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

Herausgeber: Schweizerische Gesellschaft für Bildungsforschung

Band: 3 (1981)

Heft: 1

Artikel: L'étude de généralisabilité d'un survey

Autor: Tourneur, Yvan / Cardinet, Jean

DOI: https://doi.org/10.5169/seals-786429

L'étude de généralisabilité d'un survey

Yvan Tourneur, Jean Cardinet

L'article illustre comment on peut estimer la fidélité des différentes mesures résultant d'un survey. Il utilise une analyse de variance effectuée sur des tests de mathématique de première année primaire. Après avoir décrit le plan d'observation, il montre la façon dont ont été calculées les sommes de carrés, puis les composantes de variance. Il explique les étapes successives de l'étude de généralisabilité, d'abord pour la différenciation des résultats moyens selon les domaines mathématiques, puis pour la différenciation des résultats par élève.

1. But de l'article

Dans les pages qui suivent, nous souhaitons illustrer certaines applications possibles de la théorie de la généralisabilité. Nous le ferons sur un exemple, celui des tests passés en 1976 par les enfants de Suisse romande, après une année d'un enseignement rénové de la mathématique. Nous montrerons comment les formules de généralisabilité permettent d'apprécier la fidélité des résultats d'un telle enquête.

Le problème de la fidélité des instruments de mesure se pose dans toutes les sciences; cependant, il est vite apparu comme particulièrement aigu dans les sciences sociales, en raison de la multiplicité des facteurs qui peuvent perturber les observations dans ces domaines. Les techniques d'estimation de fidélité que nous utilisons sont issues des travaux de Spearman (1910) au début du siècle en psychologie. Cet auteur, considérant chaque score observé comme la somme d'un score vrai et d'une erreur, a défini la fidélité comme le rapport de la variance des scores vrais à la variance des scores observés des personnes examinées. La corrélation entre deux prises de mesures successives en donnait une estimation utilisable. Le modèle, trop élémentaire, a dû être dépassé par la suite: il existe en effet plusieurs fidélités, autant que de définitions possibles des sources de variation à prendre en compte dans la variance-erreur. C'est pour cette raison que Cronbach, Gleser, Nanda et Rajaratnam (1972) ont développé la théorie de la généralisabilité qui permet de traiter les cas où plusieurs effets connus influencent à la fois les résultats: items, correcteurs, moments par exemple. Malheureusement, ces auteurs se sont limités aux situations où l'objet d'étude était une personne (candidat, élève, patient, etc.). L'extension récente de leur théorie à un nombre quelconque de dimensions de différenciation (Cardinet, J., Tourneur, Y. et Allal, L., 1979) ouvre de nouveaux domaines d'application, que nous allons essayer d'illustrer.

Il doit être bien clair pour le lecteur que le cas traité ne constituera qu'un prétexte, destiné à faciliter la présentation de méthodes statistiques qu'on aurait du mal à décrire dans l'abstrait. Nous montrerons étape par étape les calculs nécessaires et les informations obtenues. La place disponible ne permettra pas, cependant, de justifier les formules utilisées. Le lecteur qui souhaiterait plus d'explications pourra trouver la théorie correspondante dans une publication récente (Cardinet, Tourneur, 1979). En examinant la démarche suivie et les conclusions auxquelles on parvient par une étude de généralisabilité, chacun pourra juger de l'intérêt et des limites de cette technique dans son domaine d'application particulier.

2. Le contexte de l'étude: le survey de mathématique

Un nouveau curriculum de mathématique a été introduit dans les écoles primaires de Suisse romande en 1973. Pour faciliter son acceptation, promesse a été faite à l'époque d'en étudier scientifiquement les résultats, pour pouvoir effectuer ultérieurement les adaptations nécessaires. Un premier volet de l'évaluation a consisté à interroger les maîtres et maîtresses de première année qui appliquaient le nouveau programme. Un second volet a impliqué de mesurer les résultats des élèves, pour les mettre en rapport avec les appréciations des enseignants (Cardinet, 1977).

Les tests utilisés pour une telle enquête (appelée «survey», faute d'un terme français approprié) diffèrent des épreuves pédagogiques habituelles. Il ne s'agit pas en effet de mesurer la réussite des élèves, mais plutôt le degré d'atteinte des objectifs pédagogiques visés. Alors qu'on échantillonne d'habitude plusieurs questions pour mieux mesurer chaque élève, un survey amène à échantillonner de nombreux élèves pour mieux mesurer la réussite moyenne à chaque question. Les épreuves peuvent donc être aussi courtes qu'on le désire; les conditions n'ont pas besoin d'être identiques pour tous les élèves, ni comparables d'une classe à l'autre. La fidéilité de la mesure se marque dans ce cas par la stabilité du classement des objectifs selon le degré de maîtrise obtenu dans l'ensemble du système scolaire.
A côté de cette finalité primordiale du survey, on ne peut cependant pas ignorer d'autres informations, éventuellement intéressantes elles aussi: évolution des performances dans le temps, résultats comparés de divers groupes d'élèves, effets de méthodes pédagogiques différentes, influence de divers modes de présentation ou de correction, etc. Même si les épreuves n'ont pas été prévues pour effectuer des différenciations de ce type, il peut être utile de les réaliser, si du moins l'erreur de mesure n'est pas tellement importante qu'elle invalide à l'avance toute conclusion.
La raison d'être d'une étude de généralisabilité peut donc être d'apprécier la précision des diverses mesures que l'on pourrait tirer d'un survey, pour ne pas être tenté d'interpréter des différences trop instables, ou au contraire pour suggérer des modifications du plan d'observation qui puissent rendre de telles comparaisons suffisamment significatives.

3. L'estimation des composantes de variance

Il est rare que l'on puisse formuler des conclusions à propos d'un dispositif sur une base purement mathématique. La plupart du temps, il faut d'abord obtenir une estimation de l'importance relative des sources de variation à considérer. Ensuite, en fonction de l'ampleur des fluctuations prévisibles, on peut alors déterminer des marges d'erreur probables et modifier éventuellement le dispositif d'observation.
De ces deux étapes d'une étude de généralisabilité, la première qui conduit à estimer l'importance des sources de variance existantes a été dénommée «Etude G» par Cronbach, tandis que la seconde correspond pour lui à l'«Etude D», parce que'elle prépare le dispositif approprié aux décisions à prendre.
La suite de cet article va décrire l'étude G que nous avons effectuée. Nous présenterons d'abord les étapes que nous avons suivies pour estimer les composantes de variance qui interviennent dans ce survey. Nous en tirerons ensuite des indications sur les marges d'erreur de notre dispositif.

