Avis au Ministère de l'éducation des loisirs et des sports du Québec

Les résultats de l'échantillon d'élèves québécois du primaire ayant participé à l'enquête TIMSS 2003 en mathématiques

par

Jean-Guy Blais
Département d'administration et fondements de l'éducation
Université de Montréal

Mars 2006

ISBN : 978-2-550-50126-8 (PDF)

Dépôt légal – Bibliothèque nationale du Québec, 2007
Dépôt légal – Bibliothèque nationale du Canada, 2007

Avis au sujet des résultats de l'échantillon d'élèves québécois du primaire ayant participé à l'enquête TIMSS 2003 en mathématiques

1. Introduction

Le Québec participe régulièrement à des enquêtes internationales en éducation qui portent généralement sur différents savoirs que les élèves de l'ordre primaire ou secondaire devraient détenir ou sur différentes compétences qu'ils devraient développer. Ainsi, des échantillons d'élèves des écoles du Québec ont été soumis aux épreuves et questionnaires des enquêtes IAEP de 1988 et 1991, aux enquêtes TIMSS[1] de l'IEA de 1995, 1999 et 2003, aux récentes enquêtes PISA 2000 et 2003 de l'OCDE. De plus, des échantillons d'élèves Québécois participent régulièrement aux enquêtes pan-canadiennes PIRS du Conseil des ministres de l'éducation du Canada. Le Québec n'est donc pas en reste quant à sa participation à ce type d'enquête. De façon générale, ces enquêtes internationales ont deux objectifs : améliorer la compréhension des systèmes d'éducation et aider à mieux comprendre les causes des résultats et des différences observés entre les systèmes.

Le présent avis vise à mettre en perspective les résultats d'élèves Québécois à une de ces enquêtes, l'enquête TIMSS 2003 pour les mathématiques auprès des élèves de la quatrième année du primaire, et à examiner la comparabilité de ces résultats avec ceux de l'échantillon d'élèves ayant participé à l'enquête TIMSS de 1995. Les renseignements au sujet de l'enquête TIMSS rapportés dans cet avis proviennent des rapports sur les résultats et des rapports techniques publiés par le *TIMSS & PIRLS International Study Center* du Boston College aux États-Unis.[2] L'avis aborde différentes facettes de l'enquête TIMSS 2003 : les objectifs de l'enquête; l'échantillonnage des élèves; l'épreuve et les items; la construction des scores; le contexte d'apprentissage; la comparaison 1995-2003. Enfin, une dernière section présente brièvement différentes suggestions quant aux données qui pourraient être récoltées pour assurer un suivi et un pilotage du système d'éducation québécois. Une conclusion résume et met en perspective l'utilité et les limites de ces enquêtes.

[1] L'acronyme TIMSS signifie actuellement *Trends in International Mathematics and Science Study* et c'est l'expression qui sera utilisée dans cet avis pour décrire les enquêtes de l'IEA en mathématiques de 1995 et 2003 au primaire.

[2] Ces rapports sont disponibles sur le site «http://timss.bc.edu/» et les références complètes sont présentées à la fin de ce document.

2. Les enquêtes TIMSS

L'objectif général des enquêtes TIMSS est « d'améliorer l'enseignement et l'apprentissage en mathématiques et en science en fournissant des données au sujet de la performance des élèves selon les caractéristiques de différents curriculum, de différentes pratiques d'enseignement et de différents environnements scolaires ».

Les participants à ces enquêtes sont le plus souvent des pays, mais on y retrouve également ce qu'on appelle des juridictions, c'est-à-dire des entités politiquement constituées qui ont la responsabilité légale de leur système d'éducation. Il en est ainsi pour l'enquête TIMSS 2003 des provinces de l'Ontario et du Québec, de même que de l'état de l'Indiana.[3] Pour éviter de surcharger inutilement la suite du texte, l'expression PJ sera utilisée pour désigner les pays et juridictions ayant participé à l'enquête.

Les enquêtes TIMSS pourraient donc permettre, en principe, de situer la performance d'un échantillon d'élèves Québécois à des épreuves uniformes de mathématiques et de science en fonction de la performance d'échantillon d'élèves d'autres PJ. De plus, lorsque des enquêtes avec le même objectif se répètent dans le temps auprès de populations « semblables », comme pour les enquêtes TIMSS en 1995 et en 2003 avec des échantillons d'élèves de l'ordre primaire, on obtient deux portraits (deux instantanés) dans le temps qui à la lumière des succès ou insuccès communs aux deux portraits peuvent permettent un retour sur les programmes, les curriculums, les opportunités d'apprentissage et les séquences d'apprentissage.

3. Les élèves participants : le plan d'échantillonnage

L'enquête TIMSS 2003 est une enquête qui vise à récolter des données pour permettre «**le suivi du niveau général** de connaissance ou de compétences des élèves». Étant donné cette visée, le recours pour l'enquête TIMSS à un échantillon probabiliste d'écoles et d'élèves apparaît comme l'approche la plus efficace et la plus efficiente (en ce sens que sous certaines conditions techniques on peut obtenir des données assez précises, généralisables à la population, avec un rapport coûts/bénéfices intéressant). Cependant, au-delà de l'adéquation technique des procédures d'échantillonnage, des problèmes subsistent quant à l'équivalence des échantillons des PJ participants.

[3] On peut remarquer cependant que l'Écosse, qui n'est pas officiellement un «pays», voit ses résultats présentés dans la catégorie des pays et non des juridictions comme le Québec ou l'Ontario.

Pour l'enquête TIMSS 2003 auprès d'élèves du primaire, 26 PJ ont participé aux opérations. La population visée dans chacun de ces PJ était celle de tous les élèves inscrits au niveau qui contenait le plus grand nombre d'élèves de neuf ans au moment de l'épreuve. Pour le Québec, comme pour la grande majorité des PJ participants, ces élèves se retrouvaient à la quatrième année du primaire. Cependant, pour certains PJ ces élèves étaient plutôt en cinquième année du primaire (Angleterre, Écosse, Australie et Nouvelle-Zélande). De plus, étant donné qu'il s'agit d'un échantillonnage en grappe à deux étapes où des classes complètes sont sélectionnées (voir ci-dessous), plusieurs élèves de dix ans se retrouvaient également dans certains des échantillons.

