Analyse d'une évaluation internationale : le cas de l'enquête TIMSS 2019 grade 8 " science "

Cécile de Hosson, Nicolas Décamp, Anaïs Bret, Marion Le Cam

RDST
Recherches en didactique des sciences et des technologies

27 | 2023
L'évaluation et l'enseignement des sciences et des technologies

Analyse d'une évaluation internationale : le cas de l'enquête TIMSS 2019 grade 8 « science »

Analysis of an international assessment: the case of TIMSS 2019 eight grade science assessment

Cécile Hosson (de), Nicolas Décamp, Anaïs Bret et Marion Le Cam

Édition électronique
URL : https://journals.openedition.org/rdst/4654
DOI : 10.4000/rdst.4654
ISSN : 2271-5649

Éditeur
ENS Éditions

Édition imprimée
Pagination : 73-99
ISSN : 2110-6460

Référence électronique
Cécile Hosson (de), Nicolas Décamp, Anaïs Bret et Marion Le Cam, « Analyse d'une évaluation internationale : le cas de l'enquête TIMSS 2019 grade 8 « science » », *RDST* [En ligne], 27 | 2023, mis en ligne le 01 juillet 2023, consulté le 30 mai 2024. URL : http://journals.openedition.org/rdst/4654 ; DOI : https://doi.org/10.4000/rdst.4654

Analyse d'une évaluation internationale : le cas de l'enquête TIMSS 2019 grade 8 « science »

CÉCILE DE HOSSON
Université Paris Cité, Univ Paris Est Créteil, CY Cergy Paris Université, Univ. Lille, UNIROUEN, LDAR

NICOLAS DÉCAMP
Université Paris Cité, Univ Paris Est Créteil, CY Cergy Paris Université, Univ. Lille, UNIROUEN, LDAR

ANAÏS BRET
Direction de l'évaluation, de la prospective et de la performance - DEPP, MENJ

MARION LE CAM
Direction de l'évaluation, de la prospective et de la performance - DEPP, MENJ

RÉSUMÉ : Cet article a pour ambition d'examiner les résultats des élèves français à l'enquête internationale TIMSS science 2019 pour la classe de 4e (grade 8). Il s'agit d'apprécier les succès et les échecs des élèves français au regard de certaines caractéristiques des items de TIMSS (*Trends in International Mathematics and Science Study*) non prises en charge dans le modèle sur lequel repose le calcul des scores des élèves par l'IEA (*International Association for the Evaluation of Educational Achievement*). À partir d'un traitement statistique de type « régression linéaire multiple » des données obtenues par une analyse critériée de l'ensemble des items de l'enquête, il ressort que les élèves français sont plus performants lorsqu'il s'agit de produire une réponse à une question mettant en jeu un raisonnement, lorsque la question d'un item nécessite de faire appel aux données disponibles dans l'illustration et lorsque les items relèvent des sciences de l'univers. À l'inverse, les élèves français sous-performent lorsque les items mettent en jeu une situation qui leur est non familière ou relevant de la chimie.

MOTS CLÉS : évaluation, physique, chimie, modèle regression

ABSTRACT: The aim of this paper is to investigate the results of French students in the TIMSS science 2019 international survey for the 8th grade. The research focuses on the successes and failures of French students with respect to certain characteristics of TIMSS (Trends in International Mathematics and Science Study) items that are not taken into account in the model on which the IEA (International Association for the Evaluation of Educational Achievement) computes students' scores. A statistical processing of the data obtained by a criterion-referenced analysis of all the items in the survey, such as "multiple linear regression", showed that French students were more

successful when it came to providing an answer to a question involving reasoning, when the question of an item required the use of data available in the illustration, and when the items were related to the sciences of the universe. Conversely, French students underperformed when the items involved a situation that was unfamiliar to the students or related to chemistry.

KEYWORDS: evaluation, physics, chemistry, regression model

Introduction

Les performances des élèves français aux évaluations internationales standardisées font l'objet de diffusion et d'interprétations variées, souvent à charge (contre le système éducatif français ou contre les évaluations elles-mêmes) qui minorent, voire ignorent, ce que les évaluations évaluent vraiment et passent sous silence les inadéquations parfois criantes entre modalités d'évaluation (la manière dont les questions sont posées, les savoirs qu'elles engagent, l'environnement qui les porte, etc.) et ce à quoi les élèves ont été formés. Lors d'une précédente mission d'expertise, effectuée en 2016 par deux des auteurs de cet article au titre du CNESCO[1], il a été montré qu'il existait une distance assez grande entre l'ergonomie des items du PISA[2] « culture scientifique » 2015 (passation informatique), leur structure sémiotique (nombre important de registres mobilisés), certaines des connaissances évaluées (épistémiques) et l'environnement pédagogique usuel des élèves français en classe de sciences (Bodin *et al.*, 2016). Cette distance paraissait moins prononcée dans l'évaluation TIMSS Advanced réalisée la même année avec les élèves du lycée (grade 12) sans doute parce que les items de cette évaluation engageaient des savoirs disciplinaires identifiables (concepts et lois de la physique) contextualisés dans des situations proches de celles travaillées en classe de physique. Ce résultat n'a rien d'étonnant dans la mesure où les cadres théoriques soutenant les deux évaluations PISA et TIMSS diffèrent notamment du point de vue de la proximité de leurs items avec les savoirs scolaires des élèves. PISA science évalue la « littératie scientifique » des élèves au terme de la scolarité obligatoire, c'est-à-dire leur aptitude à mobiliser des savoirs et savoir-faire dans des situations dont le contexte se rapproche de la vie quotidienne et à forts enjeux sociétaux, sanitaires et environnementaux. TIMSS quant à elle évalue les acquis scolaires des élèves et repose sur des items mettant en jeu des savoirs et des savoir-faire supposément proches de ce qui est étudié en classe de science au terme du grade 8 (classe de 4e en France).

Dans notre étude de 2016, nous nous étions limités à l'analyse du contenu et de la forme des items constitutifs de l'évaluation TIMSS *advanced* pour en dégager quelques spécificités afin de les mettre en regard des savoirs et des savoir-faire engagés dans les programmes scolaires français. Dans cet article nous franchissons une étape supplémentaire en nous intéressant non seulement au contenu des items de l'évaluation TIMSS 2019 grade 8 pour les disciplines physique, sciences de l'univers et chimie, mais également aux performances des élèves français au regard des performances globales de l'ensemble des élèves concernés par l'évaluation. Ce travail est rendu possible grâce à un partenariat avec la DEPP[3] qui nous

1 Centre national d'étude des systèmes scolaires.

2 Programme international pour le suivi des acquis des élèves.

3 Direction de l'évaluation, de la prospective et de la performance.

a permis d'avoir accès à l'ensemble des items de l'évaluation TIMSS 2019 grade 8 pour les disciplines sus-mentionnées ainsi qu'aux scores des élèves français pour chaque item et aux scores globaux de tous les élèves. L'objectif est ici de dégager les caractéristiques des items qui sont potentiellement sources de réussite ou d'échec pour les élèves français.

Dans un premier temps, nous présentons les principes et les catégorisations sur lesquels repose l'enquête TIMSS pour l'évaluation des connaissances scientifiques des élèves de 4e (grade 8) tels que présentés dans le « cadre de l'évaluation des sciences »[4]. Dans un deuxième temps nous résumons les résultats de quelques travaux issus de la recherche en didactique des sciences engageant spécifiquement des études critiques des évaluations internationales. Dans un troisième temps, nous pointons les catégories propres à TIMSS qu'il nous semble pertinent de conserver pour notre étude (catégories de classification en termes de « domaine cognitif » et de « discipline » notamment). Celles-ci se voient enrichies par l'ajout de catégories *ad hoc* pour partie inspirées du « modèle prédictif des difficultés associées aux tâches de résolution des questions de culture scientifique », créé par Duclos et ses collaboratrices (2021) pour l'analyse des items de l'évaluation internationale PISA science de 2015, pour partie émergentes. L'ensemble des items constitutifs de l'évaluation TIMSS 2019 grade 8 pour les sciences fait l'objet, dans un quatrième temps, d'une analyse *a priori* des items *via* la grille construite à l'étape 3. Cette analyse est suivie d'une étude quantitative portant sur les résultats des élèves et visant à dégager des liens éventuels entre certaines spécificités des items et les scores des élèves français.

1. Contexte de la recherche : l'évaluation TIMSS 2019 grade 8

Depuis sa création en 1995, l'étude internationale TIMSS se donne pour objet la mesure des performances en sciences des élèves à la fin de la huitième année de scolarité obligatoire, classe de quatrième en France. En mai 2019, trente-neuf pays et sept provinces ont participé à l'enquête internationale TIMSS grade 8 organisée par l'IEA pour évaluer les acquis des élèves de quatrième en mathématiques et en sciences, sur papier ou sur ordinateur. En France, environ 3 800 élèves de quatrième âgés de 13,9 ans en moyenne et répartis dans 150 collèges ont répondu sur ordinateur à des questions relevant du domaine des sciences de la nature (physique, chimie, sciences de la vie, sciences de la Terre)[5]. Avec un score de 489 points[6], la France se situe sous la moyenne internationale des pays de l'Union européenne (UE) et de l'Organisation de coopération et de développement économiques (OCDE) (515 points). La France n'amène que 3 % de ses élèves au niveau avancé en sciences alors qu'ils sont en moyenne 10 % dans les pays de l'UE et de l'OCDE.

L'enquête TIMSS 2019 grade 8 « sciences » repose sur un « cadre » fourni par l'IEA qui fixe et explicite les contenus scientifiques évalués, désignés « domaines » (physique, chimie,

4 <https://www.education.gouv.fr/media/73339/download> (consulté le 13 août 2022). Voir également Mullis et Martin (2017).

