ANALYSE DE GÉNÉRALISABILITÉ SUR LE QUESTIONNAIRE D'ÉVALUATION DU COLLOQUE ADMEE/SSRE 2002

Le présent article s'adresse en priorité aux personnes qui ont participé au cours sur la généralisabilité qui a précédé immédiatement le colloque de l'ADMÉÉ/SSRE les 3 et 4 septembre 2002 et aux participants qui ont suivi le symposium sur le même thème lors du colloque, ainsi qu'à tous ceux qui ont eu l'occasion de s'initier d'une façon ou d'une autre à ce modèle statistique. Ce texte pourra aussi intéresser les chercheurs qui souhaiteraient avoir une rapide idée des possibilités d'utilisation de la théorie de la généralisabilité pour l'analyse et la mise au point d'instruments de mesure tels que tests, échelles d'évaluation ou questionnaires. Pour des commentaires sur les résultats de l'enquête du point de vue des responsables du colloque, on se reportera à l'article de A. Blanchet (2003).

Introduction

Pour illustrer les diverses possibilités d'analyses qu'offrent le modèle de la généralisabilité, nous partons de l'hypothèse suivante :

Il est intéressant pour de futurs organisateurs de colloques ou de congrès d'obtenir des informations
- sur la fiabilité d'un questionnaire comme celui qui a été utilisé pour le colloque de l'ADMÉÉ à Lausanne en septembre 2002 (cf. annexe 1), en fonction de différents objectifs d'évaluation, notamment
  - pour estimer le degré global de satisfaction – insatisfaction des participants (éventuellement compte tenu de certaines de leurs caractéristiques) ;
  - pour distinguer des aspects ou des domaines de l'organisation et du déroulement de la manifestation pour lesquels le degré de satisfaction est plus ou moins élevé ;
- sur d'éventuelles améliorations à apporter aux dispositifs d'évaluation visant ces objectifs.

Rappelons que la généralisabilité s'applique à des instruments de mesure comportant différents items qui s'additionnent pour donner un total (traité par le modèle comme une moyenne des items)[1]. En l'occurrence, le questionnaire

[1] Pour une présentation du modèle de la généralisabilité (hors de propos dans cet article), on se reportera aux publications suivantes : Cardinet et Tourneur, 1985 ; Bain & Pini, 1996, ainsi qu'au numéro spécial de la revue *Mesure et évaluation en éducation* de l'ADMÉÉ (Cardinet, 2003). Pour un résumé non technique de ces analyses, voir l'article de Bain (2003a).

d'évaluation peut être considéré comme évaluant par son total un certain *degré de satisfaction – insatisfaction* à l'égard du colloque, chaque item représentant un aspect (une raison partielle) de cette satisfaction – insatisfaction. Il n'est peut-être pas habituel de calculer un total sur un tel questionnaire, et cet objectif n'est généralement pas *a priori* celui des organisateurs de colloques. Pourtant, il est très probable que les responsables de telles manifestations évaluent implicitement une satisfaction globale en constatant que les profils de réponses se situent dans l'ensemble plutôt à gauche ou à droite (proches du pôle positif ou négatif) de l'échelle du questionnaire. On peut donc faire un pas de plus en quantifiant un degré de satisfaction – insatisfaction : on attribue des valeurs numériques aux modalités de réponse, situées sur une échelle d'évaluation (*rating scale*) de type Likert, et on additionne les scores de chaque item.

Pour faciliter l'interprétation des résultats, il nous a semblé préférable de coder les 4 modalités du questionnaire analysé (reproduit en annexe 1) en inversant le sens habituel (de gauche à droite) de l'échelle et en adoptant les codes suivants : *Très positive* (*Très fort*) = 4, *Plutôt positive* (*Fort*) = 3, *Plutôt négative* (*Faible*) = 2, *Très négative* (*Très faible*) = 1.

