Revue des sciences de l'éducation

La fiabilité des données d'un instrument d'observation des enseignants en classe de mathématiques

Richard Bertrand and Mariel Leclerc

Volume 10, Number 2, 1984

URI: https://id.erudit.org/iderudit/900454ar
DOI: https://doi.org/10.7202/900454ar

See table of contents

Publisher(s)

Revue des sciences de l'éducation

ISSN

0318-479X (print)
1705-0065 (digital)

Explore this journal

Cite this article

Bertrand, R. & Leclerc, M. (1984). La fiabilité des données d'un instrument d'observation des enseignants en classe de mathématiques. *Revue des sciences de l'éducation*, *10*(2), 311–329. https://doi.org/10.7202/900454ar

Article abstract

This paper describes the reliability of data obtained from "Five Minutes of Interaction", an instrument for observation. Ten separate observations were made in 30 mathematics classes at the secondary level. Following the application of generalizability theory, the study shows that for almost all the observations, the actions related to teaching could not be considered as reliable. However, for the majority, the actions related to questionning, to responding, and to retroaction in class have a high level of reliability. The paper concludes with a statement of the importance of this type of study for all those whose objective is to identify correlations between teaching and school achievement.

Revue des sciences de l'éducation, vol. X, no 2, 1984

La fiabilité des données d'un instrument d'observation des enseignants en classe de mathématiques

Richard Bertrand et Mariel Leclerc*

Résumé — Cet article traite de la fiabilité des données d'un instrument d'observation de la classe, appelé « Cinq minutes d'interaction ». Trente classes de mathématiques du niveau secondaire ont été observées à dix occasions distinctes. À l'aide de la théorie de la généralisabilité, l'étude montre que, pour la presque totalité, les pratiques d'enseignement liées à l'instruction n'ont pu être considérées comme fiables. Par contre, pour la majorité, les pratiques reliées aux questions, aux réponses et aux rétroactions en classe ont obtenu un haut niveau de fiabilité. L'article conclut sur l'importance d'une telle étude pour toute recherche visant à identifier des corrélations entre des pratiques d'enseignement et le rendement scolaire.

Abstract — This paper describes the reliability of data obtained from "Five Minutes of Interaction", an instrument for observation. Ten separate observations were made in 30 mathematics classes at the secondary level. Following the application of generalizability theory, the study shows that for almost all the observations, the actions related to teaching could not be considered as reliable. However, for the majority, the actions related to questionning, to responding, and to retroaction in class have a high level of reliability. The paper concludes with a statement of the importance of this type of study for all those whose objective is to identify correlations between teaching and school achievement.

Resumen — Este artículo trata sobre la fiabilidad de los datos de un instrumento de observación de la clase, llamado "Cinco minutos de interacción". Treinta clases de matemáticas del nivel secundario fueron observadas en diez ocasiones diferentes. Por intermedio de la teoría de la generalización, el estudio muestra que, para casi la totalidad, las prácticas de enseñanza ligadas a la instrucción no pudieron ser consideradas como fiables. En cambio, para la mayoría, las prácticas relacionadas con las preguntas, con las respuestas y con las retroacciones en clases han obtenido un gran nivel de fiabilidad. El artículo concluye con la importancia de un tal estudio en toda investigación que trata de identificar las correlaciones entre las prácticas de enseñanza y el rendimiento escolar.

Zusammenfassung — Dieser Artikel behandelt die Zuverlässigkeit der Daten eines Beobachtungsinstrumentes für die Klasse, genannt "Fünf Minuten Interaktion". Dreissig Mathematik-Klassen auf dem Niveau der höheren Schule wurden bei zehn verschiedenenen Gelegenheiten beobachtet. Mit Hilfe der Theorie von der Generalisationsmöglichkeit zeigt die Untersuchung, dass die Unterrichtspraktiken, die mit der Unterweisung zu tun haben, beinahe auf der ganzen Linie als nicht zuverlässig zu betrachten waren. Dagegen haben in der Mehrheit die Praktiken, die sich auf das Abfragen, die Antworten und die Reaktion auf die Klasse beziehen, einen hohen Grad von Zuverlässigkeit erzielt. Der Artikel schliesst mit dem Hinweis auf die Wichtigkeit einer solchen Untersuchung für jede Forschungsarbeit, die Korrelationen zwischen Unterrichtspraktiken und schulischer Leistung identifizieren will.

* Bertrand, Richard: professeur, INRS-Éducation.
Leclerc, Mariel: professeur, INRS-Éducation.

La fiabilité des données[1] provenant d'un instrument d'observation de la classe est une question propre à la recherche en enseignement. Ebel (1951) et Scott (1955) sont parmi les premiers, semble-t-il, à avoir traité cette question. Ils parlaient d'entente entre les observateurs. Cependant il est devenu d'usage de distinguer l'entente entre les observateurs de la fiabilité des données d'observation en tant que telle (Medley et Mitzel, 1958), même si on a encore tendance à confondre ces deux notions (Frick et Semmel, 1978; Rowley, 1983). Comme McGraw *et autres* (1972) l'ont indiqué, il est possible que les données d'observation ne soient pas fiables même si les observateurs s'entendent entre eux. Ainsi l'entente est insuffisante pour décréter qu'il y a fiabilité (Rowley, 1976; Mitchell, 1979).

Le manque d'entente entre les observateurs est bien sûr la première source d'erreur qu'il faut contrôler, mais ce n'est pas la plus importante (McGraw *et autres*, 1972). La source d'erreur la plus importante provient des variations des pratiques d'enseignement à l'intérieur d'une même classe, d'une occasion à l'autre (Frick et Semmel, 1978; Berliner, 1980; Webb, 1982).

Il est possible de tenir compte simultanément de ces deux sources d'erreur, limitant la fiabilité des données d'observation, en utilisant la théorie de la généralisabilité (Cronbach *et autres*, 1972).

Nous ne sommes pas les premiers à suggérer l'utilisation de la théorie de la généralisabilité pour évaluer la fiabilité de données d'observation (Lomax, 1982; Dillashaw et Okey, 1983; Smith et Teeter, 1982). De plus, il semblerait qu'il soit impératif d'évaluer cette fiabilité surtout, par exemple, quand on tente de relier des pratiques d'enseignement au rendement scolaire et aux attitudes des élèves. Il ne saurait être question en effet que des variables provenant de données d'observation auxquelles on ne peut se fier soient reliées un tant soit peu à d'autres variables (Shrout et Fleiss, 1979). Ce serait d'ailleurs ce manque de fiabilité ou tout au moins ce manque d'examen de la fiabilité des données d'observation qui empêcherait les chercheurs de trouver des relations consistantes entre les pratiques d'enseignement et le rendement scolaire (Frick et Semmel, 1978 ; Erlich et Shavelson, 1978). Berliner (1980) va même jusqu'à prétendre que de telles recherches corrélationnelles resteront primitives tant et aussi longtemps qu'on ignorera le comment et le pourquoi des variations des pratiques d'enseignement. Nous pensons que Calkins *et autres* (1977) résument bien cette question en disant que « plus les scores attribués à une dimension observable ou à un comportement seront généralisables (fiables) aux observateurs et aux occasions, plus l'ordre déterminé entre les enseignants en fonction de cette dimension ou de ce comportement sera sûr et plus il sera possible de découvrir des relations entre le rendement de l'élève et cette dimension »[2].

