

Mesure et évaluation en éducation



Un regard didactique sur les évaluations du PISA et de la TIMSS : mieux les comprendre pour mieux les exploiter

Antoine Bodin and Nadine Grapin

Volume 41, Number 1, 2018

URI: <https://id.erudit.org/iderudit/1055897ar>

DOI: <https://doi.org/10.7202/1055897ar>

[See table of contents](#)

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)

2368-2000 (digital)

[Explore this journal](#)

Cite this article

Bodin, A. & Grapin, N. (2018). Un regard didactique sur les évaluations du PISA et de la TIMSS : mieux les comprendre pour mieux les exploiter. *Mesure et évaluation en éducation*, 41(1), 67–96. <https://doi.org/10.7202/1055897ar>

Article abstract

Two international studies that aim to assess students' knowledge and skills in mathematics have won a prominent place in the global educational world: Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS). While these studies come with numerous works describing their frameworks, and while a lot of research in edumetrics, sociology and educational sciences have been published, what light the mathematics didactics can place on them? After a short outline of their frameworks, we will make use of two different approaches, both being based on a priori analysis of the tasks for analysing the content of the assessments and for reinterpreting the results. The first approach, labelled didactical and cognitive, is associated with the implicative statistical analysis and allows a new look on the PISA results as well as on comparisons between countries. The second, nested in the didactical anthropological theory, addresses mainly TIMSS content and the distribution of its assessment items in relation to a given mathematical syllabus.

Un regard didactique sur les évaluations du PISA et de la TIMSS : mieux les comprendre pour mieux les exploiter

Antoine Bodin

*Institut de recherche pour l'enseignement des mathématiques (IREM)
Université de la Méditerranée Aix-Marseille*

Nadine Grapin

*Laboratoire de didactique André Revuz
Université Paris-Est Créteil*

MOTS CLÉS : évaluations internationales, didactique des mathématiques, analyse statistique implicative, PISA, TIMSS

Deux études internationales évaluant les connaissances des élèves en mathématiques occupent une place importante dans le paysage éducatif mondial : celles du Programme international pour le suivi des acquis des élèves (PISA) et de la Trends in International Mathematics and Science Study (TIMSS). Si ces études sont accompagnées d'une littérature abondante décrivant leurs cadres, et si de nombreuses recherches en éducativité, en sociologie ou encore en sciences de l'éducation ont régulièrement été menées, quel regard la didactique des mathématiques peut-elle porter sur ces études ? Après avoir présenté succinctement les cadres de ces deux évaluations, nous exploitons deux approches didactiques distinctes, basées sur l'analyse a priori des tâches pour analyser le contenu de l'évaluation et pour réinterpréter les résultats des élèves. Couplée à une analyse statistique implicative, la première, didactique et cognitive, apporte un regard neuf sur les résultats du PISA et permet de mener des comparaisons entre pays. La seconde approche, située dans le cadre de la théorie anthropologique du didactique, s'intéresse davantage au contenu de la TIMSS et à la répartition des questions au regard d'une organisation mathématique de référence.

KEY WORDS: international studies, mathematics didactics, implicative statistical analysis, PISA, TIMSS

Two international studies that aim to assess students' knowledge and skills in mathematics have won a prominent place in the global educational world: Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS). While these studies come with numerous works describing their frameworks, and while a lot of research in edumetrics, sociology and educational sciences have been published, what light the mathematics didactics can place on them? After a short outline of their frameworks, we will make use of two different approaches, both being based on a priori analysis of the tasks for analysing the content of the assessments and for reinterpreting the results. The first approach, labelled didactical and cognitive, is associated with the implicative statistical analysis and allows a new look on the PISA results as well as on comparisons between countries. The second, nested in the didactical anthropological theory, addresses mainly TIMSS content and the distribution of its assessment items in relation to a given mathematical syllabus.

PALAVRAS-CHAVE: avaliações internacionais, didática das matemáticas, análise estatística implicativa, PISA, TIMSS

Dois estudos internacionais que avaliam os conhecimentos dos alunos em matemática ocupam um lugar importante no panorama educativo mundial: os do Programa Internacional de Avaliação de Estudantes (PISA) e os do Trends in International Mathematics and Science Study (TIMSS). Sendo estes estudos acompanhados por uma abundante literatura descrevendo os seus quadros, e que muitas investigações em edumetria, sociologia ou mesmo nas ciências da educação são realizadas regularmente, que olhar a didática das matemáticas pode apresentar sobre estes estudos? Após apresentar sucintamente os quadros dessas duas avaliações, exploramos duas abordagens didáticas distintas, baseadas na análise a priori das tarefas para analisar o conteúdo da avaliação e para reinterpretar os resultados dos alunos. Juntamente com uma análise estatística implicativa, a primeira, didática e cognitiva, traz um novo olhar para os resultados do PISA e permite comparações entre países. A segunda abordagem, situada no quadro da teoria antropológica da didática, concentra-se mais no conteúdo do TIMSS e na distribuição de questões em termos de uma organização matemática de referência.

Introduction

Depuis les premières enquêtes organisées par l'International Association for the Evaluation of Educational Achievement (IEA) il y a près de 60 ans, les études internationales centrées sur les acquis des élèves ont été en croissance continue en ce qui concerne leur nombre, le nombre de pays concernés et le nombre de rapports, d'articles et d'ouvrages de tous types publiés à leur propos.

Ces enquêtes touchent aujourd'hui un grand nombre de domaines, de la lecture à la culture numérique en passant par l'éducation civique ; elles touchent les élèves de l'enseignement élémentaire, ceux des enseignements secondaires moyen et supérieur ainsi que les adultes. Leur influence sur l'évolution des systèmes éducatifs est importante et bien documentée (Mons, 2009), tout comme celle dans des domaines a priori éloignés de l'enseignement (économie, politique, etc.).

Parmi les nombreuses enquêtes et les nombreux domaines qu'elles couvrent, nous nous limiterons, dans cet article, à étudier et à comparer le volet mathématique du Programme international pour le suivi des acquis des élèves (PISA) et de la Trends in International Mathematics and Science Study (TIMSS). De plus, nous nous limiterons à leur aspect didactique.

Alors que de très nombreuses spéculations entourent ces enquêtes et alors que de nombreuses orientations et décisions sont prises partout dans le monde en s'appuyant sur les résultats de type palmarès qu'elles produisent, très rares sont les études qui s'interrogent sur ce que mesure réellement le thermomètre utilisé (c.-à-d. sa validité, et non seulement sa fidélité)¹. Citons cependant Dupé et Olivier (2005), Ruddock, Clausen-May, Purple et Ager (2006), Bodin (2007a, 2007b), Wu (2009), Roditi et Salles (2015) ainsi que Bodin, Decamp, De Hosson et Grapin (2016).

Dans cet article, nous utilisons le terme *évaluation* pour parler du PISA et de la TIMSS, de préférence aux termes *enquête*, *étude* ou *programme* souvent utilisés, et ce, bien que le processus d'évaluation ne se limite pas au long parcours politique, technique et managérial défini par les cadres de référence, et bien qu'il finisse par faire l'objet d'interprétations et de

conclusions locales qui échappent largement aux programmes proprement dits. Responsable de l'équipe d'experts mathématiques du PISA jusqu'en 2009, Jan de Lange a pris une grande part dans l'élaboration du cadre de référence et dans le choix des questions. À propos du « choc PISA » en Allemagne, il constate :

Il semble qu'il n'y ait aucun intérêt pour le contenu de l'étude. Il n'y a pas eu de discussion sur les questions qui ont été utilisées, sur les compétences nécessaires, sur la qualité des instruments ni sur la relation avec le curriculum. Il n'y a pas eu non plus de discussion sur le fait que l'étude était comparative ou normative [par opposition à critériée], ni sur le fait qu'elle n'indiquait aucune qualité « absolue ». Ce fait souvent ignoré peut poser de sérieux problèmes d'interprétation, particulièrement en ce qui concerne les conséquences politiques (de Lange, 2007, p. 1111, traduction personnelle).

Ce jugement peut encore s'appliquer au cas français comme au cas de la plupart des autres pays, tant l'attention est attirée principalement sur les classements, et tant l'évaluation est conditionnée par des considérations d'ordre éducatif, et non didactique ou épistémologique (et ce, pour des raisons qui ne sont pas l'objet de cet article).

