Fiabilité d'un dispositif d'évaluation de l'habileté à déterminer le résultat d'une chaîne d'opérations chez des élèves québécois du secondaire

Marie-Hélène Hébert
Université du Québec à Rimouski

Pierre Valois

Gérard Scallon

Éric Frenette
Université Laval

Mots clés : approche par compétences, ressources, théorie de la généralisabilité, résolution de problèmes mathématiques

La présente étude examine la possibilité d'évaluer avec un bon niveau de fiabilité l'habileté des élèves à déterminer le résultat d'une chaîne d'opérations en respectant la priorité des opérations. Selon le devis de recherche, deux facteurs liés à l'énoncé du problème ont été manipulés : son degré de réalisme et la complétude de ses données. À la suite d'une préexpérimentation, un ensemble de 12 problèmes ont été présentés à 82 élèves québécois de première secondaire. Afin d'évaluer les qualités édumétriques du dispositif d'évaluation constitué, les données ont été analysées au moyen de la théorie de la généralisabilité. Des coefficients de généralisabilité inférieurs à la limite jugée satisfaisante révèlent que le dispositif d'évaluation n'est pas assez fiable pour assurer la différenciation des élèves en regard de leur habileté à déterminer le résultat d'une chaîne d'opérations. Les résultats démontrent la difficulté d'évaluer avec un bon niveau de fiabilité l'habileté des élèves lorsque sont manipulés deux facteurs liés à l'énoncé du problème : son degré de réalisme et la complétude de ses données.

Key words: competency-based approach, resources, theory of generalizability, mathematical problems solving

This study examines the possibility of assessing with a good level of reliability students' ability to compute a sequence of operations using the correct order of operations. Based on a research design, two factors linked to the problem

statement were manipulated: degree of realism and data completeness. Following a pretest, 12 problems were administered to 82 first grade secondary students from the province of Quebec. In order to assess the edumetric qualities of the evaluation design, the data were analyzed using generalizability theory. The generalizability coefficients were lower than the limit considered to be satisfactory, thus indicating that the evaluation design is not reliable enough to allow the differentiation of students' ability in computing a sequence of operations. The results demonstrate the difficulty to assess with a good level of reliability students' ability when two factors linked to the problem statement are manipulated: degree of realism and data completeness.

Palavras-chave: abordagem por competências, recursos, teoria da generalizabilidade, resolução de problemas matemáticos

O presente estudo examina a possibilidade de avaliar com um bom nível de habilidade os alunos para determinar o resultado de uma cadeia de operações respeitando a prioridade das operações. Segundo o desenho da investigação, dois fatores ligados ao enunciado do problema foram manipulados: o seu grau de realismo e a completude dos seus dados. Na sequência de um pré-teste, um conjunto de 12 problemas foram apresentados a 82 alunos do Quebeque do primeiro ciclo do ensino secundário. A fim de avaliar as qualidades edumétricas do dispositivo de avaliação constituído, os dados foram analisados através da teoria da generalizabilidade. Os coeficientes de generalizabilidade inferiores ao limite considerado satisfatório revelam que o dispositivo de avaliação não é suficientemente fiável para assegurar a diferenciação dos alunos em função da sua habilidade para determinar o resultado de uma cadeia de operações. Os resultados demonstram a dificuldade de avaliar com um bom nível de fiabilidade a habilidade dos alunos quando são manipulados dois fatores ligados ao enunciado do problema: o seu grau de realismo e a completude dos seus dados.

Note des auteurs – La correspondance liée à cet article peut être adressée à Marie-Hélène Hébert, Unité départementale des sciences de l'éducation, Université du Québec à Rimouski, 300, allée des Ursulines, C. P. 3300, succ. A, Rimouski (Québec) Canada, G5L 3A1, ou par courriel à l'adresse suivante : [marie-helene_hebert@uqar.ca].

Introduction

L'arrivée du concept de compétence dans le champ de l'éducation a suscité une multitude de changements en cascade au sein de l'école québécoise, notamment sur le plan de l'évaluation. Cette interdépendance entre les mutations dans les programmes d'études et celles enregistrées dans le champ de l'évaluation scolaire ne date pas d'hier.

Au Québec, avant la venue du concept de compétence à l'école, les contenus à apprendre étaient découpés et hiérarchisés dans les programmes d'études sous forme d'objectifs pédagogiques. Dans le programme de mathématique pour la première année du secondaire (MEQ, 1993), discipline d'intérêt dans la présente étude, des énoncés précisaient en termes de comportements observables et mesurables ce que les élèves devaient être capables de faire à la fin d'une séquence d'apprentissage. La nomenclature du programme de 1993 a induit certaines pratiques d'évaluation dans les classes de mathématique. Pendant près de 15 ans, la tendance en évaluation certificative était aux examens qui s'harmonisent avec des contenus formulés selon une hiérarchie d'objectifs, souvent mesurés par des tâches simples (p. ex. : items à réponse choisie ou à réponse courte).

Alors que le programme de 1993 mettait en avant des contenus formulés d'après la séquence « verbe d'action − objet », la nouvelle mouture du programme de mathématique pour l'enseignement au premier cycle du secondaire (MELS, 2006a) favorise le développement des compétences. Si le concept de compétence n'occulte pas les contenus à apprendre, ceux-ci se voient néanmoins investis d'un tout autre rôle. Ils constituent en quelque sorte un bassin de ressources[1], nécessaires au développement des compétences. Jusqu'à tout récemment, la nouvelle tendance en évaluation dans les classes de mathématique québécoises allait de pair avec la singularité d'un programme par compétences : les tâches complexes nourrissaient les jugements portés sur les compétences et contribuaient à la notation, tandis que les tâches simples informaient sur la maîtrise des ressources, sans contribution aucune au bulletin de l'élève (MELS, 2006b, 2006c).