Cet article fait état de l'étude de fiabilité que nous avons faite et qui portait sur les données d'observation de pratiques d'enseignement des mathématiques au niveau secondaire. Le but d'une telle étude est d'évaluer la fiabilité de ces pratiques d'enseignement. Cette étude fait partie d'une recherche plus vaste de l'IEA (*I*nternational

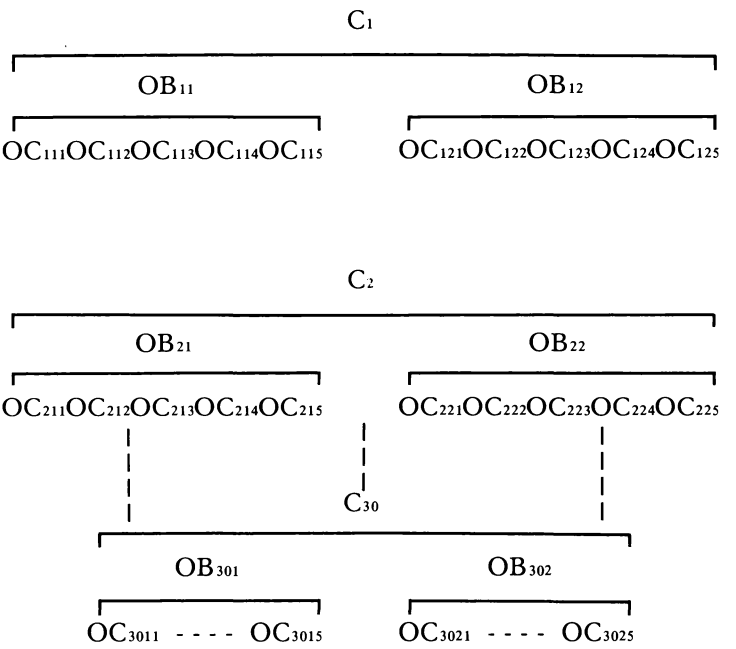
Association for the Evaluation of *E*ducational *A*chievement) dont l'objectif est d'identifier des corrélations entre des pratiques d'enseignement d'une part et, d'autre part, des effets de l'enseignement, tels le rendement scolaire et les attitudes des élèves[3]. Il comprend la présentation du dispositif d'observation choisi, la détermination du coefficient de généralisabilité, la description de la cueillette des données, les résultats de l'analyse des données et l'interprétation de ces résultats.

Le dispositif d'observation

Se basant sur les études citées ci-dessus et tenant compte d'un ensemble de contraintes, comme par exemple les coûts impliqués par le nombre des observateurs et des occasions d'observation, nous en sommes venus à adopter un dispositif qui a permis d'évaluer les variations entre les classes observées et cela, en regard de chacune des occasions d'observation.

Voici le schéma du dispositif d'observation qui a été choisi:

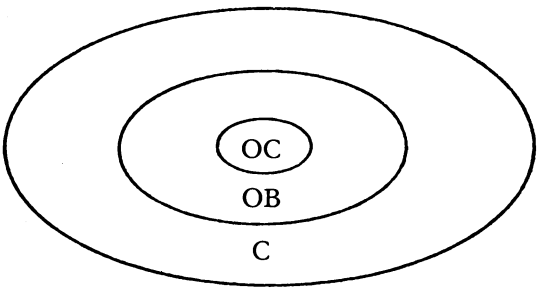
Schéma du dispositif d'observation



Chacune des 30 classes (C_1, C_2, ..., C_{30}) a été observée par deux observateurs distincts (par exemple OB_{11} et OB_{12} pour C_1), chacun observant à cinq occasions différentes (par exemple OC_{111} OC_{112} ... OC_{115} pour OB_{11}). Chaque élément de

variation du dispositif est appelé une facette. Ainsi nous distinguons trois facettes dans ce dispositif: la facette occasion (OC), la facette observateur (OB) et la facette classe (C).

Ce dispositif peut être visualisé en utilisant ce diagramme (Cronbach *et autres*, 1972, p. 37):



Comme nous le constatons, le dispositif est entièrement niché: la facette occasion (OC) est nichée sous la facette observateur (OB) elle-même nichée sous la facette classe (C). Une justification de ce dispositif s'impose. Il est fréquent en effet de rencontrer des études de généralisabilité comportant plusieurs facettes croisées, dans le but d'évaluer l'importance des interactions comme « classe x observateur », « classe x occasion ». Nous avons utilisé un dispositif niché plus simplement dès le point de départ, pour plusieurs raisons.

Tout d'abord nous ne possédions pas le temps (ni les fonds) requis pour évaluer un dispositif croisé plus complexe *avant* la cueillette des données. Même si cela semblait préférable, il n'était pas possible de procéder de cette façon. Ensuite, les considérations suivantes nous ont amenés à choisir le dispositif d'observation décrit ci-dessus.

Plaçons-nous dans la situation hypothétique où nous aurions opté pour un dispositif croisé. Considérant l'interaction « classé x occasion », comme chaque classe comporte un local distinct, il n'est pas possible, physiquement, d'observer plusieurs classes lors d'une même occasion. D'où l'impossibilité de croiser la facette « classe » avec la facette « occasion » et d'obtenir un estimé de l'interaction « classe x occasion ». Considérant ensuite les interactions « classe x observateur » et « occasion x observateur », pour pouvoir en tenir compte il aurait fallu que plusieurs observateurs soient présents dans la même classe en même temps. C'est là une procédure à la fois coûteuse et complexe. Or les données recueillies lors d'études analogues (Leclerc, Bertrand et Roberge-Brassard, 1979; Erlich et Borich, 1979) nous montrent que la variation qui résulte de ce type d'interactions est bien minime par rapport, en tout cas, à la variation causée par les occasions. Considérant enfin les problèmes d'erreurs

d'échantillonnage des estimés des composantes de variance, nous rappelons que ces problèmes peuvent devenir très aigus si on utilise un dispositif croisé avec seulement quelques niveaux par facette (Smith, 1978). Voilà, en gros, les raisons qui ont justifié le rejet d'un dispositif croisé.

Il est à noter que chaque facette de notre dispositif niché doit être perçue comme aléatoire simple, c'est-à-dire que les niveaux échantillonnés de chaque facette (les diverses occasions, observateurs ou classes) doivent être considérés comme provenant d'un ensemble infini de niveaux. En effet, les trente classes observées ne sont pas les seules qui nous intéressent. Il en est de même pour les observateurs et les occasions. Nous voulions généraliser les résultats de l'étude à plusieurs classes, observateurs ou occasions analogues.