Parallèlement, si différents auteurs (p. ex., Chevallard et Feldmann [1986], Bodin [1997], Deblois, Freiman et Rousseau [2007] ainsi que Roditi et Chesné, [2012]) ont souligné le peu de recherches en didactique des mathématiques sur l'évaluation des connaissances des élèves, force est de constater une réelle évolution ces dernières années, notamment dans les recherches françaises sur les évaluations externes. En particulier, cet article se situe, en France, dans la continuité d'une étude commandée par le Conseil national d'évaluation du système scolaire (CNESCO) (Bodin et al., 2016) et en reprend certaines parties².

Courte présentation de la TIMSS et du PISA

Le PISA ayant beaucoup emprunté à la TIMSS, l'organisation et les méthodologies de ces deux évaluations sont proches, de même que la façon dont ils présentent et diffusent leurs résultats. Toutefois, ce ne sont pas ces questions qui nous intéressent ici.

L'évaluation du PISA a lieu tous les trois ans et concerne les élèves ayant entre 15 ans et 3 mois et 16 ans et 2 mois, quelle que soit la place qu'ils occupent dans le système éducatif.

Les évaluations de la TIMSS ont lieu tous les quatre ans pour les élèves de 4^e année du primaire (*grade 4*, CM1 en France) et de 8^e année (*grade 8*, 4^e en France). Elle a aussi lieu de façon plus irrégulière pour les élèves des classes terminales du secondaire engagés dans une série à dominante mathématique (TS en France).

Les cadres de référence ont évolué au fil des années, mais nous nous référerons essentiellement à ceux définis pour les dernières évaluations : 2012 et 2015 pour le PISA, 2015 pour la TIMSS.

La TIMSS évalue les savoirs et savoir-faire mathématiques (*knowledge and skills*) à partir d'une analyse des curricula des pays participants et d'une liste de contenus organisant les notions mathématiques (le syllabus ou plan de cours). La TIMSS est d'abord un projet de recherche qui cherche à éclairer les liens entre les programmes et instructions officielles (curriculum souhaité), les pratiques d'enseignement (curriculum implémenté) et les acquis des élèves (curriculum atteint). La TIMSS tend de ce fait à poser des questions qui sont immédiatement reconnues comme des questions mathématiques.

Contrairement à la TIMSS, qui s'appuie sur les programmes enseignés, le PISA, qui est mené par l'Organisation de coopération et de développement économiques (OCDE), s'appuie sur une analyse des compétences nécessaires à l'adaptation à un monde en évolution rapide dans un contexte d'éducation tout au long de la vie. Ainsi, en 1997, l'OCDE a assigné au PISA la mission de « déterminer dans quelle mesure les élèves qui approchent du terme de leur scolarité obligatoire possèdent les savoirs et les savoir-faire indispensables pour participer à la vie de la société » (OCDE, 2005). Il ne s'agit pas de s'intéresser aux curricula ni d'évaluer par rapport à eux, mais de partir de ce qui serait reconnu comme socialement utile dans le futur. Ces évaluations sont d'abord conçues pour aider au pilotage des systèmes éducatifs dans un sens souhaité par l'OCDE, dans le cadre d'une concertation inter-États.

Le mot clé privilégié du PISA est « vie réelle », tandis que celui de la TIMSS serait simplement « mathématiques ». Le PISA évalue en fait la littératie, notion qu'il définit et organise suivant plusieurs directions : compréhension de l'écrit (*reading literacy*), littératie mathématique, littératie scientifique, puis financière, numérique (*digital*), citoyenne, etc.

Le présent article se limitant à la littératie mathématique, voici la définition qu'en donne l'OCDE :

La littératie mathématique est la capacité d'un individu à formuler, employer et interpréter des mathématiques dans une variété de contextes. Cela inclut la capacité à raisonner mathématiquement et à utiliser des concepts, des procédures, des faits et des outils mathématiques pour décrire, expliquer et prévoir des phénomènes. Elle aide les individus à reconnaître le rôle que les mathématiques jouent dans le monde, à produire des jugements bien fondés et à prendre les décisions nécessaires en citoyens constructifs, engagés et réfléchis (OECD, 2013, p. 17, traduction personnelle).

Il importe de noter ici que, contrairement à la TIMSS, la littératie mathématique du PISA n'est pas identifiable aux mathématiques scolaires telles qu'elles sont habituellement définies et enseignées. Ainsi que le souligne de Lange à propos des politiciens, ce point est en général mal compris :

Beaucoup de politiciens manquent des connaissances nécessaires pour réaliser que le PISA ne mesure qu'une partie des mathématiques, plus précisément la littératie. [...] L'ignorance, de ce fait [...], conduit à des problèmes de communication dans plusieurs pays et avec beaucoup de mathématiciens qui tendent à voir les mathématiques du PISA comme représentant la discipline mathématique (de Lange, 2007, p. 1119, traduction personnelle).

Ainsi, les philosophies sous-jacentes à la TIMSS et au PISA et leurs objectifs sont très différents, même si, dans la pratique, un certain rapprochement se fait, ne serait-ce que sous la pression des États.

Les cadres de référence

Les cadres de référence (*frameworks*) sont des documents contractuels qui lient les gestionnaires, les commanditaires et les gouvernements des pays participants. Sortes de cahiers des charges des évaluations, ils en précisent les motivations, les objectifs, les domaines et sous-domaines évalués, la façon dont sont choisies les questions et dont elles sont réparties dans les catégories. Malgré des rapprochements effectués ces dernières années, les cadres de référence de la TIMSS et du PISA restent assez différents.

Le cadre de référence de la TIMSS pour les mathématiques s'articule autour de deux domaines : les contenus (nombres, formes géométriques, représentations de données) et les compétences cognitives (connaître, appliquer, raisonner).

Le cadre du PISA, lui, s'articule autour :

1. Des processus mathématiques

Ils décrivent ce que font des individus pour établir un lien entre le contexte du problème et les mathématiques et, donc, pour résoudre le problème, ainsi que les capacités qui sous-tendent ces processus. Trois processus sont définis : formuler (des situations de façon mathématique), employer (des concepts, faits, procédures et raisonnements mathématiques) et interpréter (appliquer et évaluer des résultats mathématiques). Ces trois processus constituent ce que le PISA nomme le cycle de modélisation (voir Figure 1).

2. Des contenus mathématiques

Ils pourraient être utilisés pour le traitement des questions : variations et relations, espace et formes, quantité, incertitude et données.

3. Des contextes

Des contextes dans lesquels les questions s'inscrivent : personnels, professionnels, sociétaux et scientifiques.

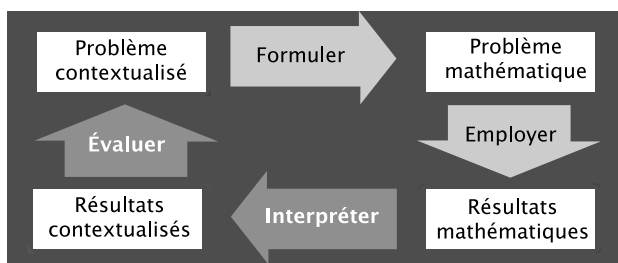


Figure 1 : Cycle de modélisation du PISA à partir de 2012 (OECD, 2013)

Après le mot *littératie*, le mot clé le plus important du volet mathématique du PISA est *modélisation*. Le cadre de référence précise que les questions d'évaluation devront être ancrées dans le monde réel et que le cycle de modélisation devra s'imposer pour leur résolution. Dans la pratique, pour des raisons liées aux problèmes de mesure et au fait que le temps moyen alloué pour répondre aux items est de l'ordre de deux minutes, chacune des questions d'évaluation est censée mobiliser de façon privilégiée l'un ou l'autre de ces processus.

Pour chacune des deux évaluations, le contenu des tests est donc construit relativement à ces cadres. Ainsi, la répartition des questions est réalisée selon les compétences ou les processus cognitifs, les contenus et les contextes. Puis, elle est soigneusement planifiée en termes de pourcentages dans les cadres de référence (OECD, 2013 ; Mullis et Martin, 2013).

Étude didactique des évaluations externes : enjeux et méthodologies

Comme nous l'avons montré précédemment, les cadres sous-tendant les évaluations définissent les domaines évalués en lien avec des savoirs mathématiques, des processus, des compétences (définies de façon différente selon l'évaluation), etc. Les questions d'évaluation sont ainsi caractérisées par leurs rapports à ces catégories et domaines et par un type de format de question (à choix multiple, ouverte, semi-ouverte). Elles sont accompagnées des consignes de codage des réponses données par les élèves.

Une étude didactique de ces évaluations vise non seulement à en étudier le contenu (tâches à réaliser, conditions de passation, format des questions, codage, etc.), mais aussi à en interpréter les résultats au filtre des programmes scolaires et des pratiques enseignantes. Nous avons retenu deux approches, toutes deux ancrées dans un cadre didactique, qui permettent une telle mise en relation et qui ont été mises à l'épreuve pour l'analyse d'évaluations externes.