Dernièrement, le ministère de l'Éducation, du Loisir et du Sport[2] publiait la Progression des apprentissages en mathématique au secondaire (MELS, 2010a), un document complémentaire au programme de 2006 qui détaille et hiérarchise le bassin de ressources en question à la façon d'un programme par objectifs. Dans la foulée de cette publication, le MELS a annoncé son intention de revenir à une approche plus « traditionnelle » de l'évaluation. Si l'importance de vérifier la maîtrise des ressources dans un contexte formatif a toujours été d'actualité, il faut reconnaître que les mutations prescrites témoignent d'un retour en force du balancier ; depuis la rentrée scolaire de 2011, il y a lieu de faire contribuer l'évaluation de la maîtrise des ressources au bulletin de l'élève, au même titre que celle des compétences (MELS, 2010b, 2011). Une manière de faire qui reconduit d'emblée le questionnement sur la fiabilité des dispositifs d'évaluation de tâches simples. Si de telles études, héritées de la psychométrie classique, avaient été quelque peu délaissées par l'arrivée massive des tâches complexes dans les écoles (Hébert, Valois, & Frenette, 2008), elles risquent de ressortir de l'ombre, au Québec du moins.

Avec pour trame de fond l'école québécoise, la présente étude a pour objectif d'examiner, au moyen de la théorie de la généralisabilité[3], la fiabilité d'un dispositif d'évaluation employé pour évaluer chez des élèves la maîtrise d'une ressource mathématique à apprendre au premier cycle du secondaire : déterminer le résultat d'une chaîne d'opérations en respectant la priorité des opérations (MELS, 2010a). La fiabilité du dispositif s'en trouverait mieux assurée s'il était démontré que le rendement des élèves n'est pas tributaire des conditions particulières des tâches choisies pour apprécier ladite ressource, comme leur degré de réalisme et la complétude de leurs données.

En mathématique, il est traditionnel pour le MELS de distinguer les problèmes selon leur contexte et leurs données (MELS, 2006a ; MEQ, 1988). Dans son fascicule consacré à la résolution de problèmes publié en 1988, le MELS indique qu'un problème est à contexte :

a) *réel*, « s'il se produit effectivement dans la réalité » (p. 26),

b) *réaliste*, « s'il est susceptible de se produire réellement » (p. 26),

c) *fantaisiste*, « s'il est le fruit de l'imagination et qu'il est sans fondement dans la réalité » (p. 27), et

d) *purement mathématique*, « s'il fait exclusivement référence à des objets mathématiques : nombres, relations et opérations arithmétiques, figures géométriques, etc. » (p. 27).

Toujours dans le même document, dont les idées maîtresses ont été reprises dans le programme de mathématique de 2006, le MELS reconnaît que les problèmes sont à données :

a) *complètes*, s'ils « présentent, de façon explicite, toutes les informations nécessaires à leur résolution » (p. 30),

b) *superflues*, s'ils « présentent, de façon explicite, certaines informations qui ne sont pas nécessaires à leur résolution » (p. 30),

c) *manquantes*, s'ils « ne présentent pas, de façon explicite, toutes les informations nécessaires à leur résolution et tels que les élèves doivent trouver eux-mêmes les informations qui manquent » (p. 31), et

d) *insuffisantes*, s'ils « ne présentent pas, de façon explicite, toutes les informations nécessaires à leur résolution et tels que les élèves ne peuvent pas trouver eux-mêmes les informations qui manquent » (p. 31).

Seuls les contextes réaliste et purement mathématique, ainsi que les données complètes et superflues sont pris en considération dans la présente étude, parce qu'ils sont d'utilisation plus fréquente dans la classe de mathématique et plus faciles à manipuler en contexte expérimental que les autres types de contextes et de données. Il s'agit, dans ce contexte, de questionner la fiabilité d'un dispositif d'évaluation de l'habileté à déterminer le résultat d'une chaîne d'opérations dans lequel les tâches à faire résoudre par les élèves sont :

a) à contexte réaliste ou purement mathématique et

b) à données complètes ou superflues.

Dans l'optique d'établir si ladite habileté est tributaire ou non des tâches choisies pour l'apprécier, une sous-question s'ajoute à celle sur la fiabilité de la mesure : y a-t-il chez les élèves une différence de rendement entre les problèmes :

a) à contexte réaliste et purement mathématique et

b) à données complètes et superflues ?

Si l'examen de la fiabilité d'un tel dispositif d'évaluation apparaît inédit, des études déjà anciennes ont montré la plus grande difficulté :

a) des contextes purement mathématiques par rapport aux contextes réalistes (Caldwell & Goldin, 1979) et

b) des données superflues par rapport aux données complètes (Li, 1990).

Méthode

Avant d'en arriver à l'étude principale proprement dite, une démarche en quatre étapes (appelée ci-après « préexpérimentation ») a conduit à la production de tâches pour apprécier chez des élèves de première secondaire (12-13 ans) l'habileté à déterminer le résultat d'une chaîne d'opérations. Des problèmes « à contexte réaliste et à données complètes » ont d'abord été construits (étape 1), puis administrés à des élèves (étape 2). Parmi ces problèmes, certains ont été retenus sur la base d'un indice de difficulté calculé d'après le modèle de Rasch pour données dichotomiques (étape 3) pour ensuite être convertis en problèmes « à contexte purement mathématique » et « à données superflues » (étape 4). Ces étapes, et leur raison d'être, sont détaillées ci-après.