La pratique de la reprise forcée d'une année scolaire (i.e. le redoublement) n'étant pas la même pour chaque PJ, il est possible que certaines classes de quatrième année accueillent un nombre non négligeable d'élèves de dix et onze ans, alors que des élèves de neuf ans sont retenus en deuxième ou troisième année (voir Brown 1999 au sujet des différences entre les pratiques de pays européens comme l'Angleterre, la Suisse et l'Allemagne, en ce qui concerne le redoublement et leurs impacts possibles sur les résultats moyens des élèves dans des enquêtes comme TIMSS).

À titre d'exemple, le tableau 1 ci-dessous présente la moyenne de l'âge des élèves de quelques PJ participants. On y observe que l'âge moyen des élèves de l'échantillon du Québec est de 10,1, alors qu'il est de 11,1 pour la Lettonie, 10,9 pour la Lituanie, 10,6 pour la Russie, 10,3 pour l'Angleterre et 9,7 pour l'Écosse. Pour ces deux derniers participants toutefois, les élèves de cet âge ont complété cinq années de scolarité.

À n'en pas douter, certaines particularités systémiques des différents PJ comme le fait d'avoir complété cinq années de scolarité plutôt que quatre années, de compter plus d'élèves âgés dans une classe, donc plus matures (ce qui n'est pas négligeable en quatrième année du primaire), et de promouvoir moins d'élèves à cause des difficultés d'apprentissage, peuvent donner un portrait difficile à comparer avec d'autres portraits où ces pratique n'ont pas cours car elles contribuent en quelque sorte à biaiser les comparaisons entre les PJ.

Tableau 1. Âge moyen pour quelques PJ

Pays/ Juridiction	Âge Moyen
Québec	10,1
Ontario	9,8
Lettonie	11,1
Lituanie	10,9
Russie	10,6
Angleterre	10,3
Écosse	9,7

Dans tous les PJ participants, les échantillons d'élèves ont été constitués selon les principes d'un échantillonnage en grappe en deux étapes. Selon cette approche, la première étape a consisté pour chaque PJ à choisir des écoles selon la technique probabiliste du choix proportionnel à la taille (après une mise en rang selon la taille, des écoles sont choisies selon une séquence déterminée, une sur dix par exemple, afin de s'assurer d'obtenir des écoles dans différentes catégories de taille d'école). La deuxième étape a consisté à choisir au hasard et selon la taille de l'école une classe d'élèves dans chacune des écoles préalablement sélectionnées lors de la première étape.[4]

Ce plan d'échantillonnage peut être qualifié de plan de base : d'abord, un choix des écoles selon la taille de l'école pour assurer la présence d'élèves provenant d'écoles de tailles différentes; ensuite le choix d'une classe complète d'élèves pour perturber le moins possible les activités régulières d'enseignement. Pour différentes raisons propres à chacun des PJ, d'autres variables ont aussi été utilisées pour ajouter un ou deux éléments de stratification. Ainsi, au Québec comme en Ontario et en Nouvelle-Zélande, un niveau supplémentaire a été ajouté et l'échantillon a été stratifié au départ pour tenir compte de la langue d'enseignement. Ensuite la méthode en deux étapes a été appliquée pour chacune des strates.

D'autres PJ ont opté pour des variables de stratification différentes. Ainsi par exemple, la Russie et l'Australie ont décidé de stratifier leur échantillon en fonction des régions, la Lituanie et la Slovénie selon la taille de l'école, le Japon et le Yémen selon la dimension rural/urbain, l'Indiana selon le réseau, l'Iran et les États-Unis selon le niveau de pauvreté.[5] D'un point de vue technique, ces différences quant au plan d'échantillonnage ne posent pas de problème puisqu'il est possible d'en tenir compte pour produire les estimations des moyennes et des écart-types des scores pour chaque PJ.

L'objectif final de l'opération d'échantillonnage était d'obtenir pour chaque PJ un échantillon d'au moins 150 écoles et d'au moins 4000 élèves. Pour le Québec, un échantillon[6] de 193 écoles et 4350 élèves a été constitué à partir d'une

[4] Dans certains PJ, on a opté pour deux classes par école, alors que d'autres PJ ont été obligés de choisir plusieurs classes dans une même école pour répondre à l'exigence d'un minimum de 4000 élèves dans l'échantillon (exigence qui n'a pas toujours été rencontrée). De plus, dans certains PJ où les élèves dans les classes sont très nombreux, un échantillon d'élèves de la classe a été prélevé.

[5] On peut noter que les échantillons n'ont pas été stratifiés selon le genre, mais que des comparaisons *a posteriori* garçons – filles sont présentées dans le rapport TIMSS. La précision des estimations des scores moyens pour les garçons et les filles pourrait donc ne pas être fondée techniquement.

[6] Pour différentes raisons, certains élèves retenus dans l'échantillon n'ont pas participé aux épreuves. Ainsi, pour le Québec, 4864 élèves ont été sélectionnés dans un premier temps, 124

population de 1879 écoles et de 98326 élèves. Comme la procédure d'échantillonnage exige dans un premier temps la sélection des écoles et qu'il est possible que certaines d'entre elles refusent de participer, des écoles de remplacement sont aussi sélectionnées pour pallier à ces refus. Ainsi, un autre indice, à part l'âge des élèves, pour l'étude de la comparabilité des échantillons est celui de la participation des écoles au premier tour (avant remplacement).

Pour le Québec, on note une participation de 97%, ce qui est très élevé et dénote une très bonne collaboration du milieu. Cependant, dans d'autres PJ plusieurs écoles n'ont pas démontré, semble-t-il, le même empressement à collaborer. Ainsi, le taux de participation des écoles a été de 52%, 54%, 64% et 70% respectivement pour les Pays-Bas, l'Angleterre, l'Écosse et les États-Unis. On peut donc s'interroger d'une part sur les motifs qui amènent un aussi grand nombre d'écoles de ces PJ à ne pas collaborer et, d'autre part, sur l'impact de ces refus sur les résultats observés.