3. 1 Choix des facettes et des niveaux

Nous avons structuré les observations en fonction des facettes suivantes:
- Domaines (D) de la mathématique: le programme de première année porte sur les quatre «avenues» suivantes: Ensembles – Relations, Numération, Opérations, Découverte de l'Espace.
- Classes (C): nous avons pu observer pour cette étude de généralisabilité vingt classes provenant de régions différentes de Suisse romande.
- Ages (A): dans chaque classe nous avons distingué deux sous-groupes d'élèves, ceux qui étaient plus jeunes que la moyenne d'âge de leur canton, et ceux qui étaient plus âgés. La différence d'âge moyen entre les deux groupes devrait être de six mois environ.
- Elèves (E): pour conserver l'équilibre du plan d'analyse de variance, nous devions avoir le même nombre d'élèves dans chaque groupe d'âge de chaque classe. Cela n'a été possible

qu'en réduisant ce nombre au minimum apparu dans l'ensemble des groupes, soit deux élèves. Ces deux élèves ont été choisis au hasard lorsque le groupe d'âge de la classe était plus nombreux. On a conservé ainsi quatre élèves par classe, soit quatre-vingts élèves au total.
- Formes (F): l'ensemble des questions a été divisé en cinq groupes (A, B, C, D et E) attribués chacun à un groupe de 4 classes différentes. Chacune des cinq formes comportait douze questions, trois par domaine.
- Séries (S): les questions d'une même forme ont encore été subdivisées en trois séries, en vue de l'étude D ultérieure où il fallait réduire au minimum la durée de la passation des tests et où chaque enfant ne devait recevoir qu'une série de quatre questions, une question par domaine. Dans l'étude G chaque enfant a reçu les trois séries d'une même forme, soit douze questions au total. Chaque question était cotée comme réussie ou non, mais avec des possibilités de degrés intermédiaires si l'enfant avait bien traité certains aspects du problème et moins bien d'autres aspects.

3.2 Plan d'observation

Le plan utilisé, relativement complexe, doit être bien explicité. Chaque enfant a répondu à trois séries de quatre questions, une question par domaine. Deux enfants jeunes et deux enfants âgés constituaient une classe. Quatre classes recevaient la même forme. Cinq formes différentes étaient expérimentées au total.
Les six facettes présentaient donc les relations suivantes:
- les domaines (D) étaient croisés avec les cinq autres facettes
- les classes (C:F) étaient croisées avec les âges; elles étaient nichées dans les formes et croisées avec les séries dans les formes; elles nichaient les élèves
- les âges (A) nichaient les élèves et étaient croisés avec les autres facettes
- les élèves (E:AC:F) étaient croisés avec les séries, mais nichés dans les formes, puisque leurs classes l'étaient aussi
- les formes (F) nichaient les séries (S:F)
- la facette Questions était confondue avec l'interaction Domaines x Séries.

Les nombres de niveaux étaient les suivants:

Domaines :	$n_d = 4$	Classses :	$n_c = 4$	Ages :	$n_a = 2$
Elèves :	$n_e = 2$	Formes :	$n_f = 5$	Séries :	$n_s = 3$

Le nombre total d'observations N est donné par le produit de tous ces niveaux:

$n_d \cdot n_s \cdot n_e \cdot n_a \cdot n_c \cdot n_f = 960$

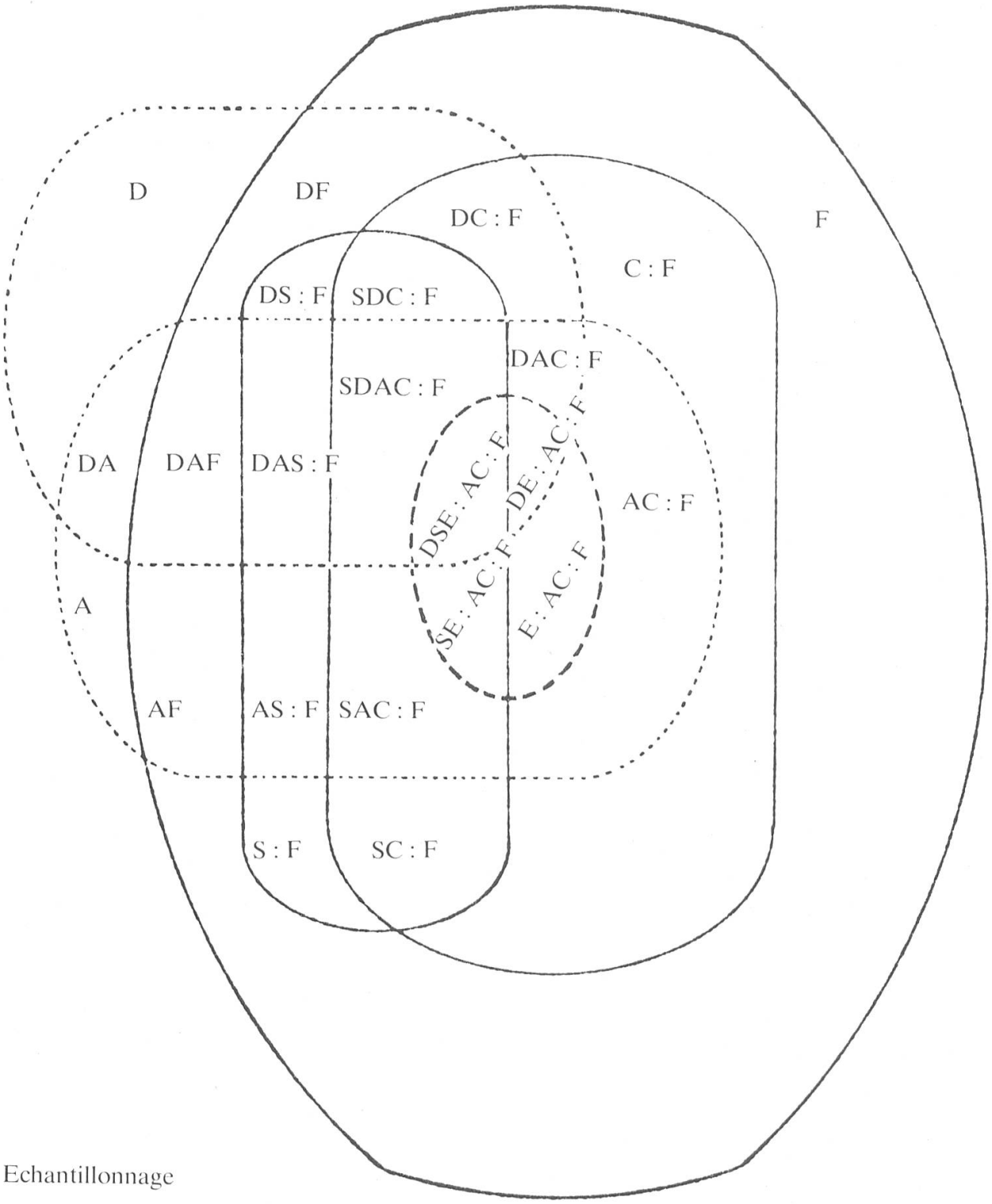
3.3 Représentation du plan d'observation

Il est possible de représenter graphiquement toutes ces relations de croisement et de nichage. Le résultat apparaît à la figure 1. On peut vérifier que 23 plages différentes apparaissent, comme une formule permet de le calculer par ailleurs. On peut faire correspondre à chacune de ces plages une partie de la somme des carrés totale, analysable par les méthodes de l'ANOVA.