En plus d'assurer la clarté des problèmes auxquels seraient confrontés les élèves dans l'étude principale, la préexpérimentation visait ultimement à garantir qu'il soit possible d'associer aux facteurs liés à l'énoncé du problème manipulés (degré de réalisme et complétude des données), plutôt qu'à des effets parasites (p. ex. : ordre de grandeur des nombres), toute différence de rendement enregistrée chez les élèves entre les problèmes :

a) à contexte réaliste et purement mathématique et

b) à données complètes et superflues.

Préexpérimentation

Construction des problèmes

Quarante problèmes « à contexte réaliste et à données complètes », modulés sous la forme d'une somme de deux produits de nombres naturels (p. ex. : $83 \times 9 + 79 \times 8$) ou de nombres décimaux (p. ex. : $2 \times 2{,}39 + 4 \times 2{,}19$), ont été construits par des membres de l'équipe de recherche. Un exemple de problème « à contexte réaliste et à données complètes » construit pour les besoins de l'étude est reproduit en annexe. Au terme d'une démarche de validation conduite par un comité d'experts formé de deux professeurs de didactique des mathématiques, d'une enseignante de mathématique et de cinq étudiants des cycles supérieurs en éducation,

28 des 40 problèmes « à contexte réaliste et à données complètes », dont 14 avec nombres naturels et 14 avec nombres décimaux, ont été retenus en vue d'une passation auprès d'élèves de première secondaire.

Passation des problèmes

Puisqu'il s'est révélé irréalisable, faute de temps, de faire résoudre par un groupe d'élèves l'ensemble des 28 problèmes initialement prévus, un devis d'échantillonnage matriciel a été mis en place. Pour ce faire, deux problèmes avec nombres naturels et deux problèmes avec nombres décimaux, choisis au hasard parmi le lot, ont été retenus pour être passés à tous les élèves. Les 24 problèmes restants ont été répartis dans trois cahiers de tests, lesquels ont été distribués aléatoirement parmi 78 élèves (41 filles et 37 garçons) scolarisés dans trois classes de mathématique d'une même école (seuil de faible revenu = 2 ; indice du milieu socio-économique = 2)[4] dont les parents avaient consenti par écrit à ce qu'ils prennent part à la préexpérimentation. Au terme de la passation qui s'est échelonnée sur trois séances de 15 minutes sous la gouverne des enseignants responsables des classes de mathématique, 12 problèmes ont été administrés à chacun des élèves : quatre problèmes communs aux trois cahiers de tests et huit problèmes spécifiques.

Les réponses aux problèmes ont été corrigées de façon dichotomique par un seul correcteur d'après le barème de notation suivant : réponse juste (1 point) ou mauvaise réponse (0 point). Compte tenu de l'empan relativement réduit des patrons de réponse possibles à des tâches simples, il a été décidé de ne pas prendre en compte la démarche des élèves dans la correction des problèmes.

Choix des problèmes retenus au terme de la passation

Les tableaux 1 et 2 reprennent, pour les 28 problèmes administrés, l'indice de difficulté et l'erreur-type du modèle de Rasch pour données dichotomiques calculés à l'aide du logiciel Winsteps (version 3.32 ; Linacre, 2001) par domaine mathématique.

Tableau 1

Rendement des élèves aux problèmes avec nombres naturels administrés lors de la préexpérimentation : indice de difficulté selon le modèle de Rasch pour données dichotomiques et erreur-type

Problème	n	Indice de difficulté	Erreur-type	Problème	n	Indice de difficulté	Erreur-type
1	25	0,56	0,51	8	24	-1,81	1,05
2	23	-0,08	0,60	9	21	0,77	0,48
3	26	0,14	0,46	10	24	0,50	0,46
4	23	-0,97	0,70	11	19	-0,28	0,60
5[a]	72	0,48	0,29	12	24	-1,70	0,81
6	23	0,16	0,54	13	25	0,51	0,54
7[a]	71	0,48	0,29	14	25	1,22	0,50

[a] Item ancré (administré à tous les élèves)

Tableau 2

Rendement des élèves aux problèmes avec nombres décimaux administrés lors de la préexpérimentation : indice de difficulté selon le modèle de Rasch pour données dichotomiques et erreur-type

Problème	n	Indice de difficulté	Erreur-type	Problème	n	Indice de difficulté	Erreur-type
15	25	0,59	0,53	22[a]	74	1,14	0,30
16	26	0,40	0,44	23	25	-1,26	0,67
17	23	-1,09	0,59	24	22	-1,05	0,60
18	21	-0,02	0,55	25	25	-0,85	0,61
19	21	0,24	0,49	26	24	0,07	0,52
20[a]	72	-0,79	0,32	27	26	0,59	0,44
21	18	-0,05	0,54	28	22	2,11	0,65

[a] Item ancré (administré à tous les élèves)

Après examen de ces statistiques, il a été possible de créer quatre regroupements de quatre problèmes aux indices de difficulté les plus proches possible (deux regroupements chez les naturels et deux chez les décimaux). Les problèmes 5, 7, 10 et 13 ont constitué le premier regroupement chez les naturels, tandis que les problèmes 2, 3, 6 et 11 ont fait partie du second. Du côté des décimaux, les regroupements ont été formés de la façon suivante : regroupement 1 pour les problèmes 18, 19, 21 et 26 et regroupement 2 pour les problèmes 17, 20, 24 et 25.