La première, que nous qualifions de «cognitive et didactique», a été développée par Gras et Bodin (2017) et a été particulièrement exploitée pour analyser le contenu des évaluations du PISA de 2003 et de 2012. La seconde, «épistémologique et didactique», a été développée par Grugeon-Allys et Grapin (2015) et a été d'abord utilisée pour l'étude du contenu d'évaluations nationales réalisées sur échantillon, puis pour l'analyse des évaluations du PISA de 2012 et de la TIMSS de 2015 (Bodin et al., 2016).

Approche cognitive et didactique

Complexité et difficulté des questions d'évaluation

Pour l'analyse de la complexité cognitive, nous avons utilisé la taxonomie Gras-Bodin. Utilisée depuis de nombreuses années (Gras, 1979 ; Gras et Bodin, 2017), et surtout validée dans le cadre de nombreuses

recherches en didactique des mathématiques, elle s'éloigne un peu de la taxonomie de Bloom. Nous ne la décrivons pas explicitement, mais pré-cisons simplement les appellations des niveaux de complexité :

- A : Connaissance et reconnaissance ;
- B : Compréhension ;
- C : Application ;
- D : Créativité ;
- E : Jugement.

Par définition, la taxonomie est hiérarchisée, c'est-à-dire qu'une question de niveau A devrait être mieux réussie qu'une question de niveau C. Classer une question au niveau C suppose en effet qu'elle met en jeu des connaissances et suppose la compréhension de celles-ci et de leurs conditions d'utilisation. Ce principe de hiérarchisation doit cependant être nuancé : en effet, il est d'autant plus facilement vérifié que les tâches concernées relèvent d'un même champ conceptuel (Vergnaud, 1990).

Le niveau de complexité taxonomique d'une question du domaine cognitif se définit à partir d'une analyse didactique de la tâche a priori. Il cherche à être aussi intrinsèque que possible, tandis que la difficulté est une donnée empirique relative à un groupe d'élèves donné et s'exprime par un taux de réussite.

Cette taxonomie est associée à l'analyse statistique implicative, méthode d'exploration des données que nous avons utilisée pour analyser les données du PISA et que nous présentons ici.

Analyse statistique implicative

Les notions de difficulté et de complexité concernent les items de façon individuelle. Les traitements de la théorie de réponses aux items ou TRI (*item response theory* ou IRT) utilisés par le PISA et par la TIMSS concernent des ensembles d'items de façon globale ; les items y perdent totalement leur individualité et les relations qu'ils « entretiennent » entre eux sont totalement ignorées. Or, les didacticiens sont justement intéressés par ces relations, en matière de contenu, mais aussi de comportement des élèves par rapport à ces items. Autrement dit, dans quelle mesure les élèves qui réussissent un item A donné ont-ils aussi tendance à réussir tel item B ?

L'analyse statistique implicative développée par Gras (1979) et ses collègues permet de répondre à cette question. Il s'agit d'une méthode d'analyse des données qui permet d'étudier l'organisation d'un ensemble de données et, particulièrement, les dépendances orientées entre variables. Développée initialement dans le cadre de la didactique des mathématiques, la théorie implicative est maintenant utilisée dans de nombreux domaines, de la psychologie à la biologie en passant par les beaux-arts.

Pour simplifier, disons que l'analyse statistique implicative procède d'une approche probabiliste : étant donné deux événements A et B (p. ex., la réussite à deux items A et B) et alors qu'on vise l'inclusion de A dans B, c'est-à-dire que les élèves qui réussissent l'item A réussissent aussi l'item B, on se demande dans quelle mesure, sous l'hypothèse d'indépendance de ces deux événements, ce qui est observé en termes de réalisation conjointe de A et de non-B peut ou non être le fruit du hasard. La probabilité d'avoir A et non-B dans ces conditions définit l'indice d'implication de A vers B. Un indice d'implication de A vers B supérieur à 0,95 signifiera alors que la probabilité d'obtenir aussi peu d'éléments de A qui ne soient pas dans B est inférieure à $1 - 0,95$, c'est-à-dire à 0,05.

L'indice d'implication est complété par un indice de confiance de l'inclusion de A dans B qui, dans le cas de variables binaires, correspond à l'estimation de la probabilité conditionnelle de B sachant A (probabilité que nous écrirons en pourcentage). Un indice d'implication élevé peut en effet être accompagné d'un indice de confiance faible, ce qui serait de moindre intérêt dans le cas de questions d'évaluation (Bodin et Giovaninni, 2017).

L'ajout de l'indice de confiance est particulièrement bien adapté aux évaluations à grande échelle. Constaté que, d'un point de vue probabiliste, la réussite à une tâche A s'accompagne d'une tendance à la réussite à une tâche B plutôt qu'à l'échec à cette même tâche est certes utile, mais constater que, de plus, dans la population observée, cela est vérifié par 80 ou 90 % des élèves sera de nature à fournir des clés utiles à l'interprétation.

L'analyse statistique implicative associée à l'indice de confiance permet de repérer des chaînes qui s'apparentent à des échelles de Guttman et qui permettent de repérer l'évolution des élèves dans le développement d'une capacité particulière et, au-delà, d'une compétence particulière, du moins, lorsque l'enquête et les données s'y prêtent.

Nous avons en particulier utilisé cette méthode avec les données élèves du volet mathématique du PISA de 2012 et avons pu non seulement étudier des organisations globales que nous n'aurions pas pu étudier avec d'autres méthodes, mais aussi comparer cette organisation entre plusieurs pays (Khaled, Bodin et Couturier, 2017). Nous présentons quelques résultats dans la section *Étude globale sur l'ensemble des questions*, mais nous voulions surtout montrer qu'il y avait là une piste qui nous semble intéressante pour aborder la complexité des données issues des évaluations internationales.

Approche épistémologique et didactique

L'approche épistémologique et didactique conduit à une méthodologie d'analyse reposant sur l'approche anthropologique du didactique (Chevallard, 1999) et développée par Grugeon-Allys et Grapin (2015). En prenant en compte les différentes étapes de la transposition didactique, elle permet non seulement de dresser un état des lieux des connaissances des élèves, mais aussi de les mettre en perspective des programmes scolaires et des pratiques enseignantes.

Pour ce faire, une organisation mathématique de référence (Bosch et Gascon, 2005) est définie, sur un domaine mathématique donné, à partir du savoir en jeu et des éléments épistémologiques sur lesquels il repose. Le domaine mathématique évalué est alors structuré par une liste de types de tâches pouvant être résolus par des techniques ; ces dernières étant justifiées par des éléments mathématiques (propriétés, théorèmes) qualifiés de « technologies » et provenant de théories mathématiques.

Montrons désormais la façon dont une telle organisation mathématique permet d'étudier le contenu d'une évaluation et d'en interpréter les résultats.

Étude du contenu de l'évaluation

L'analyse a priori de la tâche est l'élément clé de cette méthodologie. En premier lieu, il est nécessaire que la tâche proposée dans le test soit en correspondance avec l'objectif d'évaluation qui lui est assigné. (C'est la validité telle qu'elle est définie en évaluation³.) Cette remarque, valable quelle que soit l'approche considérée, n'est pas négligeable et ne va pas toujours de soi. En effet, différents paramètres, notamment le choix des valeurs de variables didactiques, peuvent venir biaiser l'objectif de la tâche. Par exemple, pour évaluer la résolution algébrique d'une équation, il faut

être vigilant à proposer une équation qui ne se résout pas simplement avec une méthode arithmétique. En complément du jeu sur les variables didactiques, les conditions de passation peuvent aussi avoir un effet sur la pertinence de la tâche relativement à son objectif. Par exemple, pour évaluer des capacités d'élèves en calcul mental, il est nécessaire de veiller au temps accordé à la réponse. (Si l'élève a le temps de poser le calcul, ce ne sont plus les mêmes connaissances qui sont alors évaluées.)