Le coefficient de généralisabilité

Afin de déterminer le coefficient de généralisabilité, il convenait d'évaluer la part de variance qui est relative à chacune des trois facettes: classe, observateur et occasion. La variance qui provient des différences entre les classes est appelée variance de différenciation (Tourneur et Cardinet, 1979, p. 179). Pour que les données soient considérées comme fiables, il importe que la variance de différenciation soit aussi grande que possible. La variance d'erreur, celle que l'on veut minimiser, provient des différences entre les observateurs d'une part et des différences entre les occasions, d'autre part.

Ainsi le coefficient de généralisabilité qui nous indique jusqu'à quel point on a observé de façon fiable les différences entre les pratiques d'enseignement de chaque classe, quels que soient les observateurs et les occasions, se lit:

$$\hat{\rho}^2_C = \frac{\hat{\sigma}^2_C}{\hat{\sigma}^2_C + \left[\frac{1}{N_{OB}}\hat{\sigma}^2_{OB} + \frac{1}{N_{OB} \times N_{OC}}\hat{\sigma}^2_{OC}\right]} = \frac{\text{variance de différenciation}}{\text{variance de différenciation} + \text{variance d'erreur}}$$

où $\hat{\sigma}^2_C$, $\hat{\sigma}^2_{OB}$ et $\hat{\sigma}^2_{OC}$ sont les estimés des composantes de variance relative respectivement aux classes, aux observateurs et aux occasions, où N_{OB} est le nombre d'observateurs par classe, où N_{OC} indique le nombre d'occasions d'observations par observateur et par classe. Pour chaque classe, on a $N_{OB} = 2$ et $N_{OC} = 5$. Ainsi le nombre total d'occasions d'observations égale 10 (N_{OB} x N_{OC}).

Les estimés $\hat{\sigma}^2_C$, $\hat{\sigma}^2_{OB}$, $\hat{\sigma}^2_{OC}$ s'obtiennent de façon conventionnelle en effectuant une analyse de la variance à trois dimensions complètement nichées avec le modèle aléatoire simple. Une fois les carrés moyens (CM) connus pour chaque facette, on obtient les estimés des composantes de variance en utilisant les formules suivantes:

$$\hat{\sigma}^2_C = \frac{1}{N_{OB} \times N_{OC}} [CM_C - CM_{OB}]$$

$$\hat{\sigma}^2_{OB} = \frac{1}{N_{OC}} [CM_{OB} - CM_{OC}]$$

$$\hat{\sigma}^2_{OC} = CM_{OC}.$$

Le coefficient $\hat{\rho}^2_C$ varie théoriquement entre 0 et 1. On comprendra que plus la composante $\hat{\rho}^2_C$ est grande par rapport à $\frac{1}{N_{OB}}\ \hat{\sigma}^2_{OB}$ et $\frac{1}{N_{OB} \times N_{OC}}\ \hat{\sigma}^2_{OC}$, plus le coefficient est grand et plus on a observé de façon fiable ou généralisable les différences entre les pratiques d'enseignement de chaque classe.

Comme l'ont suggéré Erlich et Shavelson (1978) puis Webb (1982), nous dirons que le coefficient de généralisabilité est suffisamment grand, lorsque $\hat{\rho}^2_C \geq 0{,}70$. Nous parlerons alors de fiabilité effective.

La cueillette des données

Trente classes de mathématiques, où étaient enseignés les premiers éléments de l'algèbre, du niveau du premier cycle du secondaire, en première et en deuxième années, ont été observées à dix occasions différentes, chacune selon le dispositif d'observation décrit ci-dessus. Ces classes appartenaient à treize écoles différentes de quatre commissions scolaires de la région administrative de Québec.

Les observations ont été effectuées par dix observateurs qui s'étaient entraînés[4] à maîtriser un instrument d'observation appelé « Cinq minutes d'interaction » (CMI). Les catégories de cet instrument consistent en une série de comportements liés aux interactions entre l'enseignant et les élèves. L'observateur indique à chaque cinq secondes, par un code, le contexte d'enseignement (grand groupe, petit groupe, travail individuel supervisé par l'enseignant qui agit comme moniteur, transition, enseignement individuel), la personne qui parle, celle à qui elle s'adresse et ce dont il est question (le quoi). Cinq périodes d'observation de cinq minutes chacune constituent une occasion d'observation. On trouve au Tableau 1 la feuille résumant les codes utilisés par les observateurs et à la Figure 1 la feuille de codage.

Les résultats de l'étude

Au terme de la cueillette des données, 271 variables ont été retenues pour l'étude de fiabilité. C'est-à-dire que pour chacune de ces variables on a calculé un coefficient de généralisabilité $\hat{\rho}^2_C$. Ce qui a permis d'évaluer la fiabilité des données relatives à chaque variable. Le Tableau 2 fait état des résultats de l'étude de fiabilité pour les 77 variables de l'étude dites « élémentaires », soit celles qui ont été obtenues directement. Nous parlerons plus loin des variables composites, celles qui ont été obtenues par regroupements de variables élémentaires.

Tableau 1

Les cinq minutes d'interaction (CMI): résumé des codes

Contexte

1- G Le professeur enseigne à un grand groupe.
2- P Le professeur enseigne à un petit groupe.
3- M Le professeur circule en classe et contrôle le travail individuel (il est moniteur).
4- T Le professeur et le groupe sont en transition.
5- I Le professeur et un élève discutent de choses privées et individuelles.
6- N Le professeur n'est pas en relation avec les élèves.

Qui		À qui	
7- P	Professeur	14- P	Professeur
8- E	Élève (deuxième code)	15- G	Groupe
9- V	Volontaire (élève)	16- E	Élève
10- N	Non-volontaire (élève)	17- I	Nouvel individu (élève)
11- I	Initiateur (élève)		
12- X	Quelques-uns		
13- G	Le groupe ensemble		

Quoi académique

Instruction

18- Ma Cours magistral, explication
19- Ex Exemple
20- St Structuration
21- Ai Aide à répondre correctement
22- At Attire l'attention (test, devoir, correction, vérification)

Réponse

28- Re Réponse
29- Rc Récitation
30- In Remarque, énoncé, initiation
31- Pa Ne sait pas

Question

23- Qr Requête, ordre
24- Qc Explication, complexité
25- Qf Fait, rappel de mémoire
26- Qo Opinion
27- Qe Évaluation de l'efficacité de l'enseignement

Rétroaction

32- Ap Approuve, félicite (avec P)
33- Rp Répète
34- Rd Redirige
35- Dr Donne la réponse
36- F Dit que la réponse est fausse
37- Pu Punit

Non-académique

38- Di Discipline
39- Pr Procédure
40- So Social

Autre

41- Si Silence
42- Na Inaudible

Tableau 1 (suite)

Les cinq minutes d'interaction (CMI): résumé des codes

Qualifier	
43- Nv Non-verbal	Le code *R* est utilisé dans la marge pour indiquer que la séquence précédente se répète.
44- Em Emphase	
45- P Positif	
46- N Négatif	
47- M Matériel	

La première colonne du Tableau 2 donne la définition de la variable à l'aide des codes du Tableau 1. Ainsi la variable codée « GPERe » indique, en lisant le code de gauche à droite, que dans un contexte de grand groupe (G), le professeur (P) s'adresse à un élève (E) sous forme de réponse (Re). La variable codée GPGAt précise que dans un contexte de grand groupe (G), le professeur (P) s'adresse au groupe (G) et attire son attention (At).