5 En 2019, la France participe pour la deuxième fois à TIMSS pour la classe de quatrième, la première participation étant en 1995. La France n'avait pas participé aux cycles intermédiaires pour ce niveau.

6 La méthodologie de calcul des scores repose sur le modèle de réponse à l'item – MRI dont le principe est disponible dans le rapport technique de l'évaluation TIMSS (voir Martin *et al.*, 2020).

sciences de la Terre) et « sous-domaines » scientifiques (par exemple, composition de la matière pour le domaine « chimie ») et les processus de réflexion à évaluer, désignés « domaines cognitifs » (connaître, appliquer, raisonner).

L'objectif des évaluations à grande échelle en général, et de TIMSS en particulier, est de fournir une estimation des compétences des élèves permettant des comparaisons entre les pays, et dans le temps. Cela nécessite de couvrir largement le domaine évalué. Cette couverture repose sur plusieurs centaines d'items, dont seulement une partie peut être administrée à chacun des élèves dans un temps raisonnable pour une évaluation réalisée en milieu scolaire. L'ensemble des items est ainsi réparti dans 14 *booklets* différents, chaque *booklet* comportant deux blocs d'items de science (et deux blocs d'items de mathématiques). Chaque bloc compte 12 à 18 items, et se retrouve dans deux des 14 *booklets* selon un système de rotation permettant d'assurer la comparabilité des items entre eux. L'ensemble du matériel d'évaluation (mathématiques et science) représente environ 10 h 30 de passation, mais chaque élève ne passe qu'un seul des 14 *booklets*, lors d'une séquence de 90 minutes (Mullis & Martin, 2017). Grâce à ce plan de rotation des blocs au sein des différents *booklets*, et à leur distribution sur l'ensemble des élèves, l'utilisation d'un modèle probabiliste, le modèle de réponse à l'item (MRI – voir note 6) permet de relier les items entre eux, afin que les compétences des élèves puissent être rapportées sur une échelle numérique comparable (échelle de scores), même s'ils n'ont pas tous passé exactement le même contenu de test. Lors du premier cycle de l'évaluation TIMSS en 1995, l'échelle des scores a été établie en fixant la moyenne internationale des scores à 500, et l'écart-type à 100. À chaque nouveau cycle, les scores sont distribués sur cette même échelle, permettant une comparaison temporelle.

L'IEA met à disposition du grand public 15 items dits « libérés » sur les 129 items de l'évaluation TIMSS 8 sciences. Ce sont des items qui ne seront plus utilisés dans les sessions ultérieures de l'évaluation et qui sont rendus publics. Notre collaboration avec la DEPP nous a permis d'accéder à l'ensemble des 129 items dont les 114 non libérés[7].

2. Recherche en didactique des sciences et évaluations internationales : état de l'art

Depuis leur installation dans le paysage docimologique international ces trente dernières années, les évaluations internationales (PISA, PIRLS[8], TIMSS) sont étudiées par les chercheurs et les chercheuses en éducation à l'échelle mondiale. Dans ce contexte, l'enquête PISA est certainement l'objet le plus fréquemment visité par la recherche scientifique. Il est toutefois intéressant de constater que les lignes d'analyse et les résultats de ces travaux (que ce soit sur PISA ou TIMSS) convergent vers l'idée que les performances des élèves méritent un examen approfondi des caractéristiques des questions posées, des tâches effectivement attendues des élèves (ou effectivement mobilisées) et de leur adéquation avec les *habitus*, les savoirs et les savoir-faire scolaires (acquis des élèves au regard des programmes d'ensei-

7 Ces items appartiennent aux domaines scientifiques « physique » et « chimie ». Nous avons également conservé quelques items du domaine « sciences de la Terre » lorsque ceux-ci relevaient du programme de physique (astronomie) du cycle 4. Nous les avons renommés « sciences de l'univers ».

8 *Progress in International Reading Literacy Study*.

gnement, pratiques usuelles des enseignants et des enseignantes en termes d'évaluation, par exemple). Un tel examen se veut modalisateur en ce qu'il permet de regarder « autrement » les scores des élèves à ces enquêtes. Par exemple, de nombreux articles engagent des considérations sociologiques – lien entre performance et milieu socio-économique et/ou sexe des élèves – ainsi que les travaux ciblant les liens entre performance, motivation et appétence pour les sciences (Bret *et al.*, 2016 ; Classick *et al.*, 2021 ; Duclos, 2022). Compte tenu de nos objectifs, nous faisons reposer notre état de l'art sur des travaux se distribuant selon les quatre directions suivantes :

1. Articles analysant l'adéquation entre les contenus en jeu (savoirs et savoir-faire) dans les items des évaluations pour un niveau scolaire donné et les contenus des programmes d'enseignement scientifique des pays participant pour ce même niveau (Bodin *et al.*, 2016 ; Coppens, 2012 ; Dolin & Krogh, 2010) ;
2. Articles analysant l'adéquation entre les caractéristiques du cadre conceptuel d'évaluation des enquêtes internationales (domaines cognitifs, degré de difficulté) et les caractéristiques effectives des items de ces enquêtes (Lau, 2009 ; Glynn, 2012 ; Le Hébel, Montpied & Tiberghien, 2014 ; Duclos *et al.*, 2021) ;
3. Articles analysant l'adéquation entre les compétences que les élèves utilisent lorsqu'ils construisent leur réponse et les compétences que les enquêtes internationales entendent évaluer (Harlow & Jones, 2004 ; Le Hébel, Montpied & Tiberghien, 2014) ;
4. Articles analysant les scores des élèves aux enquêtes internationales et leur niveau scolaire effectif (Wiberg & Rolfsman, 2019) et/ou les caractéristiques des items structurant les enquêtes (Bodin *et al.*, 2016).

Nous limitons notre état de l'art à quelques travaux dont l'ambition est d'une part la mise au jour d'une distance entre ce que PISA et TIMSS évaluent et ce qu'apprennent les élèves (directions 1 ci-dessus) et d'autre part, les niveaux de performance des élèves et les difficultés liées aux caractéristiques des items (direction 4 ci-dessus). Sous cette double perspective, Dolin et Krogh (2010) s'intéressent à la pertinence de PISA science 2006 vis-à-vis du système éducatif danois en analysant la correspondance entre les objectifs de PISA et les objectifs des programmes de l'enseignement scientifique obligatoire danois. Les chercheurs pointent le fait que cette démarche comparative est difficile dans la mesure où le cadre PISA engage l'évaluation d'éléments de « littératie scientifique » (capacité de l'élève à conduire une enquête scientifique, aptitude à recueillir et interpréter des données, aptitude à distinguer des questions de nature scientifique de questions ne relevant pas de la science, etc.) qui sont peu présents dans les curricula danois. À titre d'exemple, le concept de « compétence » tel que défini par PISA ne concorde pas avec la manière dont les compétences sont présentées dans le système éducatif danois. De même, les connaissances liées à la Nature de la Science occupent une place relativement faible dans le programme scolaire danois. À l'inverse, les auteurs soulignent que l'évaluation TIMSS se prête bien mieux à ce travail de comparaison : les objectifs de TIMSS rejoignant ceux de l'enseignement scientifique danois.

Mais cette proximité affichée entre les disciplines scolaires et les domaines scientifiques évalués par TIMSS « science » n'est pas pour autant garante de succès. Dans une étude conduite sur les résultats d'élèves sud-africains (grade 8) à l'évaluation TIMSS « science » 1999, Dempster et Reddy (2007) ont montré un lien fort entre le manque de « lisibilité » du texte des questions (forme sémantique, vocabulaire utilisé) et le choix de réponses incorrectes. Cette tendance était particulièrement visible chez les élèves fréquentant des établissements scolaires non anglophones. Au-delà de cette difficulté linguistique (qui a

également été pointée en France pour des élèves francophones, voir Bautier *et al.*, 2006), les chercheurs ont également constaté, cette fois pour l'ensemble des élèves, un faible recouvrement entre les contenus scientifiques évalués par TIMSS 1999 et les connaissances effectives des élèves, ainsi qu'une présence élevée d'« idées fausses » des élèves sud-africains. Ce résultat converge avec un point que nous avions soulevé à l'occasion de l'examen des items de TIMSS advanced 2015 en physique (Bodin *et al.*, 2016) : la lecture de ces différents items a montré que tester les idées fausses (les conceptions) des élèves faisait clairement partie des objectifs des concepteurs de l'enquête. Celles-ci apparaissent en effet dans un nombre non négligeable de propositions de réponse aux questions à choix multiples (nous en avions repéré des traces dans presque un quart des items, en particulier ceux relevant des domaines cognitifs « appliquer » et « raisonner »). Finalement, les résultats obtenus par Dempster et Reddy (2007) les conduisent à remettre en cause la validité de l'enquête TIMSS « science » pour l'évaluation des élèves sud-africains, remise en cause que l'on retrouve également dans le travail de Harlow et Jones (2004) en Nouvelle-Zélande pour l'enquête TIMSS 1994 et corroborée par l'étude de Wilberg et Rolfsman (2019).

Cette question de la validité de l'enquête TIMSS « science » est au cœur du travail de Glynn (2012) qui tranche, dans ses résultats, avec les réserves exprimées par d'autres chercheur.e.s. Partant du postulat selon lequel la qualité des items TIMSS influence la validité des scores des élèves à ces items, Glynn s'attache à évaluer cette qualité sur un échantillon de 20 items en mobilisant trois approches complémentaires. La première est de nature psychométrique et repose sur le modèle statistique de la réponse à l'item évoqué plus haut. La deuxième cible la formulation des items (lisibilité, vocabulaire) et la troisième, la proximité en termes de format, de contenu et de domaine cognitif, entre les intentions déclarées du cadre d'évaluation TIMSS « science » 2007 et l'expertise d'un groupe d'enseignant.e.s de sciences. L'évaluation ainsi conduite révèle que la majorité des items analysés sont de qualité élevée dans la mesure où ils présentent de bonnes caractéristiques psychométriques, des caractéristiques syntaxiques propices à une bonne compréhension par les élèves, et une bonne conformité avec le cadre d'évaluation.