3.4 Calcul des sommes de carrés

Comme nous ne possédions pas de programme d'analyse de variance capable d'effectuer directement les analyses souhaitées, nous avons utilisé la propriété d'additivité de sommes de carrés pour tirer parti d'un programme standard pour plan factoriel croisé. Nous avons effectué

Figure 1: *Représentation graphique du plan d'observation de l'étude G (les pointillés, traitillés et traits pleins traduisent le plan d'estimation).*



Echantillonnage

——— purement aléatoire

- - - - aléatoire fini

..... exhaustif

l'analyse comme s'il s'agissait d'un plan à six facteurs croisés (obtenant (2^6 - 1) sommes de carrés et autant de degrés de liberté partiels) et nous avons ensuite regroupé les sources de variance confondues, de la façon qui est indiquée au tableau 1.

Tableau 1: Sommes des carrés confondues

Somme des carrés résultante	Sommes des carrés confondues
D	D
A	A
DA	DA
F	F
DF	DF
AF	AF
DAF	DAF
S : F	S, SF
DS : F	DS, DSF
AS : F	AS, ASF
DAS : F	DAS, DASF
C : F	C, CF
DC : F	DC, DCF
AC : F	AC, ACF
DAC : F	DAC, DACF
SC : F	SC, SCF
SDC : F	SDC, SDCF
SAC : F	SAC, SACF
SDAC : F	SDAC, SDACF
E : AC : F	E, EA, EC, EAC, EF, EAF, ECF, EACF
DE : AC : F	DE, DEA, DEC, DEAC, DEF, DEAF, DECF, DEACF
SE : AC : F	SE, SEA, SEC, SEAC, SEF, SEAF, SECF, SEACF
DSE : AC : F	DSE, DSEA, DSEC, DSEAC, DSEF, DSEAF, DSECF, DSEACF

Les degrés de liberté ont été regroupés parallèlement. Les résultats apparaissent au tableau 2.

Tableau 2: Analyse de la variance pour le plan d'observation de l'étude de G

Source de variation	Somme de carrés	Degrés de liberté	Carré moyen	Composante de variance (en 10^{-5}) selon le modèle	
				aléatoire	mixte
D	5,72145	3	1,90715	623	651
A	0,09282	1	0,09282	(–16)	(– 2)
DA	0,37525	3	0,12508	55	55
F	1,16241	4	0,29060	(– 79)	(–43)
DF	4,14477	12	0,34539	126	83
AF	0,41005	4	0,10251	30	9
DAF	0,70669	12	0,05889	(–86)	(–86)
S : F	2,52408	10	0,25240	41	259
DS : F	6,89395	30	0,22979	1050	1164
AS : F	0,27405	10	0,02740	11	(–90)
DAS : F	1,32269	30	0,04408	(–403)	(–403)
C : F	2,82387	15	0,18826	30	181
DC : F	6,18491	45	0,13744	335	430
AC : F	1,54052	15	0,10270	181	294
DAC : F	3,57888	45	0,07953	54	189
SC : F	2,60004	30	0,08666	81	92
SDC : F	8,45954	90	0,09399	442	1090
SAC : F	1,67845	30	0,05595	(–509)	(–198)
SDAC : F	6,86732	90	0,07630	665	1295
E : AC : F	7,10452	40	0,17761	661	1338
DE : AC : F	12,43472	120	0,10362	1354	1354
SE : AC : F	4,61813	80	0,05772	(–132)	1443
DES : AC : F	15,12135	240	0,06300	6300	6300

Les nombres de niveaux admissibles de chaque facette sont ainsi les suivants:

Domaines : $N_d = 4$ Classes : $N_c = \infty$ Ages : $N_a = 2$

Elèves : $N_e = 10$ Formes : $N_f = \infty$ Séries : $N_8 = \infty$

3.5 Plan d'estimation

La suite des analyses dépend de la nature de l'échantillonnage utilisé pour choisir les niveaux de chaque facette.

Les quatre domaines correspondent aux quatre «avenues» du plan d'études romand. Il n'était pas possible d'en choisir d'autres et tous les domaines du programme ont été abordés. Il s'agit typiquement d'un échantillonnage exhaustif d'un univers fini. On parlera pour D de «facette fixée».

Les vingt classes observées, au contraire, ne représentent qu'une très petite partie de la population des classes de Suisse romande. Leur choix a été fait plus ou moins au hasard, sans lien en tout cas avec la performance à mesurer. On parlera pour C:F d'échantillonnage purement aléatoire.

Les deux âges, regroupant d'un côté les élèves les plus jeunes, de l'autre les plus âgés, épuisent toutes les possibilités. La facette A est donc fixée.
Les deux élèves choisis dans chacun des deux groupes d'âge de chaque classe représentent un type d'échantillonnage intermédiaire. Ils ont bien été tirés au hasard, mais dans une population très petite, que l'on peut estimer à 10 élèves environ. Il s'agit d'un échantillonnage aléatoire fini.
Les cinq formes et les trois séries de chaque forme représentent un choix tout à fait aléatoire de questions. D'autres chercheurs auraient produit d'autres questions et nous n'avons suivi aucun système pour regrouper telle question avec telle autre. La seule contrainte a consisté à choisir une question de chaque domaine pour constituer une série. On peut donc considérer qu'il existait une infinité de combinaisons de quatre questions (une par domaine). Les séries ont été tirées aléatoirement dans cette population infinie et les formes ont été constituées en même temps, par le meme tirage aléatoire, les trois premières séries tirées constituant la première forme, etc. Les facettes S:F et F sont donc échantillonnnées de façon purement aléatoire.

3. 6 Calcul des composantes de variance pour le modèle entièrement aléatoire

Plusieurs algorithmes existent pour estimer les composantes de variance en tenant compte du mode d'échantillonnage particulier de chaque facette. Nous préférons procéder en deux temps, en calculant d'abord les valeurs des composantes comme si toutes les facettes étaient aléatoires, puis en estimant les composantes mixtes à partir de ces valeurs intermédiaires. On peut ainsi utiliser des formules générales; de plus on peut mieux apprécier l'effet des décisions prises relativement au mode de tirage des niveaux observés et changer éventuellement ces décisions dans le plan d'estimation de l'étude D.
Pour permettre au lecteur intéressé de suivre notre démarche, s'il le désire, nous indiquons ci-dessous les formules de calcul pour les six effets principaux, à titre d'exemples:

$$\sigma^2 (d) = 1/ n_c n_a n_e n_f n_s \cdot \{CM (d) - \{CM (da) + CM (df)\} + CM (daf)\}$$

$$\sigma^2 (c : f) = 1/ n_d n_a n_e n_s \cdot \{CM (c : f) - \{CM (sc : f) + CM (dc : f) + CM (ac : f)\} + \{CM (sdc : f) + CM (dac : f) + CM (sac : f)\} - CM (sdac : f)\}$$

$$\sigma^2 (a) = 1/ n_d n_c n_e n_f n_s \cdot \{CM (a) - \{CM (ad) + CM (af)\} + CM (daf)\}$$

$$\sigma^2 (e:ac:f) = 1/ n_d n_s \cdot \{CM (e : ac : f) - \{CM (de : ac : f) + CM (se : ac : f)\} + CM (dse : ac : f)\}$$

$$\sigma^2 (f) = 1/ n_d n_c n_a n_e n_s \cdot \{CM (f) - \{CM (s : f) + CM (df) + CM (af) + CM (c : f)\} + \{CM (ds:f) + CM (as:f) + CM (sc:f)+ CM (daf) + CM (dc:f) + CM (ac:f)\} - \{CM (das:f) + CM (sac:f) + CM (dac:f) + CM (sdc:f)\} + CM (sdac:f)\}$$

$$\sigma^2 (s:f) = 1/ n_d n_c n_a n_e \cdot \{CM (s:f) - \{CM (ds:f) + CM (as:f) + CM (sc:f)\} + \{CM (sdc:f) + CM (das:f) + CM (sac:f) \} - CM (sdac:f)\}$$

