LA GÉNÉRALISABILITÉ :
À QUOI ÇA SERT, COMMENT ON S'EN SERT ?
EXEMPLE INTRODUCTIF

INTRODUCTION

Ce texte présente la généralisabilité (GT) à la façon d'un panorama. Sans s'arrêter sur les détails, encore moins sur les problèmes théoriques, il décrit concrètement la grande diversité d'utilisation de ce modèle pour la **mise au point de dispositifs d'évaluation**. Il se centre pour cela sur le développement d'un exemple s'inspirant de la pratique. Comme l'indique notre titre, il s'agit pour nous de fournir aux futurs utilisateurs intéressés un premier aperçu de ce que l'on peut faire à la fois avec le modèle et le logiciel *EduGf* qui lui est dédié[1]. Cette présentation constitue également une introduction rapide au maniement de cet outil informatique qui prend en charge tous les calculs.

Pour plus d'information sur le modèle et sur son mode d'emploi, le lecteur se référera aux ouvrages suivants :
Cardinet, J. & Tourneur, Y. *Assurer la mesure*. Berne : P. Lang, 1985.
Bain, D. & Pini, G. *Pour évaluer vos évaluations : généralisabilité, mode d'emploi.* Genève, Centre de recherches psychopédagogiques, Direction générale du Cycle d'orientation, 1996.
Une nouvelle version de cette dernière brochure est prévue.
On peut également s'adresser au groupe de travail *Edumétrie – Qualité de l'évaluation en éducation*, p/a Daniel Bain, route du Moulin-Roget 49, 1237 Avully ; courriel : daniel.bain@bluewin.ch

1. ÉLABORER LA PROBLÉMATIQUE

1.1 Contexte et buts de l'opération

Un groupe d'enseignant-e-s d'un cycle d'orientation (secondaire 1, niveau collège en France) projette une opération de remédiation dans le domaine de la lecture[2]. Dans un premier temps, il s'agit de mettre au point des dispositifs d'évaluation permettant d'identifier les élèves, les classes et les conduites qui font problème. Premier domaine d'exploration : l'**attitude des élèves à l'égard de la lecture (plaisir de lire)**. Les enseignant-e-s demandent à des chercheurs en sciences de l'éducation un instrument qui les aide dans leur diagnostic.

1.2. Objets / objectifs d'évaluation

Selon les demandes des enseignants et des chercheurs, il s'agit :

A) d'identifier les **élèves** qui ont développé des attitudes négatives (scores < 2.5, médian de l'échelle de 1 à 4 utilisée dans le questionnaire ; cf. figure 1) en vue d'une opération de remobilisation ou de remotivation ;

[1] Dans la présente édition de ce texte, nous avons utilisé la version EduGf 2.0 du logiciel, dont la programmation a été réalisée par la firme canadienne *Educan Inc*. avec la collaboration du groupe de travail *Edumétrie* de la Société suisse pour la recherche en éducation (SSRE).

[2] Cet exemple s'inspire de problèmes effectivement rencontrés dans notre pratique, mais ne se réfère pas à une recherche effective. De même, les données utilisées ici sont issues de la recherche (survey) PISA 2000, mais ont été aménagées pour les besoins de la démonstration. Nos résultats et conclusions n'ont donc de validité qu'à l'intérieur de cette démonstration.

B) de repérer les **classes** qui ont en moyenne les attitudes les plus positives (scores moyens > médian de l'échelle) pour analyser avec les maîtres les stratégies didactiques qui contribuent à rendre la lecture attrayante pour leurs élèves.

C) d'identifier les **conduites de lecture** (telles que décrites par les *items* du questionnaire) qui sont investies négativement par les élèves (moyennes < 2.5).

1.3 Choix de l'instrument d'évaluation

Le questionnaire suivant (figure 1) tiré de l'enquête PISA (*Programme international pour le suivi des acquis des élèves*) se prête bien aux opérations projetées[3]. Selon les présentations qu'en font D. Lafontaine (1999) ou PISA (2001), il est censé évaluer le *goût de lire*, le *plaisir de lire* ou l'*attitude à l'égard de la lecture*. Nous préférons personnellement cette dernière définition, qui couvre mieux l'ensemble des conduites présentées dans le questionnaire.

Figure 1 Questionnaire d'attitude à l'égard de la lecture

Dans quelle mesure êtes-vous d'accord avec les affirmations suivantes à propos de la lecture ? (Ne cochez qu'une seule case par ligne)

Item		Pas du tout d'accord	Pas d'accord	D'accord	Tout à fait d'accord	D.I
I1 (b)	La lecture est un de mes loisirs favoris.	£ 1	£ 2	£ 3	£ 4	1.1
I2 (c)	J'aime parler de livres avec d'autres personnes.	£ 1	£ 2	£ 3	£ 4	1.2
I3 (e)	Je suis content(e) quand je reçois un livre en cadeau.	£ 1	£ 2	£ 3	£ 4	1.3
I4 (g)	J'aime aller dans une librairie ou une bibliothèque.	£ 1	£ 2	£ 3	£ 4	1.4
I5 (a-)	Je ne lis que si j'y suis obligé(e).	£ 4	£ 3	£ 2	£ 1	2.1
I6 (f-)	Pour moi, la lecture est une perte de temps.	£ 4	£ 3	£ 2	£ 1	2.2
I7 (h-)	Je ne lis que pour trouver les informations dont j'ai besoin	£ 4	£ 3	£ 2	£ 1	2.3
I8 (i-)	Je ne peux pas rester tranquillement à lire plus de quelques minutes	£ 4	£ 3	£ 2	£ 1	2.4
I9 (d-)	J'éprouve des difficultés à finir les livres.	£ 4	£ 3	£ 2	£ 1	-

N.B. Pour les besoins des analyses ultérieures, l'ordre primitif de présentation des items *a* à *i* a été modifié dans le tableau de la figure 1 comme dans le fichier des données (cf. première colonne de la figure 1 et infra § 2.2).

[3] L'enquête internationale PISA 2000 de l'OCDE (Organisation de coopération et de développement économiques) a testé les compétences des élèves de 15 ans (nés en 1984) dans trois domaines : la lecture (littératie), les mathématiques et les sciences. Nous remercions le Consortium PISA (et plus particulièrement Christian Nidegger, SRED) d'avoir mis à notre disposition ces informations et ces données anonymisées. Sur l'élaboration du questionnaire, on lira avec intérêt l'article de D. Lafontaine : « Un goût de lire bien mesuré. Élaboration et mise à l'essai. *Mesure et évaluation en éducation, 22,* 1/1999, pp. 21-43. Sur la relation entre les résultats à ce questionnaire et la réussite en lecture dans l'enquête PISA, on consultera pour la Suisse romande le rapport coordonné par Chr. Nidegger : *Compétences des jeunes romands. Résultats de l'enquête PISA 2000 auprès des élèves de 9e.* Neuchâtel : IRDP, 2001. Voir aussi sur Internet les informations fournies par le site www.pisa.oecd.org.

2. RÉCOLTER ET PRÉPARER LES DONNÉES

2.1 Expérimentation préalable pour mettre au point les trois dispositifs correspondant aux trois objectifs ci-dessus : on décide de tester avec le questionnaire d'attitude (*9 items*) *4 classes* de *20 élèves* choisies aléatoirement parmi les *200* classes correspondant à la population cible (*univers de référence*).

2.2. Retranscription des réponses des élèves (code 1, 2 , 3 ou 4 ; cf. colonnes 3 à 6 de la figure 1) item par item sur un tableur (par exemple *Excel*). Pour les besoins des analyses, les items sont retranscrits dans un ordre différent de celui du questionnaire, afin de regrouper les items formulés positivement (b, c, e, g) et négativement (a, f, h, i, d), et les 4 classes sont ramenées à 20 élèves par élimination aléatoire des enregistrements surnuméraires.