3. Environnement théorique : le modèle de difficulté des questions de sciences

Dans un article de 2021, Duclos et ses collaboratrices proposent un modèle visant à prédire les difficultés associées aux tâches de résolution des questions des items de l'enquête internationale PISA 2015 « culture scientifique » (*scientific literacy*). Inspiré par les recherches portant sur les « modèles de difficulté de questions » (Crisp & Grayson, 2013 ; Prenzel *et al.*, 2002, citées par Duclos *et al.*, 2021), le « modèle de difficulté des questions de sciences » répond à l'exigence de comprendre quelle caractéristique ou combinaison de caractéristiques d'un item est susceptible d'augmenter sa difficulté. La structure du modèle repose sur une catégorisation tripartite : la première catégorie renvoie aux caractéristiques liées au contenu en jeu dans la tâche, la deuxième est décrite par des caractéristiques intrinsèques de l'item (sens et aspects formels de l'énoncé et des questions pour un item donné) ; la troisième est relative aux caractéristiques liées aux raisonnements et stratégies des élèves (Duclos, 2022, p. 59).

Chacune de ces catégories est subdivisée en sous-catégories auxquelles sont associées des caractéristiques suffisamment précises pour que le processus d'identification soit le moins ambigu possible. Certaines de ces catégories et sous-catégories renvoient au codage proposé dans le cadre d'évaluation de PISA, d'autres sont inspirées des résultats de recherche portant sur les difficultés des élèves dans la résolution d'une tâche en science (au-delà du cadre strict de PISA ou TIMSS). À titre d'exemple, dans la catégorie « caractéristiques intrinsèques de l'item » les chercheuses intègrent la caractéristique « illustration ». Cette décision est éclairée par les travaux de Hannus et Hyönä (1999) et de Rey-Mermet *et al.* (2019) qui montrent que les performances des élèves peuvent être impactées par la présence d'illustrations jouant un rôle central dans la tâche de résolution. Nous n'allons pas détailler ici l'ensemble des catégories et des caractéristiques du modèle de Duclos et ses collaboratrices. Dans ce qui suit, nous présentons les ajustements et les adaptations que nous avons effectués pour rendre le modèle de difficulté des questions de sciences compatible avec l'enquête TIMSS « science » 2019 et opérationnel dans sa mise en œuvre. Pour les caractéristiques conservées à l'identique, nous renvoyons le lecteur, la lectrice à Duclos *et al.* (2021). Le tableau 1 en annexe met en perspective les deux modèles : celui de Duclos et ses collaboratrices et le nôtre.

3.1. Catégorie 1 : caractéristiques liées au contenu en jeu dans la tâche

La première caractéristique – le « type de connaissance » – est reprise de Duclos *et al.* (2021). Sa présence est justifiée par le fait que les items évaluent soit des contenus de savoirs, soit des contenus procéduraux pour lesquels la mobilisation de connaissances de contenu scientifique (concepts, lois, etc.) n'est pas forcément nécessaire. Nous lui associons deux modalités : savoirs (le terme renvoie à des savoirs liés à des contenus scientifiques : concepts, lois, etc.) et savoir-faire. Les deux autres caractéristiques – « domaine scientifique évalué » et « sous-domaine scientifique évalué » – sont directement reprises du cadre de l'évaluation TIMSS. Elles se distinguent des catégories retenues par Duclos *et al.* Cette distinction est le fait de la différence liée aux spécificités des deux cadres théoriques des évaluations TIMSS et PISA. Nous n'avons, par exemple, pas retenu la caractéristique « champ d'application » dans la mesure où les items TIMSS sont très disciplinairement situés et ne renvoient pas à des situations relevant de contextes quotidiens.

3.2. Catégorie 2 : caractéristiques intrinsèques de l'item

Nous n'avons pas apporté de modification majeure à cette catégorie mais une première lecture des items TIMSS 2019 nous a permis de constater que la forme des questions des items ne reposait pas uniquement sur des QCM simples ou des questions ouvertes et incluait des propositions de réponse sous forme de tableaux croisés, de choix multiples complexes, ou une hybridation de plusieurs formes de question (par exemple : un QCM suivi d'une question ouverte). La caractéristique « illustration » a été conservée. On lui associe 8 modalités : aucune, photographie, dessin, schéma d'expérience, modèle, tableau, graphique. À cette

caractéristique est associée la caractéristique « La réponse à la question dépend des informations contenues dans l'illustration présente dans l'item » (modalités : « oui »/« non »/« sans objet » – dans le cas où aucune illustration n'est présente dans l'item). Nous n'avons pas conservé le nombre de mots présents dans l'item dans la mesure où, globalement, le texte (énoncé + questions) est plutôt court (4 lignes en moyenne, hors questions), ni la caractéristique « simulation » qui ne concerne que 2 items sur l'ensemble. La caractéristique « contexte » n'a pas été non plus retenue dans la mesure où, contrairement aux items PISA qui valorisent des situations relevant de la vie quotidienne, les items TIMSS science engagent des situations disciplinairement situées et s'ancrent dans le vécu scolaire des élèves. Enfin, la caractéristique « réponse dans l'item » n'a pas été reprise car nous n'avons pas dénombré d'item incluant cette caractéristique.

3.3. Catégorie 3 : caractéristiques liées aux raisonnements et stratégies de réponse des élèves

Cette catégorie a été modifiée pour prendre en charge les domaines et sous-domaines cognitifs évalués par TIMSS. Nous y avons ajouté deux opérations cognitives de traitement des données : une opération de nature sémiotique, une autre de nature mathématique. Nous avons également ajouté une référence aux conceptions des élèves. Pour des raisons expliquées plus haut, nous n'avons pas conservé la caractéristique « référence à la vie quotidienne » de Duclos *et al.* Nous n'avons pas non plus conservé la caractéristique « complexité cognitive » dans la mesure où cette dimension n'est pas présente dans le cadre de l'évaluation TIMSS, ni la caractéristique « projection » : on attend bien des élèves qu'ils et elles se projettent dans les situations proposées mais ces dernières sont pour la plupart conçues pour être proches de leur vécu scolaire et cela est saisi par la caractéristique « degré de familiarité de la situation » que nous ajoutons à la liste des caractéristiques de cette 3e catégorie. Sont considérées comme proches du vécu scolaire des élèves, des situations proposées par les programmes scolaires français de physique-chimie des cycles 3 et 4 (classe de 3e exclue) et de leur déclinaison dans les manuels. La caractéristique « *matching* » ou « correspondance de mots » a été conservée car elle nous a semblé pertinente au regard de certains items. Deux modalités lui sont associées : « validant »/« invalidant ». Pour Duclos et ses collaboratrices lorsque les élèves mettent en œuvre une stratégie de type « *matching* » cela consiste « à rechercher des termes communs dans la question et dans les réponses proposées » cela leur permet, par exemple, « de sélectionner une des réponses proposées dans le choix multiple en optant pour celle qui évoque un terme également contenu dans la question » (Duclos *et al.*, 2021). Par exemple, dans l'item libéré SE62032 (figure 1) l'expression « au fur et à mesure que la boule de métal refroidit » peut conduire les élèves aux choix B et D s'ils traitent les informations du registre graphique en ne tenant compte que de l'allure descendante des courbes (la stratégie de *matching* consiste ici à associer l'idée que la température diminue avec le fait que certaines courbes décroissent ; ainsi mise en œuvre, cette stratégie se révélerait invalidante). On notera au passage que l'opération de traitement sémiotique consistant, pour l'élève, à interpréter une ligne horizontale dans un graphique engageant une dimension temporelle comme l'évolution, dans le temps, d'une grandeur, est une opération difficile (voir Janvier, 1978 ; McDermott *et al.*, 1987 ; Delgado, 2020). Cet

item peut également activer une conception du type « le plus… le plus » (Tiberghien, 2004) et conduire au raisonnement suivant : plus la boule refroidit et plus sa masse diminue.

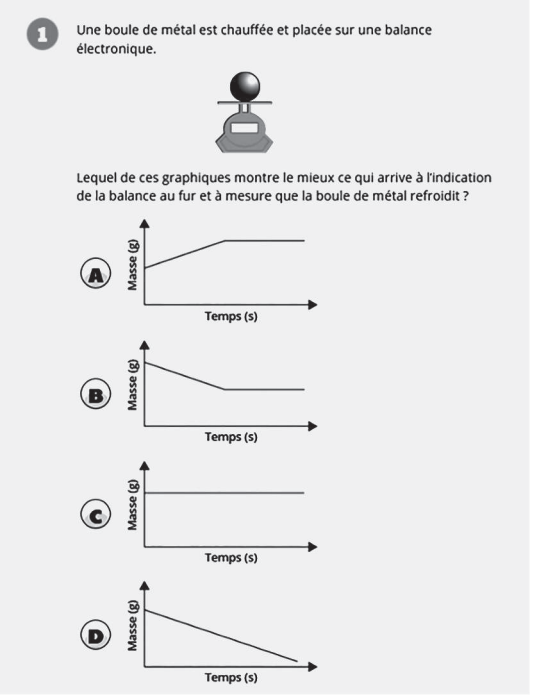


Fig. 1 : item libéré SE62032

4. Problématique et questions de recherche

Comme c'est le cas dans de nombreuses études secondaires conduites sur les évaluations internationales, notre approche vise à examiner les performances des élèves de quatrième français à l'évaluation TIMSS 2019 « science » à la lumière d'autres indicateurs que les seuls indicateurs de scores et de taux de réussite aux différents items. Le taux de réussite des élèves français à chaque item prend ici le statut de variable à expliquer ; les caractéristiques des items celui de variables explicatives. Ceci posé, nous formulons la question de recherche suivante : quelles caractéristiques des items TIMSS science 2019 sont potentiellement sources de difficultés pour les élèves français et spécifiques de ces élèves ?