Tableau 2

Étude de fiabilité des 77 variables dites élémentaires

Variable		Nombre pondéré de fréquences	Composantes de variances			Coefficient	Plan optimal	
Numéro	Définition		$\hat{\sigma}^2_C$	$\hat{\sigma}^2_{OB}$	$\hat{\sigma}^2_{OC}$	$\hat{\rho}^2_C$	N_{OB}	N_{OC}
4	GPERe	1456	11,22	0,10	16,13	0,87*		
6	GPGAt	507	0,58	3,80	6,94	0,18		
7	GPCAp, ApP	305	1,73	0,22	2,06	0,85*		
10	GPCQe	438	0,89	0,37	2,08	0,69	4 ou 3	2 3
11	GPGDr	416	0,00**	4,94	11,13	0,00		
14	GPGDi, DiN	374	3,80	0,00	6,72	0,85*		
15	GPEDi, DiN	277	0,85	0,21	2,37	0,72*		
16	GPIDi, DiN	344	0,41	0,05	2,42	0,60		
19	GPGEx	390	0,00	6,71	4,91	0,00		
21	GPGF	105	0,10	0,33	0,39	0,32		
24	GPGMa	6152	75,77	68,52	379,04	0,51		
25	GPGMaEm	408	0,00	2,45	5,10	0,00		
26	GPGMaM	11997	144,37	249,58	642,99	0,43		
27	GPGMaMNv	626	3,63	0,00	42,91	0,46		
39	GPEMaM	400	0,00	1,44	16,03	0,00		

Tableau 2 (suite)

Étude de fiabilité des 77 variables dites élémentaires

Variable Numéro	Définition	Nombre pondéré de fréquences	Composantes de variance $\hat{\sigma}^2_C$	$\hat{\sigma}^2_{OB}$	$\hat{\sigma}^2_{OC}$	Coefficient $\hat{\rho}^2_C$	Plan optimal N_{OB}	N_{OC}
40	GPGPr, PrNv	3012	10,61	11,75	111,90	0,38		
41	GPGQc	180	0,04	2,94	2,56	0,02		
42	GPEPr	349	0,00	1,10	6,07	0,00		
43	GPIPr	472	0,00	0,69	2,28	0,00		
48	GIPPr	225	0,00	0,69	2,28	0,00		
51	GPGQf	2974	27,08	10,05	50,54	0,73*		
52	GPEQf	840	4,68	2,02	15,80	0,64	5	2
53	GPIQf	1177	11,31	6,33	20,60	0,68	4	2
							ou 3	3
60	GPEMaAi	540	1,01	7,40	12,33	0,17		
61	GPGMaAi	358	0,28	1,01	18,12	0,11		
63	GPGQfAi	170	0,19	0,13	0,97	0,54		
64	GPGQr, QrAi	547	0,00	4,08	3,94	0,00		
65	GPEQr, QrAi	235	0,23	0,60	1,17	0,35		
66	GPIQr, QrAi	920	12,92	0,97	12,74	0,88*		
70	GPGRp	1247	5,18	4,36	11,19	0,61	5	2
71	GPERp	727	3,46	2,17	7,97	0,65	5	2
76	GPGSi	2456	3,11	59,12	89,75	0,07		
80	GXPSi	240	0,00	0,86	1,78	0,00		
83	GPGSo	499	2,24	4,24	21,94	0,34		
88	GPGSt	4496	11,05	17,87	50,15	0,44		
97	GEPRc	375	5,70	0,00	12,44	0,82*		
98	GEPRe	1773	24,89	6,41	36,74	0,78*		
99	GVPRe	530	1,81	2,98	5,59	0,47		
100	GNPRe	1621	21,07	2,10	26,18	0,85*		
101	GIPRe	253	0,77	0,40	1,30	0,70*		
102	GXPRe	3034	34,42	14,89	46,25	0,74*		
106	PPGMa	288	3.98	0,00	23,63	0,63		
108	PPEMa	325	0,00	4,54	24,03	0,00		
113	PPESi	203	0,00	3,08	18,64	0,00		
117	MPEPr	400	0,00	1,99	13,78	0,00		
121	MPGSi	1876	15,73	18,83	113,71	0,43		
124	MPISi	1286	38,27	15,36	33,15	0,78*		
126	MPEAp	338	1,02	0,00	4,50	0,69		
128	MPEMa	2207	19,53	49,75	153,66	0,33		

Tableau 2 (suite)

Étude de fiabilité des 77 variables dites élémentaires

Variable Numéro	Définition	Nombre pondéré de fréquences	Composantes de variance $\hat{\sigma}^2_C$	$\hat{\sigma}^2_{OB}$	$\hat{\sigma}^2_{OC}$	Coefficient $\hat{\rho}^2_C$	Plan optimal N_{OB}	N_{OC}
129	MPEMaAi	1487	0,00	91,00	117,76	0,00		
130	MPERe	1389	10,05	5,74	30,63	0,63	5	2
131	MPESt	363	0,00	1,95	4,40	0,00		
135	MPEQe	107	0,26	0,10	0,74	0,68	5	2
							ou 3	3
136	MPEQf	534	1,32	0,00	8,49	0,61		
137	MPEQr	265	0,36	0,37	2,22	0,47		
141	MEPIn	223	0,35	1,35	3,44	0,25		
142	MEPQf	802	4,34	6,37	18,58	0,46		
143	MEPQr	746	0,79	5,73	12,28	0,16		
144	MEPRe	739	2,37	0,45	14,11	0,59		
149	MPENa	376	8,55	14,37	26,19	0,47		
150	MEPNa	176	1,61	3,81	1,33	0,44		
152	TPGPr	367	2,19	2,87	13,07	0,44		
153	TPGSi	217	0,55	0,59	4,31	0,43		
156	TEPQr	176	0,00	1,71	3,33	0,00		
159	IPEAp	199	0,82	1,37	1,93	0,48		
160	IPEMa	1451	64,90	13,48	59,47	0,84*		
161	PIEQf	303	4,10	2,75	3,57	0,70*		
162	PIERe	847	9,02	13,65	16,81	0,51		
163	PIESt	177	0,97	1,91	2,77	0,44		
164	IEPIn	142	0,00	1,80	1,96	0,00		
165	IEPQf	649	9,13	7,72	13,90	0,63	4	2
166	IEPQr	426	5,57	5,30	8,18	0,62	4	2
167	IEPRe	365	5,12	2,99	4,69	0,72*		
175	IPENa	173	0,08	1,61	10,62	0,04		
176	IPEPr	184	0,06	0,74	5,22	0,06		
177	IPESi	234	0,20	4,79	5,26	0,06		
181	N	6709	113,70	131,09	613,09	0,47		

* Coefficient dépassant le seuil minimal de 0,70.