Conversion des problèmes retenus pour l'étude principale

Pour atteindre l'objectif de l'étude, il restait à ce stade à construire des problèmes à contexte purement mathématique et à données superflues. La démarche qui a conduit à la rédaction de ces problèmes est décrite ci-après pour chacun des regroupements de problèmes formés chez les naturels (2 regroupements) et chez les décimaux (2 regroupements) :

a) Un premier problème « à contexte réaliste et à données complètes » a été sélectionné aléatoirement parmi les quatre problèmes du regroupement. Il porte ci-après le nom de problème « pivot » et sert de point de référence.

b) Un deuxième problème « à contexte réaliste et à données complètes » a été choisi au hasard parmi les trois problèmes restants du regroupement pour ensuite être modifié par la transformation de son contexte d'origine en un contexte purement mathématique. Puisque les premier et deuxième problèmes étaient de niveau de difficulté voisin avant l'ajout d'un contexte purement mathématique au second, une éventuelle différence de rendement entre le problème « pivot » et le problème « à contexte purement mathématique » nouvellement créé pourra être attribuable, en grande partie du moins, au degré de réalisme du problème. Un exemple de la transformation qui s'est opérée sur le deuxième problème choisi au hasard est montré en annexe.

c) Le troisième problème « à contexte réaliste et à données complètes » pigé a été modifié par l'ajout de données superflues à son énoncé. Puisque les premier et troisième problèmes étaient de niveau de difficulté voisin avant l'ajout de données superflues au troisième, une éventuelle différence de rendement entre le problème « pivot » et le problème « à données superflues » nouvellement créé pourra être attribuable, en grande partie du moins, à la complétude des données du problème. Un exemple de la transformation qui s'est opérée sur le troisième problème pigé est montré en annexe.

d) Le quatrième et dernier problème du regroupement a été laissé de côté[5].

La répartition des problèmes retenus en prévision de l'étude principale par structure de problèmes est reproduite au tableau 3.

Tableau 3
Répartition des problèmes retenus en prévision de l'étude principale par structure de problèmes

Structure de problèmes					
« Pivot »		**« À contexte purement mathématique »**		**« À données superflues »**	
Naturels	Décimaux	Naturels	Décimaux	Naturels	Décimaux
2, 7	19, 20	11, 13	25, 26	3, 10	21, 24

Étude principale

Échantillon

Les 12 problèmes ont été administrés à 82 élèves de première secondaire (44 filles et 38 garçons) n'ayant pas pris part à la préexpérimentation, issus d'une même école (seuil de faible revenu = 2 ; indice du milieu socio-économique = 2) et scolarisés dans trois classes de mathématique distinctes. Un formulaire de consentement à l'étude a été signé par l'un des parents de chacun des élèves. Parce qu'il s'agissait de faire résoudre des problèmes dont les types de contextes et de données sont prévus dans le programme de mathématique (MELS, 2006a), les élèves ont été considérés comme familiers avec des problèmes d'une telle nature.

Collecte des données

Le mandat d'administrer les problèmes aux élèves a été confié aux enseignants responsables des classes de mathématique. Ces enseignants volontaires, au nombre de trois, ont été rencontrés au préalable afin d'uniformiser la marche à suivre. La passation des problèmes s'est déroulée sur quatre séances d'une quinzaine de minutes chacune. Les élèves ont dû résoudre l'ensemble des problèmes de façon individuelle et sans l'aide d'une calculatrice. Ils étaient libres de laisser, ou non, des traces de leur démarche. Les cahiers de tests remplis ont été récupérés auprès des enseignants au terme de la quatrième séance de passation.

Les réponses aux problèmes ont été corrigées de façon dichotomique par un seul correcteur : réponse juste (1 point) ou mauvaise réponse (0 point). À l'instar de la préexpérimentation, la démarche des élèves n'a pas été retenue comme critère d'évaluation pour apprécier la performance des élèves.

Plan d'analyse

Afin d'évaluer la fiabilité du dispositif employé pour évaluer l'habileté des élèves à déterminer le résultat d'une chaîne d'opérations lorsque sont manipulés des facteurs liés à l'énoncé du problème, les données ont été analysées en appliquant la théorie de la généralisabilité au moyen du logiciel EduG (version 5.0 ; Cardinet, Johnson, & Pini, 2010). Ce choix se justifie par la singularité du modèle statistique, qui permet notamment de déterminer si des observations sont « indépendantes » ou non du dispositif d'évaluation qui a permis de les recueillir (Pini & Hexel, 1998). Comme le requiert toute étude de généralisabilité, les données collectées ont été structurées suivant les plans d'observation, d'estimation et de mesure décrits ci-après.

Plan d'observation. Dans le contexte de l'étude, quatre facettes peuvent être distinguées : les Élèves (E), les Structures de problèmes (S), les Domaines mathématiques dans lesquels s'inscrivent les problèmes (D) et les Problèmes (P:DS). La facette Élèves comporte 66 niveaux (66 élèves ayant pris part aux quatre séances de passation[6]), la facette Structures de problèmes, trois niveaux (problèmes « pivot », « à contexte purement mathématique » et « à données superflues »), la facette Domaines mathématiques, deux niveaux (nombres naturels et nombres décimaux), et la facette Problèmes, deux niveaux (deux problèmes par croisement Structures de problèmes × Domaines mathématiques). Un résumé du plan d'observation du dispositif à l'étude est présenté au tableau 4.