Localement, chacune des tâches de l'évaluation est caractérisée par le type ou les types de tâches dont elle relève (en lien avec les techniques et les technologies), par différentes valeurs de variables selon le domaine mathématique considéré (registres de représentation en jeu, type de nombre, type de problème, etc.), par le caractère outil-objet et l'ancienneté des savoirs en jeu dans la résolution de la tâche. Pour rendre compte a priori de la complexité de la tâche, nous nous référons aux trois niveaux de mise en fonctionnement (NMF) des connaissances (Robert, 1998), à savoir :

- le niveau *technique* pour des « tâches qui amènent à des applications immédiates des connaissances, c'est-à-dire simples (sans adaptation) et isolées (sans mélanges), où seule une connaissance précise est mise en œuvre sans aucune adaptation, mis à part la contextualisation nécessaire » (p. 165) ;
- le niveau *mobilisable*, lorsque les « tâches nécessitent des adaptations de connaissances qui sont en partie au moins indiquées » (p. 166). Par exemple, le problème peut présenter des données inutiles ou, en géométrie, la configuration peut être plus complexe ;
- le niveau *disponible*, lorsque c'est à l'élève de reconnaître les connaissances à utiliser. Par exemple, en géométrie, résoudre un problème dans lequel une configuration est à reconnaître, mais n'est pas visible immédiatement.

Enfin, l'analyse a priori doit permettre de déterminer un codage adapté à l'objectif d'évaluation assigné à la tâche. Par exemple, si une tâche proposée sous la forme de question ouverte vise à évaluer la résolution de problème, une erreur liée à la modélisation du problème et une erreur de calcul (alors que le modèle mathématique utilisé est correct) ne peuvent pas être codées de la même façon ; les crédits partiels accordés à certaines réponses dans les évaluations à grande échelle sont ainsi prévus en ce sens. Dans le cas des questions à choix multiple (QCM), les distracteurs pro-

posés méritent aussi de correspondre à des «erreurs types» (conceptions erronées, raisonnement faux, etc.) afin que les réponses choisies puissent être interprétées le plus justement possible.

Globalement, ensuite, l'étude de l'ensemble du contenu du test repose sur cette analyse a priori : au regard de l'organisation mathématique de référence, il est possible de vérifier la couverture du domaine, de cibler des manques ou des redondances sur certains types de tâches ou certaines techniques et d'étudier l'équilibre du test en matière de complexité. Pour un domaine numérique ou algébrique, il est aussi possible de rechercher une hiérarchisation des techniques : par exemple, passage du comptage au calcul pour la résolution de problèmes à l'école ou de l'arithmétique à l'algébrique au collège (Grugeon-Allys et Grapin, 2015).

Interprétation des résultats

D'abord, localement, le niveau de complexité défini a priori peut être mis en perspective avec la difficulté, calculée à partir du score de réussite obtenu après la passation. Un écart entre la complexité et la difficulté, dans un sens ou dans l'autre (tâche considérée comme a priori peu complexe qui est massivement échouée ou tâche considérée comme complexe qui est massivement réussie), et notamment s'il se retrouve sur des tâches d'un même type, mérite d'être étudié d'un point de vue didactique et questionne en particulier les programmes et/ou les pratiques enseignantes.

Globalement, si le domaine est couvert par un nombre suffisant de tâches bien réparties, la présence de tâches permettant d'avoir une progressivité dans les techniques en jeu ou dans la complexité permet de déterminer des points de rupture caractérisant sur l'échelle des scores les connaissances d'un groupe relativement à un autre. Ainsi, il est possible de spécifier les connaissances manquant à un groupe, en comparaison de celui qui lui est supérieur.

Étude du contenu des évaluations du PISA de la TIMSS au filtre de ces méthodologies

Études désormais de façon locale certaines questions ou certains items, puis les tests de façon globale, et ce, avec les deux approches et de façon successive.

Selon l'approche cognitive et didactique

Étude locale de quelques questions ou items

Les exercices du PISA comportent systématiquement une première partie destinée à présenter la situation sur laquelle porteront les questions auxquelles l'élève aura à répondre. Il s'agit de mettre les questions en contexte et de les ancrer dans le monde «réel». Cette partie peut comporter un texte de présentation assez long comprenant des données surabondantes ou inutiles, comme cela est effectivement le cas en contexte non scolaire. Les questions de la TIMSS, elles, sont en général ancrées dans le monde mathématique ; leur présentation est réduite au minimum.

Débit d'une perfusion

Les perfusions intraveineuses servent à administrer des liquides et des médicaments aux patients. Les infirmières doivent calculer le débit D d'une perfusion en gouttes par minute.

Elles utilisent la formule $D = \frac{d \cdot v}{60 \cdot n}$ où d est le facteur d'écoulement en gouttes par millilitre (ml), où v est le volume (en ml) de la perfusion et où n est le nombre d'heures que doit durer la perfusion.



Question 1

Une infirmière veut doubler la durée d'une perfusion.

Décrivez avec précision la façon dont D change si n est doublé et si d et v ne changent pas.

.....

Question 2

Les infirmières doivent aussi calculer le volume v de la perfusion en fonction du débit de perfusion D .

Une perfusion d'un débit de 50 gouttes par minute doit être administrée à un patient pendant 3 heures. Pour cette perfusion, le facteur d'écoulement est de 25 gouttes par millilitre.

Quel est le volume en ml de cette perfusion ?

Volume de la perfusion: ml

Source: OCDE, 2013, p. 4

Exemple 1 : Débit d'une perfusion, PISA de 2012

Les questions du PISA étant des parties d'un exercice comprenant, sauf exception, plusieurs questions, elles ne sont pas totalement indépendantes, ce qui n'est pas sans poser de problème en ce qui concerne l'interprétation des réponses, tandis que les questions de la TIMSS sont en général indépendantes. Précisons que l'expression «en général» utilisée ci-dessus provient du fait qu'au fil des années, sous des pressions diverses, la TIMSS a tendance à se rapprocher du PISA. Ainsi, certaines questions du PISA ont été empruntées à la TIMSS, tandis que la TIMSS utilise de plus en plus souvent le modèle du PISA.

Les descripteurs donnés par le PISA pour les deux questions sont les suivants : le contenu est «variations et relations», le processus est «employer», dans un contexte «professionnel» et avec un format de «réponse construite».

Selon le PISA, l'objectif de la première question est d'«expliquer quel est l'effet produit sur la valeur du résultat, lorsqu'on double une variable dans une formule, sachant que toutes les autres variables restent constantes» (OCDE, 2013, p. 4).

La question est codée à deux niveaux :

- crédit complet si l'explication donnée décrit à la fois le sens de l'effet (augmentation ou diminution) et sa valeur (il est divisé par deux, c'est la moitié, D diminuera de 50 %, etc.);
- crédit partiel si une réponse incomplète est donnée qui indique seulement le sens de l'effet ou sa valeur, mais dont les éléments ne sont pas tous les deux incorrects (D devient plus petit [pas de valeur], il y a un changement de 50 % [pas de sens]).

Pour la seconde question, l'objectif est de «transposer une équation et y substituer deux variables par des valeurs numériques données» (OCDE, 2013, p. 5). La question est semi-ouverte, codée à un seul niveau : crédit complet si la réponse est «360 ou une solution correctement transposée avec des variables de substitution correctes» (OCDE, 2013, p. 5).

Notre analyse

Conformément aux spécifications du cadre de référence du PISA, la question est placée dans un contexte «réel» : le contexte professionnel des infirmières.

Notre analyse de la tâche est illustrée par l'organigramme de la figure 2. Pour la première question, il suffit d'appliquer une procédure en principe connue, mais qui peut cependant nécessiter une adaptation puisque la propriété *si on multiplie le dénominateur par un nombre, la fraction est divisée par ce nombre* n'est pas toujours connue sous cette forme. L'habillage n'a pu que gêner les élèves; les résultats sont faibles, bien que l'essentiel de la modélisation soit déjà fait.

Pour la seconde question, les choses sont plus simples: il suffit de remplacer les lettres par les valeurs. Le codage de cette question suppose le simple remplacement de nombres dans la formule, sans aucune réflexion. Dans ce cas, pourquoi noyer les élèves dans un flot de mesures? Cela favorise-t-il de façon certaine l'attitude scientifique souhaitée?

Compte tenu des remarques faites ci-dessus, nous avons classé les questions 1 et 2 au niveau de complexité C et au NMF mobilisable des connaissances.

