Évaluation en mathématique avec la technique des degrés de certitude

Mémoire

dans le cadre du Master of Advanced Studies pour le degré secondaire II

Auteur:

Bozorgmehr AMINIAN

Superviseur :

Prof. Dr Jean-Luc GILLES

Jury :

Melisa SHEHU

Prof. Dr Jean-Michel RIGO

Août 2023

hep/ haute école pédagogique vaud

Table des matières

1 Introduction

Ce mémoire de MAS en enseignement pour le secondaire II à la Haute École Pédagogique (HEP) de Lausanne se focalise sur l'utilisation des Degrés de certitude (DC) pour des évaluations en mathématique. Initialement conçus pour les Questions à choix multiple (QCM), les DC sont adaptés pour les Questions à réponse ouverte (QRO) dans cette étude en docimologie. Contrairement aux QCM qui peuvent être évaluées de manière binaire (correctes ou incorrectes), les réponses aux QRO sont souvent partiellement correctes ou incorrectes. La section 2 expose les principes de l'évaluation ainsi que les différences entre les QCM et les QRO.

La section 3 aborde les bases de l'évaluation avec la technique des DC. Les mesures du réalisme développées par Gilles (2002) et Prosperi (2015) sont présentées, avec leurs adaptations pour les QRO et les tests courts. En effet, la mesure probabiliste de Prosperi (2015) tend à surestimer le réalisme pour les tests comportant peu de questions, en raison de l'intervalle de confiance de Wilson (1927) utilisé par cette mesure, qui est large pour un petit échantillon. De plus, la centration utilisée par Gardner-Medwin et Gahan (2003) est adaptée à l'échelle des DC utilisée dans ce travail.

La section 4 énonce les questions de recherche que ce mémoire tente d'adresser. La conformité des ajustements apportés aux mesures du réalisme de Gilles (2002) et Prosperi (2015) est évaluée. Ces adaptations montrent une corrélation élevée avec les mesures antérieures ainsi qu'une corrélation faible avec les notes obtenues aux évaluations, ce qui encourage leur utilisation. Ensuite, ce mémoire analyse l'évolution du réalisme au fil du temps, comparant les résultats avec les études antérieures de Callender et al. (2015), Leclercq et Gilles (1994), et Miller et Geraci (2011), suggérant une augmentation du réalisme chez les élèves à faibles performances, en particulier après des retours constructifs et un entraînement à l'utilisation des DC.

Les résultats, exposés dans la section 5, démontrent que les adaptations apportées sont en accord avec la mesure du réalisme de Gilles (2002). Les analyses statistiques confirment la progression du réalisme chez les élèves aux performances faibles.

Ces résultats encouragent l'utilisation des DC dans l'évaluation des compétences mathématiques et appellent à des études plus approfondies. De plus, toutes les mesures présentées sont implémentées en *Python* et sont disponibles pour de futures évaluations dans l'enseignement et études en docimologie.

2 Évaluation

La section 2.1 aborde l'évaluation en tant qu'une des trois composantes d'un dispositif d'enseignement-apprentissage selon Anderson (2002), tandis que la section 2.2 traite de son utilisation pour l'acquisition des éléments de l'apprentissage selon la taxonomie d'Anderson. Les avantages et inconvénients des évaluations basées sur les Questions à choix multiple (QCM) sont discutés dans la section 2.3. Quant à l'utilisation des Questions à réponses ouvertes (QRO), elle est analysée en détail dans la section 2.4 ainsi que la nécessité d'utiliser les QRO en mathématiques.

2.1 Composantes d'un dispositif d'enseignement-apprentissage

L'évaluation des connaissances de l'élève occupe une place centrale dans son processus d'apprentissage. Selon Anderson (2002), un dispositif d'enseignement-apprentissage se structure en trois composantes dont il s'agit d'assurer la cohérence lors de la mise en œuvre des unités d'apprentissage. Cette préoccupation pédagogique de cohérence peut se schématiser par un triangle représentant l'alignement curriculaire entre trois composantes brièvement résumées ci-dessous :

- **Normes/objectifs.** Les normes au sein d'une unité d'apprentissage consistent en un ensemble d'objectifs d'apprentissage, aussi appelées simplement les objectifs. Ces normes sont définies, par exemple, par le Plan d'Étude Romand (PER).
- **Activité et matériel.** Dans le but de favoriser la progression de l'élève dans son apprentissage, diverses activités d'enseignement doivent être offertes. À titre d'exemple, l'enseignant peut proposer à l'élève de s'engager dans la résolution d'exercices ou de participer à la correction au tableau devant la classe. Pour faciliter la réalisation de ces activités par l'apprenant, il est essentiel de mettre à sa disposition le matériel approprié, tel que des chaises, des tables, un tableau noir, des calculatrices, etc.
- **Évaluation.** De manière informelle, l'enseignant peut fournir une rétroaction sur le travail de l'élève. Cette rétroaction peut se produire lors de la correction d'exercices ou au travers d'évaluations formatives. De plus, une unité d'apprentissage comprend généralement une ou plusieurs évaluations formelles qui peuvent être formatives ou sommatives et qui peuvent dans ce second cas qui aboutir à l'attribution d'une note reflétant les performances de l'élève.