La composante de variance selon le modèle aléatoire pour D est alors (exprimée en 10^{-5}):
$\sigma^2 (d) = 1/240 \cdot (190\,715 - 47\,047 + 5\,889) = 149\,557/240 = 623$

Pour C on obtient:

σ^2 (c:f) = 1/48 · (18 826 − 32 680 + 22 947 − 7 630) = 1463/48 = 30

La composante suivante représente un cas particulier:

σ^2 (a) = 1/480 · (9 282 − 22 759 + 5 889) = − 7 588/480 = −16

Bien qu'une composante ne puisse pas, en tant que variance, être négative, nous conservons cette valeur dans ce cas particulier car nous ne l'utilisons que comme valeur intermédiaire, pour pouvoir calculer ultérieurement une composante mixte.
Les autres valeurs des composantes aléatoires apparaissent au tableau 2.

3.7 Calcul des composantes de variance pour le modèle mixte

Pour trouver la valeur d'une composante mixte on ajoute à la composante aléatoire les parts de variance liées à ses interactions avec d'autres facettes. En appliquant un algorithme classique, on trouve, pour la facette D (la lettre M symbolisant la composante pour le modèle d'échantillonnage spécifié précédemment):

$$\begin{aligned}\sigma^2(d|M) &= \sigma^2(d) + 1/N_a \cdot \sigma^2(da) + 1/N_f \cdot \sigma^2(df) + 1/N_c N_a N_e N_f \cdot \sigma^2(de:ac:f)\\ &\quad + 1/N_f N_s \cdot \sigma^2(ds:f) + 1/n_c N_f \cdot \sigma^2(dc:f) + 1/N_a N_f \cdot \sigma^2(daf)\\ &\quad + 1/N_c N_f N_s \cdot \sigma^2(sdc:f) + 1/N_a N_f N_s \cdot \sigma^2(das:f) + 1/N_a N_c N_f \cdot \sigma^2(dac:f)\\ &\quad + 1/N_c N_a N_e N_f N_s \cdot \sigma^2(dse:ac:f) + 1/N_c N_a N_f N_s \cdot \sigma^2(sdac:f)\end{aligned}$$

En tenant compte du fait que N_c N_f et N_s sont infinis et que les composantes de variance qui sont divisées par ces termes s'annulent, deux termes seulement subsistent pour le modèle mixte considéré:

$$\begin{aligned}\sigma^2(d|M) &= \sigma^2(d) + 1/N_a \cdot \sigma^2(da)\\ &= 623 + 1/2 \cdot 55 = 650\end{aligned}$$

Les autres valeurs calculées apparaissent au tableau 2. Lorsqu'une estimation de variance est négative, on interprète généralement ce fait comme dû à des fluctuations aléatoires et la composante correspondante est alors considérée comme nulle.
En examinant ces composantes on voit que la plus importante est DSE : AC : F. Il s'agit de l'interaction des élèves et des questions, c'est-à-dire du terme d'interaction d'ordre le plus élevé. Ce résultat, bien que peu souhaitable, est dans la nature des choses: la réussite d'un élève à une question est difficilement prédictible. D'autre part, cette interaction est gonflée par le mode de cotation en vrai-faux qui ne correspond pas au modèle continu de l'ANOVA.
Vient ensuite SE:AC:F, c'est-à-dire l'interaction des élèves avec les séries, qui représente aussi une interaction d'ordre très élevé, liée aux réactions imprédictibles de chaque élève devant les questions et au mode de correction discontinu.
Les deux composantes suivantes, par ordre d'importance, représentent le niveau de compétence des élèves, soit général (E : AC : F), soit par domaine (DE : AC : F). C'est la variance que cherchent à atteindre les épreuves pédagogiques classiques.
Les trois composantes qui viennent ensuite expriment les différences dans les taux de difficulté des questions, soit pour tous les élèves (DS : F), soit selon la classe considérée (DSC : F), soit selon le groupe d'âge dans la classe (SDAC : F).

Ces sept composantes sont les seules qui dépassent .01. Peut-être peut-on considérer comme importantes encore les composantes D et DC : F qui représentent les différences de réussite selon les domaines, soit pour tous les élèves réunis (D), soit selon les classes (DC : F). On peut considérer comme négligeables les 14 autres sources de variance. Elles sont en effet du même ordre de grandeur que l'effet des séries (S : F) que les auteurs des tests considéraient au départ comme de difficulté équivalente. (Par souci d'exactitude, et puisqu' il s'agit d'un exemple, on tiendra compte néanmoins de toutes les composantes dans les calculs ci-dessous.)

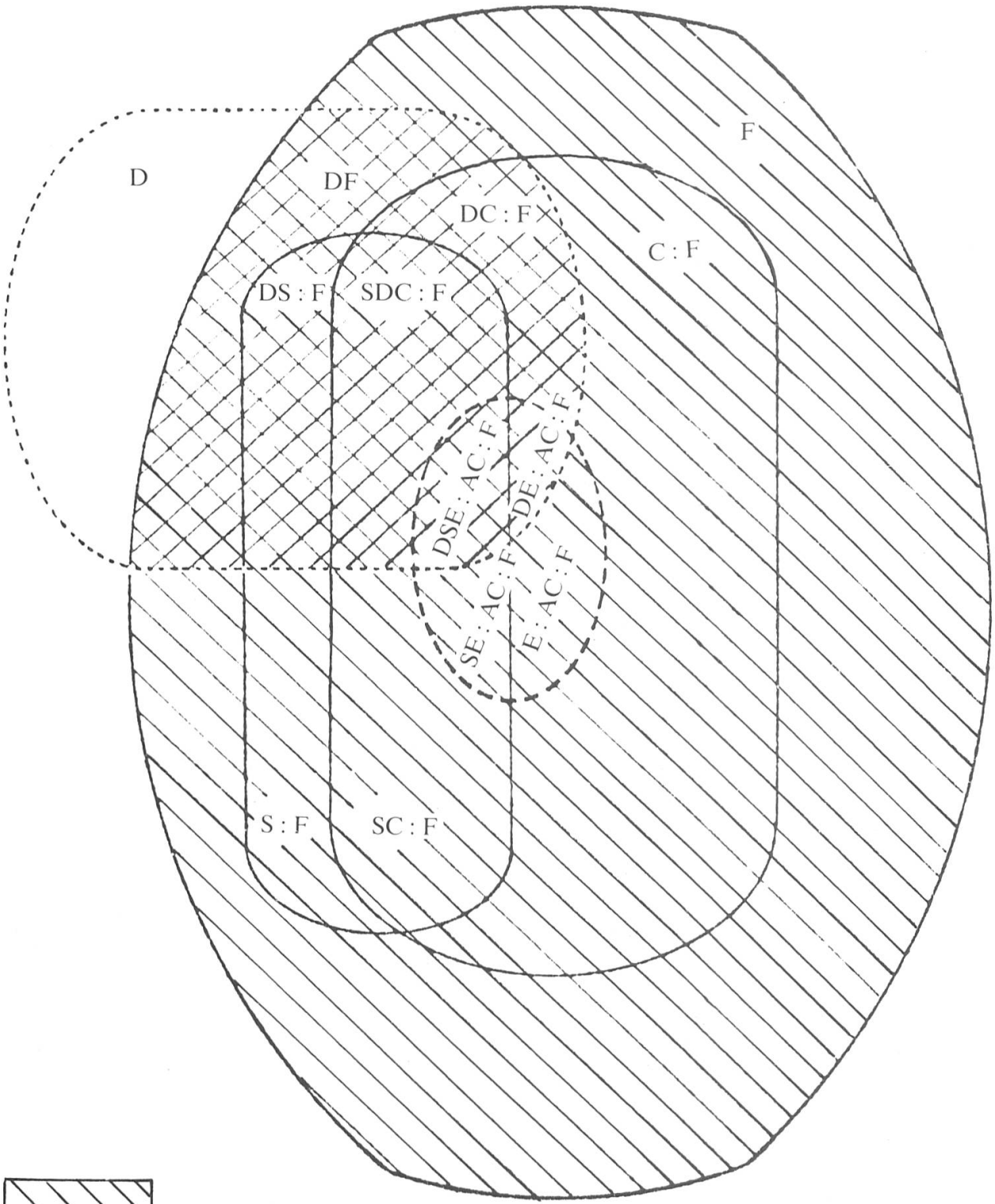
4. Les plans de mesure considérés et les formules de généralisabilité correspondantes