Pour apporter réponse à ces questions nous analysons l'ensemble des items de l'évaluation TIMSS 2019 « science » *via* l'utilisation de notre grille d'analyse. Les données de cette grille sont ensuite traitées de manière quantitative.

5. Constitution du *corpus* de données et méthodologies d'analyse

Notre *corpus* est constitué :

- de 129 items (dont 15 items dits « libérés », i.e. accessible à tous et toutes[9]) ;
  - 47 en chimie (36 % de l'ensemble des items) ;
  - 60 en physique (46 % de l'ensemble des items) ;
  - 22 en sciences de l'univers - astronomie (18 % de l'ensemble des items).
- des pourcentages de bonne réponse des élèves français pour chaque item ;
- des pourcentages moyens de réussite calculés pour l'ensemble des élèves des 39 pays et les 7 provinces participants à l'enquête pour chaque item[10] ;
- de la grille critériée complétée (voir annexe, tableaux 2, 3 et 4).

5.1. Analyse critériée des items

L'IEA fournit, pour chaque item, un codage des caractéristiques des catégories « domaine », « sous-domaine », « sujet », « domaine » et « sous-domaine cognitifs », catégories que nous avons retenues dans notre grille d'analyse. Dans un premier temps, il nous a semblé nécessaire de vérifier que nous étions en accord avec ces codages. Pour cela, les deux chercheurs en didactique (auteurs de cet article) ont analysé, à l'aide de la grille critériée, pour ces catégories, et de manière indépendante et en aveugle, les 15 items libérés. L'accord entre les deux chercheurs s'est révélé quasi parfait et a concordé avec le codage de l'IEA. Par extrapolation, nous avons conclu qu'il y avait une très bonne adéquation entre les codages choisis par les concepteurs des items TIMSS et ceux repérés par les deux chercheurs. Nous avons donc pris la décision de conserver, pour tous les items, le codage de l'IEA.

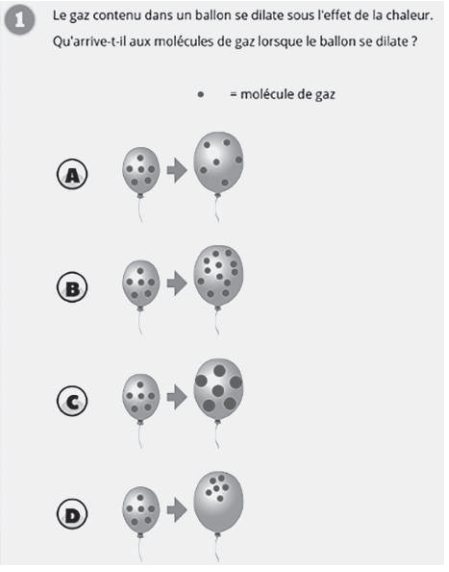
Dans un second temps, les chercheurs ont procédé au codage des caractéristiques des items de l'enquête pour l'ensemble des catégories de la grille d'analyse, à nouveau à l'aveugle et de manière indépendante. L'un des chercheurs (chercheur 1) a analysé tous les items de l'enquête ; l'autre (chercheur 2) un item sur trois. Pour confronter les deux processus de codage, nous avons calculé pour chaque catégorie le taux de concordance. Celui-ci est en moyenne de 91 %. Nous avons également évalué le degré d'accord en calculant pour chaque catégorie le coefficient kappa de Cohen. Celui-ci vaut en moyenne 0,75. On rappelle ici que selon Landis et Koch (1977), cités par Kassambara (2019), un kappa de Cohen compris entre 0,81 et 1 correspond à un accord presque parfait (c'est le cas pour six de nos catégories), un kappa compris entre 0,61 et 0,8 correspond à un accord substantiel (c'est le cas pour huit de nos catégories) et un kappa compris entre 0,41 et 0,6 correspond à un accord modéré (c'est le cas pour deux de nos catégories). Aucune de nos catégories ne correspond à un kappa inférieur à 0,4 (accords médiocres, légers ou passables selon les mêmes auteurs). Sur la base de ce résultat, le codage proposé par le chercheur 1 a été conservé.

9 <https://www.education.gouv.fr/media/73339/download> (consulté le 4 février 2022). On rappelle que tous les élèves ne répondent pas à l'ensemble des 129 items mais entre 12 et 18 items.

10 <https://timss2019.org/international-database/> (consulté le 4 mars 2022). À noter : les provinces « benchmarking participant » ne sont pas intégrées dans le calcul de moyenne internationale.

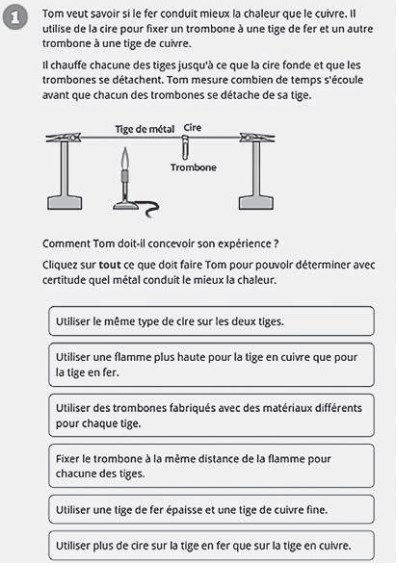
Les tableaux 1 et 2 présentent un codage complet de deux items libérés pour chacune des catégories de la grille d'analyse.

Tableau 1 : codage des caractéristiques de l'item libéré SE72200



% réussite (France) : 36 % réussite moyen : 45 %
Type de connaissance : savoir
Domaine scientifique : physique
Sous-domaine scientifique : états physiques
Format de l'item : QCM simple
Présence d'une illustration : oui (modèle)
Lien question-illustration : oui
Réponse dans l'item : non
Domaine cognitif : appliquer
Traitement sémiotique : oui (associer les boules rouges à des particules de gaz, dénombrer les boules, examiner leur répartition spatiale, examiner leur diamètre)
Matching : non
Conception : oui (l'air chaud monte/raisonnement homologique micro-macro)
Degré de familiarité : fort

Tableau 2 : codage des caractéristiques de l'item libéré SE72232



% réussite (France) : 44 % réussite moyen : 50 %
Type de connaissance : savoir-faire
Domaine scientifique : physique
Sous-domaine scientifique : conversion et transfert d'énergie
Format de l'item : QCM complexe (plusieurs choix possibles dans les réponses proposées)
Présence d'une illustration : oui (schéma d'expérience)
Lien question-illustration : oui
Réponse dans l'item : non
Domaine cognitif : raisonner
Traitement sémiotique : oui (identifier chaque élément du dispositif expérimental)
Matching : non
Conception : non
Degré de familiarité : faible

L'item libéré SE72200 (tableau 1) repose sur une illustration d'un modèle moléculaire d'un gaz par superposition d'un dessin réaliste et d'un schéma de modélisation microscopique à six boules. Pour répondre correctement à la question posée, les élèves doivent savoir que le changement de volume d'un gaz se fait à nombre de particules de ce gaz constant et qu'un gaz contenu dans une enveloppe (ici, un ballon de baudruche) occupe tout l'espace disponible dans l'enveloppe. L'élève doit également savoir que la taille des boules « particules » du dessin reste invariante au cours du changement de volume de l'enveloppe. À noter : la proposition D pourrait être interprétée sous l'autorité de l'affirmation « l'air chaud monte » traduite ici par le signifiant « regroupement des particules vers le haut du ballon », elle a d'ailleurs été choisie par 25 % des élèves ayant fourni une réponse à cette question (il s'agit de la deuxième proposition choisie par les élèves après la proposition (correcte) A. Le score des élèves français est proche du score moyen ce qui n'est plus le cas lorsque le changement physique concerne non plus un gaz mais un solide. En réponse à la question de l'item SE62004 (domaine « connaître » – figure 2) les élèves choisissent en premier l'option D (34 % contre 33 % pour l'option correcte B) et mettent probablement en œuvre un raisonnement homologique entre niveau macro et microscopique. Ce raisonnement est d'ailleurs également présent dans l'item SE72200 (tableau 1) et recueille 22 % des réponses des élèves (réponse C).

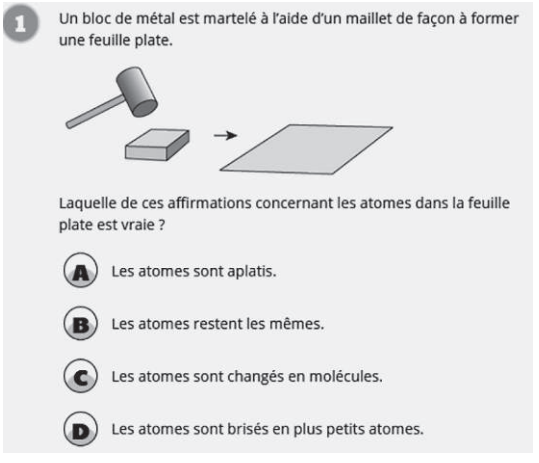


Fig. 2 : item libéré SE62004

L'item libéré SE72232 (tableau 2) évalue quant à lui la capacité des élèves à conduire une expérience scientifique. La situation présentée, bien qu'inhabituelle, fait l'objet d'une question de nature méthodologique expérimentale. Pour y répondre correctement (options A et D), les élèves doivent mobiliser l'idée suivante : « pour voir l'effet d'un paramètre sur une expérience, il est nécessaire que l'ensemble des autres paramètres reste constant ». Ils doivent également identifier, dans l'illustration de l'expérience, les éléments présents dans le texte et repérer (dans le texte) que l'expression « *tout* ce que doit faire… » (c'est nous qui soulignons) signifie que plusieurs réponses sont possibles. Le score des élèves français (44 % de réponses correctes, i.e. pourcentage d'élèves ayant choisi les options A et D) est proche du score moyen.