Tableau 2 Extrait du tableau de données Excel

Classe	Élève	I1	I2	I3	I4	I5	I6	I7	I8	I9
1	**1**	2	2	3	2	2	3	2	3	2
1	**2**	3	1	3	3	2	4	3	2	4
1	**3**	1	1	1	1	1	1	1	4	4
1	**4**	4	3	4	4	4	4	4	4	4
(…)										
2	**1**	4	3	3	3	4	4	4	4	3
2	**2**	3	2	4	3	4	4	4	3	2
2	**3**	1	2	2	1	3	3	2	1	2
2	**4**	2	3	2	3	3	3	3	2	3

Après suppression des deux premières colonnes et de la première ligne (identifiant les classes, élèves et items ; plages grisées dans le tableau 2), le fichier est enregistré sous le format : *Texte (séparateurs : tabulations)*. Il sera importé par le logiciel à une étape ultérieure (cf. infra § 3.2).

2.3 Autres possibilités pour saisir les données

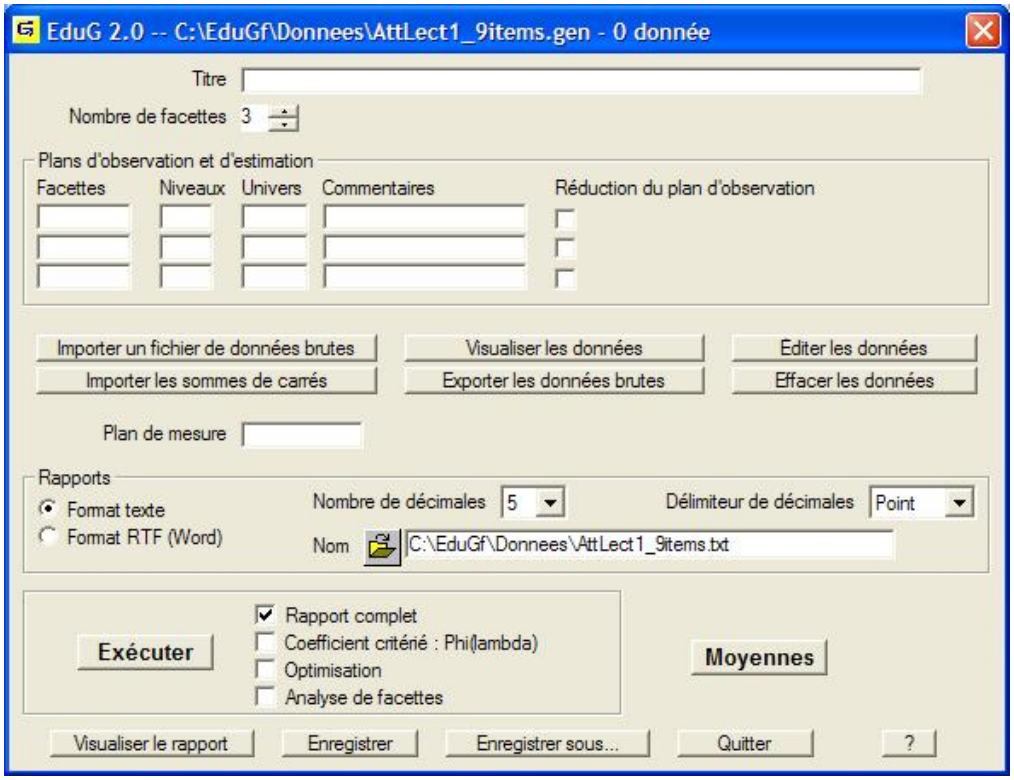
- Les logiciels de traitement de textes comme *Word* permettent aussi de créer un tableau et de transformer ce tableau (*Convertir tableau en texte*) en un vecteur de données segmenté par un séparateur à choix (choisir de préférence le tabulateur).
- Le programme (version 2.0 ; cf. figure 3) propose une routine (bouton Éditer les données) permettant de saisir les réponses dans l'ordre indiqué par les facettes. Cette routine de saisie est pratique pour des données peu nombreuses et si l'on n'envisage pas de modifications importantes de celles-ci. Nous aurions tendance sinon à en déconseiller l'utilisation.
- Le programme comporte également une routine permettant de saisir des sommes de carrés telles qu'elles apparaissent dans les tableaux de résultats de certaines recherches. Nous n'en dirons rien ici.

3. PRÉPARER ET EXÉCUTER L'ANALYSE DE GÉNÉRALISABILITÉ (GT)

Dans les analyses qui suivent nous nous intéresserons à l'objectif d'évaluation A défini ci-dessus : **repérer les élèves ayant développé des attitudes négatives à l'égard de la lecture.**

L'introduction des paramètres et des informations nécessaires est guidée par le logiciel ; cf. la *fenêtre de pilotage* du programme *EduGf2.0* (figure 3).

Figure 3 Fenêtre de pilotage[4]



3.1 Définir les plans d'observation et d'estimation

Les champs à remplir (nombre de facettes : 3) sont reproduits au tableau 4.

Tableau 4 Plans d'observation et d'estimation pour la différenciation des élèves

Facettes	Niveaux	Univers	Commentaires
C	4	200	Classes
E:C	20	INF	Elèves par classe
I	9	INF	Items

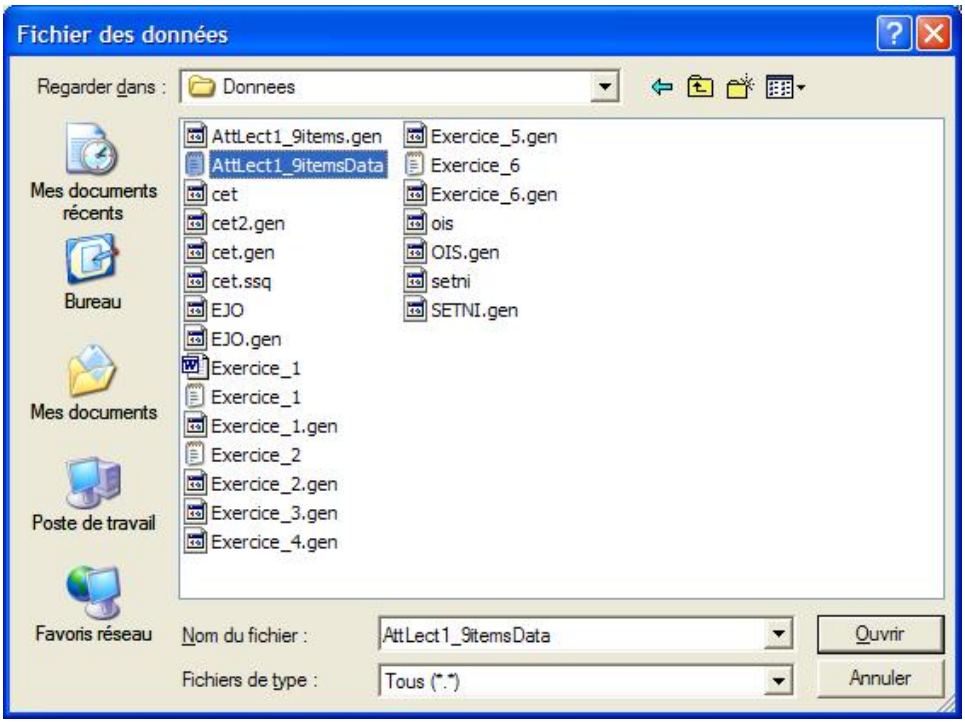
Facettes = facteurs pris en considération ; dans E:C, les deux-points signifient *inclus (ou nichés) dans ;* ne pas insérer d'espace dans cette expression (E : C) ; les facettes non nichées sont considérées implicitement comme *croisées* (cf. à ce sujet le *Mode d'emploi* cité dans l'introduction).
Niveaux = modalités pour chaque facette
Univers : « réservoir » ou « pool » dans lequel sont puisées les données (*niveaux*) de chaque *facette*. Si le nombre de niveaux univers est considéré si grand qu'il n'est pas possible d'en trouver le nombre exact, on dira que la facette est *aléatoire infinie* (indiqué par INF dans le champ *ad hoc*) ; c'est ici le cas des Elèves et des Items. Si le nombre de niveaux univers est supérieur au nombre de niveaux observés tout en étant fini, la facette est dite *aléatoire finie* ; c'est le cas des 4 classes tirées au sort dans un ensemble de 200 classes. Si le nombre de niveaux univers égale le nombre de niveaux observés, c'est que la facette est *fixée* ; on ne pourra donc pas généraliser les résultats relatifs à cette facette ; ce sera le cas plus loin, au § 8.3, d'une facette *Dimension de formulation* groupant séparément les items positifs et négatifs.