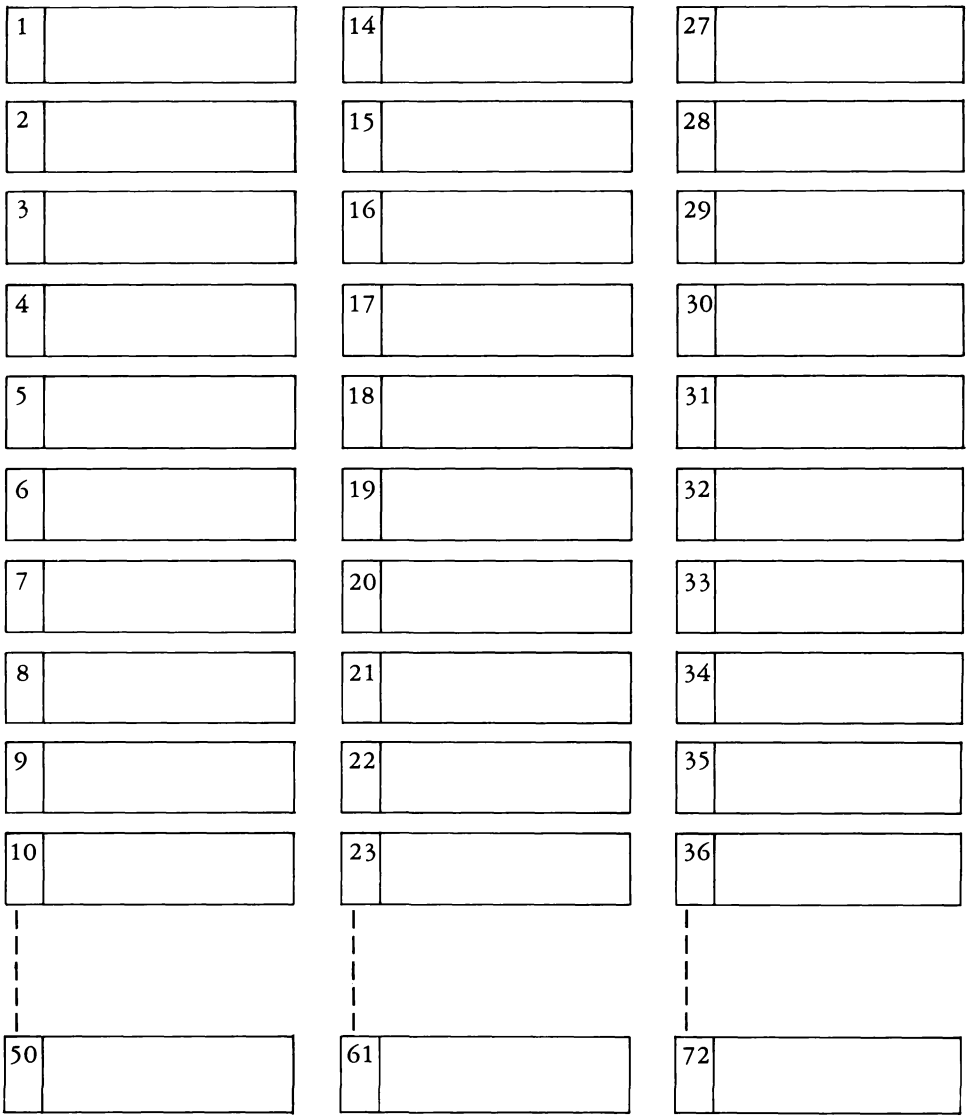
** Les composantes de variance négatives sont remplacées par «0,00».

La seconde colonne du Tableau 2 présente le nombre pondéré[5] de fréquences enregistrées pour chaque variable. Les trois autres colonnes donnent les estimés des composantes de variance utilisés dans le calcul du coefficient de généralisabilité que

l'on retrouve dans la sixième colonne. Les deux dernières colonnes font référence à ce que nous avons appelé le plan optimal. Lorsque le coefficient $\hat{\rho}^2_C$ n'atteint pas le seuil requis de 0,70 en utilisant le dispositif décrit plus haut ($N_{OB} = 2$, $N_{OC} = 5$), on

Figure 1

Feuille de codage



peut faire varier le nombre d'observateurs et le nombre d'occasions puis calculer un nouveau $\hat{\rho}^2_C$ avec $N_{OB} = 3$ et $N_{OC} = 3$, $N_{OB} = 4$ et $N_{OC} = 2$, ou encore $N_{OB} = 5$ et $N_{OC} = 2$. S'il arrive que ce nouveau $\hat{\rho}^2_C$ dépasse le seuil de 0,70 et que le produit du nombre d'observateurs et du nombre d'occasions reste en deçà de 10, le nombre initial d'occasions, alors on enregistre les N_{OB} et N_{OC} utilisés dans ce nouveau dispositif ou plan optimal. Ce nouveau dispositif pourra par la suite être employé lors d'une étude ultérieure. Le nouveau $\hat{\rho}^2_C$ simule en fait ce qu'on serait en droit de s'attendre si on utilisait le nouveau dispositif avec d'autres données dans des conditions analogues. Par exemple la variable 10, notée GPGQe, donne $\hat{\rho}^2_C = 0{,}69$. Mais si on avait utilisé $N_{OB} = 4$ observateurs et $N_{OC} = 2$ occasions par observateur ou alors $N_{OB} = 3$ observateurs et $N_{OC} = 3$ occasions par observateur (plutôt que le dispositif employé où $N_{OB} = 2$ et $N_{OC} = 5$), on aurait obtenu un $\hat{\rho}^2_C \geq 0{,}70$. Cependant, pour la variable 6 GPGAt, où $\hat{\rho}^2_C = 0{,}18$, aucun plan optimal n'existe tel que défini plus haut.

Discussion de la fiabilité des variables élémentaires

L'importance du nombre de fréquences

En faisant la somme des nombres pondérés des fréquences de toutes les variables du Tableau 2, on se rend compte que les variables du contexte « grand groupe » ont été observées dans l'ensemble deux fois plus fréquemment que les variables de tous les autres contextes réunis. On peut interpréter ce résultat comme suit : les 30 groupes-classes observés se trouvent en moyenne deux fois plus souvent dans un contexte de grand groupe que dans l'un ou l'autre des autres contextes. En fait, si on répartit le temps total d'observation en classe dans les six contextes sur la base des nombres pondérés de fréquences, on obtient :

(1)	Grand groupe	67%
(2)	Petit groupe	1%
(3)	Moniteur	17%
(4)	Transition	1%
(5)	Individuel	6%
(6)	Non impliqué	8%

On remarque aussi que les variables GPGMa et GPGMaM drainent à elles seules plus de 23% du total des fréquences d'observation. Ainsi le professeur, en contexte de grand groupe, s'adresse au groupe de façon magistrale avec ou sans matériel pour plus de 23% du temps total observé.