Pour les besoins de notre étude et pour satisfaire aux contraintes du modèle et du logiciel de généralisabilité, nous avons par ailleurs sélectionné dans le fichier des données constitué par l'*Unité de recherche pour le pilotage des systèmes pédagogiques*[2] (URSP, Lausanne) :

- les 8 items à choix multiple auxquels la grande majorité des répondants avaient répondu ; nous avons ainsi laissé de côté les deux questions souvent négligées par les participants (nombreuses non-réponses), soit les 2 items relatifs respectivement à la traduction des interventions et aux posters ;
- les questionnaires remplis complètement en ce qui concerne les 8 items retenus ; lorsqu'une seule réponse manquait, nous avons remplacé la donnée manquante par une valeur approximée à partir d'une régression multiple sur les 7 autres réponses ; nous avons ainsi retenu dans un premier temps les réponses de 93 participants.

Pour affiner et compléter notre étude, nous avons distingué deux parties (deux dimensions de contenu) dans le questionnaire, soit

- les 4 premiers items retenus, qui portent sur l'organisation matérielle du congrès : forme, modalités d'organisation, déroulement, intendance ;
- les 4 derniers, relatifs à la thématique et à sa présentation sous différentes formes de communication : conférences, ateliers et symposiums,

l'objectif étant de vérifier le rôle que pourrait jouer cette facette *Dimensions du questionnaire* dans la généralisabilité des dispositifs étudiés. Au départ, nous avons opéré cette distinction avec une intention méthodologique, à titre de

[2] Nous remercions l'URSP, et plus particulièrement son directeur, M. Alex Blanchet, d'avoir mis à notre disposition les données de cette enquête. L'analyse que nous faisons d'un échantillon de ces données, aménagées par nous (cf. infra), n'engage cependant que nous.

démonstration, pour illustrer les possibilités du modèle de la généralisabilité en complexifiant les *facettes* (facteurs) du dispositif d'évaluation ; d'où un certain arbitraire dans cette dichotomie. On verra à l'analyse qu'une telle différenciation a une certaine pertinence. Nous indiquons en annexe 1, sur le questionnaire (colonne de droite), la nouvelle numérotation des items retenus.

En outre, il nous a paru intéressant de contrôler l'influence que pouvaient exercer sur la mesure la connaissance et la pratique que les participants avaient du thème traité par le congrès. Nous avons donc regroupé les répondants selon qu'ils déclaraient *avoir participé ou non à une démarche qualité dans l'un de leurs emplois* (cf. dernier item du questionnaire). Le modèle exigeant un plan équilibré, nous avons constitué, de façon aléatoire, deux groupes égaux de 41 participants ; plus précisément, nous avons conservé les 41 questionnaires émanant de personnes sans expérience de la démarche qualité, tandis que dans l'autre groupe (n = 52) nous avons éliminé au hasard 11 enregistrements. Des questionnaires récoltés nous avons donc finalement conservé 82 ; un rapide contrôle sur la moyenne des appréciations par item nous a montré que cet échantillon pouvait être considéré globalement comme représentatif de l'ensemble de départ.

Les *plans d'observation et d'estimation* qui résultent de notre démarche préparatoire sont présentés dans le tableau 1 ci-dessous. Nous avons considéré les participants comme constituant un *échantillon aléatoire* de l'ensemble, estimé *infini*, de personnes acceptant de répondre à un tel questionnaire à la fin d'un colloque de ce genre. On pourrait admettre que c'est un hasard d'avoir choisi cet échantillon et qu'il est donc extrait de la population des gens qui participent aux colloques de ce type, qui, comme on le sait, sont innombrables (sic !). On peut également postuler que les questions ont le même statut : elles sont extraites du grand nombre de questions (cf. contenus et formulations) que l'on aurait pu poser à l'occasion d'un tel colloque.

Tableau 1 Plans d'observation et d'estimation

Facettes	Niveaux	Univers	Nom
E	**2**	**2**	**expérience (1=oui, 2=non)**
P:E	**41**	**INF**	**participants**
D	**2**	**2**	**dimensions du questionnaire**
I:D	**4**	**INF**	**items**

Les données ainsi définies et organisées vont nous permettre d'analyser la fiabilité de quatre différents dispositifs d'évaluation, portant sur quatre objets d'études et visant à mesurer :

1. le degré de satisfaction des *participants* (P:E) ;
2. l'influence de l'*expérience* d'une démarche qualité sur cette satisfaction (E) ;

3. les différences de satisfaction selon les divers domaines abordés par les huit *items* (I:D) ;
4. les différences selon les deux parties ou *dimensions* du questionnaire : organisation matérielle et modalités de communication (D).