À travers ces exemples nous entrevoyons la possibilité d'éclairer les performances des élèves en les mettant en relation avec certaines des caractéristiques des items.

5.2. Analyse des données de l'ensemble de la grille critériée

À partir du codage des items de TIMSS 8 selon la grille critériée, nous obtenons une base de 129 items, avec 12 variables catégorielles décrivant les caractéristiques des items, ainsi que deux variables continues, le taux de réussite moyen international et le taux de réussite des élèves français. Notre objectif est d'identifier, parmi les caractéristiques des items, celles qui sont susceptibles d'avoir un effet sur la réussite des élèves, et en particulier des élèves français. Nous cherchons à estimer ce que l'on désigne par « effet pur » par opposition à « l'effet brut » qui ne rendrait compte que de l'écart entre le pourcentage de réussite des élèves français et le pourcentage de réussite moyen de l'ensemble des élèves. D'après Martin (2017), « L'effet pur permet d'estimer le rôle d'une variable ou d'une modalité compte tenu de toutes les modalités, c'est-à-dire une fois contrôlés et neutralisés les effets des autres modalités. Il exprime l'effet d'une modalité/variable sur une autre, "toutes choses égales par ailleurs" » (Martin, 2017, p. 114).

Dans un premier temps, nous réalisons deux modèles statistiques (M) de régression linéaire multiple qui vont nous permettre d'établir une relation entre une variable à expliquer (ici, le taux de réussite moyen international pour le premier modèle M1, le taux de réussite des élèves français pour le modèle M2) et un ensemble de variables dites explicatives (ici, les caractéristiques des items), susceptibles d'influencer la variable à expliquer. Le premier modèle permet d'identifier les caractéristiques des items ayant un lien avec la réussite des élèves de façon générale, tandis que le deuxième permet d'apprécier ce qui se passe plus précisément concernant les élèves français. La comparaison des résultats de ces deux modèles permet déjà de renseigner sur d'éventuelles spécificités des élèves français.

On peut cependant anticiper que ces deux modèles ont un pouvoir explicatif assez faible, puisqu'ils n'intègrent pas de mesure de la difficulté des items, qui reste le principal facteur de détermination de la réussite, d'autant que l'enquête TIMSS intègre, par construction, des items de niveaux de difficulté variés. Nous réalisons donc un troisième modèle de régression (M3) reprenant le modèle M2, auquel on ajoute parmi les variables explicatives le taux de réussite moyen international. On peut en effet considérer que le taux de réussite moyen international est un bon indicateur du niveau de difficulté d'un item. Moins l'item est réussi en moyenne au niveau international, plus il est difficile, et inversement. L'ajout de cette variable dans le modèle permet de mesurer l'impact des caractéristiques des items sur le taux de réussite des élèves français « à niveau de difficulté équivalent ».

L'intérêt de l'usage de la méthode statistique de régression linéaire est de pouvoir apprécier les effets de chacune des modalités des variables explicatives à l'aide d'un coefficient dit de régression qui « mesure l'ampleur du rôle que joue la modalité sur la variable à expliquer » (Martin, *op. cit.*, p. 116) et de produire une modélisation de la variable à expliquer par la contribution des modalités des variables explicatives. Pour les variables de nature catégorielle, le coefficient de régression peut prendre des valeurs positives ou négatives ; il est calculé à partir d'une modalité de référence qu'il est possible de fixer à l'avance. Dans la mesure où nous nous intéressons aux difficultés des élèves, notre modèle de régression prend pour référence pour chaque variable la modalité qui nous semble *a priori* la plus

favorable pour les élèves. Par exemple, pour la variable « familiarité », la modalité « forte » sera prise pour référence. On s'attend dans ce cas à une valeur négative du coefficient de régression associé à la modalité « faible » de la variable « familiarité ». Pour les variables numériques, le coefficient de régression s'interprète comme l'évolution de la variable à expliquer si la variable à expliquer augmente d'une unité. Au calcul du coefficient de régression vient s'ajouter celui d'une p-value qui renseigne sur la significativité de l'effet de la modalité de la variable explicative (par rapport à la modalité de référence pour une variable catégorielle) sur la variable à expliquer. Par convention, nous considérerons qu'une modalité influence significativement la variable à expliquer si la valeur de la p-value associée à cette modalité est inférieure à 0,1.

D'un point de vue méthodologique, nous allons réaliser une régression descendante dite « pas à pas » (Gunst & Mason, 1979). Cela consiste à effectuer dans un premier temps une régression prenant en compte toutes les variables explicatives puis à affiner le modèle obtenu en supprimant de manière itérative les variables sans effet sur le modèle une par une. L'ordre selon lequel s'opère la suppression d'une variable dépend de la p-value renvoyée par un test d'analyse des variances ANOVA[11] sur le résultat de la régression. Pour une régression donnée sur n variables explicatives, la variable affectée de la plus grande p-value à l'issue de l'ANOVA est supprimée ; on passe alors à la régression suivante pour n-1 variables explicatives. Le processus s'arrête lorsque toutes les variables explicatives restantes ont un effet significatif sur la variable à expliquer.

À chaque étape il est possible d'évaluer le pouvoir explicatif du modèle en examinant la valeur du coefficient de détermination linéaire de Pearson, noté R^2 qui est une mesure de la qualité de la prédiction d'une régression linéaire. Dans le cas de la régression linéaire multiple, il est d'usage de regarder le R^2 ajusté. Plus la valeur de ce coefficient est proche de 1 et plus la qualité de la prédiction de la régression est élevée.

L'ensemble de ces manipulations statistiques sont effectuées avec le logiciel *R*.

6. Résultats

6.1. Résultats des analyses du modèle M1 : régression sur le taux de réussite international moyen

L'opération de traitement pas à pas permet de faire émerger un modèle parcimonieux dans lequel une variable semble significativement impacter le taux de réussite international moyen : le format de la question. L'ajout des autres caractéristiques ne modifie pas ce modèle et peut donc être considéré comme sans effet sur le taux de réussite des élèves à l'item. Par exemple, un *matching* invalidant et/ou la présence d'une situation susceptible d'activer une conception erronée n'impactent pas significativement la proportion d'élèves qui répondent correctement à la question posée. Il est probable que ceci soit dû au fait que la proportion des items présentant ces caractéristiques sont finalement peu nombreux (respectivement 24 et 21 items sur 129, soit 17 % et 15 %).

11 *Analysis of variance*.

Tableau 3 : résultat de la régression linéaire multiple M1

Variable	Modalité	Coefficient	Test
Format de la question	Choix simple	(ref)	(ref)
	Choix multiple	-4,521	p = 0,66 (ns)
	Hybride	-21,92	p = 1,82 e-08
	Tableau croisé	-10,27	p = 0,16 (ns)
	Réponse ouverte	-10,95	p = 0,007

La désignation (ref) renvoie au fait que la modalité est prise pour référence.
La désignation (ns) renvoie au fait que la modalité n'a pas d'effet significatif sur le modèle.

Les effets de chacune des modalités des variables retenues sont appréciés à l'aide du coefficient dit « coefficient de régression » et de la p-value associée. Ici (tableau 3), seule la variable « format de la question » semble impacter de manière significative les taux de réponses correctes des élèves. Spécifiquement, la proportion d'élèves, tous pays confondus, qui répondent correctement à la question posée diminue significativement pour les items dont la réponse n'implique pas de choisir une proposition de réponse dans un QCM simple (modalité prise pour référence) ou multiple. Lorsque la question implique une réponse en deux temps (format hybride associant QCM et justification) le taux de réussite diminue de près de 22 points de pourcentage, et ce coefficient est très significatif (p-value = 1,82e-08). Un tel résultat ne signifie pas nécessairement que la seule difficulté rencontrée par les élèves concerne le format de la question. On peut en effet faire l'hypothèse que les autres difficultés se compensent entre l'ensemble des pays concernés par l'enquête, raison pour laquelle les autres caractéristiques des items ne ressortent pas dans ce modèle. Une autre hypothèse pourrait être que les QCM ont été conçues pour être plus faciles en majorité que les questions ouvertes ou hybrides. Le lien entre le format de la question et le taux de réussite de l'ensemble des élèves à l'échelle internationale pourrait s'avérer artificiel dans la mesure où le niveau de difficulté des items n'est ici pas contrôlé.

6.2. Résultats des analyses du modèle M2 : régression sur le taux de réussite des élèves français

Le pourcentage de bonnes réponses des élèves français est significativement influencé par quatre variables explicatives : le format de la question, comme dans le modèle 1, le lien illustration-question et le degré de familiarité de l'item avec le vécu scolaire des élèves, mais aussi les types de connaissance évalués (tableau 4, page suivante).

Tableau 4 : résultat de la régression linéaire multiple M2

Variable	Modalité	Coefficient	Test
Types de connaissance	Savoir-faire Savoirs	(ref) -10,39	(ref) p = 0,059
Format de la question	Choix simple Choix multiple Hybride Tableau croisé Réponse ouverte	(ref) -1,142 -21,01 -7,44 -13,47	(ref) p = 0,92 (ns) p = 1,01 e-05 p = 0,41 (ns) p = 0,0077
Lien illustration-question	Oui Non	(ref) -7,99	(ref) p = 0,038
Degré de familiarité	Forte Faible	(ref) -11,09	(ref) p = 0,004

Comme au niveau international, les questions impliquant une réponse ouverte ou une réponse hybride ont un moins bon taux de réussite que lorsque la question engage un QCM simple (p-value resp. 0,0077 et p = 1,01 e-05) et les écarts sont à peu près semblables au niveau international (environ -13 points de pourcentage pour les réponses ouvertes et -21 points pour les réponses hybrides). Lorsque la question d'un item nécessite de recourir à des informations présentes dans une illustration, la proportion d'élèves français qui répondent correctement à la question est plus importante que si l'illustration ne contient pas des informations nécessaires pour répondre à la question (p-value = 0,038). Ensuite, si la situation présentée aux élèves fait faiblement écho à des situations scolaires familières alors la proportion d'élèves qui répondent correctement à la question posée diminue de manière très significative (p-value = 0,004). Enfin, la proportion d'élèves français qui répondent correctement à une question diminue significativement (p-value =0,059) lorsque la question fait appel à un savoir par rapport à une question engageant un savoir-faire (situation prise pour référence dans le modèle de régression).