Même si certains auteurs remarquent que les variables de fréquences peu nombreuses sont plus sujettes à un manque de fiabilité (Berliner, 1980 ; Erlich et Shavelson, 1978), ce résultat s'éloigne de ce que nous avons obtenu. En effet, si on met en relation les colonnes titrées « nombre pondéré de fréquences » et « coefficient $\hat{\rho}^2_C$ » du Tableau 2, on s'aperçoit que les coefficients élevés ne correspondent pas[6]

nécessairement à un nombre pondéré de fréquences élevé. De même les coefficients faibles ne correspondent pas nécessairement à un nombre pondéré de fréquences faible. D'ailleurs Rowley (1976) prétend qu'il est bien possible d'obtenir des données fiables à partir de variables de fréquences peu nombreuses.

Les trois niveaux de fiabilité

Nous avons convenu de présenter trois niveaux de fiabilité. Le premier niveau est celui qui correspond au seuil généralement accepté dans le cas d'un coefficient tel que $\hat{\rho}^2_C$, soit 0,70. Nous identifions ce niveau comme étant celui de la fiabilité effective. Le deuxième niveau caractérise les variables dont le $\hat{\rho}^2_C$ est inférieur à 0,70 mais dont on a trouvé un plan optimal. Ce niveau réfère à ce que nous appelons la fiabilité potentielle. Le troisième niveau comprend les variables dont le $\hat{\rho}^2_C$ est inférieur à 0,70 et dont nous n'avons pas de plan optimal. On parle alors de non-fiabilité. Ainsi nous obtenons:

Niveau 1 (fiabilité effective): $\hat{\rho}^2_C \geq 0{,}70$
Niveau 2 (fiabilité potentielle): $\hat{\rho}^2_C < 0{,}70$, avec plan optimal
Niveau 3 (non-fiabilité): $\hat{\rho}^2_C < 0{,}70$, sans plan optimal.

L'analyse des résultats

À l'aide du Tableau 2 on voit que quinze variables seulement obtiennent un coefficient de généralisabilité dépassant le seuil 0,70 et sont donc classées au niveau 1 de fiabilité. Neuf variables ont un plan optimal et sont de niveau 2. Les 53 autres variables sont de niveau 3 puisqu'aucun plan optimal n'a pu être produit.

On retrouve au Tableau 3 la répartition des 77 variables dans les six contextes d'enseignement retenus dans l'étude suivant leur niveau de fiabilité. Aucune des variables des contextes 2, 4 ou 6 n'obtient un niveau de fiabilité 1 ou 2. Il faut bien dire cependant que ces contextes comprennent fort peu de variables. Par ailleurs

Tableau 3

Répartition des 77 variables élémentaires selon le contexte et le niveau de fiabilité

Niveau de fiabilité \ Contexte	1	2	3	4	5	6	Total
1. $\hat{\rho}^2_C \geq 0{,}70$	11	0	1	0	3	0	15
2. $\hat{\rho}^2_C < 0{,}70$ avec plan optimal	5	0	2	0	2	0	9
3. $\hat{\rho}^2_C < 0{,}70$ sans plan optimal	25	3	14	3	7	1	53
Total	41	3	17	3	12	1	77

27% des variables du contexte 1, 6% des variables du contexte 3 et 25% des variables du contexte 5 obtiennent le niveau 1 de fiabilité. Au total 19% des 77 variables élémentaires obtiennent le premier niveau de fiabilité et 12% le second niveau. Ces faibles pourcentages peuvent paraître inusités. Ils sont pourtant conformes à beaucoup d'études similaires (Berliner, 1980; Shavelson et Atwood-Russo, 1977; Erlich et Shavelson, 1978; Erlich et Borich, 1979; Calkins *et autres*, 1977).

Le Tableau 4 regroupe les mêmes 77 variables mais cette fois selon le niveau de fiabilité et la catégorie relative à l'enseignement: quoi académique, non académique, non impliqué (contexte 6), autre. Selon ce tableau, il appert que les sous-catégories question, réponse et rétroaction comprennent proportionnellement plus de variables fiables que la sous-catégorie instruction ou les catégories non académique, non impliqué et autre. Fait à remarquer, la sous-catégorie instruction, qui comprend des variables de fréquences très nombreuses, ne présente qu'une seule variable de niveau 1 (toutes les autres sont de niveau 3), soulignant ainsi la grande instabilité des variables de cette sous-catégorie. Par ailleurs au moins la moitié des variables des sous-catégories question et réponse obtiennent une fiabilité de niveau 1 ou 2.

Tableau 4

Répartition des 77 variables élémentaires selon les catégories et le niveau de fiabilité

Catégorie / Niveau de fiabilité	Quoi académique				Non académique	Non impliqué	Autre	Total
	Instruction	Question	Réponse	Rétroaction				
1. $\hat{\rho}^2_C \geq 0{,}70$	1	3	7	1	2	1	0	15
2. $\hat{\rho}^2_C < 0{,}70$ avec plan optimal	0	6	1	2	0	0	0	9
3. $\hat{\rho}^2_C < 0{,}70$ sans plan optimal	16	9	5	4	9	9	1	53
Total	17	18	13	7	11	10	1	77

Discussion de la fiabilité des variables composites

Analyse des résultats

On retrouvera dans Leclerc *et autres* (1983, p. 131) un tableau faisant état des résultats de l'étude de fiabilité pour les 194 variables composites. Ce tableau est construit selon le modèle du Tableau 2. Même si, faute d'espace, nous ne pouvons le représenter ici, nous donnons au Tableau 5 la synthèse de ces résultats suivant le contexte et le niveau de fiabilité. Comme ces variables ont été « composées » à partir

de variables élémentaires, il ne sera donc pas étonnant de rencontrer des résultats similaires à ceux qui ont été obtenus précédemment. Nous commenterons brièvement ces résultats.

Tableau 5

Répartition des 194 variables composites selon le contexte et le niveau de fiabilité

Niveau de fiabilité \ Contexte	1	2	3	4	5	6	Inter-contextes	Total
1. $\rho^2_{\hat{C}} \geq 0{,}70$	33	0	4	0	8	0	12	57
2. $\rho^2_{\hat{C}} < 0{,}70$ avec plan optimal	8	0	3	0	2	0	7	20
3. $\rho^2_{\hat{C}} < 0{,}70$ sans plan optimal	50	16	17	10	3	0	21	117
Total	91	16	24	10	13	0	40	194

Tout comme pour les variables élémentaires, aucune des variables des contextes 2 ou 4 n'obtient un niveau de fiabilité 1 ou 2. Or 36% des variables du contexte 1, 17% des variables du contexte 3 et 62% des variables du contexte 5 obtiennent le niveau 1 de fiabilité: des résultats sensiblement plus intéressants que ceux relatifs aux variables élémentaires. Par ailleurs 30% des variables inter-contextes (les 6 contextes regroupés) obtiennent également un coefficient supérieur à 0,70. Au total 29% des variables composites peuvent être considérés comme fiables au premier niveau et 10% au second niveau. Ainsi, 77 variables composites, soit 40%, obtiennent le niveau 1 ou 2 de fiabilité. Les 117 autres variables composites, c'est-à-dire 60%, sont considérées non fiables.