[4] Pour afficher cette *Fenêtre de pilotage*, dans l'écran de départ du logiciel, ouvrir le menu *Fichier* et cliquer sur *Nouveau*.

3.2 Importer les données du fichier *Texte* contenant les données en appuyant sur le bouton : Importer un fichier de données brutes

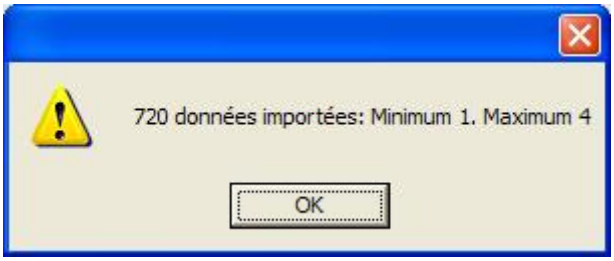
Dans la fenêtre ouverte (figure 5), chercher et enregistrer le fichier concerné, en l'occurrence *AttLect1_9itemsData.txt* dans le dossier *Donnees*.

Figure 5 Fenêtre pour la recherche du fichier de données et son enregistrement



Le logiciel enregistre les données en confirmant pour contrôle le nombre total de données (4 x 20 x 9 = 720) ainsi que les valeurs minimales et maximales (figure 6).

Figure 6 Fenêtre de confirmation de l'enregistrement des données à partir d'un fichier texte



3.3 Définir le plan de mesure (= ce que l'on veut mesurer / avec quoi)

dans la plage de saisie prévue à cet effet (figure 7) sous la forme : **EC/I** (ou CE/I)[5], soit

- **à gauche de la barre de fraction** la ou les facettes faisant partie de la *face de différenciation* (ce que l'on veut mesurer) ; ici, les Elèves inclus dans les Classes (**EC**) ; N.B. Une facette nichée (incluse) doit être accompagnée de sa facette nichante (incluante) ; on supprime dans ce cas les deux-points entre les deux facettes ;

[5] A l'intérieur de chacune de deux faces de différenciation et d'instrumentation (donc de chaque côté de la barre de fraction), l'ordre des facettes n'a pas d'importance.

- **à droite de la barre de fraction** la ou les facettes utilisées pour réaliser la mesure (= *face d'instrumentation*), ici les Items (**I**) mesurant l'attitude à l'égard de la lecture.

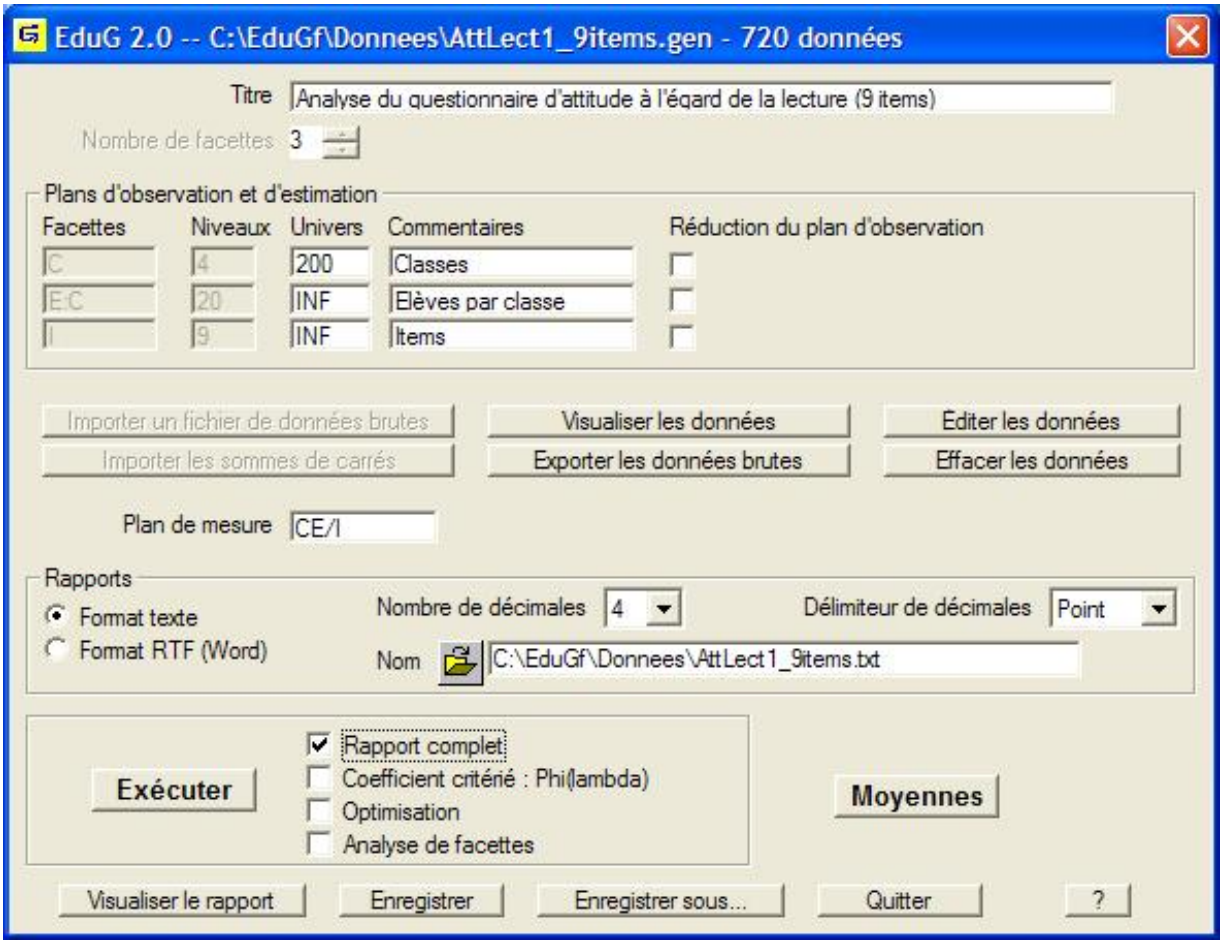
N.B. : ne pas insérer d'espace dans la formule.

3.4 Définir le nom et le format du fichier de sortie pour les rapports (= pour les résultats ; format *texte* ou *Word RTF*) ; modifier si nécessaire le **nombre et le délimiteur de décimales** (choisir par exemple 4 décimales et le point décimal ; cf. figure 7).