1. Mesurer le degré de satisfaction des répondants ; différencier les participants (plan de mesure : EP/DI)

Pourrait-on distinguer de façon fiable différents degrés de satisfaction, en admettant que cela puisse être un des objectifs (éventuellement dérivé) de l'instrument ? Dans le cas d'un tel colloque, la question de la fiabilité de la mesure se poserait notamment si l'on cherchait à identifier un sous-ensemble de participants présentant un degré de satisfaction moindre, par exemple inférieur à 3.0 (appréciation *plutôt positive*) sur l'échelle du questionnaire.

Tableau 2 Analyse de généralisabilité pour le plan de mesure EP/DI[3]

Sources de variance	Variance de dif.	Sources de var.	Var. d'err. relative	Var. d'err. absolue	%
E	0.004195				
P:E	0.051823				
		D		0.000000	0
		I:D		0.002509	7.7
		ED	0.000000	0.000000	0
		EI:D	0.000390	0.000390	1.2
		PD:E	0.000000	0.000000	0
		PI:ED	0.029684	0.029684	91.1
Totaux	0.056018		0.030074	0.032584	
Ecarts types	0.236682		0.173420	0.180510	

Coefficient de généralisabilité	relatif	absolu
	0.650673	0.632245

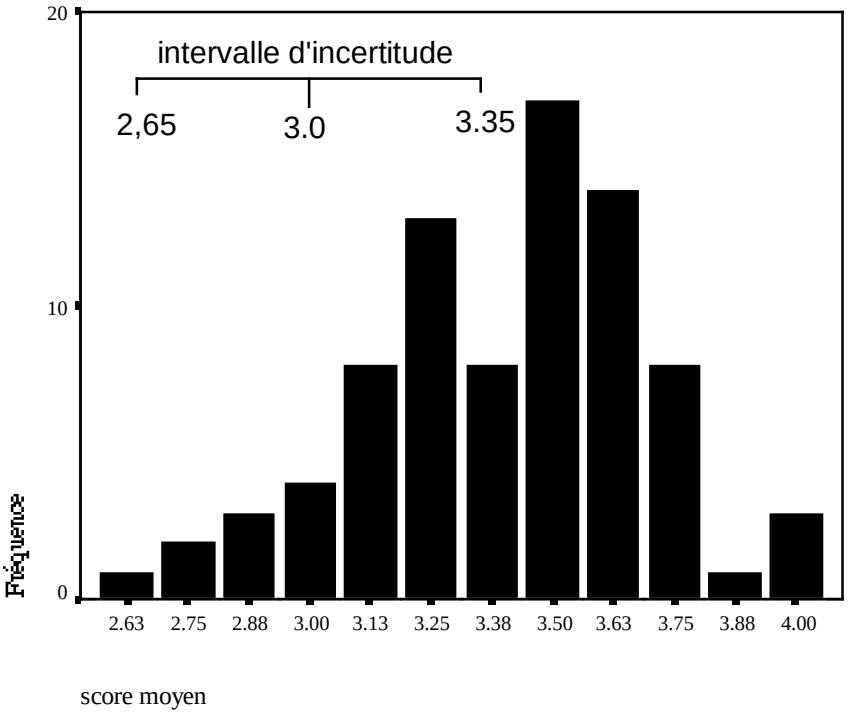
Sans entrer dans le détail des analyses[4], nous constatons que le dispositif ne permet pas tel quel d'assurer la mesure envisagée : les *coefficients de généralisabilité relative et absolue* (resp. 0.65 et 0.63) sont en dessous du seuil de 0.80 généralement considéré comme satisfaisant. Si l'on calcule l'*intervalle de confiance* (ou d'incertitude) sur la mesure (en l'occurrence à partir de l'écart-type de l'*erreur absolue* : 1.96 * 0.18051 = 0.35) et qu'on l'applique au seuil de 3.0, on s'aperçoit qu'on ne peut pas identifier avec suffisamment d'assurance un

[3] La facette P étant incluse dans la facette E, cette dernière figure sur la face de différenciation (à gauche de la barre de fraction) dans le plan de mesure.
[4] Cf. en annexe 2 l'analyse de variance, valable pour les 4 dispositifs testés dans cet article, et pour les 4 analyses de généralisabilité fournies ci-dessous.

groupe de participants moins satisfaits que les autres (appréciation moyenne < 3.0) : à une exception près (score de 2.625), tous les participants se situent au-dessus de la limite inférieure de l'intervalle d'incertitude 2.65 (3.0 – 0.35 ; cf. infra figure 3).