Tableau 4
Dispositif d'évaluation : plan d'observation

Facettes	Niveaux observés	Nombre de niveaux
E	E_1 à E_{66}	66
S	« Pivot », « À contexte purement mathématique » et « À données superflues »	3
D	Nombres naturels et nombres décimaux	2
P:DS	P_1 et P_2	2

Puisque chaque élève, en résolvant chacun des problèmes, a rencontré chacune des structures de problèmes et chacun des domaines mathématiques, la facette Élèves a été « croisée » avec les facettes Problèmes (E × P:DS), Structures de problèmes (E × S) et Domaines mathématiques (E × D). Étant donné que chacune des structures de problèmes a fait intervenir chacun des domaines mathématiques, la facette Structures de problèmes a elle aussi été considérée comme « croisée » avec la facette Domaines mathématiques (S × D). Enfin, puisque chaque problème s'est inscrit à la fois dans une structure de problèmes particulière et dans un domaine mathématique particulier, la facette Problèmes a été « nichée » dans l'interaction entre les facettes Structures de problèmes et Domaines mathématiques. Les données pertinentes à l'étude ont été traitées selon le plan d'observation E(P:DS) illustré à la figure 1.

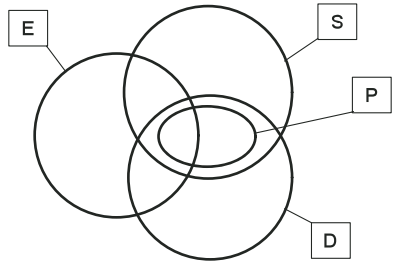

Figure 1 ***Illustration du plan d'observation du dispositif d'évaluation***

Plan d'estimation. En présumant que les 66 élèves ont été choisis parmi des milliers d'élèves et que les 12 problèmes ont été puisés à même des milliers de problèmes, les facettes E et P:DS ont été qualifiées d'« aléatoires infinies ». Puisqu'il existe d'autres structures de problèmes[7] et d'autres domaines mathématiques[8] susceptibles d'être enseignés aux élèves de première secondaire, il a été décidé de considérer les facettes S et D comme « aléatoires finies ». Le tableau 5 rapporte ces informations.

Plan de mesure. Dans l'optique de différencier les élèves quant à leur habileté à déterminer le résultat d'une chaîne d'opérations au moyen des structures de problèmes (S), des domaines mathématiques (D) et des problèmes (P:DS), le plan de mesure E/DPS a été retenu.

Tableau 5
Dispositif d'évaluation : plan d'estimation

Facettes	Niveaux observés	Niveaux admissibles	Type
E	66	Infini	Aléatoire infinie
S	3	16	Aléatoire finie
D	2	4	Aléatoire finie
P:DS	2	Infini	Aléatoire infinie

Résultats

Statistiques descriptives

Les tableaux 6 et 7 indiquent le rendement des élèves aux problèmes administrés lors de l'étude principale par structure de problèmes et par domaine mathématique. Il est à remarquer que les taux de réussite aux problèmes sont particulièrement faibles. Un seul problème, parmi les 12 administrés, avoisine les 80 % de réussite.

Tableau 6
Rendement des élèves aux problèmes avec nombres naturels administrés lors de l'étude principale par structure de problèmes

Structure de problèmes	n	Taux de réussite
« Pivot »		
Problème 2	73	0,795
Problème 7	76	0,566
« À contexte purement mathématique »		
Problème 11	76	0,579
Problème 13	78	0,692
« À données superflues »		
Problème 3	73	0,247
Problème 10	73	0,096

Tableau 7
Rendement des élèves aux problèmes avec nombres décimaux administrés lors de l'étude principale par structure de problèmes

Structure de problèmes	n	Taux de réussite
« Pivot »		
Problème 19	73	0,589
Problème 20	78	0,590
« À contexte purement mathématique »		
Problème 25	73	0,603
Problème 26	73	0,397
« À données superflues »		
Problème 21	78	0,372
Problème 24	78	0,372

Des analyses (*test des signes*) conduites avec le logiciel SPSS (version 18) montrent que les problèmes à contexte réaliste ne sont pas plus difficiles à résoudre pour les élèves que les problèmes à contexte purement mathématique (nombres naturels : z (69) = 0,000 ; p = 1,000 ; nombres décimaux : z (67) = –1,739 ; p = 0,082). À l'inverse, il apparaît que les problèmes à données superflues sont plus difficiles à résoudre pour les élèves que les problèmes à données complètes (nombres naturels : z (71) = –6,742 ; p = 0,000 ; nombres décimaux : z (71) = –3,714 ; p = 0,000)[9].

Étude de généralisabilité

Le tableau 8 synthétise les résultats de l'étude de la généralisabilité du dispositif d'évaluation au moyen du logiciel EduG.

Les coefficients de généralisabilité relatif (0,57) et absolu (0,48) n'atteignent pas le seuil traditionnellement requis (0,80 ; Bain & Pini, 1996)[10]. Un examen des sources de variation indique que c'est l'interaction Élèves × Problèmes (EP :DS) qui concourt le plus à l'erreur de mesure dans l'appréciation de l'habileté des élèves à déterminer le résultat d'une chaîne d'opérations. En comparant son importance relative à celle des autres sources de variation, les résultats du tableau 8 montrent que sa valeur (0,01571) représente 79 % de la variance d'erreur relative et 54 % de la variance d'erreur absolue. La source d'erreur la plus importante après

l'interaction EP:DS est l'effet principal Structures de problèmes (S: 26% de la variance d'erreur absolue), suivi de l'interaction Élèves × Structures de problèmes (ES: 15% de la variance d'erreur absolue). En poursuivant l'étude de généralisabilité dans le but d'améliorer le dispositif d'évaluation, on estime que l'administration additionnelle de 66 problèmes[11] aux élèves aurait été nécessaire pour l'obtention d'une valeur acceptable du coefficient relatif (dans le cas du coefficient absolu, la passation d'une centaine de problèmes supplémentaires n'aurait même pas suffi). À l'opposé, une analyse de facettes a montré que l'élimination d'un domaine mathématique particulier ou d'une structure de problèmes particulière n'aurait pas conduit à de meilleurs coefficients de généralisabilité pour le dispositif d'évaluation.