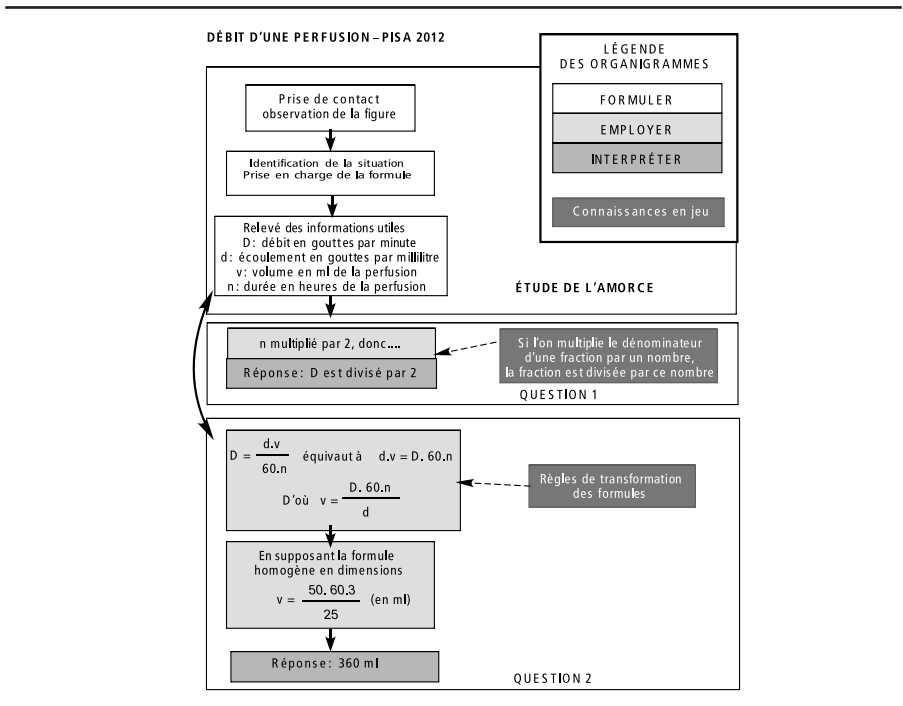
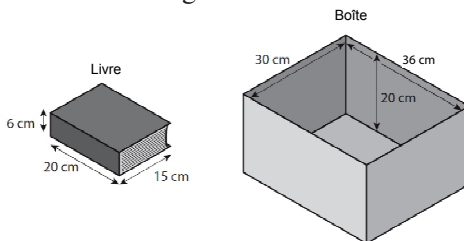


Figure 2: Organigramme de raisonnement – Débit d'une perfusion

François range ses livres dans une boîte rectangulaire.



Tous ses livres ont la même taille.

Quel est le plus grand nombre de livres qui peuvent tenir dans la boîte?

Réponse:

Source: Mullis et Martin, 2013, p. 135, traduction personnelle

Exemple 2: Rangement de livres, TIMSS de 2011, grade 8

Cette question, sans données inutiles et située dans un contexte concret et familier, aurait pu être utilisée dans le PISA. Toutefois, dans les questions du PISA, les données parasites sont générales et voulues pour refléter la complexité du réel, tandis que, dans la TIMSS, elles sont exceptionnelles.

Il s'agit bien d'une question de géométrie (organisation spatiale). C'est bien, comme le suggère le classement proposé par la TIMSS, le raisonnement qui constitue l'essentiel du processus de résolution. Ce raisonnement suppose un minimum de tâtonnement, avant de s'apercevoir qu'il est possible de placer les livres debout dans la boîte et, finalement, d'en mettre 12 sans espace restant.

L'organigramme (voir Figure 3) montre que les connaissances nécessaires sont réduites au minimum et que ce n'est pas du côté des connaissances qu'il faut chercher les raisons de la faible réussite de cet item, dans la plupart des pays.

Nous avons expérimenté une question semblable à l'Institut de recherche sur l'enseignement des mathématiques (IREM) de l'Université de Franche-Comté, à Besançon, dans les années 1980. Nous avons constaté que nombre d'élèves la traitaient en utilisant une procédure non pas géométrique, mais plutôt calculatoire: calcul du volume de la boîte ($21\,600\text{ cm}^3$), puis de celui d'un livre ($1\,800\text{ cm}^3$), puis quotient du premier volume par le second (12).

Puisque les élèves ont accès à leur calculatrice, il est probable que cette démarche a été utilisée par de nombreux élèves. Or, cette méthode n'est pas correcte : elle donnerait en effet le même résultat si les dimensions des livres n'étaient pas des diviseurs opportunément choisis des dimensions de la boîte.

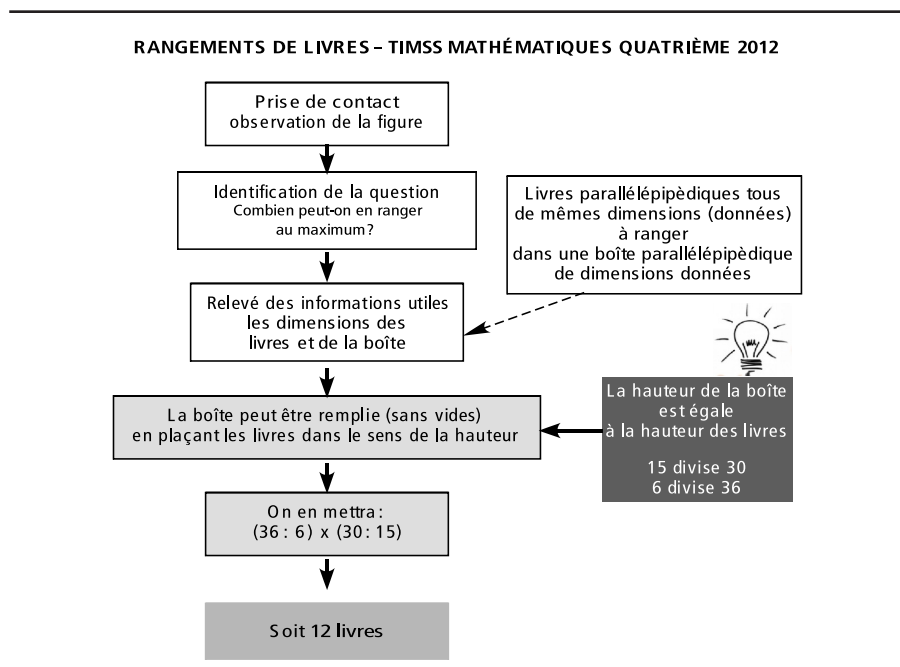


Figure 3 : Organigramme de raisonnement – Rangement de livres

Le codage ne s'intéressant qu'à la réponse, il n'est pas possible de savoir si cette réponse exacte correspond à une démarche correcte. D'ailleurs, il est possible que le niveau de réussite assez faible à cette question soit justement dû à l'utilisation de cette méthode incorrecte. En effet, elle est beaucoup plus propice aux erreurs de calcul que la méthode illustrée par l'organigramme proposé.

Étude globale sur l'ensemble des questions

L'analyse statistique implicative a permis une mise à plat des relations de dépendance des questions du PISA de 2012 en ce qui concerne la réussite des élèves. Rappelons que nous travaillons ici sur les données brutes des réponses des élèves⁴, et non sur des résultats agglomérés, comme cela est habituellement le cas.

Nous avons pu extraire des chaînes de questions fortement liées d'un point de vue implicatif. Par exemple, à partir des 21 questions de la catégorie *Espace et formes*, un graphe qui traduit l'organisation des réponses des élèves par rapport à ces questions est obtenu. Bien sûr, un graphe différent est obtenu pour chaque pays. La figure 4 reproduit le graphe obtenu pour la France, avec un seuil d'implication de 0,99 et un seuil de confiance de 80 %. Le graphe n'est pas très lisible dans ce document, mais nous avons mis en évidence une chaîne d'implications de six questions à laquelle nous nous sommes particulièrement intéressés.

Les pourcentages de réussite de ces six questions croissent de 6 à 75 % dans le sens des flèches. L'ordre reste le même à quelques permutations près pour les autres pays étudiés (Allemagne, Japon et Italie) : cela est général et est dû à la façon dont les questions sont calibrées et sélectionnées en utilisant l'analyse des réponses aux items sur les données obtenues lors d'une préexpérimentation. Un indice de différenciation dans le fonctionnement des items selon les pays (*differential item functioning* ou DIF) est en effet calculé et les questions qui dysfonctionnent de ce point de vue sont éliminées. Malgré ces précautions, certaines paires d'items peuvent avoir, selon les pays, des taux de réussite ordonnés différemment, et ce, de façon statistiquement significative, ce qui nous amène à nous poser des questions sur les différences curriculaires ou autres qui peuvent expliquer ces différences.