Quelles améliorations devrait-on / pourrait-on alors apporter au dispositif, si l'on jugeait important d'évaluer plus finement un degré global de satisfaction ? Pour répondre à cette question, il s'agit d'abord d'identifier les principales causes de cette médiocre généralisabilité. Une première raison est tout à l'honneur des organisateurs du congrès : les appréciations se regroupent assez étroitement (écart-type de seulement 0.29) autour d'une moyenne élevée : 3.4 sur un maximum de 4. La plupart des participants ont en effet coché à la plupart des questions les modalités *Très positive* (*Très fort*) = 4, *Plutôt positive* (*Fort*) = 3. La *variance de différenciation* (au numérateur de la formule de la généralisabilité) est donc relativement faible.

Figure 3 Distribution des scores moyens et intervalle d'incertitude



Dans ces conditions, une distinction fine entre les degrés de satisfaction ne serait possible que si l'erreur de mesure était relativement faible également, ce qui n'est pas le cas. La principale source d'erreur (relative et absolue) est constituée par l'interaction entre les facettes (facteurs) Participants et Items

(PI:ED = 91% de la *variance d'erreur absolue*). Elle traduit notamment une appréciation très variable d'un répondant à l'autre selon les items[5].

On peut soupçonner que le caractère relativement hétérogène des questions posées[6] explique la difficulté à obtenir une mesure fiable du degré global de satisfaction. De ce point de vue, certains items contribueraient-ils plus que d'autres à l'erreur, par exemple l'item 2.1, qui exprime une opinion globale : *Quel a été votre intérêt pour la thématique de la qualité?* Gagnerait-on à le supprimer ? Pour pouvoir le vérifier, nous avons appliqué une *analyse de facette* aux huit items retenus, sans considérer leur regroupement dans les 2 dimensions distinguées par rapport au contenu des questions : le logiciel que nous avons utilisé (*EduG*) ne permet en effet pas d'effectuer une telle analyse sur une *facette nichée* (incluse) dans une autre. Cette analyse (non reproduite ici) montre clairement que l'on ne peut incriminer un item plutôt qu'un autre ; la suppression de la question 2.1, par exemple, n'apporterait quasiment aucune amélioration à la fiabilité du dispositif (gain de 0.01 sur chacun des coefficients de généralisabilité).

Tableau 4 Analyse d'optimisation pour le plan de mesure EP/DI

	Plan original		Opt 1		Opt 2		Opt 3	
	Niv.	Univ.	Niv.	Univ.	Niv.	Univ.	Niv.	Univ.
E	2	2	2	2	2	2	2	2
P	41	INF	41	INF	41	INF	41	INF
D	2	2	2	2	2	2	2	2
I	4	INF	8	INF	9	INF	10	INF
Observations	656		1312		1476		1640	
Coeff. Rel.	0.650673		0.788373		0.807357		0.823216	
Coeff. Abs.	0.632245		0.774694		0.794585		0.811250	
Var. Err. Rel.	0.030074		0.015037		0.013366		0.012030	
Err. Typ. Rel.	0.173420		0.122626		0.115613		0.109680	
Var. Err. Abs.	0.032584		0.016292		0.014482		0.013034	
Err. Typ. Abs.	0.180510		0.127640		0.120340		0.114165	

La seule amélioration envisageable serait de minimiser l'erreur due à l'interaction PI:ED en augmentant le nombre d'items tirés du même *univers* de contenu. La procédure d'*optimisation* montre qu'il faudrait plus que doubler la longueur du questionnaire, en ajoutant 5 à 6 items par dimension (tableau 4) pour atteindre le seuil de 0.80 (généralisabilité absolue). Sans chercher si une telle solution serait pratiquement envisageable, on en conclura plutôt que les questionnaires d'évaluation habituellement utilisés en fin de colloque (avec peu d'items pour ne pas décourager les répondants) sont probablement trop courts et

[5] Rappelons que cette interaction (Participants x Items) est confondue avec la variance d'erreur résiduelle provenant de toutes les sources d'influence non identifiées.