En résumé, on peut dire que la question de l'âge des élèves et celle de la faible collaboration des écoles au premier tour mettent en perspective les difficultés que les responsables ont rencontrées dans la mise en place d'échantillons assez équivalents pour permettre des comparaisons raisonnables entre les résultats des PJ. À la lumière des propos de Brown (1996 et 1999) sur le sujet et de ce qui est écrit dans les rapports TIMSS, des réserves peuvent être émises quant au succès de cette opération pour l'enquête 2003 en mathématiques (la même remarque vaut évidemment pour la partie touchant les sciences).

L'ensemble des commentaires de cette section met ainsi en relief les limites des indices de précision comme l'erreur standard d'estimation et les intervalles de confiance qui en découle, car ceux-ci tiennent la route pour faire des comparaisons seulement si les échantillons sont comparables, ce que ne confirment pas tout à fait les informations disponibles au sujet des élèves réellement échantillonnés.

4. Les caractéristiques de l'épreuve de 2003

L'épreuve 2003 de mathématiques touchait cinq domaines : les nombres, la mesure, la géométrie, l'analyse des données, les formes et relations. L'épreuve comptait en tout 161 items répartis dans ces cinq domaines selon les proportions illustrées au tableau 2 ci-dessous. La plus grande part revient au domaine des nombres qui accapare 40% des items et des points, alors qu'aucun autre domaine ne prend plus de 20% du total d'items et de points. La partie qui

---

élèves ont été exclus et 390 ne se sont pas présentés le jour de l'épreuve (4864 – 124 – 390 = 4350).

contribue le moins à l'ensemble de l'épreuve étant celle sur l'analyse des données avec 17 items et 10% des points attribués.

Environ 60% des items sont à réponse choisie alors que les items dits de «résolution de problème» constituent le 40% restant. De plus, 37 items ont été retenus de l'épreuve de 1995 pour analyser les différences observées avec les résultats de l'opération 2003. Ces items communs aux deux opérations sont tous des items à réponse choisie. [7]

Cependant, il n'est pas tout à fait exact de dire que l'épreuve comptait 161 items. En effet, la structure des épreuves TIMSS a été conçue en s'inspirant de celles des enquêtes NAEP aux États-Unis et IAEP de 1988 et 1992. Avec cette structure, chaque élève n'a pas à répondre aux 161 items de l'épreuve, ce qui prendrait beaucoup de temps et d'énergie, il ne répond qu'à un nombre raisonnable d'items étant donné le temps réservé à l'opération. Les items sont répartis dans 14 blocs d'items qui eux-mêmes sont répartis dans 12 cahiers contenant deux sections de trois blocs d'items de façon à obtenir un ou deux blocs d'items de mathématiques dans chaque section d'un cahier. Un des douze cahiers de l'épreuve est ensuite attribué au hasard à chaque élève.

Tableau 2. Répartition des items et des points accordés en fonction du domaine

Domaine	Nombre d'items	% du nombre total d'items	Nombre de points	% points
Nombres	63	39	68	40
Formes et relations	24	15	25	15
Mesure	33	20	33	20
Géométrie	24	15	25	15
Analyse des données	17	11	18	10

Les blocs d'items ont été constitués de façon à contenir entre 10 et 13 items chacun et ils ont été assemblés en respectant l'équilibre entre les domaines, les processus cognitifs et le type d'items. Ainsi, dans chaque bloc on compte en moyenne 8 ou 9 items à réponse choisie, 4 ou 5 items à réponse courte, 0 ou 1 item à réponse construite plus élaborée.

[7] Cela peut rendre difficile la comparaison 1995-2003 si, pour différentes raisons, les élèves de 2003 ont été soumis durant leur parcours scolaire à moins d'épreuves contenant des items à réponse choisie, ainsi que le suggèrent plusieurs des dimensions de la réforme au primaire qui visent l'évaluation des apprentissages et des compétences.

Le temps total réservé au passage des épreuves TIMSS 2003, mathématiques **ET** sciences, a été de 72 minutes. Cette période était divisée en deux séances de 36 minutes avec une pause entre les deux. Les blocs d'items de mathématiques ayant été répartis dans les deux séances (un ou deux blocs par séance selon le cahier), on peut estimer que l'épreuve de mathématiques a été d'une durée variant entre 24 et 48 minutes selon le cahier attribué. Durant cette période, les élèves ont du répondre à entre 20 et 39 items de mathématiques.

La structure des cahiers et la répartition des items dans ceux-ci peuvent laisser perplexe la personne non initiée et soulever des interrogations légitimes sur la possibilité de comparer les résultats d'élèves qui n'ont pas répondu aux mêmes items. La réponse à ces interrogations est plutôt technique et se base sur ce qu'on appelle **l'appariement** des scores aux épreuves. À l'aide de modèles de mesure et de techniques statistiques,[8] les résultats à des épreuves différentes mais contenant des items communs sont placés sur la même échelle de scores, permettant la production de moyennes et d'écarts-types pour les scores de chaque PJ participant qui, **sous certaines conditions**, peuvent être considérés comme étant sur la même échelle de mesure.

Au-delà du fait que le tout a été réalisé selon les règles de l'art reconnues, il faut retenir de ces opérations d'appariement des scores aux épreuves qu'il y a plusieurs voies techniques qui sont possibles et que la voie empruntée par les responsables de l'enquête n'est qu'une voie parmi d'autres. De plus, les scores estimés ne proviennent pas uniquement des réponses aux items des épreuves, mais font aussi usage des données récoltées avec le questionnaire aux élèves (la technique des «valeurs plausibles» est utilisée, voir la section suivante). Évidemment, dans une telle situation il est difficile de comparer les résultats de l'opération avec ceux qui seraient obtenus avec une autre méthode d'estimation (par exemple, tout simplement en comptant le nombre de bonnes réponses) et il faut faire confiance à la méthode d'appariement utilisée (i.e. aux modèles de mesure et aux outils statistiques).