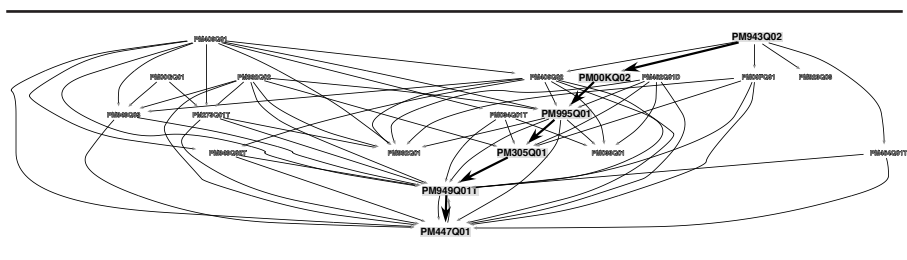


Figure 4 : Graphe des 21 questions de la catégorie *Espace et formes* pour la France et repérage d'une chaîne d'implications

Toutefois, même lorsque l'ordre des difficultés d'un ensemble d'items est commun à deux pays, une chaîne observée dans un pays ne se retrouve pas nécessairement dans d'autres pays. Par exemple, par rapport à la chaîne présentée ci-dessus, il n'existe que des sous-chaînes de trois questions dans les cas allemand et italien et de deux questions dans le cas japonais.

Cela nous amène à étudier la façon dont les autres questions s'éloignent de la chaîne initiale et à nous poser d'autres questions de nature didactique.

Pour l'étude globale des questions, nous avons mis au point un indicateur de densité d'implication basé sur la proportion de couples d'items ayant à la fois un indice d'implication supérieur à 0,99 et un indice de confiance supérieur à 80 % (indices très élevés).

Le tableau 1 donne la valeur de cet indice obtenu pour chacun des pays étudiés et pour ces pays pris deux à deux.

Tableau 1
Indicateurs de densité d'implication > 0,99 et de confiance > 80 %, par pays et inter-pays

	France	Allemagne	Italie	Japon
France	0,117	0,066	0,074	0,058
Allemagne		0,129	0,080	0,063
Italie			0,213	0,078
Japon				0,140

Ainsi, la France et l'Allemagne partagent environ la moitié de leurs relations les plus fortes en matière de niveaux d'implication et de confiance. La France partage environ 60 % de ses relations fortes avec l'Italie (0,074/0,117), tandis que l'Italie ne partage qu'un peu plus du tiers de ses relations fortes avec la France (0,074/0,213). Ces proportions sont sensiblement les mêmes en ce qui concerne le rapport France-Japon.

Indépendamment des niveaux de réussite, nous pouvons donc conclure à une proximité plus grande des organisations de connaissances révélées par le volet mathématique du PISA entre la France et l'Allemagne qu'entre la France et l'Italie ainsi qu'entre la France et le Japon.

La valeur prise par cet indicateur dans le cas italien est environ le double de la valeur qu'il prend dans les trois autres pays. Cela indique une dispersion nettement moins grande des connaissances et des savoir-faire des élèves italiens par rapport à l'ensemble des questions posées que dans le cas des élèves des trois autres pays.

La place disponible pour cet article ne permet pas de développer davantage ces apports de l'analyse statistique implicative à l'exploration des données des évaluations internationales. Pour cela, nous renvoyons le lecteur à Khaled et ses collaborateurs (2017). La recherche se poursuit tant pour les pays cités ici que pour d'autres pays, et une étude analogue est prévue avec les données de la TIMSS.

Selon l'approche épistémologique et didactique

La méthodologie employée par Roditi et Salles (2015) pour catégoriser les items du PISA de 2012 reprend une partie des éléments décrits dans l'approche épistémologique et didactique. En réalisant une analyse a priori du savoir en jeu dans chacun des items de cette évaluation, ils ont catégorisé les items selon le caractère outil ou objet des contenus mathématiques en jeu et selon le niveau de mise en fonctionnement des connaissances. Les résultats de leur étude montrent que la répartition des items selon les NMF et les domaines n'est pas équilibrée. Ainsi, les items du champ *quantité* relèvent majoritairement d'un NMF technique, alors que ceux du champ *espace et formes* nécessitent majoritairement l'introduction d'intermédiaires. L'étude des résultats au filtre de cette catégorisation permet par exemple de montrer que «les filles sont d'autant plus en difficulté par rapport aux garçons que le niveau de mise en fonctionnement des connaissances est un niveau exigeant» (p. 252) ou encore que «les différences de réussite selon les catégories socio-professionnelles restent stables lorsque le niveau de mise en fonctionnement des connaissances augmente» (p. 253). Nous ne reprenons pas ici l'ensemble des résultats de cette étude, mais souhaitons en montrer l'intérêt pour décrire le contenu, mais aussi pour en interpréter les résultats.

Nous allons désormais exemplifier, sur l'évaluation de la TIMSS, la mise en œuvre d'une analyse épistémologique et didactique telle que nous l'avons décrite précédemment. La TIMSS visant à évaluer les connaissances des élèves au regard des programmes scolaires des différents pays, l'utilisation d'un cadre épistémologique et anthropologique pour analyser le contenu de l'évaluation et pour interpréter les résultats est particulièrement adaptée. La définition de l'organisation mathématique de référence étant assez indépendante des curricula nationaux, elle apparaît donc comme un référent permettant d'étudier le contenu de l'évaluation, sans prendre en compte les spécificités de chacun des pays. Au regard de ce référent, il sera alors possible d'interpréter les résultats obtenus dans différents pays en lien avec les curricula.

Nous consacrons ainsi les deux prochaines sections à l'analyse du contenu de l'évaluation de la TIMSS *grade 4* au filtre de la méthodologie décrite précédemment. Après avoir illustré la façon dont nous menons l'analyse a priori sur quelques exemples de tâches, nous décrivons le contenu global du test en mathématiques. Nous avons choisi de nous concentrer sur le domaine des nombres (entiers, fractions et décimaux) en exploitant la définition de l'organisation mathématique de référence définie par Grapin (2015) pour le domaine des nombres entiers. Dans ce domaine, trois sous-domaines sont définis : la numération, le calcul et la résolution de problèmes.

Analyse a priori de questions d'évaluation

Nous avons choisi des questions permettant d'illustrer le contenu du test et appartenant donc aux différents sous-domaines. La tâche suivante, *Le nombre de Noémie* (voir Figure 5), a été proposée pour évaluer les connaissances des élèves en numération.

Noémie a créé cette énigme sur un nombre à 4 chiffres :

Le chiffre des centaines est 7.

Le chiffre des milliers est plus grand que le chiffre des centaines.

Le chiffre des unités est plus petit que le chiffre des centaines.

Quel est le nombre de Noémie ?

(A) 2708

(B) 4733

(C) 8726

(D) 9718

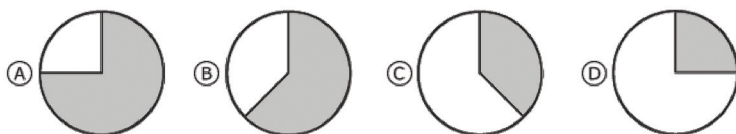
Figure 5 : Le nombre de Noémie

Il s'agit ici d'une tâche qui évalue uniquement l'aspect positionnel de la numération écrite chiffrée, et non l'aspect décimal. De plus, dans les quatre choix de réponse, le chiffre 7 est en position de centaines, ce qui amène à ne devoir distinguer que le chiffre des milliers et celui des unités, et à prendre en compte la comparaison avec 7. En France, les savoirs relatifs à l'écriture chiffrée débutent au CP et sont donc considérés comme

anciens au CM1 (*grade 4*). Malgré la prise en compte des contraintes sur les chiffres des milliers et des unités et la comparaison qu'elles engendrent, nous considérons qu'il s'agit d'un NMF technique des connaissances.

Les changements de registre de représentation font l'objet de différentes questions en numération, par exemple le passage d'une écriture fractionnaire à une écriture à virgule ou la reconnaissance d'écritures chiffrées à partir de décomposition en unités de numération. Nous avons choisi d'illustrer ces questions relatives au changement de registre avec l'exemple suivant (voir Figure 6) parce qu'il nous permettait d'illustrer une tâche plus complexe que celles habituelles pour ce type de tâche.

A. Lequel des cercles ci-dessous a les $\frac{3}{8}$ de sa surface grisés ?



B. Explique ou montre pourquoi ta réponse est correcte.

Source : Bodin, Decamp, De Hosson et Grapin (2016)

Figure 6 : Huitièmes d'un disque

Les fractions étant introduites en France en CM1, cet exercice ressemble beaucoup à ceux présents dans les manuels de cette année d'études, avec néanmoins une différence de taille : le partage en huitièmes n'est pas matérialisé sur les différents choix de réponse. Dans le partage des disques, différentes procédures sont à la disposition de l'élève : faire apparaître le partage en huitièmes pour se ramener à une situation connue, procéder par élimination des différentes propositions de réponses (la première est assez facilement identifiable à $\frac{3}{4}$, la dernière à $\frac{1}{4}$ et la deuxième est supérieure à $\frac{1}{2}$). Quelle que soit la procédure choisie, la résolution de l'exercice demande des adaptations, alors que le savoir en jeu (la représentation de fractions) est récent : nous considérons donc que le NMF des connaissances de l'élève doit être mobilisable pour résoudre cette tâche.