Intégrée dans un dispositif d'enseignement où la cohérence avec les objectifs visés et les activités d'apprentissage est respectée, l'évaluation joue un rôle crucial dans le processus de formation de l'élève. Conscient des enjeux liés à l'évaluation, nous avons choisi de centrer cette étude sur le champ de la docimologie et plus particulièrement sur l'utilisation de la technique des degrés de certitude en évaluation des apprentissages.

2.2 Taxonomie

L'objectif d'une évaluation est de déterminer dans quelle mesure l'apprenant a atteint les objectifs visés. Ceux-ci peuvent être catégorisés selon la taxonomie révisée de Bloom par Anderson et ses collaborateurs (Anderson et al., 2000). Selon cette taxonomie, un objectif d'apprentissage se présente sous la forme *Sujet-Verbe-Objet* (SVO), où le « sujet » est généralement l'apprenant. Le « verbe » représente un processus cognitif (*se souvenir, comprendre, appliquer, analyser, évaluer*), et l'« objet » est une forme de connaissance (*factuelle, conceptuelle, procédurale, métacognitive*). Par exemple, prenons l'objectif défini dans le plan d'études de la Direction Générale de l'Enseignement Postobligatoire[1] (DGEP) pour la première année d'une école de culture générale : « *évaluer une expression littérale* ». Dans ce cas, le processus cognitif consiste à appliquer une connaissance procédurale en remplaçant les paramètres dans l'expression littérale par des valeurs numériques. Afin de favoriser un apprentissage efficace, l'enseignant doit offrir à l'élève la possibilité d'utiliser divers processus cognitifs en relation avec les différentes formes de connaissance de la taxonomie. De plus, il doit être en mesure de comprendre le fonctionnement de ces processus cognitifs, en partie grâce aux évaluations. En règle générale, une évaluation se compose d'une série de questions et de réponses attendues. Deux types de modalités de questionnement sont présentés dans les sections suivantes : les Questions à choix multiple (QCM) et les Questions à réponse ouverte (QRO).

2.3 Question à Choix Multiple

Une Question à choix multiple (QCM) se compose d'une consigne et d'une amorce accompagnée d'un certain nombre de propositions de réponses (Leclercq, 1986, p. 13). Dans le cas d'une QCM à réponse unique, une seule des propositions est la réponse correcte. Si plusieurs des réponses proposées sont correctes, on parle alors de QCM à réponse multiple. Une QCM est une modalité de questionnement « fermée », car l'élève doit choisir sa réponse parmi les propositions fournies. Pour des informations plus détaillées, on peut se référer aux travaux de Leclercq (1986), Leclercq et Gilles (1995) ainsi qu'à ceux de Gilles (2002).

Dans le contexte académique, les QCM sont fréquemment utilisés comme méthode d'évaluation. En effet, leur correction peut être automatisée grâce à des logiciels, ce qui permet un gain de temps considérable. De plus, la détermination de la réponse correcte est objective, évitant ainsi les biais potentiels du correcteur (comme abordé dans la section 2.4). En outre, les QCM exigent moins de temps pour répondre, car il n'est pas nécessaire de rédiger une réponse détaillée comme dans le cas des Questions à réponse ouverte (QRO) (comme expliqué dans la section 2.4). Cela signifie qu'il est possible de poser

[1] Plan d'étude DGEP, École de Culture Générale, https://www.vd.ch/themes/formation/formations-gymnasiales/ecole-de-culture-generale

un plus grand nombre de questions et, par conséquent, d'évaluer l'élève sur un plus grand nombre d'objectifs d'apprentissage.

Pourtant les QCM comportent un certain nombre d'écueils (Gilles, 2002, p. 29) :