Finalement, on peut également se demander si le temps imparti était suffisant pour permettre aux élèves de démontrer leurs compétences (entre 24 et 48 minutes pour l'épreuve de mathématiques et un peu plus d'une minute par item). À ce sujet, les auteurs des rapports TIMSS 2003 mentionnent que pour certains PJ et pour certains blocs d'items passés à la fin de chacune des deux séances, plusieurs items ont été classés non-atteints[9], i.e. que le temps alloué n'était pas suffisant pour répondre à tous les items du dernier bloc de chaque séance. Ces items ont du recevoir un traitement séparé lors de l'analyse pour la construction des scores.

[8] Des modèles de la théorie des réponses aux items, à un, deux ou trois paramètres, et des modèles de régression.
[9] Selon le cahier, autour de 10% des items d'un cahier.

5. La construction des scores

Étant donné la complexité du design de l'épreuve et le désir de présenter les résultats sur la même échelle de mesure, les responsables ont eu recours à trois modèles de mesure et à la méthodologie dite des «valeurs plausibles» pour construire les scores des élèves des PJ participants. L'échelle de mesure intègre ainsi les cotes attribuées pour les items de l'épreuve[10] et les réponses au questionnaire sur les antécédents des élèves (le *background* des élèves). La méthodologie utilisée demande une compréhension avancée de la modélisation de la mesure[11] et elle n'est pas fréquente dans les opérations à grande échelle. À notre connaissance elle n'a été utilisée que pour les enquêtes NAEP aux États-Unis à partir de la fin des années 1980 et pour l'enquête IAEP de 1992. [12]

L'échelle des scores produite avec cette méthodologie est arbitraire en ce sens que l'origine de l'échelle et les unités de mesure sont fixées arbitrairement à des valeurs constantes. Il faut ouvrir une parenthèse ici pour préciser que la réalité d'une échelle arbitraire n'est pas unique à ce type d'enquête car même lorsque les scores prennent des valeurs entre 0 et 100, l'échelle est tout aussi arbitraire et elle est choisie d'abord pour sa simplicité plutôt que pour sa validité. C'est ainsi que pour les résultats de l'enquête TIMSS 2003, la moyenne a été fixée à 500 et l'écart-type à 100. D'autres points de référence arbitraires sont possibles, plaçant ainsi les différences observées dans une situation ou la perception de la grandeur des nombres peut être importante.

À titre d'exemple, on peut dire que les différences entre les PJ peuvent prendre une autre couleur si les résultats sont ramenés sur une échelle avec une moyenne de 50 et un écart-type de 10. On obtient alors un portrait qui n'est pas différent numériquement mais qui place les choses en perspective. En effet la moyenne arrondie des élèves de l'échantillon du Québec est de 51 en 2003, moyenne qui lorsqu'elle est comparée à la moyenne arrondie de PJ semblables ne semble plus si éloignée comme on peut le voir au tableau 3 ci-dessous. Le portrait pourrait être encore plus susceptible d'influencer les perceptions si les résultats étaient ramenés sur une échelle avec une moyenne de 5 par exemple. Les différences observées entre les filles et les garçons prennent aussi une autre dimension si les valeurs de l'échelle sont divisées par 10. Les moyennes pour le Québec sont respectivement de 51 pour les garçons et 50 pour les filles. En extrapolant aux populations de garçons et de filles, on peut avancer qu'il n'y a pas de différence réelle pour ce qui a trait au genre des élèves.

---

[10] On peut noter que les réponses à 159 items ont été utilisées car deux items ont été mis de côté après l'analyse d'items.

[11] Voir le chapitre 11 du rapport technique de TIMSS à ce sujet.

[12] Pour l'enquête IAEP de 1992, on peut consulter le rapport de Blais 1992. On peut aussi ajouter l'enquête PIRLS de l'IEA sur la littéracie qui a adopté la même approche.

Tableau 3 : Scores moyens de certains PJ lorsque la moyenne des scores est fixée à 50 et l'écart-type à 10

PJ	Score moyen
Belgique	55
Pays-Bas	54
Hongrie	53
Ontario	51
Québec	51
Australie	50
Nouvelle-Zélande	49
Écosse	49
Norvège	45

En ajoutant à ce score moyen tout ce qui peut avoir un impact sur l'erreur d'échantillonnage réelle et tout ce qui peut contribuer à l'imprécision des valeurs numériques obtenues, imprécision qui a été illustrée dans les sections précédentes et qui n'est pas très bien documentée dans les rapports TIMSS, on peut penser que les **vraies** différences entre plusieurs des moyennes des PJ sont assez minimes, et souvent inexistantes. Par exemple, qui pourrait affirmer qu'il existe une différence réelle entre le Japon et Taipei lorsque les moyennes non arrondies des scores de leurs élèves sur cette nouvelle échelle sont respectivement 56,5 et 56,4 ! Ou encore, quelle est la vraie différence entre le score moyen de 55 de l'échantillon des élèves de la Belgique et celui de 51 des élèves du Québec (sans compter qu'ils n'ont pas eu les mêmes «opportunités d'apprentissage» comme la section suivante l'illustre) ?

Ainsi, les inférences que l'on aimerait faire sur les différences entre les moyennes ou encore sur les rangs de ces seules moyennes doivent être faites avec prudence car les données telles que récoltées et transformées ne nous permettent pas d'avancer avec une assurance à toute épreuve dans cette direction.

Lorsqu'on examine la position respective de la moyenne de chacun de PJ sur l'échelle des scores on observe qu'il y en a quatre qui sont nettement détachés des autres et qui obtiennent des résultats moyens plus bas que 40 (Tunisie, Philippines, Iran, Maroc). Dans le jargon statistique, ces résultats moyens peuvent être considérés comme des valeurs extrêmes qui influencent indûment la moyenne et l'écart-type de l'ensemble des données. Si les scores des élèves de ces PJ étaient exclus des calculs, le portrait serait passablement différent et les valeurs de toutes les statistiques s'en trouveraient influencées. Il s'agit cependant du lot de ce type d'enquête, chaque PJ y retrouve d'autres PJ «comparables», mais y retrouve également des entités auxquelles il ne peut en

toute bonne foi se comparer. Cependant, les modèles de mesure, les statistiques et les calculs ne font pas la différence entre ce qui est comparable et ce qui ne l'est pas. Cette tâche revient aux responsables des systèmes d'éducation de chacun des PJ. La prudence est donc encore une fois de mise lorsqu'on commente les rangs et les écarts à la moyenne générale car ceux-ci sont dépendants des caractéristiques des PJ participants et, évidemment, de la performance des élèves de l'ensemble des PJ aux épreuves. Une comparaison adéquate demande plus qu'un regard sur le score moyen ou sur le rang relatif, elle demande un regard sur les items et sur le lien des items avec le programme et avec ce qui a été enseigné. On y reviendra à la section suivante.