En calcul, le test propose des tâches classiques de calcul posé, où l'élève doit donner le résultat (p. ex., calculer la somme de $3\,126 + 845 + 72$ ou le produit de 27×43). La consigne accompagnant l'item ne mentionne pas explicitement la nécessité de poser le calcul et laisse donc la possibilité à l'élève de calculer différemment. La calculatrice n'est pas autorisée. La technique opératoire de la multiplication posée est enseignée en France au CE2 (*grade 3*) et celle de l'addition débute au CP (*grade 1*). Il s'agit donc de tâches mettant en jeu des savoirs anciens relevant d'un NMF technique des connaissances; ce type de calcul posé correspond à une application directe d'un algorithme dans la taxonomie Gras-Bodin (niveau A).

Pour conclure cette analyse a priori, nous pouvons comparer les NMF que nous avons attribués (au regard des programmes français) avec les processus définis par la TIMSS pour chacun des items :

Tableau 2
*Classification des tâches analysées en fonction des NMF
et des processus de la TIMSS*

Tâche de la TIMSS	NMF	Processus de la TIMSS
Le nombre de Noémie	Technique	Raisonner
Huitièmes d'un disque	Mobilisable	Connaître
Somme de trois entiers	Technique	Connaître
Produit de deux entiers	Technique	Connaître

Ainsi, dans le tableau 2, la hiérarchie dans les NMF n'est pas toujours en correspondance avec celle des processus de la TIMSS, en particulier pour les huitièmes d'un disque.

Analyse de l'ensemble des tâches du test

L'analyse a priori ayant été menée sur l'ensemble des tâches relevant du domaine des nombres (qui représente plus de la moitié du test), il est possible de décrire le contenu du test sur ce domaine et d'étudier la répartition selon les types de tâches et de complexité (Bodin et al., 2016).

La répartition par sous-domaine (numération, calcul et résolution de problèmes [additifs, multiplicatifs] et production d'expressions algébriques) est équilibrée, avec environ $\frac{1}{3}$ de tâches pour chacun d'eux. Une étude plus

approfondie nous montre que, au regard d'une organisation mathématique de référence, le nombre sous son aspect cardinal (p. ex., avec des problèmes de dénombrement) est très peu évalué. L'aspect ordinal, la numération et la résolution de problèmes sont quant à eux globalement couverts par les types de tâches présents dans l'évaluation.

Concernant le domaine du nombre, 40% des tâches sont d'un NMF technique, 20% d'un NMF disponible et 12% d'un NMF mobilisable. Les autres tâches (28%) sont considérées comme hors programme pour la France (p. ex., la production d'expressions algébriques ou le calcul de la somme de deux fractions).

Il reste de nombreux points que nous aurions pu aborder pour chacune de ces évaluations, dont la répartition des questions en fonction de leur format, les NMF selon les différentes catégories, les corrélations observées entre les catégories selon les pays à l'année d'études des élèves, etc. Nous n'avons pas non plus montré l'intérêt de ces approches pour interpréter les résultats des élèves. Nous renvoyons pour cela le lecteur à notre étude complète faite pour le CNETCO⁵.

Conclusion

Nous avons cherché à montrer, au fil de cet article, l'intérêt de mener des analyses didactiques du contenu des évaluations de la TIMSS et du PISA pour décrire le contenu avec un autre point de vue que celui choisi dans le cadre de référence, mais aussi pour mener d'autres études, comme celle produite avec l'analyse statistique implicite. Au-delà de la diffusion des questions d'évaluation et des résultats obtenus par les élèves français, il serait nécessaire de faire une analyse plus fine des productions des élèves pour accéder à leurs démarches et ainsi travailler avec les enseignants et les formateurs pour dépasser des constats basés uniquement sur des pourcentages de réussite. Par ailleurs, la possibilité offerte pour exploiter des données nationales en comparaison de celles d'autres pays ouvre aussi l'enseignant à d'autres pratiques et approches conceptuelles des notions à enseigner.

Comme nous l'avons indiqué en introduction, l'influence des enquêtes internationales sur les politiques éducatives a été étudiée par de nombreux chercheurs, mais qu'en est-il de leur influence sur les pratiques effectives des enseignants de mathématiques? Plus largement, de quelle façon les

évaluations externes standardisées et/ou certificatives (examens) influencent-elles les pratiques des enseignants? De Lange (2007) a montré que les résultats du PISA avaient amené certains pays à modifier le contenu de certaines parties de leurs curricula, notamment celle relative à la résolution de problèmes. Plus largement, différentes études recensées dans Suurtamm et ses collaborateurs (2016) ont identifié des changements dans les pratiques liées au type de format de questions (p. ex., plus de QCM) ou allant vers l'enseignement pour l'examen (*teaching to the test*).

Une évolution se dessine peu à peu dans le mode de passation des tests du PISA. Alors que, jusqu'en 2012, toutes les passations du volet littéracie mathématique étaient du type papier crayon, les questions du même volet du PISA de 2015 (qui formaient un sous-ensemble des questions de 2012) ont été passées sur écran, ce qui n'est pas sans poser de questions sur l'interprétation qu'il faut donner aux différences constatées entre 2012 et 2015. En 2018, l'évaluation du PISA se passera entièrement sur écran ; pour la TIMSS, en 2019, ce sera aussi le cas pour la moitié des pays. Si l'intérêt de l'abandon du papier crayon lors des évaluations standardisées au profit des nouvelles technologies s'explique par un traitement plus simple des réponses des élèves et par une prise en compte de l'évolution de l'enseignement (du moins, cela est-il supposé), cela entraîne de multiples questions quant à la validité du contenu de l'évaluation. La prise en main et l'accès aux fonctionnalités de l'instrument (application ou logiciel dédié sur tablette ou ordinateur) demandent un temps de formation pour l'élève et, par la suite, une certaine autonomie qui ne va pas de soi. En France, si les élèves ont accès de plus en plus à du matériel informatique au sein de leur établissement scolaire, l'utilisation de tels supports pour des évaluations reste peu fréquente. Comment les élèves abordent-ils alors une telle évaluation? Quel est l'impact de ce nouveau support sur l'activité de l'élève? Il nous semble important de considérer de telles questions dans un cadre didactique pour déterminer ce qui est effectivement évalué.

Enfin, comme nous l'avons montré dans cet article avec les analyses statistiques faites à partir des résultats du PISA, l'exploitation secondaire des données de telles enquêtes constitue une base potentiellement riche, mais encore très peu utilisée dans le champ de la didactique. Même s'il est regrettable de ne pas pouvoir accéder à une partie plus importante des items pour le PISA, nous soulignons tout de même que la TIMSS libère

davantage de questions et que le descriptif qui en est fait sur un plan mathématique est beaucoup plus évocateur que celui du PISA puisqu'il est rédigé selon des objectifs en lien avec des savoirs mathématiques.

Les évaluations externes se multipliant, il serait dommage de ne pas les exploiter autrement qu'à des fins seules de pilotage du système. Nous avons tenté de montrer, au fil du texte, l'intérêt de les analyser dans un cadre didactique et les résultats potentiels que cela pouvait produire, que ce soit pour les chercheurs, les formateurs d'enseignants ou les enseignants eux-mêmes. Pour que ces évaluations puissent être davantage comprises en termes d'enjeu et exploitées pour leurs résultats et leurs contenus, la diffusion de leurs résultats ne peut se limiter au pointage de la hausse ou de la baisse du score d'un pays. Pour les enseignants et les formateurs, ce sont des analyses complémentaires sur le contenu du test qui leur permettront de situer les questions d'évaluation dans leur enseignement et, éventuellement, de le faire évoluer. C'est ce que nous avons cherché en partie à faire à travers cet article.