Déterminer, comme nous venons de le faire, les sources de variation principales qui affectent la réussite à une épreuve de mathématique, ne répond pas à toutes les questions que l'on peut se poser à propos de la mesure de cette performance. Par exemple, des différences minimes, comme celles qui apparaissent entre les quatre domaines étudiés, peuvent-elles être mesurées avec une précision suffisante si le nombre d'observations est suffisamment élevé? Le supplément d'information éventuellement nécessaire pour stabiliser l'estimation des différences entre les niveaux de D doit-il être obtenu par l'augmentation du nombre de séries ou plutôt du nombre d'élèves?
Pour traiter des problèmes de ce genre, il faut attribuer aux facettes des rôles qui n'apparaissaient pas dans les plans d'observation ni d'estimation. Certaines facettes doivent être distinguées des autres comme constituant l'objet de la mesure, ce qui implique que les autres facettes deviennent les instruments de cette mesure. Ainsi, dans sa finalité première, un survey vise à contrôler des apprentissages. Ces connaissances constituent l'objet à mesurer et les élèves deviennent les instruments de ce contrôle.
Selon le problème que l'on se pose, les mêmes composantes de variance peuvent contribuer soit à la variance vraie entre objets de la mesure, soit à la variance erreur. Par exemple, les variations entre élèves, intéressantes pour les tests habituels, sont une source d'erreur à réduire quand on planifie un survey.
Un plan de mesure définit le rôle de chaque facette dans la mesure. Il répartit d'abord les facettes en deux groupes, selon qu'elles appartiennent à l'objet d'étude (face de différenciation, ou aux instruments de mesure (face d'instrumentation). Il précise ensuite si elles sont sources de fluctuations aléatoires ou non, d'après leur mode d'échantillonnage, défini précédemment.
Tout plan de mesure peut être décrit de la façon suivante:
M ($D^R/D^F/I^F/I^R$) où M signale qu'il s'agit d'un plan de mesure, D^R symbolise la liste des facettes de différenciation aléatoires, D^F la liste des facettes de différenciation fixées, et où similairement I^F et I^R situent les facettes d'instrumentation fixées et aléatoires.

4. 1 La comparaison entre les domaines

Nous allons décrire d'abord les procédures de calcul applicables lorsque la différenciation est recherchée entre domaines. Le but est alors d'estimer la marge d'erreur à prévoir lorsqu'on compare les taux de réussite des différents niveaux de cette facette D.

Figure 2: *Diagramme du plan de mesure M (−/D/A/E,C,F,S)*



régions correspondant à la variance erreur absolue

régions correspondant à la variance erreur relative

On veut tirer des conclusions qui soient stables, en l'occurence évaluer si les différences de réussite aux quatre objectifs se maintiennent pour d'autre échantillons aléatoires de formes, de séries, de classes et d'élèves. Les facettes F, S : F, C : F et E : AC : F sont donc les facettes de généralisation. L'influence de l'âge (facette A) par contre peut être laissée de côté. En effet, autant d'enfants d'âge inférieur à la moyenne que d'enfants plus âgés interviennent à tous les niveaux de D, comme aussi dans toutes les interactions avec D; les effets de l'âge sont donc annulés. Il reste cinq facettes actives dont les caractéristiques essentielles apparaissent à la figure 2. Ce diagramme diffère de la figure 1 par le fait que la facette A est absente et qu'il ne reste plus que 13 plages différentes. Ces plages correspondent pourtant à 13 composantes de variance calculées dans le tableau 2. Le plan de mesure peut être symbolisé par M (−/D/A/E,C,F,S).

La variance des erreurs relatives est estimée en calculant une somme pondérée de composantes de variance. Pour savoir quelles composantes de variance interviennent, on examine la figure 2. Toutes les interactions de la facette D, la face de différenciation, avec les facettes de généralisation F, S : F, C : F et E : AC : F sont situées dans l'intersection de la surface D et de la réunion des surfaces F, S : F, C : F et E : AC : F.

Chaque composante ainsi identifiée entre dans l'erreur relative avec un coefficient qui est l'inverse du «nombre effectif» de niveaux des facettes recouvrant la plage. On appelle «nombre effectif» celui sur lequel est calculée la moyenne étudiée. Par exemple, la variance σ^2 (DF) est divisée par le nombre de formes utilisées dans l'expérience, à savoir 5. De même, la plage DSC : F est contenue dans la surface D, F, S : F et C : F; la variance correspondante σ^2 (dsc : f) doit donc être divisée par n_d, n_s, n_c, n_f, mais $n_d = 1$, puisqu'on étudie l'erreur sur la moyenne d'un objectif. En appliquant ces règles, on trouve la valeur de la variance des erreurs relatives pour la moyenne d'un objectif:

$$\begin{aligned}\sigma^2(\delta_d) &= \sigma^2(df)/n_f + \sigma^2(ds:f)/n_s n_f + \sigma^2(dc:f)/n_c n_f \\ &\quad + \sigma^2(dsc:f)/n_s n_c n_f + \sigma^2(de:ac:f)/n_e n_a n_c n_f \\ &\quad + \sigma^2(dse:ac:f)/n_s n_e n_a n_c n_f\end{aligned}$$

Exprimée en 10^{-5}, cette valeur est de:

$$\sigma^2(\delta_d) = 16{,}600 + 77{,}600 + 21{,}500 + 18{,}166 + 16{,}925 + 26{,}250 = 177$$

Nous obtenons ainsi la variance erreur relative sur le taux moyen de réussite pour un domaine. A partir de cette valeur, on peut calculer la «marge d'erreur» sur les domaines. On calcule pour cela la variance erreur sur la différence de deux moyennes (ici de deux taux de réussite). La marge d'erreur sur D est donc définie par l'écart-type de la différence de deux moyennes de D calculées grâce au dispositif.