En terminant cette section, on peut aussi se demander si les résultats sont fidèles et possèdent une cohérence interne suffisamment élevée pour que l'on puisse faire confiance à la précision des résultats. L'information disponible à cet égard est donnée par la valeur d'un coefficient de consistance interne bien connu, le coefficient alpha de Cronbach. Même si plusieurs valeurs sont présentées dans les rapports 2003, il apparaît que les plus intéressantes sont celles qui ont été calculées pour chacun des cahiers et chacun des PJ. Ainsi, on peut observer que toutes les valeurs médianes du coefficient alpha sont inférieures à 0,90. Par exemple, la médiane des valeurs du coefficient alpha pour les 12 cahiers pour les élèves du Québec est de 0,84. La valeur de cette médiane n'est pas très élevées,[13] surtout qu'il y a six cahiers pour lesquels la valeur du coefficient est inférieure à 0,84 (mais on ne sait pas quelles sont les valeurs précises). Il est assez évident que certains des cahiers produisent des scores plus instables sur lesquels on ne peut se fier et que cette situation est différente pour chacun des PJ participants. Il s'agit encore une fois d'un indice utile pour estimer l'imprécision des mesures présentées.

6. Le contexte d'apprentissage

Lorsqu'on examine les résultats des élèves d'un PJ participant à une enquête comme celle de TIMSS, ce n'est pas suffisant d'utiliser des procédures d'échantillonnage adéquates techniquement et de mettre en œuvre des modèles de mesures sophistiqués. En effet, au-delà des dimensions techniques, il faut se demander si les items retenus pour faire partie des épreuves représentent des tâches que les élèves sont capables d'affronter; en d'autres mots, si les élèves ont vu la matière à laquelle les items de l'épreuve font référence. Dans le jargon de la recherche sur ce sujet, c'est ce qu'on définit par «les opportunités d'apprentissage». Il s'agit en fait d'une question centrale qui touche la validité des données et des inférences possibles. Dans le passé et dans le cadre d'enquêtes internationales comme TIMSS, les différences quant au degré de couverture du curriculum prévu, i.e le curriculum réellement abordé en classe,

[13] Généralement, on aimerait obtenir des valeurs qui sont plus élevées que 0,90.

ont été citées comme étant les facteurs qui influenceraient le plus les différences de performance aux épreuves entre les PJ (Burstein, 1992; Husen, 1967; McLean 1996).

Les items qui composent les épreuves d'enquêtes comme celles de TIMSS ne sont qu'un des ensembles d'items de l'univers des items qu'il serait possible d'imaginer et sont le fruit de négociations et de concessions entre les PJ qui participent. Il est possible que ces négociations aboutissent à une épreuve contenant des items qui sont plus ou moins reliés au curriculum prescrit et à celui réellement abordé dans un PJ particulier, affectant ainsi la validité des scores, des comparaisons et des conclusions que l'on pourrait tirer à partir de statistiques utilisant ces scores.

Le tableau 4 ci-dessous présente les résultats des élèves du Québec pour chacun des domaines ciblés par l'épreuve de mathématiques, quatrième année, pour TIMSS 2003. Lorsqu'on examine les résultats en parallèle avec la variable «opportunités d'apprentissage» (OA), on constate d'abord que le domaine ou on retrouve le plus d'items et le plus de points, celui des «nombres», ne serait couvert qu'à 67% dans les classes participantes[14]. D'un autre côté, le domaine des «formes et relations» est celui qui serait le plus couvert, avec environ 83% des élèves qui auraient abordé les items de ce domaine. On observe aussi que même si selon les enseignants le domaine de la géométrie a été le moins couvert, la moyenne des résultats des élèves est la plus élevée à 52.

Tableau 4. Opportunités d'apprentissage, moyenne et % de points selon le domaine, élèves du Québec

Domaine	OA : %	Moyenne	% points
Nombres	67	50,8	40
Formes	83	49,9	15
Mesure	70	50,4	20
Géométrie	59	52,2	15
Données	69	50,6	10

[14] Cette information a été fournie par l'enseignant lorsqu'il a répondu au questionnaire qui lui était destiné et elle n'a pas fait l'objet d'une vérification ultérieure.

Ces données sont intéressantes certes, mais si elles ne sont pas comparées avec ce qu'on observe pour d'autres PJ, il demeure difficile de conclure qu'elles pourraient être à la source des différences observées pour les scores moyens. Les tableaux 5 et 6 ci-dessous nous permettent de nous livrer à un petit exercice de comparaison entre différents PJ qui ont des caractéristiques les rapprochant. Au tableau 5, on constate en comparant le Québec à l'Ontario, aux États-Unis et à la Norvège, qu'il y a des différences importantes quant à ce qui est déclaré par les enseignants comme «opportunités d'apprentissage». Par exemple, pour l'Ontario et les Etats-Unis et toujours d'après les déclarations des enseignants, un nombre plus élevé d'élèves aurait eu l'occasion de se frotter aux contenus de l'épreuve et ce pour chaque domaine. Pour les élèves de l'échantillon des États-Unis le domaine de «l'analyse des données» est celui qui a été déclaré le plus couvert, par environ 90% des élèves, et le résultat moyen pour ce domaine est le plus élevé à 54,9 (ou 549 selon l'échelle de mesure adoptée). En ce qui concerne la Norvège, on ne peut s'empêcher de remarquer que mis à part le domaine de la mesure, il semble y avoir une bonne partie des élèves qui n'a pas eu l'occasion de voir en classe la matière faisant l'objet de l'épreuve (autour de 78% pour ce qui est de la mesure, un peu plus de 50% pour les nombres, les formes et les données et 32% pour la géométrie). Il est donc légitime à mon avis de se demander si pour la Norvège cette dimension ne serait pas la meilleure source d'explication des résultats plus «faibles» obtenus par les élèves échantillonnés.