Pour résumer, les élèves français ont plus de chance de répondre correctement à la question posée dans un item TIMSS8 si celle-ci engage un savoir-faire, fait écho à une situation scolaire familière de l'élève, engage un QCM simple et un lien entre la question posée et une illustration. La comparaison des résultats des deux modèles indique que les élèves français sont en difficulté sur des questions engageant un savoir lié à un contenu disciplinaire de physique ou de chimie (concept, lois, modèle, etc.), faiblement familières de leur vécu scolaire et pour lesquelles les informations présentes dans l'illustration ne sont pas nécessaires pour répondre à la question.

Comme attendu, les modèles 1 et 2 ont un pouvoir explicatif assez faible avec des R^2 ajustés de 27 % et 24 % respectivement. Chacun de ces modèles permet d'expliquer moins de 30 % de la variance observée dans les taux de réussite aux items. En effet, celle-ci dépend avant tout du niveau de difficulté des items, ce qui n'est pas pris en compte ici.

Afin d'affiner l'analyse, le modèle 3 (tableau 5) permet d'expliquer le taux de réussite des élèves français en fonction des caractéristiques des items, et en contrôlant le niveau de difficulté des items par le biais du taux de réussite moyen international que nous ajoutons à notre traitement statistique par régression linéaire multiple en tant que nouvelle variable explicative.

6.3. Résultats des analyses du modèle M3 : régression sur le taux de réussite des élèves français et contrôle du niveau de difficulté des items

Tableau 5 : résultat de la régression linéaire multiple M3

Variable	Modalité	Coefficient	Test
Pourcentage total	Sans objet	1,05	P < 2e-16
Domaine scientifique	Sciences de l'univers Chimie Physique	(ref) -6,09 -3,15	(ref) p = 0,01 p = 0,168
Domaine cognitif	Raisonner Connaître Appliquer	(ref) -4,87 -4,94	(ref) p = 0,021 p = 0,012
Lien illustration-question	Oui Non	(ref) -3,47	(ref) p = 0,041
Degré de familiarité	Forte Faible	(ref) -4,69	(ref) p = 0,005

Le modèle 3 a un pouvoir explicatif plus important (R^2 ajusté = 79%)

On constate que le taux de réussite des élèves français est principalement fonction du niveau de difficulté de l'item, ici capté par le pourcentage de réussite global de l'ensemble des élèves (p-value < 2e-16). L'évolution des performances des élèves français suit la même tendance que les performances de l'ensemble des élèves au niveau international : à une augmentation d'un point de pourcentage du taux international correspond une augmentation d'environ 1,05 point de pourcentage du taux français.

Lorsqu'on contrôle le niveau de difficulté des items, on constate que le format de question et le type de connaissance ne sortent plus significativement. On pourrait conclure qu'à niveau de difficulté équivalent au niveau international, les élèves français réussissent aussi bien les questions ouvertes ou hybrides que les QCM. Mais il faut nuancer ce résultat car on a vu dans le modèle M1 que le taux de réussite global intègre cette caractéristique des items et il est possible que les QCM aient été conçues pour être plus faciles que les questions ouvertes ou hybrides et ce, de manière générale pour l'ensemble des élèves. Aussi, le lien illustration-question et le degré de familiarité sont toujours présents, mais les écarts sont moins importants que dans le modèle M2. À niveau de difficulté équivalent au niveau international, le taux de réussite des élèves français diminue de presque 7 points de pourcentage lorsque la question n'est pas familière. Enfin, dans ce troisième modèle, les domaines scientifiques et cognitifs sortent significativement. À niveau de difficulté équivalent, les élèves français réussissent moins bien en physique et surtout en chimie qu'en sciences de l'univers, et dans les domaines connaître et appliquer que dans le domaine raisonner. Ces constats invitent toutefois à la prudence. D'une part, un modèle de régression donne une idée des relations de corrélation et non de causalité qui peuvent exister entre une variable à expliquer et des variables explicatives. D'autre part, nos variables explicatives ne sont peut-être pas complètement indépendantes les unes des autres (en particulier le taux de réussite global dépend peut-être de certaines de nos autres variables ce qui pourrait expliquer que le format de l'item, présent dans le modèle M1, disparaisse du modèle M3).

7. Discussion

Cet article avait pour ambition d'examiner les résultats des élèves français à l'enquête internationale TIMSS science 2019 pour la classe de 4e (grade 8). Il s'agissait d'une part de proposer un éclairage didactique des items de l'enquête par l'application d'une version adaptée du « modèle de difficulté des questions de sciences » (Duclos *et al.*, 2021) et d'autre part, d'apprécier les succès et les échecs des élèves français au regard de certaines caractéristiques des items non prises en charge dans le modèle MRI sur lequel repose le calcul des scores des élèves par l'IEA. Le grain d'analyse choisi se voulait plus fin que celui choisi par l'IEA (qui propose un calcul de score unidimensionnel) et davantage porté par les résultats des travaux de recherche en didactique des sciences.

D'un traitement statistique de type « régression linéaire multiple » des données obtenues par une analyse critériée de l'ensemble des items de l'enquête, il ressort que les élèves français sont plus performants lorsqu'il s'agit de produire une réponse à une question mettant en jeu un raisonnement. Ce résultat peut peut-être s'expliquer par le fait que l'enseignement des sciences en cycle 4 en France s'incarne dans la mise en œuvre de démarches d'investigation, dans des activités de résolution de tâches complexes, dans l'analyse documentaire. Ces trois entrées pédagogiques répondent aux exigences curriculaires de faire entrer les élèves « dans une relation scientifique avec les phénomènes naturels [...] en adoptant une posture scientifique faite d'attitudes (curiosité, ouverture d'esprit, remise en question de son idée, exploitation positive des erreurs, etc.) et de capacités (observer, expérimenter, mesurer, raisonner, modéliser ; etc.) ». Ce constat est sans doute à rapprocher du fait que la réussite des élèves est favorisée lorsque la question d'un item nécessite de faire appel aux données disponibles dans l'illustration.

Les contre-performances des élèves en chimie sont associées à de très faibles scores aux contenus « propriétés des acides et des bases », « tableau périodique des éléments », « transformation chimique » et « structure de la matière ». La réussite des élèves à ces items est de 25 points inférieure à la réussite moyenne de tous les participant.e.s à l'enquête. Plusieurs éléments sont susceptibles d'expliquer ce constat. Il y a d'abord une suspicion d'inadéquation entre certains savoirs chimiques évalués par TIMSS et la disponibilité de ces savoirs chez les élèves de la classe de 4e. Usuellement, les savoirs sur les acides et les bases sont abordés en fin de cycle 4 (i.e. en classe de troisième), mais nous pourrions ajouter les savoirs liés à la description de la constitution de l'atome et de la structure interne du noyau qui, selon le *BOEN* n° 31 (2020) « peut être réservée à la classe de 3e ». Il n'est, dès lors, pas étonnant que le score de réussite des élèves français à l'item SE72103 portant sur les propriétés des acides et des bases soit de 20 points inférieur au score moyen (30 % contre 50 %).

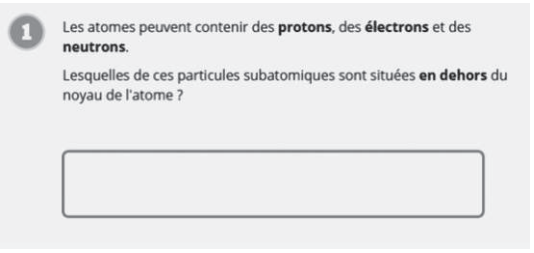


Fig. 3 : item libéré SE72013
On notera que le terme « subatomique » ne fait pas partie du vocabulaire courant de l'enseignement de la chimie au collège.

D'ailleurs, un écart semblable est constaté pour les items engageant l'exploitation d'extraits du tableau périodique des éléments. L'item SE72110 (figure 4) ne recueille, par exemple, que 15 % de réponses correctes (contre 33 % pour l'ensemble des élèves tout pays confondu). La tâche attendue des élèves repose sur la mobilisation d'un savoir *a priori* non disponible (le principe de l'organisation des éléments dans le tableau et sa lecture en ligne, de gauche à droite et de bas en haut). Pour les items relevant de ces savoirs, nous avons systématiquement choisi la modalité « faible » de la catégorie « familiarité de la situation ».