La marge d'erreur relative sur D, qui est designée par ErR (D), vaut donc $\sqrt{2\sigma^2(\delta_d)}$, soit 0,0595. Nous admettrons qu'une marge de 0,05 représente le maximum acceptable, maximum qui serait donc ici dépassé. Une marge d'erreur de 0,05, en effet, signifie que la différence entre deux scores devrait être supérieure à 0,10 (approximativement) pour qu'on puisse conclure, avec seulement 5% de chance de se tromper, que les deux scores représentent des niveaux de réussite réellement différents.

La variance erreur absolue inclut toutes les composantes ci-dessus, plus les composantes qui sont spécifiques aux facettes de généralisation. Ces composantes spécifiques se trouvent à la figure 2, à l'extérieur de la surface occupée par D. Chaque composante de variance est divisée par le produit du nombre de niveaux observés de facettes de son indice total (à condition de remplacer par 1 le nombre de niveaux des facettes de différenciation).

$$\sigma^2(\Delta_d) = \sigma^2(\delta_d) + \sigma^2(f)/n_f + \sigma^2(e:ac:f)/n_e n_a n_c n_f + \sigma^2(se:ac:f)/n_s n_e n_a n_c n_f + \sigma^2(s:f)/n_s n_f + \sigma^2(sc:f)/n_s n_c n_f + \sigma^2(c:f)/n_c n_f$$

Exprimée en 10^{-5} la variance des erreurs absolues est de:

$$\sigma^2(\Delta_d) = 177 + 0 + 16{,}725 + 6{,}012 + 17{,}266 + 1{,}533 + 9{,}050 = 228$$

La marge d'erreur absolue ErX(d), vaut 0,0675.

En examinant ces résultats, quelques conclusions méritent d'être formulées:

1. Les estimations de marges d'erreur indiquent que le dispositif utilisé fournit une précision à la limite de l'acceptabilité pour les comparaisons entre objectifs, qu'elles soient relatives ou absolues. Il faut que la différence entre les taux de réussite de 2 objectifs dépasse 0,1166 (soit 1,96 × ErR (d)) pour qu'on conclue avec moins de 5 chances sur 100 de se tromper que les taux relatifs de maîtrise de ces 2 objectifs sont différents. On tirera une conclusion similaire en ce qui concerne leurs maîtrises absolues si la différence observée dépasse 0,1323 (soit 1,96 · ErX (d)).
2. Les mesures définitives ne sont pas prises à partir de l'étude de généralisabilité, mais bien à partir de l'étude de décision dont on parlera au point suivant. En fait, le but principal de l'étude G est d'estimer la grandeur des diverses composantes de variance. Elle sert en quelque sorte d'enquête-pilote pour déterminer le nombre d'observations qui seront nécessaires pour atteindre la précision souhaitée. Il n'y a donc pas lieu de s'inquiéter si l'erreur paraît trop grande dans l'étude G. Des résultats que nous venons d'obtenir, on peut conclure qu'il faudra davantage d'observations dans l'étude D mais qu'on s'approche déjà de l'ordre de grandeur nécessaire si l'on veut pouvoir interpréter une différence avec une marge d'erreur de .05.
3. Il est intéressant d'exprimer la précision de la mesure en utilisant le coefficient de généralisabilité qui est le rapport de la variance vraie des objectifs d'étude à la variance observée de ces mêmes objets d'étude.

Le calcul de ces rapports pour le score relatif, puis pour le score absolu, donnent deux coefficients de généralisabilité:

$$\hat{\rho}^2(\delta_d) = \sigma^2(d)/(\sigma^2(d) + \sigma^2(\delta_d)) = 651/(651 + 177) = 0{,}786$$

$$\hat{\rho}^2(\Delta_d) = \sigma^2(d)/(\sigma^2(d) + \sigma^2(\Delta_d) = 651/(651 + 228) = 0{,}741$$

Les valeurs obtenues confirment les indications tirées du calcul des marges d'erreur: la précision peut être améliorée, par exemple en multipliant le nombre de classes échantillonnées pour le survey.

4. 2 La comparaison entre les élèves

Le but de l'analyse à ce niveau est d'estimer la marge d'erreur qui existe lorsque l'on compare les scores de deux élèves.

La comparaison peut concerner des élèves qui appartiennent à des populations différentes. Par exemple, on peut différencier tous les élèves quels que soient leur âge, leur classe scolaire, ou la forme du test qu'ils ont reçue. Dans ce cas, la face de différenciation est composée des facettes E : AC : F, C : F, A et F alors que la facette S est facette de généralisation. Deux plans de mesure différents sont à considérer, selon que l'on compare les résultats pour la moyenne des quatre domaines, ou pour chaque domaine séparément.
La comparaison peut au contraire concerner un sous-ensemble de la population d'élèves: à partir des données fournies par le plan d'observation un maître désire comparer entre eux les élèves d'une même classe et de même âge qui ont reçu les mêmes séries. Dans ce troisième plan de mesure, la face de différenciation est constituée par la seule facette E : AC : F et, à condition que l'on considère D comme facette de contrôle, la seule facette de généralisation est encore S.
Ces trois exemples sont, de toute évidence, bien distincts. Nous allons décrire les procédures de calcul pour chaque cas.

4.2.1 Différenciation entre les scores moyens de tous les élèves

Rappelons d'abord comment les facettes du plan d'observation se répartissent lorsque l'on choisit comme objet d'étude la note moyenne obtenue par chaque élève du groupe total.
La mesure utilisée est bien ici la moyenne par élève. Comme cette moyenne est calculée à partir des notes obtenues à trois séries, chacune d'elles contrôlant la maîtrise des mêmes quatre domaines fixés, nous souhaitons généraliser la valeur de cette moyenne à tous les échantillons aléatoires de séries analogues, susceptibles d'être distribués aux élèves. La facette de généralisation est S tandis que D constitue une facette de contrôle.
Les élèves à comparer appartiennent à des groupes d'âges (A) et à des classes (C) différents. Les facettes E : AC : F, A et C : F font donc partie de la face de différenciation, de même que la facette F, puisque les élèves et les classes sont nichés dans les formes. On peut schématiser le plan de mesure de la manière suivante: M (E, C, F/A/D/S).
Les marges d'erreur qui apparaissent à la première rangée du tableau 3 indiquent que le dispositif utilisé n'assure pas une précision suffisante pour comparer les moyennes des élèves pris individuellement. La différence observée entre les scores de deux élèves devrait atteindre 0,22 approximativement (près du quart de la marge totale de variation) pour que l'on puisse conclure à une différence réelle entre leurs niveaux absolus de réussite en mathématique! Néanmoins, comme nous le précisions plus haut, l'étude de décision sert à accroître la précision de la mésure: dans le cas présent, puisqu'on généralise sur la seule facette S, il suffirait d'augmenter le nombre de séries par élève pour diminuer la marge d'erreur et atteindre la précision voulue.
Si l'on désirait optimiser le plan de mesure pour une différenciation de tous les élèves les uns par rapport aux autres, il faudrait aller plus loin et rectifier le défaut évident du plan que l'on vient d'examiner, qui est de présenter aux diverses classes des épreuves différentes. Il faudrait croiser les formes et les séries avec les élèves. Le fait d'avoir situé dans notre exemple la variance de F sur la face de différenciation et d'avoir en même temps considéré la variance de S comme de la variance erreur semble effectivement illogique, mais ne fait que traduire un défaut du plan expérimental lui-même qui confond «facilité de la forme» et «capacité de l'élève». Les formules de généralisabilité permettent au moins d'en apprécier la gravité: comme la variance de F est nulle, on peut penser que le défaut du dispositif est sans conséquence pratique en ce qui concerne la précision de la mesure.