Tableau 8
Étude de généralisabilité du plan de mesure E\|DPS

Sources de variation	Explication	Variance vraie	Variance d'erreur			
			Relative	%	Absolue	%
E	Variation entre les élèves	0,02630				
S	Variation entre les structures				0,00745	26
D	Variation entre les domaines				0,00000	0
P:DS	Variation entre les problèmes				0,00081	3
ES	Interaction E × S		0,00426	21	0,00426	15
ED	Interaction E × D		0,00000	0	0,00000	0
EP:DS	Interaction E × P:DS		0,01571	79	0,01571	54
DS	Interaction D × S				0,00074	3
EDS	Interaction triple E × D × S		0,00000	0	0,00000	0

Note. Coefficient de généralisabilité relatif: 0,57; absolu: 0,48

Discussion et conclusion

La présente étude a examiné la fiabilité d'un dispositif d'évaluation employé pour évaluer chez des élèves l'habileté à déterminer le résultat d'une chaîne d'opérations en respectant la priorité des opérations lorsque sont manipulés deux facteurs liés à l'énoncé du problème: son degré de réalisme (réaliste ou purement mathématique) et la complétude de ses données (complètes ou superflues).

À la lumière des résultats obtenus, il n'y a pas lieu de recommander le dispositif d'évaluation dans son état actuel pour évaluer avec une fiabilité certaine l'habileté des élèves à déterminer le résultat d'une chaîne d'opérations. Les résultats observés aux 12 problèmes sont davantage le reflet de fluctuations aléatoires, introduites par le dispositif d'évaluation, que du véritable potentiel des élèves à déterminer le résultat d'une chaîne d'opérations.

La mise en évidence d'une interaction statistique Élèves × Problèmes, commune à plusieurs dispositifs d'évaluation (Lane, Liu, Ankenmann, & Stone, 1996; Randhawa & Hunter, 2001; Shavelson, Baxter, & Gao, 1993), soulève une inquiétude au sujet de la méthodologie de l'évaluation : le rendement des élèves varierait d'un problème à l'autre, et ce, indépendamment de leur habileté à déterminer le résultat d'une chaîne d'opérations et du degré de difficulté des problèmes à résoudre. Dans ce contexte, il devient alors souhaitable, comme l'a suggéré l'étude d'optimisation, de recourir à un nombre plus élevé de problèmes que n'en contient le dispositif d'évaluation pour évaluer l'habileté des élèves à déterminer le résultat d'une chaîne d'opérations. C'est d'ailleurs ce que dicte la théorie de l'échantillonnage : plus l'échantillon de problèmes à faire résoudre par les élèves est grand, plus la somme des erreurs aléatoires de mesure qui affectent leur rendement devrait tendre vers zéro.

En somme, la question qui se pose est la suivante : l'habileté des élèves à déterminer le résultat d'une chaîne d'opérations en respectant la priorité des opérations est-elle indépendante des conditions particulières des problèmes ? Les résultats de l'étude invitent à la prudence : il s'agirait plutôt d'une habileté limitée à des contextes particuliers, comme l'a montré l'étude de généralisabilité par la mise en évidence d'une interaction Élèves × Problèmes et d'un effet principal Structures de problèmes. Le rendement des élèves dépendant des problèmes dans lesquels ils sont placés, il devient inévitable de considérer les tâches retenues pour évaluer l'habileté des élèves à déterminer le résultat d'une chaîne d'opérations. Le phénomène, documenté en éducation notamment (p. ex. : Linn & Burton, 1994), est connu sous le nom de spécificité de la tâche (*task specificity*).

Les conclusions de l'étude conduisent à formuler des recommandations à l'intention des enseignantes et des enseignants qui œuvrent dans le domaine de la mathématique. D'abord, étant donné que les problèmes à données superflues ont été plus difficiles à résoudre pour les élèves que les problèmes à données complètes, il y a lieu d'interpréter ce résultat comme

un point valable pour encourager l'entraînement des élèves dans la résolution de problèmes d'une telle nature. Ensuite, en raison du degré de fiabilité limité du dispositif d'évaluation, et ce, malgré un contrôle très serré dans la rédaction des problèmes à faire résoudre par les élèves, il convient de suggérer que le personnel enseignant ait recours au plus grand nombre de problèmes qu'il soit possible d'administrer dans le contexte de leur classe pour évaluer l'habileté des élèves à déterminer le résultat d'une chaîne d'opérations en respectant la priorité des opérations.

Les limites engendrées par nos choix méthodologiques se situent à deux niveaux. Dans le but d'attribuer d'éventuelles différences au niveau du rendement des élèves aux facteurs liés à l'énoncé du problème manipulés, il a été décidé de former quatre regroupements de problèmes de niveau de difficulté voisin pour ensuite créer trois structures de problèmes qui ont servi à l'étude principale. Une limite de l'étude tient au fait que les indices de difficulté qui ont servi à la composition des regroupements n'étaient pas identiques chez les problèmes d'un même regroupement. Par conséquent, il est possible d'envisager que les variations, ou absence de variation, enregistrées sur le plan du rendement des élèves aux problèmes (a) à contexte réaliste et purement mathématique et (b) à données complètes et superflues ne soient pas à cent pour cent expliquées par les facteurs manipulés liés à l'énoncé du problème.