3.5 Choisir l'analyse désirée en cochant l'option *Rapport complet* (= rapport sur l'analyse de la variance et de la généralisabilité ; option choisie dans la fenêtre de pilotage de la figure 7). Autres options disponibles : *Coefficient critérié*, *Optimisation*, *Analyse de facettes* ou encore *Moyennes* (cf. infra).

La *fenêtre de pilotage*, à ce stade, se présente comme dans la figure 7.

Figure 7 *Fenêtre de pilotage après saisies des paramètres nécessaires pour l'analyse de GT*



3.6 Exécuter l'analyse de GT
en cliquant sur le bouton *ad hoc*

4. LIRE / INTERPRÉTER LES RÉSULTATS DE L'ANALYSE

La **stratégie de lecture** de ces résultats variera selon le chercheur et sa familiarité avec le modèle.

Un « parcours du débutant » est proposé ici : des résultats globaux à leur analyse plus détaillée[6].

Tableau 8 Extrait du rapport complet (sortie en format Word RTF)

(…)

Coefficients de Généralisabilité
Plan de mesure: CE/I

Sources de variance	Variance de dif.	Sources de var.	Var. d'err. relative	Var. d'err. absolue	%
C	0.1107				
E:C	0.3956				
		I		0.0136	20.2
		CI	0.0018	0.0018	2.6
		EI:C	0.0520	0.0520	77.2
Totaux	0.5064		0.0538	0.0674	
Ecarts types	0.7116		**0.2319**	**0.2595**	

Coefficient de généralisabilité	relatif	absolu
	0.9040	**0.8826**

Les résultats qui nous intéressent dans un premier temps (en gras dans le tableau 8) sont ceux qui répondent globalement à nos questions de départ :

1. l'estimation globale de la fiabilité ou *généralisabilité* du dispositif et

2. le calcul d'un *intervalle de confiance ou d'incertitude* autour d'un seuil (score médian de l'échelle : 2.5) afin de repérer dans la distribution des résultats (scores moyens, cf. figure 10) les élèves ayant une attitude nettement négative.

N.B. Le logiciel calcule systématiquement des *scores moyens* = scores totaux / nombre d'items ; l'élève no 1 de la classe 1 (cf. tableau 2 supra) qui a un total de 21 points obtient ainsi un score moyen de 2.3333 (21 pts / 9 items)

4.1. **Considérer les coefficients de généralisabilité** (*rhô carrés :* ρ^2 ; varient de 0 à 1) qui, selon deux perspectives différentes, donnent une **évaluation globale de la fiabilité du dispositif d'évaluation** :
- le **coefficient de généralisabilité relatif** est utilisé s'il s'agit seulement de **hiérarchiser ou classer** les niveaux de la facette de différenciation = ici classer les élèves selon leur attitude plus ou moins favorable ; il ne convient pas dans le cas considéré ;
- le *coefficient de GT absolu* quand il s'agit non seulement de classer les élèves, mais de **situer** leur niveau d'attitude **sur l'échelle des scores** (1 à 4).

C'est ce **coefficient de généralisabilité absolu** que nous prendrons en considération dans le cas qui nous intéresse étant donné la façon dont nous avons posé l'objectif de l'analyse (repérer les élèves dont le score est inférieur au médian de l'échelle).

Figure 9 Formule pour le calcul du coefficient de généralisabilité

$$\text{Coefficient de généralisabilité}_{\text{rel. ou abs.}} = \frac{\text{var. de différenciation}}{\text{var. différ.} + \text{var. d'erreur}_{\text{rel. ou abs.}}}$$

[6] Dans cette présentation introductive, nous laisserons de côté l'analyse de variance qui prépare l'analyse de généralisabilité.

Interprétation de la formule de la figure 9 :

Le coefficient de généralisabilité représente la *proportion de variance vraie dans la variance observée* (variance de différenciation + variance d'erreur). **La généralisabilité d'un dispositif d'évaluation est considérée comme satisfaisante si ce coefficient est ≥ 0.80** ; ce qui est le cas dans notre exemple (cf. tableau 8) pour les coefficients relatif (**0.9040**) et absolu (**0.8826**).

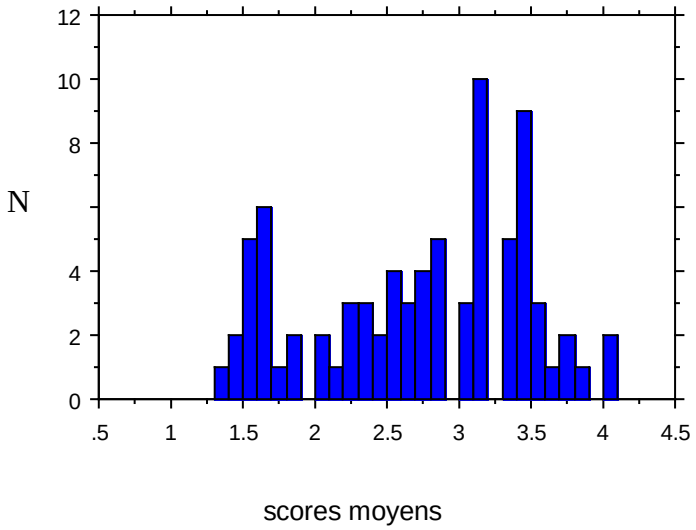
4.2. **Calculer l'intervalle de confiance** (ou d'incertitude ; marge d'erreur) autour du **seuil** choisi à partir de *l'écart type de l'erreur absolue*[7], selon la formule :

intervalle de confiance = ±1.96 * écart type de l'erreur absolue

soit : ±1.96 * 0.2595 = ± **0.5086** pour $\alpha = 0.05$

Seuil choisi : 2.5 (médian de l'échelle). Grosso modo, **entre 2 et 3 (2.5 ± 0.51)**, il s'agirait, avant de prendre une décision, de recueillir des informations complémentaires telles que l'avis des maîtres… et des élèves. Autres stratégies possibles :
- si l'on veut n'ouvrir le programme de remotivation qu'aux élèves qui ont une nette probabilité d'être motivés négativement, on n'acceptera que ceux qui ont un score moyen inférieur à 2 points (2.5 – 0.5).
- si l'on veut éviter de laisser de côté des élèves motivés négativement, on déplacera le seuil plus haut et on retiendra tous ceux qui ont un score moyen inférieur ou égal à 3 points (2.5 + 0.5086 = 3.0086)[8].

Figure 10 Distribution de scores moyens pour l'ensemble des 40 élèves testés



[7] Comme il s'agit de situer les résultats des élèves sur l'échelle des scores, on choisit l'écart type de l'*erreur absolue*.

[8] Dans ce cas, comme on le constate sur l'histogramme de la figure 10, la proportion d'élèves concernés serait très importante !

5. AFFINER L'ANALYSE

5.1 Examiner les sources de différenciation dans le rapport d'analyse de la GT.

Tableau 11 Composition de la variance de différenciation (cf. tableau 8, col. 1 et 2)

Sources de variance	Variance de dif.
C	0.1107
E:C	0.3956
(…)	
Totaux	0.5064

Environ un cinquième de la variance de différenciation est dû à la facette Classe (tableau 11 : 0.1107 / 0.5064 = 0.2186) ; l'attitude à l'égard de la lecture est donc liée en partie à l'appartenance à telle classe plutôt qu'à telle autre.
Pour vérifier cette conclusion, on peut calculer les *moyennes* par classe au moyen de la fonction *ad hoc* du logiciel : dans la *fenêtre de pilotage* (figure 7, en bas), cliquer sur le bouton Moyennes et sélectionner la facette C dans la fenêtre qui s'ouvre (figure 12).