Tableau 5. Opportunités d'apprentissage (OA) et moyenne selon le domaine, élèves du Québec, de l'Ontario, des États-Unis et de la Norvège

	Québec		Ontario		États-Unis		Norvège	
Domaine	**OA : %**	**Score moyen**	**OA : %**	**Score moyen**	**OA : %**	**Score moyen**	**OA : %**	**Score moyen**
Nombres	67	50,8	75	49,4	83	51,6	54	44
Formes	83	49,9	83	51,3	89	52,4	53	43,9
Mesure	70	50,4	86	51,2	81	50	78	47,5
Géométrie	59	52,2	75	53,5	74	51,8	32	47,8
Données	69	50,6	91	54,4	90	54,9	54	47,9

Le tableau 6 présente les valeurs estimées pour les «opportunités d'apprentissage» et le score moyen pour le domaine des «nombres» pour 12 PJ ayant des similarités économiques et sociales avec le Québec. Aux deux extrémités, ce qui est observé pour la Belgique et la Norvège confirmerait en quelque sorte le lien entre les «opportunités d'apprentissage» et les scores à l'épreuve. Cependant, pour les Pays-Bas la situation serait différente et n'irait pas dans la direction d'un lien fort entre les «opportunités d'apprentissage» et les scores.

Tableau 6. Opportunités d'apprentissage et score moyen arrondi pour le domaine des «nombres» et quelques PJ

PJ	OA : %	Score moyen arrondi
Belgique	93	55
Italie	88	50
Angleterre	87	52
États-Unis	83	52
Nouvelle-Zélande	76	47
Ontario	75	49
Australie	74	48
Hongrie	68	52
Québec	67	51
Écosse	67	47
Pays-bas	63	54
Norvège	54	44

Étant donné les informations que l'enquête TIMSS a permis de colliger sur ce qui est identifié comme les «opportunités d'apprentissage» et les disparités entre ce qui est rapporté à ce sujet pour chacun de PJ participants, il est bien évident que le sens général à donner aux résultats doit être revu à la lumière de la partie de l'enquête qui était au programme officiel de mathématiques de quatrième année, et, surtout, à la lumière ce qui a réellement été vu par les élèves en salle de classe.

7. Les niveaux de performance[15]

Les niveaux de performance ne sont pas absolus, ils sont déterminés empiriquement à partir des résultats des élèves de chacun des PJ. Si pour quelques raisons que ce soit les résultats de certains PJ sont enlevés ou ajoutés, il est fort possible que cela change le portrait de la situation, remettant en perspective les réussites ou les échecs de l'un et l'autre, de même que les points d'ancrage de ces niveaux de performance.

Lorsque la distribution des élèves du Québec est examinée en fonction des quatre niveaux de performance, on constate que les élèves se répartissent selon une distribution semblable à celle observée pour les échantillons d'élèves de la Nouvelle-Zélande, de l'Australie et de l'Écosse. Toutefois, cette comparaison est

[15] Ce qu'on appelle les *benchmarks* dans les rapports TIMSS.

assez limitée s'il n'est pas possible d'étudier les différences en fonction des items qui sont associés à chacun des niveaux de performance.

Grâce au dévoilement de quelques items qui seraient représentatifs de chacun des niveaux de performance, il est possible de pousser l'analyse un peu plus loin. Les items retenus pour cette analyse sont des items se situant aux niveaux 625, 475 et 400 de l'échelle et ils sont présentés aux tableaux 7, 8 et 9 ci-dessous.

Le tableau 7 présente un item qui demande de traduire la fraction 7/10 en représentation décimale. Il s'agit d'un item à réponse choisie du domaine des nombres se situant au niveau le plus élevé de 625. Les taux de réussite de l'item apparaissent assez faibles pour le Québec, la Hongrie et la Norvège, mais la variable «opportunités d'apprentissage» présente aussi les valeurs les plus faibles pour ces trois PJ. Les taux de réussite plus élevés de l'Ontario, des Philippines et de l'Italie sont également en relation avec des valeurs plus élevées pour l'opportunité d'apprentissage.

Le tableau 8 présente un item du domaine de l'analyse des données qui a été situé au niveau intermédiaire, i.e. à un score de 475. En principe donc, cet item est plus «facile» que le précédent qui se situerait au niveau de performance avancé. Les taux de réussite sont plus élevés que pour l'item précédent, sauf pour les Philippines. Si on compare le Québec et l'Ontario, les taux de réussite respectifs sont très élevés, mais les valeurs pour les «opportunités d'apprentissage» sont très différentes avec 69% et 91% respectivement.

Pour terminer, le tableau 9 permet d'observer les taux de réussite pour un item du niveau 400 du domaine des nombres. En principe cet item est un item «facile» et, si la hiérarchie entre les items tient, on devrait obtenir des taux de réussite plus élevés que pour l'item précédent qui se situe au niveau 475 du domaine de l'analyse des données. La tâche à exécuter pour cet item consiste à réaliser la multiplication 15 x 9. On observe que les taux de réussite sont moyens, sauf pour la Hongrie qui affiche un taux de 84% et la Norvège qui affiche un taux de 30%. On constate aussi que pour ces deux PJ les valeurs pour la dimension des «opportunités d'apprentissage» sont assez semblables avec 68% et 54%, des valeurs qui sont du même ordre de grandeur que ce qu'on observe pour le Québec avec 67% mais qui affiche un taux de réussite de 66%.

On peut certainement se demander ce qui fait que les taux de réussite diffèrent à ce point pour un item qui est classé comme assez facile. Peut-être qu'un examen plus poussé du programme de quatrième année amènerait le constat que le principe des multiplications comme 15 x 9 n'est pas abordé au même moment dans le cursus scolaire de chacun de ces PJ. En bout de ligne cependant, on ne le sait pas vraiment et, sans un examen approfondi du lien entre les items et les curriculums enseignés, on ne le saura probablement jamais.