Nous ne poursuivons pas davantage la discussion relative aux variables composites. Nous avons regroupé des variables élémentaires dans le but de trouver un plus grand nombre de variables dites fiables. Ces variables pourront alors être mises en corrélation avec les résultats que les élèves de nos trente classes ont obtenus d'une part au test d'initiation à l'algèbre et d'autre part au questionnaire d'attitudes face aux mathématiques.

Si on fait la somme des 77 variables composites et des 24 variables élémentaires qui ont obtenu le niveau 1 ou 2 de fiabilité, on est amené à retenir 101 variables en tout pour l'étude corrélationnelle.

Comparaison avec d'autres études

On peut penser comparer ces résultats avec ceux d'autres études, même si les définitions des catégories des divers instruments d'observation comparés ne concordent pas toujours.

Erlich et Shavelson (1978) mentionnent, par exemple, que neuf des dix variables relatives à la présentation du contenu par l'enseignant ont été jugées non fiables. Ce résultat rappelle celui de Shavelson et Dempsey-Atwood (1976), mais aussi le nôtre. En effet seize des dix-sept variables élémentaires liées à l'instruction et les variables composites d'enseignement magistral ne peuvent être considérées comme fiables.

En outre les études de Erlich et Shavelson (1978), Erlich et Borich (1979) puis Calkins *et autres* (1977) soulignent un certain nombre de variables fiables, relatives aux questions et aux rétroactions (« feedback »). Ce résultat s'apparente à ce que nous avons obtenu.

Conclusion

L'affirmation que nous reprenions à notre compte au début de cet article, à savoir qu'une très bonne entente entre les observateurs n'était pas une condition suffisante pour s'assurer de la fiabilité des données, a été confirmée à nouveau dans cette étude. Nous avons trouvé peu de variables fiables, même si le niveau d'entente entre les observateurs était élevé.

Nous avons trouvé également qu'un nombre de fréquences élevé n'est pas un gage de fiabilité, comme en font foi les variables relatives à l'instruction. Un résultat qui corrobore empiriquement les dires de Rowley (1976).

Si on combine ensemble les résultats pour les variables élémentaires et composites, il appert que c'est le contexte 5 relatif à l'enseignement individualisé qui a donné proportionnellement le meilleur taux de variables fiables : 44% des variables de ce contexte atteignent le niveau 1 de fiabilité et 16% le niveau 2. Par contre *aucune* des variables des contextes 2 ou 6 ne s'est révélée fiable au cours de notre étude.

Il y a peu d'enseignement en petit groupe (contexte 2) au cours de la leçon de mathématiques, semble-t-il. Et les quelques pratiques d'enseignement reliées à ce contexte sont apparues instables. Il en est de même pour les transitions (contexte 4) dans une classe de mathématiques. Est-ce qu'il faut redéfinir ces contextes? Est-ce qu'ils sont observables de la façon dont nous avons procédé? Il reste à trouver des réponses à ces questions. Il en est sans doute ainsi du contexte 6 qui ne comprenait qu'une variable.

Il nous paraît inconcevable qu'il y ait autant de variables non fiables dans notre étude, même si plusieurs études, comme on l'a déjà indiqué, présentent des résultats analogues. La recherche en enseignement a à résoudre ce problème. À notre avis les solutions doivent être cherchées autant du côté méthodologique que du côté théorique.

Les questions d'ordre méthodologique ne manquent pas. Par exemple, le dispositif d'observation était-il adéquat? L'échantillonnage des occasions d'observation ne gagnerait-il pas à être plus structuré de manière à obtenir une meilleure

représentativité du temps d'enseignement? On sait cependant que cette dernière question ne peut être résolue facilement. Les imprévus causés par la température, les maladies des observateurs ou des enseignants, sont des facteurs qui peuvent perturber considérablement les prévisions concernant les périodes d'observation.

La combinaison du nombre d'observateurs et du nombre d'occasions peut-elle être améliorée? Théoriquement oui, sans doute. Le calcul du plan optimal montre en effet que si on avait eu un dispositif de cinq observateurs et deux occasions par observateur au lieu de deux observateurs et cinq occasions par observateur, on aurait pu s'attendre à une augmentation de près de 40% du nombre de variables de niveau 1 de fiabilité. Est-ce là pourtant un dispositif viable? Les enseignants vont-ils accepter que cinq personnes distinctes viennent observer dans leur classe (même si les cinq personnes n'observent pas en même temps)?

Du côté théorique, il nous apparaît également qu'il reste beaucoup de progrès à faire. A-t-on utilisé toutes les facettes susceptibles d'influencer la fiabilité des données? Il est bien possible que certaines facettes, fort importantes au niveau de la variation des observations, aient été complètement ignorées. McGraw *et autres* (1972) de même que Calkins *et autres* (1977) ont déjà soulevé cette question. Ces auteurs condamnent en particulier l'ignorance de la facette « situation » lors de l'élaboration d'un dispositif d'observation. Cette facette fait référence par exemple à l'année scolaire, au jour de la semaine ou de l'année, au moment de la journée, à la matière enseignée, à la présentation d'un nouveau contenu ou à la révision d'un contenu. Cependant l'inclusion d'une telle facette ne devrait pas compliquer outre mesure le dispositif d'observation. Sinon le coût de la recherche risque de croître de façon démesurée.

Une fois définie la situation, il faut sans doute penser connaître de façon exhaustive tous les contextes possibles que l'on rencontre lors d'une leçon de mathématiques, c'est-à-dire connaître chacune des activités vraiment distinctes: par exemple la présentation d'un objectif ou d'un concept, l'explication et l'information à propos de cet objectif, la motivation des élèves, les revues, les exercices individuels, les rétroactions, les corrections, les transitions, la procédure, la discipline. Car le contexte donne sens, selon nous, aux rôles distincts des acteurs en présence.

L'instrument d'observation que nous avons utilisé tenait compte de 6 contextes et des rôles des acteurs à l'intérieur de ces contextes. Si nous pensons qu'il faut redéfinir avec plus de minutie qu'on ne l'a fait chacun des contextes, c'est-à-dire passer de six contextes à dix ou quinze contextes par exemple, il nous apparaît important aussi de ne pas limiter les observations aux rôles des acteurs seulement, comme nous l'avons fait à l'aide de l'instrument d'observation que nous avons utilisé, mais d'observer en plus et précisément le contenu de leurs interactions, c'est-à-dire les messages qui sont échangés entre eux dans un contexte donné, lors d'une situation précise. Quelle conception de l'enseignement pourra intégrer tous ces éléments? Et comment, à partir de cette conception, monter un dispositif

d'observation et définir un instrument d'observation praticable? Il faudrait tout de même que cela se fasse!