[6] Confirmé par une analyse factorielle en composantes principales sur les 8 items : le facteur général ne représente que le tiers de la variance totale et le coefficient de Kaiser-Meyer-Olkin mesurant l'adéquation de l'échantillonnage n'est que de 0.67.

souvent trop hétérogènes par leur contenu pour mesurer de façon fiable un *taux global* de satisfaction (habituellement, ce n'est pas non plus leur objectif).

En revanche, en l'état, le questionnaire autorise tout de même un constat intéressant et encourageant : il permet de vérifier qu'en moyenne, les appréciations individuelles des participants peuvent être situées avec une fiabilité suffisante dans la zone positive de l'échelle du questionnaire, au-dessus du seuil de 3 sur 4. Ceci est attesté par le calcul du *coefficient critérié Phi(lambda) de Brennan*. Comparant la moyenne générale de 3.4 au seuil-critère de 3.0, le calcul de ce coefficient donne une valeur nettement supérieure à 0.80, soit 0.87.

2. Mesurer la différence de satisfaction entre deux catégories de participants (expérience *vs* inexpérience de la démarche qualité) ; différencier les niveaux d'expérience (plan de mesure : E/PDI)

Comme on pouvait s'y attendre, dans notre échantillon, la satisfaction des participants ayant déjà pratiqué professionnellement la démarche qualité est moindre (moyenne 3.3) que celle des répondants profanes dans le domaine (3.5). Mais, pour ce plan de mesure, la *variance de différenciation* est trop faible par rapport à l'erreur dont elle est affectée pour qu'on considère que le dispositif permet d'assurer une telle mesure. C'est ce qu'atteste le *coefficient de généralisabilité relative* de 0.64, donc nettement inférieur à la valeur de référence 0.80 (tableau 5).

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

Sources de variance	Variance de dif.	Sources de var.	Var. d'err. relative	Var. d'err. absolue	%
E	0.004195				
		P:E	0.001264	0.001264	25.9
		D		0.000000	0
		I:D		0.002509	51.3
		ED	0.000000	0.000000	0
		EI:D	0.000390	0.000390	8
		PD:E	0.000000	0.000000	0
		PI:ED	0.000724	0.000724	14.8
Totaux	0.004195		0.002378	0.004887	
Ecarts types	0.064771		0.048766	0.069911	

Coefficient de généralisabilité	relatif	absolu
	0.638221	0.461892

Cela signifie notamment que le programme et les prestations offertes ont su satisfaire, à quelques nuances près, à la fois les participants expérimentés et les inexpérimentés. Pour différencier avec plus de fiabilité les deux groupes, il

faudrait en particulier plus d'homogénéité dans les réponses aux items (composante d'erreur I:D ; cf. tableau 5) et dans les appréciations des participants de chaque catégorie (composante P:E). Une *optimisation* du dispositif dans cette perspective de mesure apparaît quasiment impossible : il faudrait un nombre invraisemblable d'items (plusieurs centaines)… et susciter les réponses d'un plus grand nombre de participants (au moins 130 pour atteindre un coefficient relatif de 0.80). Pratiquement, on ne voit guère quels seraient la possibilité et surtout l'intérêt d'une telle opération.

3. Mesurer le taux de satisfaction selon les aspects de l'organisation et du déroulement du colloque investigués ; différencier les items du questionnaire (plan de mesure : DI/EP)

Dans la plupart des colloques, le questionnaire d'évaluation finale vise avant tout à mettre en évidence les aspects de l'organisation et du déroulement de la manifestation qui ont, selon les participants, plus ou bien marché. Avec parfois l'intention de faire mieux la prochaine fois dans les domaines qui ont reçu le taux de satisfaction le plus faible. Ce faisant, on postule qu'on peut différencier de façon fiable (généralisable) les valeurs moyennes obtenues pour les différents items du questionnaire. Toutefois, il vaudrait mieux le vérifier avant de décider qu'on a moins bien réussi (ou qu'on devrait faire mieux) dans tel domaine que dans tel autre.