Réception : 17 septembre 2017

Version finale : 2 février 2018

Acceptation : 13 mars 2018

NOTES

1. Dans une étude recensant les articles de recherche portant sur le PISA publiés entre 2000 et 2015, Kyriakides (2016) a identifié 808 articles en langue anglaise publiés dans des revues à comité de lecture. Le fait que 40 de ces articles ont été publiés entre 2013 et 2015 témoigne de l'intérêt croissant des chercheurs pour ces évaluations. Toutefois, parmi ces articles, très rares sont ceux qui s'intéressent de près à leurs cadres de référence, et encore plus rares sont ceux qui s'intéressent directement au questionnement cognitif sur lequel ces évaluations s'appuient.
2. L'accès à la totalité des items du test a été rendu possible dans le cadre d'une étude préalable menée pour le CNESCO par les deux auteurs de cet article. Les questions du test de la TIMSS publiées dans ce rapport sont des questions libérées.
3. La validité est une notion classique en évaluation, mais les approches psychodidactiques, en particulier celles de Vantourout et Goasdoué (2014), sont de nature à renouveler cette notion en la faisant sortir du champ de la psychométrie pour l'intégrer à la réflexion didactique.
4. Le programme construit par Khaled et ses collègues (2017) utilise les spécificités de l'évaluation par cahiers tournants pour construire les matrices d'implication de l'ensemble des questions du PISA ou de tout sous-ensemble de questions.
5. Cette étude peut être téléchargée à l'adresse <https://antoine-bodin.com/question-generales>.

RÉFÉRENCES

- Bodin, A. (1997). L'évaluation du savoir mathématique. *Recherches en didactique des mathématiques*, 17(1), 49-96.
- Bodin, A. (2007a). Dissonances et convergences évaluatives – De l'évaluation dans la classe aux évaluations internationales : quelle cohérence ? *Bulletin de l'APMEP*, 474, 47-79. Repéré à www.apmep.fr/IMG/pdf/Coherence_article_BV_V5.pdf
- Bodin, A. (2007b). What does PISA really assess? In S. Hopman, G. Brinek, & M. Retzl (Eds.), *PISA according to PISA* (pp. 21-57). Wien, Austria: Lit Verlag.
- Bodin, A., Decamp, N., De Hosson, C. et Grapin, N. (2016). *Comparaison des cadres de référence et des instruments d'évaluation des enquêtes PISA et TIMSS*. Paris : CNESCO. Repéré à www.cnesco.fr/fr/comparaison-pisa-timss
- Bodin, A. et Giovaninni, M.L. (2017, octobre). Lecture didactique des résultats d'une recherche évaluative à la lumière de l'ASI. Dans J.-C. Régnier, R. Gras, R. Cousturier et A. Bodin (dir.), *Points de vue conceptuels, applicatifs et métaphoriques* (pp. 416-436). Actes du 9^e Colloque international de l'analyse statistique implicite, Besançon, France : Université Bourgogne Franche-Comté. Repéré à <http://sites.univ-lyon2.fr/asi9/pub/ISBN9782956204510ASI9.pdf>

- Bosch, M. et Gascon, J. (2005). La praxéologie comme unité d'analyse des processus didactiques. Dans A. Mercier et C. Margolinas (dir.), *Balises pour la didactique des mathématiques* (pp. 197-122). Grenoble: La Pensée Sauvage.
- Chevallard, Y. (1999). L'analyse des pratiques enseignantes en théorie anthropologique du didactique. *Recherches en didactique des mathématiques*, 19(2), 221-266.
- Chevallard, Y. et Feldmann, S. (1986). *Pour une analyse didactique de l'évaluation*. Marseille: IREM d'Aix-Marseille.
- Deblois, L., Freiman, V. et Rousseau, M. (2007, juin). Les résultats des élèves aux tests internationaux et leur possible influence sur les thèmes de recherche. *Actes du colloque du Groupe de didactique des mathématiques du Québec* (pp. 135-145), Rimouski, Québec: UQAR. Repéré à <http://turing.scedu.umontreal.ca/gdm/documents/ActesGDM2007.pdf>
- de Lange, J. (2007). Large-scale assessment and mathematics education. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1111-1142). Reston, VA: National Council of Teachers of Mathematics.
- Dupé, C. et Olivier, Y. (2005). Ce que l'évaluation PISA 2003 peut nous apprendre. *Bulletin de l'APMEP*, 460, 626-644.
- Grapin, N. (2015). *Étude de la validité de dispositifs d'évaluation et conception d'un modèle d'analyse multidimensionnelle des connaissances des élèves de fin d'école* (Thèse de doctorat, Université Paris-Diderot, Paris). Repéré à <https://tel.archives-ouvertes.fr/tel-01294076/document>
- Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques* (Thèse de doctorat). Université de Rennes 1, Rennes.
- Gras, R. et Bodin, A. (2017, octobre). L'ASI, analyseur et révélateur de la complexité cognitive taxonomique. Dans J.-C. Régnier, R. Gras, R. Couturier et A. Bodin (dir.), *Points de vue conceptuels, applicatifs et métaphoriques* (pp. 125-139). Actes du 9^e Colloque de l'analyse statistique implicative, Besançon, France: Université Bourgogne Franche-Comté. Repéré à <http://sites.univ-lyon2.fr/asi9/pub/ISBN9782956204510ASI9.pdf>
- Grugeon-Allys, B. et Grapin, N. (2015, janvier). Validité d'une évaluation externe: complémentarité des approches. Dans A.-C. Mathé et E. Mounier (dir.), *Actes du séminaire national de didactique des mathématiques* (pp. 13-26). Paris: IREM, Université Paris-Diderot. Repéré à <https://hal.archives-ouvertes.fr/hal-01317134/document>
- Khaled, H., Bodin, A. et Couturier, R. (2107, octobre). L'analyse statistique implicative (ASI) appliquée aux données de Programme international pour le suivi des acquis des élèves (PISA). Dans J.-C. Régnier, R. Gras, R. Couturier et A. Bodin (dir.), *Points de vue conceptuels, applicatifs et métaphoriques* (pp. 437-456). Actes du 9^e Colloque de l'analyse statistique implicative, Besançon, France: Université Bourgogne Franche-Comté.
- Kyriakides, L. (2016, December). *A synthesis of studies using PISA data: Implications for research, policy and practice*. Paper presented at the PISA 2016 Seminar, Oxford University Center for Educational Assessment, Oxford. <http://bit.ly/Kyriakides>
- Mons, N. (2009). Effets théoriques et réels des politiques d'évaluation standardisée, *Revue française de pédagogie*, 169, 99-140. doi: 10.4000/rfp.1531

- Mullis, I., & Martin, M. (Eds.) (2013). *TIMSS 2015 Assessment Frameworks*. Chestnut Hill, MA: Boston College/IEA. Retrieved from https://timssandpirls.bc.edu/timss2015/downloads/T15_Frameworks_Full_Book.pdf
- Organisation de coopération et de développement économiques [OCDE] (2005). *La définition et la sélection des compétences clés : résumé*. OECD Publishing. Repéré à <https://www.oecd.org/pisa/35693273.pdf>
- Organisation de coopération et de développement économiques [OCDE] (2013). *PISA 2012: unités libérées PISA mathématiques*. OECD Publishing.
- Organisation for Economic Co-operation and Development [OECD] (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf
- Robert, A. (1998) Outil d'analyse des contenus mathématiques à enseigner au lycée et à l'université, *Recherches en didactiques des mathématiques*, 8(2), 139-190.
- Roditi, E. et Chesne, J.-F. (2012). Un point de vue didactique sur les questions d'évaluation en éducation: les compétences et le socle commun. Dans A. Robert, J. Penninckx et M. Lattuati, *Une caméra au fond de la classe de mathématique* (3^e partie, chap. 4). Besançon : France : Presses universitaires de Franche-Comté.
- Roditi, E. et Salles, F. (2015). Nouvelles analyses de l'enquête PISA 2012 en mathématiques : un autre regard sur les résultats. *Éducation et formation*, 86-87, 235-258. Repéré à http://cache.media.education.gouv.fr/file/revue_86-87/56/9/depp-2015-EF-86-87-nouvelles-analyses-enquete-PISA-2012-mathematiques_424569.pdf
- Ruddock, G., Clausen-May, T., Purple, C., & Ager, R. (2006). *Validation study of the PISA 2000, PISA 2003 and TIMSS-2003 international studies of pupil attainment*. London: DfES.
- Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., Silver, E., Ufer, S., & Vos, P. (2016). Assessment in mathematics education. In G. Kayser (Ed.), *ICME 13 Topical Surveys* (pp. 1-38). London: Springer Open.
- Vantourout, M. et Goasdoué, R. (2014). Approches et validité psycho-didactiques des évaluations. *Éducation et formation*, e-302. Repéré à <https://halshs.archives-ouvertes.fr/halshs-01239551>
- Vergnaud, G. (1990). La théorie des champs conceptuels. *Recherches en didactique des mathématiques*, 10(2.3), 133-170.
- Wu, M. (2009). *A critical comparison of the contents of PISA and TIMSS mathematics assessments*. Melbourne: University of Melbourne Assessment Research Centre.