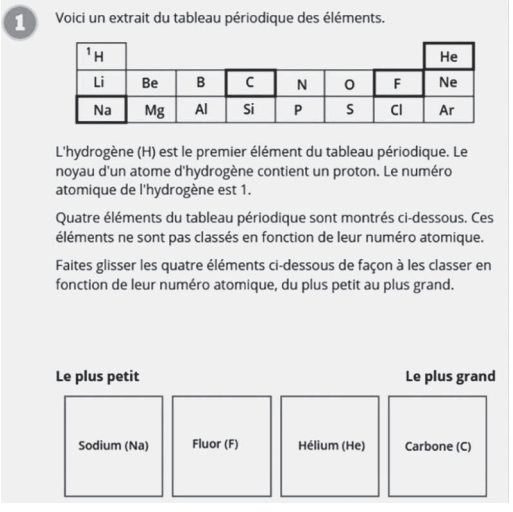


Fig. 4 : item libéré SE72110

Un autre élément d'explication s'identifie par les difficultés mises au jour dans les travaux de recherche en didactique de la chimie. La défaillance des élèves français en chimie a été pointée par Laugier et Dumon en 2004 à partir d'une évaluation du niveau des élèves de troisième français réalisée en 1995 par le ministère de l'Éducation nationale (Murat & Jouvanceau, 1997). Entre autres difficultés, les chercheurs soulignent la faible capacité des élèves de collège à distinguer une transformation chimique d'une transformation physique. Ils indiquent par exemple que les élèves « considèrent comme physique tout ce qui se transforme "naturellement" et comme chimique tout ce qui est provoqué par l'homme »

(Laugier & Dumon, 2004, p. 62). Ce résultat vient corroborer les travaux de Solomonidou et Stavridou (1994) qui indiquent que « tout ce qui est chimique est exclu de la nature » et que « même le bois qui brûle est considéré comme un phénomène physique par une grande partie des élèves [français] » (*op. cit.*, p. 77). Or, 35 items relevant des domaines scientifiques chimie et physique engagent les élèves à opérer une telle distinction. Cette tâche est d'autant plus difficile que l'introduction de ces deux types de transformation ne s'inscrit pas dans la même temporalité curriculaire : les transformations physiques sont introduites dès le cycle 3 et les transformations chimiques sont tout juste abordées en classe de quatrième. Dans d'autres items du domaine scientifique « chimie », les élèves sont invités à distinguer différentes entités chimiques à partir de leur formule brute ou à dénombrer le nombre d'atomes constitutifs d'une molécule dont la formule brute est donnée. Là encore, les travaux de recherche en didactique de la chimie montrent que les tâches d'interprétation des formules chimiques sont difficiles pour les élèves (Fillon, 1997 ; Canac & Kermen, 2016), et la difficulté est certainement renforcée par le fait « qu'il existe des contraintes institutionnelles, épistémologiques et didactiques qui mettent les enseignants en difficulté au moment de l'introduction des formules chimiques, faute de ressources adaptées » (Canac & Kermen, 2020, p. 70).

On rappelle qu'à l'inverse, les items relevant des sciences de l'univers sont mieux réussis que les autres, ce qui peut s'expliquer par une familiarisation précoce des élèves français avec les savoirs relevant de l'astronomie (dès le cycle 3).

Conclusion et perspectives

Examiner les 129 items de l'enquête TIMSS 2019 en science à la lumière d'une adaptation de la grille critériée proposée par Duclos et ses collaboratrices nous a permis de mettre du relief sur certaines des caractéristiques des items qui semblent impacter les performances des élèves français au regard des performances moyennes de l'ensemble des élèves testés. En cela, la recherche présentée ici vient nourrir et corroborer les résultats des études secondaires conduites sur les évaluations internationales, et sur TIMSS en particulier. Comme cela a été pointé par Dempster et Reddy (2007), une part non négligeable des items de TIMSS 2019 « science » engagent des savoirs non disponibles chez la plupart des élèves français de la classe de 4e et des situations qui leur sont faiblement familières.

Notre travail ouvre plusieurs pistes de recherche pour des études ultérieures. La première de ces pistes concerne les sous-performances des élèves français en chimie dont les causes méritent certainement d'être davantage explorées. On pourrait approcher cette perspective de plusieurs manières : en nous intéressant, par exemple, au degré d'affinité des enseignant.e.s de physique-chimie français pour la discipline « chimie » (Alturkmani, Trouche & Morge, 2018), ou au niveau d'expertise des enseignant.e.s en chimie en comparant les notes de physique et de chimie obtenues par les candidat.e.s au CAPES de physique-chimie.

La deuxième piste s'inscrit en prolongement des travaux de Le Hebel (2021) qui s'est attachée à utiliser la grille critériée de Duclos *et al.* sur les items de l'évaluation PISA 2015 « culture scientifique ». Il pourrait être désormais intéressant de comparer les résultats obtenus pour les deux enquêtes PISA et TIMSS et de dégager ainsi des traits communs et des spécificités à chacune des enquêtes pour apprécier, plus finement encore, les connais-

sances et les compétences des élèves français en sciences. Rappelons toutefois que les deux enquêtes ne poursuivent pas le même objectif : l'ambition du PISA « culture scientifique » est de s'extraire des contraintes telles que l'évaluation de connaissances liées aux contenus disciplinaires alors que TIMSS cible ces connaissances. Notre grille a été conçue pour prendre en charge cette spécificité TIMSS et son utilisation s'est accompagnée d'une analyse didactique des items : repérage de situations susceptibles d'activer des conceptions d'élèves, propositions d'explications des sous-performances des élèves français aux items relevant de la chimie, mise en valeur des difficultés des élèves face à des items nécessitant de faire appel à des connaissances liées aux contenus disciplinaires propres à la physique et à la chimie, difficultés associées à certaines spécificités sémiotiques des illustrations présentes dans les items, etc. Un projet de mise en perspective des deux enquêtes PISA et TIMSS nécessiterait certainement une harmonisation des deux grilles de manière à ce que puissent être prises en charge les spécificités didactiques des items constitutifs de PISA. Mais conservées en l'état, certaines caractéristiques pourraient être rapprochées. Par exemple, nous avons montré que les élèves français surperforment aux items nécessitant la mise en œuvre d'un raisonnement. Qu'en est-il pour PISA ? Les élèves français surperforment-ils aux items de complexité cognitive (DOK[12]) de niveau 3 et/ou 4 « réflexion stratégique » et/ou « réflexion étendue » ?

Enfin, il pourrait être fructueux d'analyser de façon plus approfondie la manière dont les élèves comprennent et résolvent les questions de TIMSS en conduisant quelques entretiens d'explicitation avec des élèves de la fin de la classe de 4e. Notre recherche ne donne en effet aucune indication sur les types de raisonnements mis en œuvre, sur les indices retenus par les élèves pour résoudre les tâches demandées, sur les cheminements suivis. Cette troisième piste pourrait se voir nourrie par la démarche suivie par Duclos (2022) dans son travail doctoral portant sur l'évaluation PISA dans lequel on perçoit clairement l'intérêt de « faire parler » les élèves lorsqu'ils et elles s'attachent à répondre aux questions des items PISA. On voit, par exemple, que derrière certaines réponses correctes se cachent des raisonnements erronés, ou encore, que le passage de l'oral à l'écrit apparaît comme une difficulté à ne pas négliger (certains élèves pouvant produire des réponses correctes à l'oral et ne pas être en capacité de transcrire ces réponses à l'écrit – ce qui pourrait expliquer, au moins en partie, les résultats obtenus *via* le modèle de régression M2).

Cécile de Hosson
cecile.dehosson@u-paris.fr

Nicolas Décamp
nicolas.Décamp@u-paris.fr

Anaïs Bret
anais.bret@education.gouv.fr

Marion Le Cam
marion.lecam@education.gouv.fr

12 *Depth of knowledge.*

Bibliographie

ALTURKMANI M. D., TROUCHE L. & MORGE L. (2018). Étude des liens entre affinités disciplinaire et didactique et travail de l'enseignant : le cas d'un enseignant de physique-chimie en France. *Recherches en didactique des sciences et des technologies (RDST)*, nº 17, p. 129-157.

BAUTIER É., CRINON J., RAYOU P. & ROCHEX J.-Y. (2006). Performances en littéracie, modes de faire et univers mobilisés par les élèves : analyses secondaires de l'enquête PISA 2000. *Revue française de pédagogie*, nº 157, p. 85-101.

BODIN A., HOSSON C. de, Décamp N. & GRAPIN N. (2016). *Comparaison des évaluations PISA et TIMSS : comprendre les évaluations internationales*. Rapport du CNESCO.

BRET A., KESKPAIK S., ROUSSEL L. & VERLET I. (2016). Les élèves de 15 ans en France selon PISA 2015 en culture scientifique : des résultats toujours marqués par de fortes inégalités. *Note d'information, MENESR-DEPP*, décembre.

CANAC S. & KERMEN I. (2020). Design of a didactical resource to introduce chemical formulas in secondary school. *Enseñanza de las Ciencias. Revista de investigación y experiencias didácticas*, vol. 38, nº 2, p. 65-82.

CANAC S. & KERMEN I. (2016). Exploring the mastery of French students in using basic notions of the language of chemistry. *Chemistry Education Research and Practice*, vol. 17, nº 3, p. 452-473.

CLASSICK R., GAMBHIR G., LIHT J., SHARP C. & WHEATER R. (2021). PISA 2018 additional analyses: what differentiate disadvantaged pupils who do well in PISA from those who do not? in *National Foundation for Educational Research*. En ligne : <https://files.eric.ed.gov/fulltext/ED612577.pdf> (consulté le 24 août 2022).

COPPENS N. (2012). L'évaluation de la culture scientifique des élèves français de quinze ans dans PISA 2009. *Recherches en éducation,* nº 14, p. 51-64.

CRISP V. & GRAYSON R. (2013). Modelling question difficulty in an A level physics examination. *Research Papers in Education*, vol. 28, nº 3, p. 346-372.

DELGADO I. (2020). *L'utilisation d'un logiciel de géométrie dynamique comme une stratégie possible pour surmonter des difficultés dans l'interprétation des représentations graphiques x(t) de la cinématique classique*. Thèse de doctorat, université de Paris.

DEMPSTER E. R. & REDDY V. (2007). Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education*, vol. 91, nº 6, p. 906-925.

DOLIN J. & KROGH L. B. (2010). The relevance and consequences of PISA science in a Danish context. *International Journal of Science and Mathematics Education*, vol. 8, nº 3, p. 565-592.