Tableau 6 Analyse de généralisabilité pour le plan de mesure DI/EP[7]

Sources de variance	Variance de dif.	Sources de var.	Var. d'err. relative	Var. d'err. absolue	%
		E		0.000000	0
		P:E		0.000632	17.6
D	0.044583				
I:D	0.020075				
		ED	0.000000	0.000000	0
		EI:D	0.000000	0.000000	0
		PD:E	0.000065	0.000065	1.8
		PI:ED	0.002896	0.002896	80.6
Totaux	0.064658		0.002961	0.003593	
Ecarts types	0.254279		0.054411	0.059938	

Coefficient de généralisabilité	relatif	absolu
	0.956217	0.947362

[7] La facette Items étant incluse dans la facette Dimension, cette dernière figure sur la face de différenciation, soit à gauche de la barre de fraction dans la formule du plan de mesure.

Dans le cas du questionnaire analysé ici, on constate que ce type d'évaluation apparaît fiable (tableau 6) : les *coefficients de généralisabilité relative et absolue* sont nettement supérieurs à 0.80 (0.96 et 0.95). L'*analyse de généralisabilité* (tableau 6, colonne *Variance de dif.)* et une rapide inspection des moyennes aux différents items (tableau 7) montrent aussi que le facteur Dimension du questionnaire (organisation générale *vs* modalités de communication) joue un rôle non négligeable dans les différences entre items : ce facteur rend compte en effet des deux tiers de la variance de différenciation (0.044583/0.064658) ; la satisfaction exprimée est un peu moins élevée quand on considère les différentes formes de communication : conférences, ateliers, symposiums. Nous y reviendrons au chapitre suivant.

Tableau 7 Moyennes aux 8 items du questionnaire, regroupés selon deux dimensions

D dimensions du quest.	I items	Moyennes	Questions sur
1	1	3.536585	la forme du congrès
1	2	3.768293	l'organisation du congrès
1	3	3.646341	le déroulement du congrès
1	4	3.536585	l'intendance
2	1	3.378049	l'intérêt du répondant pour la thématique
2	2	3.304878	les conférences
2	3	3.000000	les ateliers
2	4	3.060976	les symposiums
Moyenne générale :		3.403963	

Si l'on veut comparer les items les uns aux autres, il faut tenir compte d'un intervalle d'incertitude, calculé à partir de l'écart-type de l'erreur considérée[8]. Comme on cherche à situer les valeurs moyennes sur l'échelle du questionnaire (1 à 4), on considérera pour cela l'erreur absolue. Il faudra alors une différence d'au moins 0.17 (1.96 * √2 * 0.059938) pour qu'on considère une supériorité ou infériorité dans les taux de satisfactions manifestés à l'égard de tel aspect de l'organisation par rapport à tel autre.

On peut finalement distinguer différentes plages (pas vraiment disjointes) dans les degrés de satisfactions. En tête, les appréciations relatives à l'organisation du congrès et à son déroulement (3.8-3.6) ; relativement proches, les évaluations portant sur sa forme et son intendance (3.5). Puis, un peu en dessous de la moyenne générale, l'intérêt pour la thématique et les conférences (3.4-3.3). Les deux autres formes de communications proposées, ateliers et symposiums, ont eu un peu moins de succès auprès des répondants ; elles se

[8] Dans le cas de la comparaison de deux items, on utilise la formule : 1.96 * √2 * écart-type de l'erreur considérée.

détachent quelque peu des autres domaines évalués (3.1-3.0), tout en restant en moyenne dans la zone positive. Elles devraient faire l'objet d'une analyse complémentaire, à partir des commentaires à ce sujet, afin de saisir les principales réserves des participants (cf. à ce sujet Blanchet, 2003).

4. Mesurer les différences d'appréciation entre les deux parties du questionnaire ; différencier les dimensions du questionnaire (plan de mesure : D/EPI)

La répartition des items analysée à l'instant révèle une différence d'appréciation entre les deux parties ou dimensions du questionnaire. Dans notre échantillon, les moyennes pour ces deux parties sont respectivement de 3.62 et 3.19. Les aspects généraux de l'organisation (aspects matériels ou administratifs, forme, déroulement, intendance) semblent l'emporter sur l'intérêt de la thématique et les différentes modalités de communication. Mais le dispositif permet-il une mesure fiable de cette différence d'appréciation ? L'*analyse de généralisabilité* l'atteste par des *coefficients de généralisabilité relative et absolue* supérieurs au seuil de référence de 0.80, soit respectivement 0.88 et 0.87 (tableau 8).