Figure 12 Fenêtre de sélection des moyennes à calculer

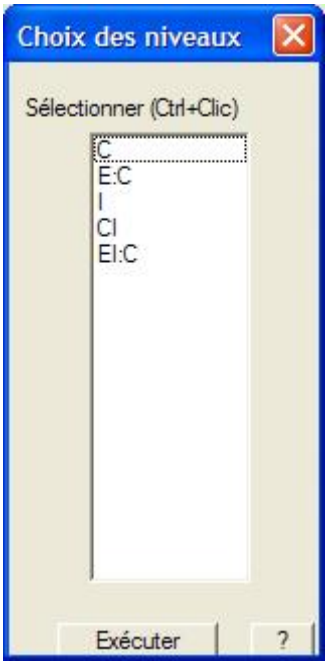


Tableau 13 Moyennes des scores par classe

Moyenne générale: 2.7194

C classes	Moyennes
1	2.5000
2	**3.2389**
3	2.7167
4	2.4222

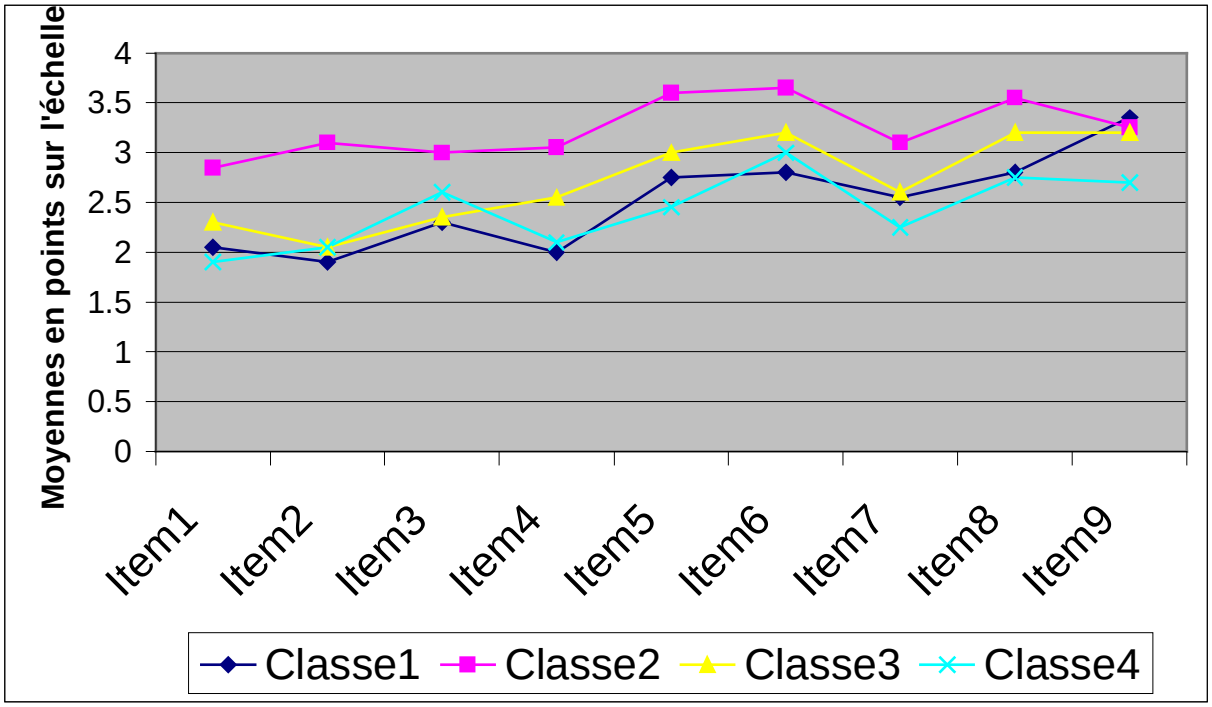
On constate (tableau 13) que la classe 2 se détache nettement des autres.

5.2 Examiner les sources de variance d'erreur
ici d'*erreur absolue* :

Tableau 14 Variances d'erreur (extrait de l'analyse de GT, tableau 8 col. 5 à 8)

Sources de var.	Var. d'err. relative	Var. d'err. absolue	%
I		**0.0136**	**20.2**
CI	0.0018	**0.0018**	**2.6**
EI:C	0.0520	**0.0520**	**77.2**

Graphique 15 Moyennes par classe et par item (illustration de l'interaction CI)



Comme c'est souvent le cas, la variance d'erreur la plus importante (tableau 14) correspond à *l'interaction Elèves x Items* (**EI:C**, 77.2%) : compte tenu de leur niveau moyen d'attitude, les élèves diffèrent les uns des autres dans leurs réactions aux différents items (ils les interprètent plus ou moins différemment) ; source d'erreur généralement difficile à réduire.
En revanche, l'*interaction Classes x Items CI* est très faible (2.6%). Cela signifie que les classes, compte tenu de leur niveau moyen d'attitude, réagissent de façon presque identique face aux différents items. La routine (bouton) Moyenne permet de calculer les moyennes par classe et par item en sélectionnant **CI** dans la fenêtre de *Choix des niveaux* (cf. figure 12). Cette interaction faible se traduit graphiquement par des profils de moyennes par classes et par items relativement parallèles (graphique 15).

6. AMÉLIORER / OPTIMISER LE DISPOSITIF D'ÉVALUATION

6.1 ***Analyse de facettes*** : elle vise à repérer dans les facettes aléatoires d'instrumentation (ici la facette Items) les niveaux qui contribuent à restreindre la généralisabilité, pour les analyser et éventuellement les éliminer. Quand on coche dans la *fenêtre de pilotage* (figure 7 supra) la case *Analyse de facettes*, la fenêtre suivante (figure 16) apparaît ; on y sélectionne la facette I avant de cliquer sur OK.

Figure 16 Choix de la facette à analyser

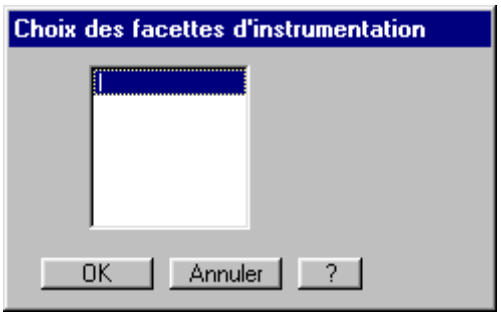


Tableau 17 Résultats de l'analyse de la facette Items

Facette	Niveau	Coef. rel.	Coef. abs.
I	1	0.8790	0.8571
	2	0.8939	0.8727
	3	0.9007	0.8755
	4	0.8907	0.8661
	5	0.8855	0.8584
	6	0.8876	0.8665
	7	0.8899	0.8632
	8	0.8947	0.8717
	9	0.9126	**0.8921**

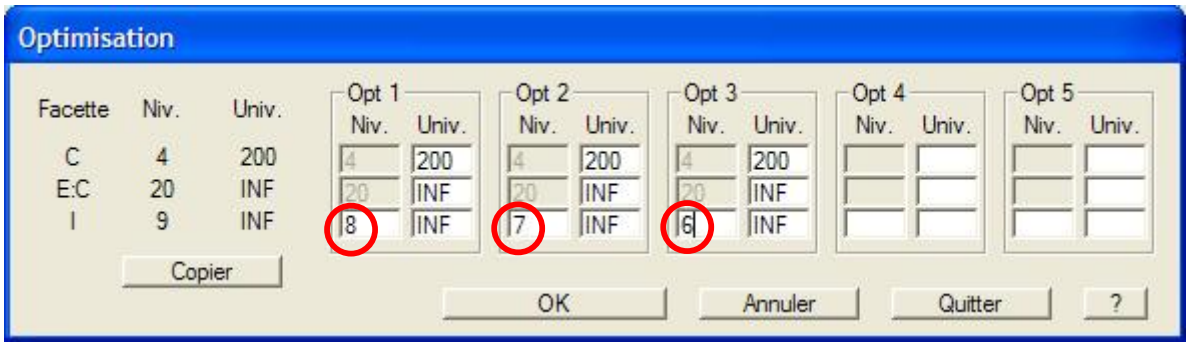
Les niveaux 1 à 9 de la facette I correspondent aux 9 items du questionnaire dans leur ordre remanié.

