. Quelques élèves furent exclus au hasard, pour équilibrer les effectifs masculins et féminins passant chaque cahier (le nombre d'élèves par cahier allait de 223 à 237). Les détails de l'analyse sont présentés au tableau 5.

Tableau 5
Dispositif d'évaluation en 2000 : plans d'observation et d'estimation

Facettes	*Niveaux*	*Univers*	*Nom*	*Réduction*
C	15	INF	Cahiers	aucune
I:C	4	INF	Items	aucune
S	2	2	Sexe	aucune
E:SC	100	INF	Élèves	aucune

Le tableau 6 fournit les résultats de l'analyse de variance pour les données de l'enquête 2000. On a 60 items au total relatifs aux Types et Grandeur de Nombres (mélange équilibré des niveaux D et E), distribués de façon égale dans les 15 cahiers, soit quatre items de ce thème par cahier. L'échantillon d'élèves conservé comprenait 100 garçons et 100 filles.

Tableau 6
Analyse de variance pour le modèle mixte (E:SC) x (I:C) pour le thème Types et Grandeur de Nombres au degré deuxième secondaire en 2000 (15 cahiers, contenant chacun quatre items sur ce sujet, avec 200 élèves par cahier)

Source de variation	*Somme des carrés*	*d. l.*	*Carré moyen*	*Estimation composante de variance*
Cahiers (C)	159,9097	14	11,4221	0,0028
Sexes (S)	1,1603	1	1,1603	0,0001
Sexes x Cahiers (SC)	6,6297	14	0,4735	0,0002
Items dans les Cahiers (I:C)	404,2750	45	8,9839	0,0442
(Sexes x Items) dans Cahiers (SI:C)	7,8800	45	0,1751	0,0004
Élèves dans (Sexes x Cahiers) (E:SC)	937,8350	2970	0,3158	0,0439
Élèves dans (Sexes x Cahiers) x (Items dans Cahiers) (E:SC) x (I:C) plus l'erreur résiduelle	1249,3450	8910	0,1402	0,1402

En retournant au tableau 4, on voit que le Plan II prédisait une variance d'erreur absolue de 0,0010 pour une telle stratégie d'échantillonnage des élèves et d'administration des items. (Le Plan II prévoyait 250 élèves par cahier, alors que nous n'en n'avons que 200 dans cette analyse.) En fait, la variance d'erreur absolue pour les données de 2000 est de 0,0009, ce qui donne une erreur type de mesure de 0,031, et un intervalle de confiance à 95% de ± 6 points pour cent autour de la moyenne de facilité observée, égale à 65% pour les items relatifs aux Types et Grandeur de Nombres (niveaux D et E confondus).

Selon le tableau 4, si l'on réduisait le nombre d'items à 30 au total (deux par cahier), il était prévu que la variance d'erreur augmenterait à 0,0017 et que l'intervalle de confiance à 95% serait de ± 8,1 points pour cent. Effectivement, quand on a analysé séparément les 30 items de niveau D et les 30 items de niveau E, on a obtenu comme résultat une variance d'erreur de 0,0016 dans le premier cas, et de 0,0018 dans le second, avec des intervalles de confiance à 95% de ± 7,8 et de ± 8,3 points pour cent autour de la valeur moyenne des taux de facilité des items.

Perspectives d'avenir

Comme on vient de le voir, dans l'enquête sur les mathématiques de 2000, aux divers contenus (thèmes) correspondaient à chaque niveau le même nombre d'items. C'est probablement la meilleure stratégie à adopter à l'avenir, étant donné que la variation dans les réponses aux items restera probablement du même ordre de grandeur pour les autres thèmes. Cette hypothèse sera, bien sûr, contrôlée par d'autres analyses et, si c'est nécessaire (et réalisable pratiquement), le dispositif d'observation sera adapté en conséquence pour la prochaine enquête sur les mathématiques en 2004. En plus, d'autres évaluations des dispositifs d'enquête seront effectuées pour les autres matières scolaires que suit le programme AAP, soit la langue maternelle, la science et les études sociales, du fait que la façon de présenter les résultats subit une transformation semblable dans ces domaines aussi.

Les résultats des analyses envisagées intéresseront évidemment les personnes engagées dans la planification et l'exploitation des enquêtes de l'AAP, mais ils auront des implications également dans d'autres programmes d'enquêtes ailleurs dans le monde, y compris les programmes d'enquêtes comparatives internationales.

NOTES

1. Précédemment Directrice technique adjointe de l'APU (Assessment of Performance Unit) du Programme d'enquête en sciences du Royaume-Uni ; actuellement Conseillère technique du Programme d'évaluation des acquis d'Écosse (Assessment of Achievement Programme : AAP).
2. En fait, la division en contenus utilisée pour cette enquête comprenait des « catégories » et des « sous-catégories » qui, à l'exception de quelques sous-catégories, correspondaient aux « Thématiques » et aux « Thèmes » des Directives nationales.

RÉFÉRENCES

AAP (2001). Assessment *of Achievement Programme : Report of the Sixth Survey of Mathematics*. Edinburgh : Scottish Executive Education Department.

Bell, J.F. (1985). Generalizability theory : The software problem. *Journal of Educational Statistics, 10*, 19–29.

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City : The American College Testing Program.

Brennan, R.L. (2001). *Generalizability theory*. New York : Springer-Verlag.

Cardinet, J. & Allal, L. (1983). Estimation of generalizability parameters. In L.J. Fyans (Ed.), *Generalizability theory : Inferences and practical applications*. New Directions for Testing and Measurement, no. 18. San Francisco : Jossey-Bass.

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

Cardinet, J., Tourneur, Y. & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18*, 183-204.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements : theory of generalizability for scores and profiles*. New York : Wiley.

Johnson, S. (1989). *National Assessment : the APU Science Approach*. London : HMSO.

Johnson, S. (1997). *The 1994 AAP Mathematics Survey : An evaluation based on analysis of pupil response data*. Consultancy report submitted to the Scottish Office Education Department.

Johnson, S. & Bell, J. (1985). Evaluating and predicting survey efficiency using generalizability theory. *Journal of Educational Measurement, 22*, 107-119.

Robertson, I.J., Meechan, R.C., Clarke, D., Moffat, J. & McCormick, E. (1995). *Assessment of Achievement Programme : Report of the Fourth Survey of Mathematics*. Volume 1. Glasgow : Faculty of Education, University of Strathclyde.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability theory : a primer*. Newbury Park, Ca : Sage Publications.

SOED (1991a). *National Guidelines : English Language 5-14*. Edinburgh : Scottish Office Education Department.

SOED (1991b). *National Guidelines : Mathematics 5-14*. Edinburgh : Scottish Office Education Department.

SOED (1993). *National Guidelines : Environmental Studies 5-14*. Edinburgh : Scottish Office Education Department. [Revised in 2000]

SOEID (1999). *National Guidelines : Mathematics 5-14 Level F*. Edinburgh : Scottish Office Education and Industry Department.

Scottish Executive (2000). *Environmental Studies : Society, Science and Technology. 5-14 National Guidelines*. Glasgow : Learning & Teaching Scotland. [Revision of 1993 Guidelines]

Verhelst, N.D. (2000). *Estimating variance components in unbalanced designs*. (R&D Notities 2000-1). Arnhem : CITO.