Même si la planification d'une étude de fiabilité relative à un instrument d'observation pose un grand nombre de problèmes d'ordre méthodologique et théorique, nous pensons cependant que de telles études doivent se faire et être nombreuses. Car bien qu'elles soient onéreuses, ce n'est qu'à ce prix que nous pourrons mieux évaluer la fiabilité des pratiques d'enseignement et avoir une chance de trouver des relations stables entre ces pratiques, le rendement scolaire et les attitudes des élèves.

NOTES

1. On entend par « données », les fréquences d'une pratique d'enseignement pour une occasion d'observation.
2. « To the extent that scores for a behavioral dimension are generalizable over raters and occasions, rank orderings among teachers on that dimension will be consistent and the likelihood of discovering relationships between pupil achievement and that dimension enhanced ».
3. Voir Leclerc, M., R. Bertrand, E. Maunsell, D. Rhéaume, Étude de la classe et de son environnement, Phase corrélationnelle, IEA, *Document R-159*, INRS-Éducation, Québec, 1983.
4. L'entraînement a duré environ 60 heures. Il s'est terminé au moment où les observateurs sont arrivés à s'entendre entre eux d'une façon satisfaisante (tests passés par les observateurs nécessitant un niveau d'entente égal ou supérieur à 80%). Au milieu, à peu près, de la cueillette des données d'observation, qui s'est échelonnée de la fin de janvier 1981 à la fin d'avril de la même année, les observateurs ont dû vérifier par un test leur niveau de stabilité dans leur entente, qui devait rester égal ou supérieur à 80%.
5. Pour chaque occasion d'observation, il aurait dû y avoir théoriquement 300 fréquences enregistrées pour l'ensemble des variables, puisque l'observateur devait enregistrer une fréquence à toutes les cinq secondes pendant cinq fois cinq minutes. Or en réalité on compte plutôt de 250 à 360 fréquences par occasion. Nous avons pondéré le nombre de fréquences pour chaque variable en utilisant le rapport:

$$\text{nombre pondéré de fréquences pour la variable par occasion et par classe} = \frac{\text{nombre brut de fréquences de la variable par occasion et par classe}}{\text{nombre brut de fréquences pour toutes les variables par occasion et par classe}} \times 300$$

6. Un coefficient de corrélation de Pearson calculé à partir de ces deux colonnes nous donne $r = 0{,}14$ pour $n = 77$.

RÉFÉRENCES

Berliner, D.C., Studying instruction in the elementary classroom in *Issues in Microanalysis*, R. Dreeben et J.A. Thomas (éds), Cambridge: Ballinger, 1980.

Calkins, D, G.D. Borich, M. Pascone et C.L. Kugle, Generalizability of teacher behaviors across classroom observation systems dans *Classroom observation data: Is it valid? Is it generalizable?*, G. Borich *et autres* (éds), Austin: The University of Texas at Austin, 1977.

Cronbach, L.J., G.C. Gleser, H. Nanda et N. Rajaratnam, *The dependability of behavioral measurements*, New York: John Wiley & Sons, Inc., 1972.

Dillashaw, F.G. et J.R. Okey, Effects of a modified mastery learning strategy on achievement, attitudes and on-task behavior of high school chemistry students, *Journal of Research in Science Teaching*, vol. 20, no 3, 1983, p. 203-211.

Ebel, R.L., Estimation of reliability of ratings, *Psychometrika*, vol. 16, 1951, p. 407-424.

Erlich, O. et G. Borich, Occurence and generalizability of scores on a classroom interaction instrument, *Journal of Educational Measurement*, vol. 16, no 1, 1979, p. 11-18.

Erlich, O. et R.J. Shavelson, The search for correlations between measures of teacher behavior and student achievement: measurement problem, conceptualization problem, or both? *Journal of Educational Measurement*, vol. 15, no 2, 1978, p. 77-89.

Frick, T. et M.I. Semmel, Observer agreement and reliabilities of classroom observational measures, *Review of Educational Research*, vol. 48, no 1, 1978, p. 157-184.

Leclerc, M., R. Bertrand et J. Roberge-Brassard, Étude de fiabilité d'un instrument d'observation des comportements de l'élève en classe, *Revue des Sciences de l'Éducation*, vol. V, no 3, 1979, p. 359-372.

Leclerc, M., R. Bertrand, E. Maunsell et D. Rhéaume, Étude de la classe et de son environnement, Phase corrélationnelle, IEA, *Document R-159*, Québec: INRS-Éducation, 1983.

Lomax, R.G., An application of generalizability theory to observational research, *Journal of Experimental Education*, vol. 51, no 1, 1982, p. 22-30.

McGraw, B., J.L. Wardrop et M.A. Bunda, Classroom observation schemes: where are the errors?, *American Educational Research Journal*, vol. 9, no 1, 1972, p. 13-27.

Medley, D.M. et H. Mitzel, Application of analysis of variance to the estimation of the reliability of observations of teachers' classroom behavior, *Journal of Experimental Education*, vol. 27, no 1, 1958, p. 23-35.

Mitchell, S.K., Interobserver agreement, reliability, and generalizability of data collected in observational studies, *Psychological Bulletin*, vol. 86, no 2, 1979, p. 376-390.

Rowley, G., The reliability of observational measures, *American Educational Research Journal*, vol. 13, no 1, 1976, p. 51-59.

Rowley, G., Reliability and other indicators of data quality in observational research, La Trobe University, Australie, 1983, 39 p.

Scott, N.A., Reliability of content analysis: The case of nominal scale coding, *Public Opinion Quaterly*, vol. 19, 1955, p. 321-325.

Shavelson, R. et N. Atwood-Russo, Generalizability of measures of teacher effectiveness, *Educational Research*, vol. 19, no 3, 1977, p. 171-183.

Shavelson, R. et N. Dempsey-Atwood, Generalizability of measures of teaching behavior, *Review of Educational Research*, vol. 46, no 4, 1976, p. 553-611.

Shrout, P.E. et J.L. Fleiss, Intraclass correlations: uses in assessing rater reliability, *Psychological Bulletin*, vol. 86, no 2, 1979, p. 420-428.

Smith, P.L., Sampling errors of variance components in small sample multifacet generalizability studies, *Journal of Educational Statistics*, vol. 3, no 4, 1978, p. 319-346.

Smith, P.L. et P.A. Teeter, *The use of generalizability theory with behavioral observation*, Communication présentée au Annual meeting of the American Educational Research Association, New York, avril 1982.

Tourneur, Y. et J. Cardinet, *Analyse de variance et théorie de la généralisabilité: guide pour la réalisation des calculs*, Service d'étude des méthodes et des moyens d'enseignement, Mons, Belgique: Université de l'État à Mons, 1979.

Webb, N.M., *Generalizability of classroom processes: Taking into account the correlations among observations*, Communication présentée au Annual meeting of the American Educational Research Association, New York, avril 1982.