Tableau 8 Analyse de généralisabilité pour le plan de mesure D/EPI

Sources de variance	Variance de dif.	Sources de var.	Var. d'err. relative	Var. d'err. absolue	%
		E		0.000000	0
		P:E		0.000632	9.8
D	0.044583				
		I:D	0.005019	0.005019	77.9
		ED	0.000000	0.000000	0
		EI:D	0.000000	0.000000	0
		PD:E	0.000065	0.000065	1
		PI:ED	0.000724	0.000724	11.2
Totaux	0.044583		0.005807	0.006439	
Ecarts types	0.211146		0.076206	0.080245	

Coefficient de généralisabilité	relatif	absolu
	0.884753	0.873794

On peut donc s'attendre à ce que, lors d'un autre ou d'un prochain colloque, un même dispositif permette de vérifier la présence ou l'absence d'un décalage du même ordre entre les deux parties du questionnaire.

Résumé et brève conclusion

Comme on le constate souvent avec des dispositifs analogues[9], un instrument d'évaluation n'est pas « bon à tout faire ». Avec un questionnaire semblable à celui analysé dans cet article, on peut s'attendre à ce qu'il ne se prête guère à évaluer de façon fiable le degré global de satisfaction des participants. Celui que nous avons examiné n'avait d'ailleurs pas été prévu à cet effet. Si c'était vraiment le but principal de l'opération, on risquerait de se heurter au même problème que celui rencontré dans notre étude : le fait que les jugements des répondants peuvent être relativement hétéroclites, c'est-à-dire différents d'un aspect évalué à l'autre. Cette interaction Participants x Items se manifeste par des profils de réponses très divers, même quand on fixe les dimensions du questionnement (ici : qualité de l'organisation générale *vs* modalités de communication). En outre, la mesure fiable de différences de satisfaction globale entre des catégories de participants supposerait, comme nous l'avons vu, des réactions plus contrastées entre les groupes et plus d'homogénéité de réactions aux divers items. Ce ne sera probablement le cas que si certains groupes se sentent plus spécifiquement et systématiquement frustrés, ce qu'un organisateur ne peut guère souhaiter et qu'il cherchera activement à prévenir !

En revanche, le questionnaire semble se prêter relativement bien à une des utilisations très habituelles d'une enquête organisée à la fin d'un colloque : identifier les aspects de l'organisation qui ont plus ou moins satisfait les participants. Comme on l'a constaté dans notre exemple, les domaines suscitant un peu plus souvent des réserves sont les modalités d'organisation qui échappent en partie aux principaux responsables d'un colloque ou d'un congrès parce que ces derniers doivent en déléguer partiellement la gestion à certains participants : nous faisons allusion en l'occurrence à la mise sur pied des symposiums et des ateliers.

Le questionnaire d'évaluation proposé à la fin d'un colloque pourrait tenir du rituel social, fournissant un reflet gratifiant aux organisateurs (qui ont généralement bien mérité cette gratification) ou un exutoire aux récriminations de participants frustrés. Nous espérons avoir montré grâce à l'analyse de la généralisabilité qu'il peut également fonctionner, à certaines conditions, comme un instrument de mesure, utile à l'occasion pour mieux apprécier le fonctionnement d'une manifestation ou éventuellement tirer des conclusions pour un prochain colloque.

[9] Cf. par exemple Bain, 2003b, ainsi que la base de données rassemblée par l'auteur et le groupe Edumétrie de la SSRE (à disposition sur demande à daniel.bain@bluewin.ch ou à l'adresse : route Moulin-Roget 49, CH-1237 Avully.

Daniel Bain

49, route du Moulin-Roget
CH-1237 Avully
daniel.bain@bluewin.ch

Janvier 2003 / révision mai 2004

Références bibliographiques

Bain, D. (2003a). Le questionnaire d'évaluation d'un colloque : instrument rituel ou instrument de mesure ? *Bulletin de l'ADMÉÉ-Europe* n° 2002/3 et 2003/1, 9-10.