Les valeurs fournies correspondent aux **coefficients de GT si on supprime l'item en question** ; les items (niveaux) qui font problème sont donc ceux dont les coefficients sont élevés !

L'élimination de l'item 9 (*J'éprouve des difficultés à finir les livres*) permettrait d'améliorer un peu le coefficient de GT absolu, qui passerait ainsi de 0.8826 à 0.8921. Cet item exprime probablement plus une difficulté qu'une attitude ; la réponse dépend peut-être du genre de texte…

6.2 ***Plans d'optimisation*** : ils permettent d'estimer les résultats de modifications apportées au *plan d'observation* (consistant souvent à changer le nombre de niveaux des facettes d'instrumentation) ou au *plan d'estimation* (en modifiant l'univers de référence).
- Quand le coefficient de GT considéré est trop faible, on cherche généralement à augmenter le nombre de niveaux de la ou des facettes d'instrumentation. On élargit ainsi la base d'observation et diminue l'importance de l'erreur.
- Quand le coefficient est très élevé, comme ici, on vérifie si l'on peut faire l'économie d'un certain nombre de niveaux = d'items.

Figure 18 Plans d'optimisation



Dans la fenêtre *Optimisation*, qui s'ouvre quand on coche la case *ad hoc*, un bouton (Copier) permet de recopier les plans d'observation et d'estimation, ce qui permet de modifier ce qui nous intéresse (figure 18), en l'occurrence le nombre d'items ; N.B. les niveaux des facettes de différenciation sont non modifiables.

Après avoir cliqué sur le bouton Exécuter, on obtient les résultats illustrés par le tableau 19.

Tableau 19 Plan d'optimisation : extrait des résultats

	Plan original		Opt 1		Opt 2		Opt 3	
	Niv.	**Univ.**	**Niv.**	**Univ.**	**Niv.**	**Univ.**	**Niv.**	**Univ.**
C	4	200	4	200	4	200	4	200
E	20	INF	20	INF	20	INF	20	INF
I	9	INF	8	INF	7	INF	**6**	INF
Observations		720		640		560		480
Coeff. Rel.		0.9040		0.8933		0.8798		0.8626
Coeff. Abs.		0.8826		0.8698		0.8539		**0.8336**
Var. Err. Rel.		0.0538		0.0605		0.0692		0.0807
Err. Typ. Rel.		0.2319		0.2460		0.2630		0.2840
Var. Err. Abs.		0.0674		0.0758		0.0866		0.1010
Err. Typ. Abs.		0.2595		0.2753		0.2943		0.3179

On constate qu'il serait possible de raccourcir le questionnaire (par ex. si on avait d'autres questions à poser lors de la même passation), le coefficient de GT absolue estimé pour le plan réduit à 6 items (**0.8336**) étant encore supérieur à 0.80.

6.3 *Réduction du plan* : à titre de vérification, on peut estimer ce que donnerait une réduction du plan d'observation si l'on supprimait 3 items de façon ciblée (vs aléatoire) grâce à la routine *ad hoc*. Quand on coche (en haut de la *fenêtre de pilotage*, cf. figure 7) dans la colonne *Réduction du plan d'observation* la case figurant à côté de la facette *Items*, la fenêtre de la figure 20 apparaît. On y sélectionne (par Ctrl+Clic) les trois niveaux (items) à enlever, on clique sur OK, on coche (au bas de la fenêtre de pilotage) *Rapport complet* et on appuie sur Exécuter.

Si l'on supprime les items 2, 6 et 9 (en fonction de leur contenu et pour conserver 3 items positifs et 3 négatifs), l'analyse de GT donne les résultats du tableau 21.

Figure 20 Fenêtre de sélection des niveaux à supprimer

Tableau 21 Extrait de l'analyse de GT pour le plan réduit à 6 items

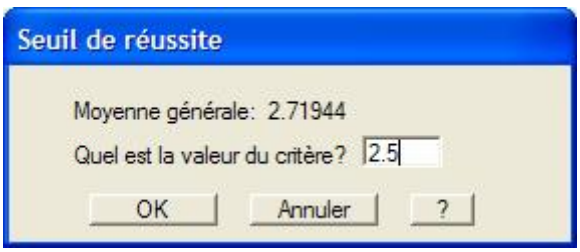
Coefficient de généralisabilité	relatif	absolu
	0.8853	**0.8663**

Le coefficient de GT *absolu* du plan réduit n'est inférieur que de peu à celui du plan complet (0.8826), notamment parce qu'on a supprimé l'item 9 qui fonctionnait différemment des autres (cf. supra au § 6.1 l'analyse de facettes).

6.3 Situer le score moyen par rapport à une norme : calcul d'un coefficient critérié

La généralisabilité fournit en outre une réponse à une autre question intéressante si l'on cherche à évaluer le *niveau global* d'attitude dans l'ensemble testé : le dispositif permet-il d'estimer si un certain standard d'attitude a été atteint en moyenne ? En d'autres termes, compte tenu des erreurs d'échantillonnage, peut-on mesurer avec fiabilité la distance entre le niveau moyen d'attitude observé (sur 40 élèves et 9 items = **2.7194**; cf. tableau 13) et un seuil déterminé, par exemple le médian de l'échelle utilisée par le questionnaire : **2.5** ? Le logiciel offre la possibilité de calculer un **coefficient critérié** (phi(lambda) de Brennan et Kane, à évaluer comme un coefficient de généralisabilité absolue). Pour ce faire, cocher dans la dernière section de la fenêtre de pilotage la case à côté de *Coefficient critérié : Phi(lambda*). Dans la fenêtre qui apparaît, inscrire le seuil ou critère visé (figure 22) et cliquer sur le bouton OK .

Figure 22 Fenêtre d'enregistrement du seuil à tester



Appuyer sur Exécuter (fenêtre de pilotage, figure 7 en bas); le logiciel affiche la valeur du coefficient sous la forme suivante :

Figure 23 Résultats du calcul du coefficient critérié

Moyenne générale pour les niveaux traités: 2.7194
Variance d'échantillonnage de la moyenne générale pour les niveaux traités: 0.0206

Phi(lambda) (**Seuil:2.5**) = **0.8880**

On peut donc conclure que l'instrument et le dispositif d'évaluation utilisés permettent de situer de façon fiable sur le pôle positif de l'échelle (scores > 2.5) l'attitude moyenne mesurée par le questionnaire sur l'échantillon d'élèves testé.

EVALUER LES ATTITUDES DES ÉLÈVES : RÉSUMÉ ET CONCLUSION

Le dispositif d'évaluation mis en place, centré sur le questionnaire emprunté à PISA 2000, fournit une évaluation fiable des différences d'attitudes entre les élèves (**ρ^2 absolu** = **0.88** > 0.80). L'analyse attire cependant l'attention sur le fait que ces différences sont en partie dépendante de l'appartenance à telle ou telle classe.