DUCLOS M. (2022). *Influence du contexte socio-économique et du niveau scolaire sur la réalisation de tâches en sciences : liens entre caractéristiques des tâches et compréhension des élèves*. Thèse de doctorat, ENS de Lyon.

DUCLOS M., LE HEBEL F., TIBERGHIEN A., MONTPIED P. & FONTANIEU V. (2021). Élaboration d'un modèle de difficulté de questions évaluant la culture scientifique des élèves. *Éducation et didactique*, vol. 15, nº 3, p. 103-131.

FILLON P. (1997). Des élèves dans un labyrinthe d'obstacles. *Aster*, nº 5, p. 113-141.

GLYNN S. M. (2012). International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items. J*ournal of Research in Science Teaching,* vol. 49, nº 10, p. 1321-1344.

GUNST R. F. & MASON R. L. (1979). Some considerations in the evaluation of alternate prediction equations. *Technometrics*, vol. 21, nº 1, p. 55-63.

HANNUS M. & HYÖNÄ J. (1999). Utilization of illustrations during learning of science textbook passages among low-and high-ability children. *Contemporary Educational Psychology*, vol. 24, nº 2, p. 95-123.

HARLOW A. & JONES A. (2004). Why students answer TIMSS science test items the way they do. *Research in Science Education*, vol. 34, nº 2, p. 221-238.

JANVIER C. (1978). *The Interpretation of Complex Cartesian Graphs Representing Situations-Studies and Teaching Experiments*. Thèse de doctorat, Université de Nottingham.

KASSAMBARA A. (2019). *Inter-Rater Reliability Essentials: Practical Guide In R*. Édition 1. independently published, ISBN-10 : 1707287562.

LANDIS J. R. & KOCH G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, vol. 1, n°33, p. 159-74.

LAU K. C. (2009). A critical examination of PISA's assessment on scientific literacy. *International Journal of Science and Mathematics Education*, vol. 7, n°6, p. 1061-1088.

LAUGIER A. & DUMON A. (2004). L'équation de réaction : un nœud d'obstacles difficilement franchissable. *Chemistry Education Research and Practice*, vol. 5, n° 1, p. 51-68.

LAVEAULT D. & GRÉGOIRE J. (1997). *Introduction aux théories des tests en sciences humaines*. Bruxelles : De Boeck.

LE HEBEL F. (2021). *Contexte culturel et compréhension de la culture scientifique par les élèves*. Habilitation à diriger les recherches, Lyon : ENS de Lyon.

LE HEBEL F., MONTPIED P. & TIBERGHIEN A. (2014). Which effective competencies do students use in PISA assessment of scientific literacy? In C. Bruguière, A. Tiberghien & P. Clément (dir.), *Topics and Trends in Current Science Education,* Dordrecht : Springer, p. 273-289.

MARTIN O. (2017). *L'analyse quantitative des données*. Paris : Armand Colin.

MARTIN M. O., VON DAVIER M. & MULLIS I. V. S. (éd.). (2020). *Methods and Procedures: TIMSS 2019 Technical Report*. Retrieved from Boston College, TIMSS & PIRLS International Study Center. En ligne : <https://timssandpirls.bc.edu/timss2019/methods> (consulté le 4 mars 2022).

MCDERMOTT L. C., ROSENQUIST M. L. & VAN ZEE E. H. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics*, vol. 55, n°6, p. 503-513.

MEN (Ministère de l'Éducation nationale) (2020). *Bulletin officiel de l'Éducation nationale*, n° 31.

MULLIS I. V. S. & MARTIN M. O. (éd.). (2017). *TIMSS 2019 Assessment Frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center. En ligne : <http://timssandpirls.bc.edu/timss2019/frameworks/> (consulté le 4 mars 2022).

MURAT F. & JOUVANCEAU P. (1997). *Évaluation pédagogique en fin de troisième générale et technologique 1995*. Paris : MENESR-DEPP.

PRENZEL M., HÄUßLER P., ROST J. & SENKBEIL M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, vol. 30, n°2, p. 120-135.

REY-MERMET A., GADE M., SOUZA A. S., VON BASTIAN C. C. & OBERAUER K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General*, vol. 148, n°8, p. 1335-1372. En ligne : <https://doi.org/10.1037/xge0000593>.

SOLOMONIDOU C. & STAVRIDOU H. (1994). Les transformations des substances, enjeu de l'enseignement de la réaction chimique. *Aster*, n° 18, p. 75-95.

TIBERGHIEN A. (2004). Causalité dans l'apprentissage des sciences. *Intellectica*, vol. 38, n° 1, p. 69-102.

WIBERG M. & ROLFSMAN E. (2019). The association between science achievement measures in schools and TIMSS science achievements in Sweden. *International Journal of Science Education*, vol. 41, n°16, p. 2218-2232.

Annexes

Tableau 1 : mise en perspective du modèle de difficulté des questions de sciences (Duclos *et al*., 2021) et de son adaptation pour l'analyse des items TIMSS8

Modèle de Duclos *et al*.		**Modèle adapté pour la présente étude**	
Catégorie 1 : caractéristiques liées aux contenus en jeu dans la tâche			
Types de connaissances	Connaissances scientifiques Connaissances procédurales Connaissances épistémiques	Types de connaissances	Savoirs Savoir-faire
Systèmes	Physique Chimie Biologie Sciences de la Terre et de l'univers	Domaines scientifiques*	Physique* Chimie* Sciences de l'univers*
Types de compétences	Expliquer Évaluer et concevoir Interpréter	Voir catégorie 3	
Champs d'application	Frontières de la science et de la technologie Qualité de l'environnement Ressources naturelles Risques Santé et maladies	Sous-domaine scientifique évalué*	Composition de la matière* Propriétés de la matière* Transformations chimiques* États physiques et transformations de la matière* Conversion et transfert d'énergie* Lumière et son* Électricité et magnétisme* Mouvement et forces* Terre dans le système solaire et l'univers*
Catégorie 2 : caractéristiques intrinsèques de l'item			
Format de l'item	Choix multiple simple Choix multiple complexe Réponse ouverte construite	Format de l'item	Choix simple Choix multiple Hybride (QCM puis question ouverte) Tableau croisé Réponse ouverte
Longueur du texte	44 à 136 mots 137 à 195 mots 196 à 244 mots 245 à 417 mots	Non pertinent (textes courts)	

Modèle de Duclos *et al.*		**Modèle adapté pour la présente étude**	
Type d'illustration	Aucune Graphique Photo/dessin Schéma Tableau et multiples illustrations	Type d'illustration	Aucune Graphique Photo Dessin Schéma d'expérience Modèle Tableau
Simulation	Oui Non	Non pertinent (seulement 2 items engagent une simulation)	
Contextes	Global Global-personnel Sociétal Sociétal-global Sociétal-personnel	Non pertinent (le contexte des items relève du vécu scolaire des élèves)	
Réponse dans l'item	Oui Non	Non pertinent pour les items de l'évaluation	
Dépendance aux ressources	Oui Non	Fusionné avec la catégorie suivante	
Lien question-illustration	Oui Non	Lien question-illustration	Oui Non
Catégorie 3 : caractéristiques liées aux raisonnements et aux stratégies de réponse des élèves			
Complexité cognitive (DOK à 4 niveaux)	Remémoration Utilisation de savoirs conceptuels Réflexion stratégique Réflexion étendue	Domaine cognitif*	Connaître* Appliquer* Raisonner*
		Autre opération cognitive	Traitement sémiotique Traitement mathématique Aucun
		Conception	Oui Non
Matching	Aidant Invalidant	Matching	Aidant Invalidant
Projection	Directe Indirecte	Non pertinent (le contexte des items relève du vécu scolaire des élèves)	
Niveau de référence à la vie quotidienne		Degré de familiarité de la situation (vécu scolaire)	Faible Fort

Les caractéristiques suivies d'un * sont directement reprises du cadre de l'évaluation TIMSS sciences. Les autres ont été soit reprises de Duclos *et al.*, soit induites lors d'une première analyse *a priori* des items.

Tableau 2 : dénombrement des caractéristiques des items relevant de la catégorie 1

Caractéristiques et sous caractéristiques	Description	Nombre items
Types de connaissances	Savoirs	113
	Savoir-faire	16
Domaine scientifique	Chimie	47
	Physique	60
	Sciences de l'univers	22
Sous-domaine scientifique évalué	Composition de la matière	11
	Propriétés de la matière	25
	Transformations chimiques	11
	États physiques et transformations de la matière	12
	Conversion et transfert d'énergie	8
	Lumière et son	7
	Électricité et magnétisme	11
	Mouvement et forces	22
	Terre dans le système solaire et l'Univers	8

Tableau 3 : dénombrement des caractéristiques des items relevant de la catégorie 2

Caractéristiques et sous caractéristiques	Description	Nombre items
Format de l'item	Choix simple	64
	Choix multiple	3
	Hybride (QCM puis question ouverte)	31
	Tableau croisé	7
	Réponse ouverte	24
Type d'illustration	Photographie	3
	Dessin	22
	Schéma d'expérience	27
	Modèle	16
	Aucune	35
	Tableau	16
	Graphique	10
La réponse à la question dépend des informations contenues dans l'illustration présente dans l'item	Oui	56
	Non	29

Tableau 4 : dénombrement des caractéristiques des items relevant de la catégorie 3

Caractéristique et sous caractéristiques	Description	Nombre items
Domaines cognitifs	Connaître Appliquer Raisonner	41 42 36
Autres opérations cognitives	Traitement sémiotique Traitement mathématique Aucune	66 2 61
Matching	Invalidant Aidant Aucun	24 10 95
Conception	Oui Non	21 108
Degré de familiarité de la situation	Faible Fort	50 79