Une deuxième limite, liée à celle qui vient d'être énoncée, concerne le faible échantillon d'élèves retenu en préexpérimentation. Bien que Linacre (1994) ait avancé que le modèle de Rasch puisse offrir des estimés assez stables lorsqu'employé avec des échantillons de l'ordre de 50 participants, il y a lieu de croire qu'un nombre supérieur d'élèves aurait mené à des estimations plus précises des indices de difficulté (c.-à-d. à des erreurs-types plus faibles).

Malgré ses limites, l'étude ouvre la voie à un champ de recherche dans lequel d'autres chercheurs pourraient s'investir pour assurer la fiabilité des dispositifs d'évaluation de tâches simples, et même de tâches complexes, en mathématique. Parmi les avenues de recherche à explorer, il pourrait s'agir de poursuivre des efforts empiriques afin de déterminer l'ensemble des conditions nécessaires pour l'obtention de dispositifs d'évaluation fiables, dont le nombre minimal de tâches à y inclure. Par ailleurs, il serait souhaitable de reprendre l'étude en exploitant d'autres facteurs liés à l'énoncé du problème laissés en plan. Par exemple, les contextes réels, ainsi que les données manquantes et insuffisantes.

Au final, malgré la simplicité apparente des tâches et le respect de prescriptions ministérielles, l'étude a montré la difficulté d'évaluer avec un bon niveau de fiabilité une ressource mathématique à apprendre au premier cycle du secondaire. Que dire alors de l'évaluation des compétences par les tâches complexes ?

NOTES

1. Différentes terminologies sont possibles pour qualifier le contenu du bassin de ressources : a) savoirs, savoir-faire, savoir-être ; b) connaissances déclaratives, procédurales et conditionnelles ; c) connaissances, habiletés, stratégies et savoir-être, etc.
2. Depuis 2005, le ministère de l'Éducation du Québec (MEQ) est dénommé ministère de l'Éducation, du Loisir et du Sport (MELS).
3. Pour une présentation détaillée de la théorie de la généralisabilité, le lecteur est invité à se référer aux numéros spéciaux de la revue Mesure et évaluation en éducation (vol. 26, nos 1-2).
4. Publiés par le MELS, le seuil de faible revenu (SFR) et l'indice du milieu socio-économique (IMSE) donnent un aperçu de la situation socio-économique des milieux de provenance des élèves. Le rang décile 1 caractérise les écoles les plus favorisées et le rang décile 10, les plus défavorisées.
5. Dans le mémoire (Hébert, 2006) qui a inspiré le présent article, ce problème a servi à analyser un troisième facteur lié à l'énoncé des problèmes, soit sa « définition » (problème bien défini vs mal défini). Parce que ce facteur ne procède pas des prescriptions du MELS, il n'est pas considéré ici.
6. Pour l'étude de généralisabilité, un échantillon final de 66 élèves (et non de 82) a été conservé aux fins d'analyse en raison de l'absence d'élèves à l'une ou l'autre des quatre séances de passation.
7. En considérant deux à deux les facteurs liés à l'énoncé du problème formulés par le MELS pour le degré de réalisme et la complétude des données (pour un total de 16 possibilités) : problèmes « à contexte réaliste et à données manquantes », problèmes « à contexte fantaisiste et à données insuffisantes », etc.
8. D'après le programme de mathématique (pour un total de quatre possibilités) : nombres naturels, nombres décimaux, factions et nombres entiers.
9. Le test des signes a été privilégié en raison de la non-normalité et de l'asymétrie des distributions considérées. Parce que des tests multiples ont été effectués avec les mêmes données, la correction de Bonferroni, qui ramène le seuil de signification à 0,01 (0,05/4 tests), a été appliquée. À noter que des conclusions similaires ont été enregistrées lorsqu'il s'est agi de considérer les tests alternatifs suivants : le test t pour échantillons appariés et le test de Wilcoxon pour échantillons appariés. Par ailleurs, comme l'incidence d'effets parasites, tel l'ordre de grandeur des nombres, n'a pas été contrôlée pour le domaine mathématique, il y a lieu d'éviter toute comparaison entre le rendement des élèves aux problèmes avec nombres naturels et avec nombres décimaux.

10. Comme l'avancent Bain et Pini (1996), les deux coefficients de généralisabilité « correspondent à deux façons de poser le problème de l'évaluation » (p. 33). On parle de coefficient relatif (et donc de variance d'erreur relative) lorsqu'il s'agit d'estimer si certains élèves sont plus ou moins performants que d'autres, tandis qu'on parle de coefficient absolu (et donc de variance d'erreur absolue) lorsqu'il s'agit de comparer les taux de réussite des élèves à un seuil donné (p. ex. : 60 %).
11. Évidemment, il s'agit là d'une approximation théorique qui gagnerait à être démontrée sur le plan empirique.

RÉFÉRENCES

Bain, D., & Pini, G. (1996). *Pour évaluer vos évaluations. La généralisabilité : mode d'emploi.* Genève, Suisse : Centre de recherches psychopédagogiques.

Caldwell, J. H., & Goldin, G. A. (1979). Variables affecting word problem difficulty in elementary school mathematics. *Journal for Research in Mathematics Education*, *10*(5), 323-336. doi: 10.2307/748444

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG.* New York, NY: Routledge.