Tableau 3: Différenciation des élèves selon la population de référence

Population de référence	Différenciation considérée	Plan de mesure	Variance d'erreur ($x10^{-5}$)		Marge d'erreur		Variance de différenciation ($x10^{-5}$)	Coefficient de généralisabilité des mesures	
			relative	absolue	relative	absolue		relatives	absolues
le groupe total	les moyennes des élèves calculées sur l'ensemble des objectifs	M (E, C, F/A/D/S)	598	598	0,1094	0,1094	1822	–	0,75
le groupe total	les moyennes des élèves calculées pour chaque objectif	M (E, C, F/A, D/–/S)	3881	3881	0,2786	0,2786	4395	–	0,53
les élèves d'une classe ayant le même âge	les moyennes des élèves calculées sur l'ensemble des objectifs	M(E/–/D/S)	481	598	0,0981	0,1094	1338	0,74	0,69

4.2.2 Différenciation entre les scores de tous les élèves pour les différents domaines

Il est normal en pédagogie que l'on s'intéresse à la maîtrise des objectifs par les élèves qui ont suivi un certain curriculum.
Dans le cas précédent, l'objet d'étude était constitué par la performance moyenne des élèves, obtenue sur l'ensemble des quatre domaines. Dans le cas présent, on vise à estimer la marge d'erreur pour des comparaisons entre scores obtenus par chaque élève aux quatre objectifs ou domaines.
Par rapport au cas précédent, d'importantes modifications sont apportées à la composition des variances d'erreur relative ou absolue. Toutes les composantes liées à D deviennent variance active et vont être réparties soit dans la variance de différenciation, soit dans la variance erreur.
On s'aidera de la figure 1 pour repérer les composantes d'erreur: elles figurent toutes dans l'ovale de la facette S.
La variance d'erreur absolue est confondue avec la variance d'erreur relative, puisque la facette de généralisation S est nichée dans la face de différenciation.

$$\sigma^2(\delta_{e,\,c,\,f,\,a,\,d}) = \sigma^2(\Delta_{e,\,c,\,f,\,a,\,d})$$
$$= 1/n_s \{\sigma^2(s:f) + \sigma^2(sc:f) + \sigma^2(as:f) + \sigma^2(sac:f) + \sigma^2(se:ac:f)$$
$$+ \sigma^2(dse:ac:f) + \sigma^2(dsac:f) + \sigma^2(dsa:f) + \sigma^2(ds:f) + \sigma^2(dsc:f)\}$$

On obtient la valeur suivante pour la variance erreur:

$$= 1/3 \cdot (259 + 92 + 0 + 0 + 1443 + 6300 + 1295 + 0 + 1164 + 1090)$$
$$= 3881 \cdot (10^{-5})$$

La variance erreur est importante, de même que les marges d'erreur relative ou absolue (deuxième rangée du tableau 3). Tel quel, le dispositif utilisé est totalement inadéquat en ce qui concerne la comparaison d'élèves en fonction de leur maîtrise des objectifs. La différence entre ces taux de maîtrise de deux domaines devrait dépasser .56 approximativement (plus de la moitié cette fois de la marge de variation possible) pour que l'on conclue à une différence réelle entre les niveaux de réussite.
En examinant les composantes qui entrent dans les formules de la variance erreur, on s'aperçoit que l'interaction Elèves × Questions (DSE : AC : F puisqu'il n'y a qu' une question par objectif et par série) est la source de variance erreur la plus importante. Pour accroître la précision de la différenciation des taux de réussite par élève et par domaine, il serait nécessaire de modifier le dispositif en introduisant un ou plusieurs des changements suivants:

1) augmenter le nombre de séries proposées à chaque élève (on réduit du même coup toutes les composantes d'erreur)
2) augmenter le nombre de questions qui dans chaque série évaluent la maîtrise d'un domaine
3) augmenter l'homogénéité des niveaux de difficulté d'une série à l'autre.

On peut estimer la fidélité du dispositif actuel. La variance de différenciation s'obtient en totalisant les composantes de variance pour toutes les plages à l'extérieur de S: on trouve $4395 \cdot 10^{-5}$. La variance erreur étant de $3881 \cdot 10^{-5}$, la fidélité de mesures absolues est de 0,53, ce qui est manifestement insuffisant.

4.2.3 Différenciation entre les élèves de même âge à l'intérieur d'une classe

Le plan de mesure que nous allons traiter maintenant ne touche qu'une partie du plan d'observation initial. Il s'agit du plan M (E/–/D/S).
On veut en effet différencier au sein de chaque classe les élèves de même âge qui ont reçu les mêmes séries. Comme ces élèves appartiennent à la même classe, la facette classe n'est plus pertinente, ni la facette F, puisque les élèves que l'on compare ont reçu aussi la même forme. Pour le même motif (comparaison au sein d'un même âge), la facette A doit être aussi supprimée. Le plan de mesure ne mentionnera donc plus les facettes A, C et F sur la face de différenciation. Comme la comparaison se fait à partir de la moyenne des domaines, la facette D est facette de contrôle, source de variance passive. Seule la facette S est facette de généralisation.
La variance de différenciation ne comprend que σ^2 (e : ac : f). Les composantes de la variance de généralisation sont toutes situées à l'intérieur de la facette de généralisation S. Pour la variance d'erreur relative la situation est simple:

$$\sigma^2 (\delta_{e:ac:f}) = 1/n_s \quad \sigma^2 (se : ac : f) = 1/3 \cdot 1443 = 481\ (10^{-5})$$

Pour la variance d'erreur absolue, on est obligé de tenir compte de tout ce qui modifie la difficulté des séries pour les élèves, c'est-à-dire la variance de S et de l'interaction de S avec A, avec C et avec AC.

$$\begin{aligned}\sigma^2 (\Delta_{e:ac:f}) &= 1/n_s \{\sigma^2 (se : ac : f) + \sigma^2 (s : f) + \sigma^2 (as : f) \\ &\quad + \sigma^2 (sc : f) + \sigma^2 (sac : f)\} \\ &= 1/3 \cdot (1443 + 259 + 0 + 92 + 0) \\ &= 598(10^{-5})\end{aligned}$$