Tableau 7. Exemple d'un item du domaine des nombre au niveau 625

Niveau : 625
Domaine : nombres (voir p.87 dans le rapport TIMSS 2003)
Description : identification de la représentation décimale d'une fraction avec un dénominateur de 10.

Lequel des choix suivants représentent 7/10 ?

a) 70
b) 7
c) 0.7
d) 0.07

PJ	Taux de réussite	Score moyen pour la section sur les nombres	OA : %
Québec	26%	508	67
Ontario	47%	494	75
Italie	58%	502	88
Philippines	49%	380	95
Hongrie	17%	524	68
Norvège	17%	440	54

Tableau 8. Exemple d'un item du domaine de l'analyse des données au niveau 475

Niveau : 475
Contenu : analyse de données (voir p.96 dans le rapport TIMSS 2003)
Description : Compléter un diagramme en barres à partir d'un problème présenté avec des mots (*word problem*)

PJ	Taux de réussite	Score moyen pour la section sur l'analyse des données»	OA : %
Québec	83%	506	69
Ontario	85%	544	91
Italie	71%	497	83
Philippines	29%	384	72
Hongrie	84%	513	74
Norvège	75%	479	54

Tableau 9. Exemple d'un item du domaine des nombres au niveau 400

Niveau : 400
Contenu : Nombre (p.99 dans le rapport TIMSS 2003)
Description : Multiplication d'un nombre de deux chiffres par un nombre d'un chiffre.

15X9=

PJ	Taux de réussite à l'item	Score moyen pour la section sur les nombres	OA : %
Québec	66%	508	67
Ontario	54%	494	75
Italie	75%	502	88
Philippines	59%	380	95
Hongrie	85%	524	68
Norvège	30%	440	54

8. Comparaisons 2003 – 1995

Plusieurs différences existent entre l'opération de 2003 et celle de 1995. Pour le Québec, la première de ces différences se situe au niveau des caractéristiques de l'échantillon des élèves participants. Contrairement à ce qui s'est produit en 2003, le Québec n'a pas participé à l'enquête de 1995 en tant que PJ à part entière. En effet :

> «L'échantillon d'élèves de chacune de ces provinces a été choisi de façon à représenter une population d'élèves de la province correspondante. Au Québec, par contre, l'échantillon d'élèves n'a pas été choisi pour représenter la population d'élèves de cette province mais bien pour compléter la population d'élèves canadiens.»[16]

Ainsi, pour l'échantillon canadien TIMSS 1995 des élèves de quatrième année, on relève à travers le Canada la participation de 390 écoles et de 8 400 élèves environ, mais parce que ce n'est pas indiqué dans les rapports on ignore combien d'écoles et d'élèves québécois ont participé à l'enquête de 1995. On

[16] Rapport: Bulletin statistique de l'éducation, no 6 août 1998, direction des statistiques et des études quantitatives.

peut penser cependant que ces nombres ne sont pas très élevés puisqu'il n'y avait aucun impératif à «bien» représenter la population des élèves.

Si pour le Québec, le nombre d'élèves de 4e année de l'échantillon non représentatif de 1995 était, disons, égal à 500 et qu'en comparaison, l'échantillon représentatif de 2003 a été constitué par plus de 4000 élèves, les comparaisons entre les résultats des deux opérations sont assez fragiles, car au départ le potentiel de distorsion des résultats de 1995 est assez élevé. En fait, on peut même penser que pour le Québec les résultats de 2003 sont déjà au départ plus «précis» et donnent un portrait plus proche de la réalité que les résultats de 1995.

Une deuxième différence entre les opérations de 1995 et 2003 est celle qui touche la composition des épreuves. La répartition des items selon les domaines est assez différente et il n'y pas en 1995 de domaine dominant comme celui des nombres pour l'épreuve de 2003. Il y a un domaine de plus en 1995 et les items et les points sont répartis plus également en fonction des domaines. Malgré tout, selon les principes de la validité «apparente», ces deux épreuves visent des compétences générales semblables, même si pour certaines compétences particulières le lien apparaît être plus ténu.

Sur les 161 items disponibles pour former les cahiers de l'opération de 2003, 37 items étaient des items présents dans les cahiers lors de l'opération de 1995 (voir tableau 10 ci-dessous). Ces items représentent 22% du nombre de points en 2003. La comparaison entre les scores des échantillons d'élèves aux deux épreuves reste ainsi délicate, particulièrement pour les domaines des formes, de la géométrie et de l'analyse des données, car on y retrouve respectivement 2, 4 et 4 items communs aux deux épreuves. Tous les items retenus pour faire partie des deux épreuves étaient des items à choix de réponse.

Tableau 10. Items communs aux cahiers des opérations 2003 et 1995.

Domaine	**Nombre d'items communs**
Nombres	19
Formes	2
Mesure	8
Géométrie	4
Données	4

Finalement, comme cela a déjà été mentionné, l'échelle des scores dépend toujours des PJ qui ont participé. Cela est d'autant plus vrai lorsque les scores issus de deux opérations éloignés dans le temps sont appariés pour être placés sur une échelle commune. Un total de 17 PJ ayant participé aux opérations de 1995 ont aussi participé aux opérations de 2003 et c'est à partir des résultats des élèves de ces PJ participants que les scores de 1995 et 2003 ont été mis sur la même échelle.

L'échelle de référence reste toujours arbitraire et toute autre échelle pourrait être utilisée à condition que la transformation conserve les propriétés de l'échelle. Ainsi, si les scores sont placés sur une autre échelle en les divisant par 10, on obtient pour le Québec une moyenne de 55 en 1995 et une moyenne de 51 en 2003 : selon ce qui est dit dans les rapports la différence est statistiquement significative, mais étant donné ce qui a été dit à propos de l'équivalence des échantillons et des différences entre les épreuves, il est important de se demander si la différence est pertinente pour porter un jugement d'ensemble sur les différences systémiques entre ces deux moments ?