Comme souhaité, le dispositif permet de repérer les élèves qui développent une hostilité marquée à l'égard de la lecture en les situant sur le pôle négatif de l'échelle d'attitude. **L'intervalle de confiance** autour du seuil de référence choisi (médian de l'échelle = 2.5 points) est d'environ **0.5 points**. Selon le type d'erreur qu'on accepte de commettre, on retiendra comme candidats à une opération de remotivation à la lecture les élèves ayant un score moyen inférieur à 2 points (2.5 – 0.5) ou à 3 points (2.5 + 0.5), ou on prendra des informations complémentaires dans la zone 2-3 points.

Une **analyse de facette** montre qu'on aurait intérêt à **supprimer l'item 9** qui fonctionne un peu différemment des autres (il porte sur la difficulté à lire plutôt que sur la réticence à lire) ; le coefficient rhô carré absolu s'améliore légèrement, passant de 0.88 à **0.89**. Il serait aussi possible de **raccourcir le questionnaire** en se limitant **à 6 items**. Si l'on supprime par exemple items 2, 6 et 9 (questions c, d et f), le dispositif d'évaluation conserve un degré de fiabilité tout à fait satisfaisant (**0.87**).

Le calcul d'un **coefficient critérié** permet d'attester avec une fiabilité suffisante (**0.88**) que la **moyenne des évaluations** faites par les élèves (2.72) **se situe au-dessus du médian de l'échelle d'attitude** du questionnaire (> 2.5), donc dans une zone d'opinion positive.

On relèvera la richesse des analyses qu'autorise le modèle de la généralisabilité, notamment par rapport au modèle classique d'analyse des tests. Mais cette originalité se marque particulièrement dans les analyses présentées ci-dessous en prolongement à notre exemple introductif.

8. PROLONGEMENTS : AUTRES ANALYSES SUR LES MÊMES DONNÉES

Nous ne donnerons ici qu'un aperçu des résultats d'autres analyses sur les mêmes données pour illustrer les possibilités du modèle de la généralisabilité.

8.1 Évaluer les classes pour repérer celles qui manifestent en moyenne des résultats positifs (scores moyens > 2.5).

Plan de mesure : C/EI

La *face de différenciation* (ce qu'on cherche à mesurer) est représentée par la facette Classes tandis que les Elèves et les Items constituent les deux *facettes d'instrumentation*, c'est-à-dire les moyens permettant d'estimer le niveau des différentes classes.

Tableaux 24 et 25 Différenciation des classes : coefficients de généralisabilité et Moyennes par classe

Coefficient de généralisabilité	relatif	absolu
	0.8209	**0.7458**

C classes	Moyennes
1	2.5000
2	**3.2389**
3	2.7167
4	2.4222

Commentaire : Le *coefficient de généralisabilité absolue* est celui qui nous intéresse puisque nous voulons situer les classes sur l'échelle du test, par rapport au seuil de 2.5 ; ce coefficient (**0.7458**) n'atteint pas tout à fait le seuil de 0.80. On obtiendrait un résultat un peu meilleur, mais encore insuffisant, en supprimant l'item 9 (ρ^2 abs. = 0.7664 ; résultat fourni par une *analyse de facette*). Il faudrait ajouter une douzaine d'items au questionnaire pour atteindre le seuil de 0.80 (ρ^2 abs. = 0.8012 avec 21 items ; résultat fourni par une *analyse d'optimisation*). L'*intervalle de confiance* dans ce dernier cas serait de 0.32 (1.96 * 0.1658); dans notre échantillon, une seule classe (no 2, cf. tableau 25) pourrait être considérée avec une bonne probabilité comme ayant un score moyen supérieur au médian de l'échelle (limite considérée : 2.5 + 0.32 = 2.82).

8. 2 Évaluer les conduites de lecture (ou représentations) **sous-jacentes aux items** pour repérer celles qui sont investies négativement par les élèves (scores moyens < 2.5).

Plan de mesure : I/CE

Tableau 26 Différenciation des items : coefficients de généralisabilité

Coefficient de généralisabilité	relatif	absolu
	0.9259	**0.7442**

Commentaire : Le *coefficient de généralisabilité absolue* (0.7442) est de nouveau inférieur au seuil de 0.80. Différentes possibilités d'amélioration peuvent être explorées :

- avec l'*analyse de facettes* (tableau 27 ci-dessous) : en éliminant une classe comme la classe no 2 qui se présente comme « atypique » ; mais encore faudrait-il identifier

les raisons de son fonctionnement particulier (c'est celle qui obtient le score moyen le plus élevé !) ;

Tableau 27 Analyse de la facette Classe

Facette	Niveau	Coef. rel.	Coef. abs.
C	1	0.90604	0.61679
	2	0.92937	**0.88717**
	3	0.86623	0.55833
	4	0.90197	0.68125

- en calculant différents *plans d'optimisation*, on constate qu'il faudrait augmenter le nombre de classes à 6 au moins (ρ^2 abs. = **0.8147**) ; ce serait la solution la plus judicieuse ; l'*intervalle de confiance* serait alors de 0.33 (1.96 * 0.1667) et aucun item ne pourrait être considéré de façon suffisamment fiable comme situé sur le pôle négatif de l'échelle (2.5 – 0.33 = 2.17 ; tous les items ont des moyennes supérieures à cette valeur ; cf. infra tableau 28).
Pour des raisons évidentes, dans ce cas, on renoncera à vérifier ce que donnerait une *optimisation* par augmentation du nombre d'élèves par classe !

8.3 Modifier le plan d'observation en introduisant une nouvelle facette

L'analyse des moyennes par item (tableau 28) montre par ailleurs une **influence** importante de la **formulation des questions** sur les réponses. Les items dont la formulation est négative par rapport au trait à évaluer (le « plaisir de lire » ; par exemple : *Pour moi, la lecture est une perte de temps*) reçoivent en moyenne une évaluation plus favorable que les autres items : les élèves semblent hésiter à approuver, et surtout à approuver fortement (*tout à fait d'accord*), des propositions qu'ils perçoivent comme « socialement » ou « scolairement incorrectes ».

Tableau 28 Moyennes par item (effet de la formulation positive ou négative)

I+ Items +	Moyennes	I- Items -	Moyennes
1	2.27500	5	2.95000
2	2.27500	6	3.16250
3	2.56250	7	2.62500
4	2.42500	8	3.07500
		9	3.12500

Ce constat incite à regrouper, dans une nouvelle phase de l'analyse[9], les 8 premiers items[10] dans deux *niveaux* (modalités) d'une nouvelle *facette* D (Dimension ou direction de formulation) pour contrôler l'influence de ce facteur sur la fiabilité du dispositif d'évaluation. Dans ce cas, il s'agira d'une *facette fixée* : on ne peut chercher à généraliser au-delà des deux niveaux définis : formulations positive et négative des items ; le nombre de niveaux de l'*univers* de référence est identique à celui des niveaux observés (= 2). Les nouveaux *plans d'observation* et *d'estimation* comportent alors 4 *facettes* (cf. tableau 29). Une nouvelle analyse doit alors être construite, impliquant l'importation du fichier correspondant au plan d'observation défini et comportant donc 640 données (4 classes x 20 élèves x 2 dimensions x 8 items).

[9] Cette nouvelle analyse suppose un nouveau fichier ne comprenant que 8 colonnes (correspondant aux 8 items) et la définition de nouveaux plans d'observation, d'estimation et de mesure.

[10] Nous avons vu ci-dessus que l'on améliore la généralisabilité du dispositif en supprimant l'item 9.

Tableau 29 Plans d'observation et d'estimation avec 4 facettes

Facettes	Niveaux	Univers	Commentaires
C	4	200	Classes
E:C	20	INF	Elèves par classe
D	2	2	Dimension de formulation
I :D	4	INF	Items par dimension

Commentaires:
- la nouvelle facette D est fixée (cf. supra) ; de ce fait, elle n'induira pas d'erreur de mesure sur la face d'instrumentation ;
- les Items sont nichés (inclus) dans la facette Dimensions ; on fait l'hypothèse qu'ils sont extraits aléatoirement de 2 pools d'items pratiquement infinis.