Hébert, M.-H. (2006). *L'interaction statistique Élèves × Tâches comme source d'erreur de mesure en évaluation des apprentissages dans une approche par compétences* (Mémoire de maîtrise inédit). Université Laval, Québec.

Hébert, M.-H., Valois, P., & Frenette, É. (2008). *La validation d'outils alternatifs d'évaluation des apprentissages : progressiste ou rétrograde ?* Repéré à https://plone2.unige.ch/admee08/communications-individuelles/v-a1/v-a1-1

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *33*(1), 71-92. doi: 10.1111/j.1745-3984.1996.tb00480.x

Li, F. L. N. (1990). The effect of superfluous information on children's solution of story arithmetic problems. *Educational Studies in Mathematics, 21*(6), 509-520. doi: 10.1007/BF00315942

Linacre, J. M. (1994). Sample size and item calibration (or person measure) stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (2001). *Winsteps (version 3.32) [Computer Software].* Chicago, IL: Winsteps.com.

Linn, R. L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice*, *13*(1), 5-8. doi: 10.1111/j.1745-3992.1994.tb00778.x

Ministère de l'Éducation du Québec (1988). *Guide pédagogique. Primaire. Mathématique. Résolution de problèmes. Orientation générale (Fascicule K).* Québec, QC : Gouvernement du Québec.

Ministère de l'Éducation du Québec (1993). *Programme d'études. Mathématique 116. Enseignement secondaire.* Québec, QC : Gouvernement du Québec.

Ministère de l'Éducation, du Loisir et du Sport (2006a). *Programme de formation de l'école québécoise. Enseignement secondaire, premier cycle.* Québec, QC : Gouvernement du Québec.

Ministère de l'Éducation, du Loisir et du Sport (2006b). *L'évaluation des apprentissages au secondaire. Cadre de référence.* Québec, QC : Gouvernement du Québec.

Ministère de l'Éducation, du Loisir et du Sport (2006c). *L'évaluation des compétences disciplinaires et la place des connaissances. Questions et éléments de réponse. Principales références dans les encadrements ministériels.* Québec, QC : Gouvernement du Québec.

Ministère de l'Éducation, du Loisir et du Sport (2010a). *Progression des apprentissages au secondaire. Mathématique.* Repéré à http://www.mels.gouv.qc.ca/progression/secondaire/mathematique/

Ministère de l'Éducation, du Loisir et du Sport (2010b). *Évaluation des connaissances – Un nouveau bulletin unique qui reflétera mieux l'évaluation des connaissances des élèves.* Communiqué de presse du 11 juin 2010. Repéré à http://www.mels.gouv.qc.ca/salle-de-presse/communiques-de-presse/detail/article/emevaluation-des-connaissancesem-un-nouveau-bulletin-unique-qui-refletera-mieux-levalua-1/

Ministère de l'Éducation, du Loisir et du Sport (2011). *Cadre d'évaluation des apprentissages. Mathématique. Enseignement secondaire. 1er et 2e cycle.* Repéré à https://www7.mels.gouv.qc.ca/dc/evaluation/index.php?page=recherche

Pini, G., & Hexel, D. (1998). La théorie de la généralisabilité appliquée à un instrument de mesure des attitudes face à l'apprentissage d'une langue étrangère. *Éducation et recherche*, *20*(2), 289-302.

Randhawa, B. S., & Hunter, D. M. (2001). Validity of performance assessment in mathematics for early adolescents. *Canadian Journal of Behavioural Science*, *33*(1), 14-24. doi: 10.1037/h0087124

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232. doi: 10.1111/j.1745-3984.1993.tb00424.x

Date de réception : 5 mars 2010

Date de réception de la version finale : 27 janvier 2014

Date d'acceptation : 28 janvier 2014

ANNEXE

Exemple de problème « à contexte réaliste et à données complètes »

Accompagnant sa maman bien-aimée au marché *Gingras Fruits & Légumes*, Orianne met dans son panier 2 kg de fèves jaunes à 2,39$ le kilogramme et 4 kg d'asperges à 2,19$ le kilogramme. Le marché terminé, quelle somme d'argent Orianne remettra-t-elle au caissier ?

Exemple de problème « à contexte purement mathématique »

Problème initial

Autrefois un traversier et son capitaine faisaient la navette entre Deschambault et Lotbinière. En ces beaux jours, il en coûtait 0,50$ par adulte et 0,35$ par enfant pour la traversée du fleuve St-Laurent. Si, lors d'une traversée, un maximum de 21 adultes et de 7 enfants étaient autorisés à monter sur son engin flottant, quel était alors le revenu du capitaine ?

Problème transformé

Si je fais la somme du produit de la multiplication de 0,50 par 21 et du produit de la multiplication de 0,35 par 7, qu'est-ce que j'obtiens ?

Exemple de problème « à données superflues »

Problème initial

Elliot se rend à la *Société canadienne des postes*. Si l'envoi d'un paquet coûte 3,56$ et qu'on réclame 0,49$ pour l'envoi d'une lettre, combien d'argent Elliot devra-t-il débourser au comptoir postal s'il compte poster 3 paquets et 10 lettres ?

Problème transformé

Elliot se rend à la *Société canadienne des postes*. Si l'envoi d'un paquet coûte 3,56$ et qu'on réclame 0,49$ pour l'envoi d'une lettre, combien d'argent Elliot devra-t-il débourser au comptoir postal si, en plus d'y récupérer les 2 lettres que lui a envoyées sa maman, il compte y poster 3 paquets et 10 lettres ?