Bain, D. (2003b). Généralisabilité et séquence didactique : illustration et défense d'un modèle à vocation édumétrique. *Mesure et évaluation en éducation, 26* (1-2), 19-36.

Bain, D. & Pini, G. (1996). *Pour évaluer vos évaluations - La généralisabilité: mode d'emploi.* Genève : Centre de Recherches Psychopédagogiques/DIPCO, Direction générale du Cycle d'orientation.

Blanchet, A. (2003). Compte rendu sur les évaluations du colloque de l'ADMEE 2002. *Bulletin de l'ADMÉÉ-Europe n° 2002/3 et 2003/1*, 2-3.

Cardinet, J. (Ed.) (2003). Numéro spécial sur la Généralisabilité : « Que valent nos mesures ? ». *Mesure et Évaluation en Éducation, 26*, (1-2).

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

ANNEXE 1

Évaluation du colloque de l'ADMEE / congrès SSRE

Lausanne, septembre 2002

Les organisateurs souhaitent connaître votre avis sur le colloque à propos des points suivants:

Quelle est votre appréciation sur:	Très positive	Plutôt positive	Plutôt négative	Très négative	*No d'item*
• la forme du congrès (équilibre entre conférences et ateliers/symposiums?	£	£	£	£	*1.1*
• l'organisation du congrès (inscription, accueil, documentation)?	£	£	£	£	*1.2*
• le déroulement du congrès (informations, salles, horaire)?	£	£	£	£	*1.3*
• la traduction des interventions?	£	£	£	£	- *
• l'intendance (pour les pauses, repas, moments festifs) ?	£	£	£	£	*1.4*

	Très fort	Fort	Faible	Très faible	
Quel a été votre intérêt pour la thématique de la qualité?	£	£	£	£	*2.1*

Globalement, quelle est votre appréciation sur:	Très positive	Plutôt positive	Plutôt négative	Très négative	
• les conférences?	£	£	£	£	*2.2*
• les ateliers?	£	£	£	£	*2.3*
• les symposiums?	£	£	£	£	*2.4*
• les posters?	£	£	£	£	- *

En quoi les travaux du colloque ont-ils modifié votre connaissance de la problématique «qualité» et votre avis à son égard?

* Item non retenu dans les analyses de généralisabilité (nombreuses non-réponses) ; la numérotation a été ajoutée par nous.

Je suis de nationalité:

Je suis professeur d'université, chercheur, formateur d'enseignants, enseignant primaire, enseignant secondaire, responsable scolaire, autre:
(soulignez ce qui convient)

J'ai déjà participé à une démarche qualité dans l'un de mes emplois:

Oui £ Non £

Les résultats de cette évaluation seront publiés dans le bulletin de l'ADMEE

ANNEXE 2

Analyse de variance

Plans d'observation et d'estimation			
Facettes	**Niveaux**	**Univers**	**Nom**
E	**2**	**2**	**expérience (1=oui, 2=non)**
P:E	**41**	**INF**	**participants**
D	**2**	**2**	**dimensions du quest.**
I:D	**4**	**INF**	**items**

Sources de var.	**S.C.**	**D.L.**	**C.M.**	**Comp. aléat.**	**Comp. mixtes**	**Espér. mixtes**	**%**	**Erreurs types**
E	3.660061	1	3.660061	0.002955	0.008390	0.004195	1.1	0.010791
P:E	52.164634	80	0.652058	0.046532	0.051823	0.051823	14.0	0.013852
D	31.172256	1	31.172256	0.083730	0.089166	0.044583	12.1	0.077869
I:D	11.301829	6	1.883638	0.016954	0.020075	0.020075	5.4	0.011873
ED	2.318598	1	2.318598	0.010871	0.010871	0.002718	0.7	0.011644
EI:D	2.960366	6	0.493394	0.006242	0.006242	0.003121	0.8	0.006029
PD:E	22.384146	80	0.279802	0.010582	0.010582	0.005291	1.4	0.011574
PI:ED	113.987805	480	0.237475	0.237475	0.237475	0.237475	64.3	0.015297