Sur la base des niveaux *plans d'observation et d'estimation*, on peut définir quatre *plans de mesure* visant à différencier tour à tour les Elèves, les Classes, les Dimensions et les Items.

Nous en donnons ci-dessus, en bref, les principaux résultats.

8.3.1 Evaluer les Elèves

Plan de mesure : **EC/DI**[11]

Noter que la *face d'instrumentation* (à droite de la barre de fraction) est composée de la Dimension de formulation et des Items qu'elle inclut.

Tableau 30 Coefficients de généralisabilité pour le plan de mesure EC/DI

Coefficient de généralisabilité pour EC/DI	relatif	absolu
	0.9183	**0.9128**
Cf. coefficients de GT pour EC/I (9 items; tableau 8 supra)	0.9040	0.8826

On constate une légère amélioration des coefficients de GT par rapport au dispositif d'évaluation précédent (tableau 8), amélioration attribuable notamment à la suppression de l'item 9 et à la neutralisation des erreurs dues à la formulation (positive – négative) des items du fait que la facette D (sur la face d'instrumentation) est fixée.

8.3.2 Evaluer les Classes

Plan de mesure : **C/EDI**

La *face d'instrumentation* est composée dans ce cas des Elèves ainsi que de la Dimension de formulation et des Items inclus dans ses deux niveaux.

Tableau 31 Coefficients de généralisabilité pour le plan de mesure C/EDI

Coefficient de généralisabilité pour C/EDI	relatif	absolu
	0.8372	**0.8171**
Cf. coefficients de GT pour C/EI (9 items; tableau 22 supra)	0.8209	0.7458

[11] Rappelons qu'à l'intérieur de chacune de deux faces de différenciation et d'instrumentation (donc de chaque côté de la barre de fraction), l'ordre des facettes n'a pas d'importance.

La modification du dispositif d'évaluation améliore nettement le *coefficient de généralisabilité absolue*, neutralisant notamment une partie de l'erreur due à la facette Items (cf. les différences d'évaluation selon que les items sont formulés positivement ou négativement ; tableau 28). L'*intervalle de confiance* est alors de 0.34 ; une seule classe (no 2, cf. tableau 25) pourrait être considérée avec une bonne probabilité comme ayant un score moyen supérieur au médian de l'échelle (limite considérée : 2.5 + 0.34 = 2.84).

8.3.3 Evaluer les Items

Plan de mesure : **ID/EC**

La *face de différenciation* (à gauche de la barre de fraction) est composée dans ce cas des Items et de la Dimension de formulation dans lesquels ils sont inclus ; les « instruments » d'évaluation sont alors les Elèves inclus dans les Classes.

Tableau 32 Coefficients de généralisabilité pour le plan de mesure DI/EC

Coefficient de généralisabilité pour DI/EC	relatif	absolu
	0.9443	**0.7056**
Cf. coefficients de GT pour I/EC (9 items; tableau 22 supra)	0.9259	0.7442

Dans ce cas, le nouveau dispositif n'améliore pas le coefficient de généralisabilité absolue, sa modification n'affectant pas les erreurs absolues. La principale de ces erreurs reste les différences moyennes d'attitudes entre Classes (cf. § 8.2). La diminution des deux coefficients tient en particulier à la suppression de l'item 9, qui apportait sa contribution à la variance des Items sur la face de différenciation (c'était un des items qui recueillaient les évaluations les plus élevées ; cf. tableau 28)

8.3.4 Evaluer la Dimension de formulation

Plan de mesure : **D/IEC**

La question correspondant à ce *plan de mesure* consiste à se demander s'il est possible de généraliser les conclusions tirées à partir des différences entre dimensions. En d'autres termes, va-t-on retrouver des différences analogues entre groupes d'items formulés positivement et négativement dans d'autres questionnaires d'attitudes construits selon le même principe ?
La *face d'instrumentation* (à droite de la barre de fraction) est composée dans ce cas des Items et des Elèves dans les Classes.

Tableau 33 Coefficients de généralisabilité pour D/IEC

Coefficient de généralisabilité	relatif	absolu
	0.8874	**0.6125**

On peut conclure que la formulation négative des items, dans un questionnaire de ce type, induit des évaluations supérieures à celles d'items formulés positivement (coefficient de GT relatif = **0.8874**), sans qu'on puisse situer de façon fiable sur l'échelle d'évaluation la moyenne des items pour chacune des deux formulations (coefficient absolu = **0.6125** < 0.80).

8.3.5 Conclusion sur la modification des plans d'observation et d'estimation

Cet aménagement du dispositif de mesure se révèle favorable quand on cherche à évaluer les élèves et particulièrement les classes, neutralisant une partie de l'erreur absolue affectant les Items. Il atteste en outre une influence de la formulation sur le niveau de l'évaluation, à prendre en compte dans la construction du questionnaire.

PROLONGEMENTS : RÉSUMÉ ET CONCLUSION

L'originalité du modèle de la généralisabilité (GT) tient notamment au fait qu'il permet d'évaluer la fiabilité de dispositifs que ne peuvent traiter d'autres modèles. Les analyses révèlent en outre qu'avec un même dispositif (un instrument de mesure avec des caractéristiques particulières appliqué à un échantillon d'une certaine structure) **le degré de généralisabilité varie selon l'objet de mesure** : une évaluation peut être fiable quand elle porte sur des élèves et l'être moins (voire pas du tout) quand on estime des moyennes de classes ou d'items (comparer les coefficients absolus dans les analyses de GT des tableaux 8, 24 et 26)

Dans cette dernière section, la GT nous a permis de contrôler la qualité docimologique de notre dispositif lorsqu'il s'agit d'évaluer les **Classes** et les **Items** opérationalisant diverses conduites de lecture. Dans les deux cas, le **coefficient relatif supérieur à 0.80** a montré qu'il était possible de **hiérarchiser de façon fiable les objets d'évaluation**, classes ou items, mais **pas de les situer** avec une assurance suffisante **sur l'échelle de mesure du questionnaire**. Le détail des analyses a suggéré pour chaque cas des solutions différentes pour **améliorer le dispositif en réduisant les facteurs d'erreur absolue** qui affectent la mesure.

S'il s'agit de différencier les **classes**, par exemple pour distinguer celles qui se situent nettement sur le pôle positif de l'échelle (score moyen > 3), la solution la plus judicieuse consisterait à supprimer l'item 9 et à définir une facette Dimension (ou direction) de formulation dont les deux modalités regrouperaient en nombres équivalents d'une part les items positifs et d'autre part les items négatifs. En introduisant une telle facette fixée, on neutralise une partie de la variance d'erreur absolue.

D'autre part, une évaluation fiable de l'attitude des élèves à l'égard des diverses **conduites de lecture**, opérationalisées par les **items**, supposerait que l'on augmente l'échantillon de test à au moins 6 classes pour atteindre un coefficient absolu satisfaisant et situer les conduites de façon généralisable sur les degrés de l'échelle du test.

L'introduction dans le dispositif de la facette **Dimension de formulation** et son analyse attestent enfin qu'on peut mesurer de façon fiable l'influence de cette variable sur les différences moyennes d'évaluation (cf. **coefficient relatif = 0.8874**) selon que les items sont formulés positivement ou négativement. Un tel facteur devrait donc être pris en compte lors de l'élaboration d'un questionnaire analogue à celui analysé ici.

D. Bain

avril 2004