Pour produire une comparaison vraiment pertinente, il faudrait examiner les items communs aux deux épreuves et comparer leur contenu, comparer les programmes et les objectifs d'apprentissage des deux époques, comparer leurs liens avec ce qui a été réellement enseigné et, finalement, comparer les résultats aux items communs aux deux épreuves en tenant compte de ce qui a été réellement enseigné.

9. Suggestions pour le pilotage du système

1. Pousser les analyses TIMSS plus loin pour illustrer le lien avec le curriculum enseigné et pour déterminer la réussite respective 1995-2003 aux items communs. Préparer les analyses en vue de l'opération TIMSS 2007 prévoir l'examen des données récoltées en lien avec les différences entre le curriculum prescrit, celui au programme, et le curriculum enseigné.

2. Utiliser ce qui est déjà en place dans le système et créer une base de données longitudinale permettant de suivre les résultats des élèves sur une longue période. Par exemple, en utilisant le code permanent des élèves et en se centrant sur l'orthographe et la syntaxe, il est possible de relier les résultats aux épreuves de français écrit de la fin de la dernière année du primaire, de la fin de la cinquième année du secondaire et de la fin du parcours au collégial.

3. Développer des capacités de recherche fortes pour faire un suivi des constats et pour publiciser ces constats en fonction du curriculum et des programmes. Il serait ainsi possible de réaliser des études par échantillonnage aux deux ans en lecture, écriture, mathématiques et sciences, avec les élèves de la 3e année du primaire et de la 2e année du secondaire.

10. Conclusion

La participation aux enquêtes internationales en éducation demande un bon investissement en temps et en argent pour les PJ. Il est donc important de tirer les bonnes conclusions au sujet des résultats des élèves échantillonnés et surtout pour ce qui est des liens entre ces résultats et sur ce qui se passe en salle de classe.

Si les seules inférences qui sont faites sur les connaissances et les compétences des élèves le sont à partir des moyennes générales ou encore à partir du rang respectif de ces moyennes, il est évident que l'objectif général de l'enquête, représenté ci-dessous, n'est pas atteint et que le portrait sera toujours tronqué.

> «…d'améliorer l'enseignement et l'apprentissage en mathématiques et en science en fournissant des données au sujet de la performance des élèves selon les caractéristiques de différents curriculum, de différentes pratiques d'enseignement et de différents environnements scolaires »

Les conclusions à tirer et les comparaisons qu'il est possible d'effectuer sont toujours tributaires de plusieurs assurances à donner au sujet de la représentativité de l'échantillon des élèves, au sujet de la comparabilité des échantillons entre les PJ, au sujet de la représentativité des items de l'épreuve, au sujet des limites des modèles de mesure et, finalement, au sujet des opportunités d'apprentissage. Comme cela a été illustré dans cet avis, plusieurs points restent en suspens quant à ces dimensions de l'enquête TIMSS 2003 pour le Québec mais aussi pour plusieurs PJ participants, et il donc est important de garder en tête que ces points en suspens ont un impact sur les constats potentiels.

Les comparaisons entre deux moments dans le temps sont possibles, 1995-2003 par exemple, mais la performance «normale» des élèves n'est pas connue car l'importance de l'erreur de mesure n'est pas connue et elle pourrait l'être uniquement en répétant souvent l'opération. Avec deux points d'ancrage il y a peu d'information pour dire si les élèves régressent ou s'améliorent parce qu'on sait peu de chose sur ce qui prévaut «habituellement», c'est-à-dire qu'il y a peu de choses connues sur les variations normales à long terme et, donc, sur la

distribution de la moyenne des résultats d'un échantillon d'élèves. En conséquence, une moyenne supérieure pour une année donnée pourrait très bien être une valeur extrême par rapport à la distribution et une valeur plus basse pour une autre année être la moyenne «normale» pour la population.

Si l'examen des différences entre deux moments dans le temps se limite à faire une soustraction de deux moyennes au lieu d'un examen attentif du contenu des items communs, des taux de succès respectifs et des opportunités d'apprentissage, l'opération n'atteint pas les buts fixés et elle prête flanc à des critiques rapides parce les données pertinentes à une compréhension de l'évolution de la situation ne sont pas disponibles ou accessibles. Cependant, il existera toujours des limites méthodologiques à ces comparaisons entre cohortes. Certains l'oublient souvent, une relation statistique entre deux «variables» n'est pas synonyme d'une relation de cause à effet entre ces deux variables. Ainsi, attribuer la différence entre les résultats observés aux épreuves de 1995 et 2003 à la réforme de l'éducation au Québec constitue un saut conceptuel que les caractéristiques des données ne permettent pas de faire à mon avis. La situation pourrait être différente si l'opération TIMSS 2007 au Québec est bien planifiée et si elle l'est en fonction d'une étude des impacts possibles de la réforme sur les compétences en mathématiques et en science des élèves de l'ordre primaire.

11. Références

Blais, J.-G. (1992). IAEP Technical Report: Volume 2. Princeton, NJ: Educational Testing Service. 110 pages, août.

Brown, M. (1999). Problems of interpreting international comparative data. Oxford studies in comparative education, vol. 9, no 1, 183-205.

Burstein, L. (1992). The IEA study of mathematics III: student growth and classroom process. Oxford: Pergamon press.

Husén, T. (1967). International study of achievement in mathematics, vol. I and II. Stockholm: Almquist & Wiksell.

McLean, L. (1996). Large scale assessment programmes in different countries and international comparisons. Dans H. Goldstein et T. Lewis: Assessment: problems, developments and statistical issues, 189-207. Londres : Wiley.

Martin, M.O., Kelly D.L. (dir. publ.) (1996). TIMSS Technical Report. Chesnut Hill, MA: TIMSS and PIRL International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Beaton, A.E., Chrostowski, S.J. (1997). Mathematics achievement in the primary years: IEA Third International Mathematics and science study. TIMSS International Study Center, Boston College, USA.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J. Chrostowski, S.J. (2004). TIMSS 2003 International Mathematics Report. TIMSS and PIRL International Study Center, Boston College, USA.

Martin, M.O., Mullis, I.V.S., E.J. Chrostowski, S.J. (dir. publ.) (2004). TIMSS 2003 Technical Report. Chesnut Hill, MA: TIMSS and PIRL International Study Center, Boston College.