La marge d'erreur relative (troisième rangée du tableau 3) est la plus faible parmi celles que nous avons observées pour la comparaison des performances moyennes par élève. On peut encore la réduire en augmentant le nombre de séries par élèves ou en diminuant l'hétérogénéité des séries.
On voit cependant que, pour différencier les moyennes d'élèves, la marge d'erreur n'est pas très différente si l'on applique la même forme (erreur relative) ou si l'on applique des formes différentes selon les élèves (erreur absolue). Ceci confirme la conclusion du premier cas ci-dessus.
La marge d'erreur absolue est aussi la même à l'intérieur d'une classe, ou pour l'ensemble de tous les élèves. La fidélité de la mesure, par contre, est différente, puisque la variance de différenciation n'est pas la même. Dans une classe et un groupe d'âge, elle est égale à σ^2 (e : ac : f), c'est-à-dire $1338 \cdot 10^{-5}$. Dans l'ensemble de tous les élèves, elle est la somme des composantes de variance liées aux quatre facettes E, C, A et F et à leurs interactions:

$$\begin{aligned}\sigma^2 (\tau) &= \sigma^2 (e : ac : f) + \sigma^2 (c : f) + \sigma^2 (a) + \sigma^2 (f) + \sigma^2 (ac : f) + \sigma^2 (af) \\ &= 1338 + 181 + 0 + 0 + 294 + 9 = 1822\end{aligned}$$

La fidélité des différenciations à l'aide de formes différentes, ou au contraire d'une forme unique à l'intérieur de la classe, est respectivement de $1338 / (1338 + 598) = 0{,}69$

et $1338 / (1338 + 481) = 0{,}74$

La fidélité des différenciations effectuées à l'aide de formes différentes à l'intérieur de la population totale est de: 1822 / (1822 + 598) = 0,75. Dans aucun cas, on n'atteint une fidélité suffisante. Il faudrait donc se garder d'utiliser une des formes du survey pour évaluer un élève.

4.3 Autres comparaisons envisagées

Il serait possible, (et nous l'avons fait dans un document de travail qui peut être fourni sur demande, Tourneur et Cardinet, 1980), de calculer des marges d'erreurs pour d'autres comparaisons: entre classes, entre groupes d'âges, entre formes et entre séries. Ces comparaisons peuvent porter sur la moyenne des quatre domaines, ou sur chaque domaine séparément. En faisant varier encore le mode d'échantillonnage (lorsque cela un sens dans la réalité), on voit qu'on arrive à un très grand nombre de possibilités, qu'il serait fastidieux d'énumérer, et surtout de traiter de façon exhaustive.
Pour ne pas allonger ce texte, nous n'aborderons pas non plus les adaptations possibles du plan d'observation, qui formeraient la suite logique d'une étude de généralisabilité. Nous renvoyons plutôt les lecteurs intéressés au document de travail sus-mentionné.

5. La portée de cet exemple

Ce n'était pas le but des pages précédentes de fournir des plans de mesure utilisables pour d'autres surveys et encore moins de présenter des résultats particuliers concernant le survey de 1976. Notre intérêt était purement méthodologique. Certaines conclusions de portée plus générale se dégagent cependant de cet exemple.
On ne peut manquer d'être étonné du nombre énorme de possibilités d'analyses qui apparaissent lorsqu'on examine toutes les combinaisons possibles de facettes de différenciation et d'instrumentation, ainsi que toutes les possibilités de transformation des plans d'observation et d'estimation qui précèdent le choix des plans de mesure. Les cas qui ont été présentés n'épuisent pas, et de loin, tous les plans de mesure et d'optimisation concevables.
Cette richesse insoupçonnée suggère plusieurs directions de recherche. On pourrait envisager d'abord des démarches plus systématiques pour la recherche des plans d'optimisation. Nous avons pu, en particulier, utiliser le calcul différentiel pour déterminer des dispositifs qui minimisent certaines marges d'erreurs. L'étude a priori des plans de mesure, sur la base des formules de la théorie de la généralisabilité, paraît également prometteuse.
On se rend compte, d'autre part, de la complexité réelle sous-jacente à de nombreuses expériences pédagogiques entreprises sans analyse préalable des plans de mesure. L'explicitation des facettes cachées, des effets d'interaction et des variances confondues, entre autres, aiderait certainement à l'interprétation ou à la critique des résultats observés dans ces expériences.
On s'aperçoit enfin que les dispositifs classiques proposés dans les manuels de statistique ne représentent qu'une petite partie de tous les dispositifs utilisables. Les plans expérimentaux proposés ont dû être choisis essentiellement en fonction de la possibilité de tester les effets principaux par des tests de F valides; ils ne tiennent pas compte, par contre, de considérations métrologiques qui pourraient amener à préférer d'autres dispositifs. La technique des diagrammes employée dans cet article permet de maîtriser des plans de mesure déjà relativement complexes et des formules générales sont maintenant à disposition pour l'ensemble des plans équilibrés. On peut donc s'attendre à un enrichissement à l'avenir du nombre des dispositifs expérimentaux utilisables pour la recherche pédagogique.

Untersuchung zur Generalisierbarkeit einer Erhebung

Der Artikel veranschaulicht, auf welche Weise die Zuverlässigkeit der verschiedenen Messungen eingeschätzt werden kann, die sich aus einer Erhebung ergeben. Er bedient sich einer Va-

rianzanalyse, die im ersten Primarschuljahr an Mathematiktests durchgeführt wurde. Nach der Beschreibung des Beobachtungsplans wird die Art und Weise beschrieben, wie die Quadratsummen und die Varianzkomponenten errechnet wurden. Er schildert sodann die aufeinanderfolgenden Etappen bei der Untersuchung der Generalisierbarkeit, zuerst im Hinblick auf die Differenzierung der durchschnittlichen Resultate im mathematischen Bereich, sodann im Hinblick auf die Differenzierung der Schülerresultate.

The generalizabilitiy study of a survey

This paper shows how the reliability of the various measures derived from a survey can be estimated. It takes as a basis an analysis of variance applied to tests of mathematics given at the end of the first year of schooling. The observation design is described; the procedure to compute sums of squares and estimates of the components of variance is presented. The successive steps of the generalizability study are then explained, concerning on the one hand the differentiation of achievements for the various domains of content and on the other hand the differentiation of the results of the pupils.

BIBLIOGRAPHIE

Cardinet J., Tourneur Y. & Allal L.: The symmetry of generalizability theory: applications to educational measurement. Journal of Educational Measurement, 1976, 13, p. 119-135.
Cardinet J.: L'évaluation du nouvel enseignement de la mathématique en première année primaire: présentation de l'expérience romande. Neuchâtel, Institut Romand de Recherches et de Documentation Pédagogiques, IRDP/R, 77.17, 1977.
Cardinet J., Tourneur Y. & Allal L.: Extension of generalizability theory and its applications in educational measurement. Neuchâtel, Institut Romand de Recherches et de Documentation Pédagogiques, IRDP/ 79.06, 1979.
Cardinet J., Tourneur Y.: Analyse de variance et théorie de la généralisabilité: Guide pour la réalisation des calculs. Neuchâtel, Institut Romand de Recherches et de Documentation Pédagogiques, IRDP/R 79.10, 1979.
Cronbach L., Gleser G., Nanda J. & Rajaratnam N.: The dependability of behavioral measurement: theory of generalizability for scores and profiles. New York, Wiley, 1972.
Spearman C.: Correlations calculated from faulty data. British Journal of Psychology, 1910, 3, 271-295.
Tourneur Y., Cardinet J.: Une étude de généralisabilité pour la planification d'un survey. Mons, Université de l'Etat, Document SEMME, no. 800.525/CT/6, mai 1980.