- **Le piège de la parcellisation des connaissances.** Dans le cas où les tests comportent de nombreuses QCM, il est crucial de maintenir une certaine cohérence entre les questions, ce qui peut parfois être négligé. L'absence de lien entre les questions pourrait en effet encourager l'apprentissage basé sur la simple mémorisation.
- **Le danger de la mémorisation des réponses incorrectes.** Les élèves ont souvent accès aux anciens examens, mais pas nécessairement à la correction officielle. Lorsque j'étais étudiant, nous nous organisions en groupes pour retenir différentes questions de l'examen. À la fin de l'examen, nous nous réunissions et partagions les questions que nous avions mémorisées, afin de les transmettre aux étudiants de l'année suivante. Cela était une pratique courante, parfois accompagnée des réponses que les élèves-correcteurs avaient fournies. Il est important de noter que cela ne relève pas de la fraude et les enseignants ne pourraient empêcher ce genre d'initiatives. Cependant, dans le cas d'une évaluation utilisant des QCM, il devient problématique qu'une réponse incorrecte soit perpétuée de cette manière. En fournissant une correction officielle, l'enseignant peut prévenir ce risque.
- **Le risque élevé de fraudes.** Il peut être relativement facile de commettre de la fraude lors d'un test utilisant des QCM. En effet, il suffit de mémoriser une séquence de nombres (ou de lettres) correspondant aux réponses. De plus, copier une réponse à une QCM sur un camarade est plus simple que pour une Question à réponse ouverte (QRO). En effet, il est aussi plus facile pour le correcteur de détecter les tentatives de fraude pour une QRO. Par exemple, si deux dissertations sont identiques, cela devient évident qu'il y a eu fraude. Pour minimiser les risques de fraude, il est possible de modifier l'ordre des questions et des réponses dans chaque évaluation qui fait appel aux QCM.
- **Les QCM ne permettent pas de mesurer tous les types de performances.** Les Questions à choix multiple (QCM) ont une limite en ce qui concerne l'évaluation de certains éléments de la taxonomie d'Anderson. Elles sont efficaces pour évaluer le processus cognitif de se souvenir de connaissances factuelles. Cependant, elles semblent moins adaptées pour vérifier d'autres processus cognitifs. Bien sûr, certains QCM peuvent nécessiter la mobilisation de plusieurs processus cognitifs, par exemple, appliquer une méthode de résolution d'équations linéaires pour choisir la réponse correcte parmi les options proposées. Néanmoins, contrairement aux Questions à réponse ouverte (QRO) (comme expliqué dans la section 2.4), une QCM ne permet pas de vérifier le raisonnement de l'élève ni le choix hasardeux (ou non) de sa réponse. Les évaluations utilisant la technique des Degrés de certitude (DC) offrent la possibilité d'élargir le champ de l'évaluation (comme expliqué dans la section 3).

2.4 Question à Réponse Ouverte

Les évaluations utilisant des Questions à réponse ouverte (QRO) sont le plus couramment employées à l'école primaire et au niveau secondaire. Elles permettent à l'enseignant d'évaluer un éventail plus large de niveaux cognitifs que les QCM. Elles sont dites « ouvertes » car elles offrent aux élèves la possibilité de développer leurs réponses. Cela facilite la vérification de la capacité à appliquer une connaissance procédurale, par exemple, en résolvant une équation et en présentant toutes les étapes de la résolution dans sa réponse. Les réponses ouvertes peuvent varier en longueur. On parle de Questions à Réponse ouverte courte (QROC) lorsque la réponse est un nombre ou consiste en quelques mots. Si une réponse nécessite plusieurs calculs ou quelques phrases, on parle de Questions à réponse ouverte moyenne (QROM). Ces types de questions sont fréquents en mathématiques à l'école. Enfin, les Questions à réponse ouverte longue (QROL) exigent des réponses plus élaborées, comme la démonstration d'un théorème ou une dissertation.

Les QRO présentent également un certain nombre d'écueils (Gilles, 2002, p. 23) :

- **Le manque de concordance intra et inter-correcteurs dans la correction des réponses ouvertes.** Dans le milieu académique, surtout lorsqu'il s'agit de corriger un grand nombre de copies, plusieurs correcteurs sont souvent impliqués dans le processus de correction. Cependant, cette approche peut introduire un biais inter-correcteur. Agazzi (1967) a identifié et expliqué ce biais en utilisant la notion de proportion de consensus entre six correcteurs. Il a examiné l'accord ou le désaccord entre ces correcteurs quant à l'admission ou au refus des candidats à l'examen du baccalauréat dans différentes matières. Les résultats de cette analyse pour les domaines des mathématiques et de la philosophie sont présentés dans le Tableau 1. Les résultats révèlent que, dans tous les cas, un consensus absolu n'est pas atteint entre les correcteurs. Il semble que les mathématiques soient moins touchées par le biais inter-correcteur que la philosophie, probablement parce que la matière est moins sujette aux interprétations subjectives des correcteurs et peut-être aussi en raison de l'utilisation plus répandue et aisée d'échelles descriptives en évaluation pour les mathématiques.

Tableau 1

Pourcentage de consensus.

Domaine	Tous les correcteurs ont admis ou refusé
Mathématique	64%
Philosophie	19%

Note. Pourcentage de consensus (de six correcteurs) quant à l'admission ou le refus à l'examen du baccalauréat (Agazzi, 1967)

En ce qui concerne les biais intra-correcteurs, Leclercq (1986) identifie trois catégories distinctes : ceux attribuables uniquement au correcteur lui-même (tels que la moyenne et la variance des notes qu'il attribue), ceux issus des interactions entre l'enseignant et l'élève (tels que l'effet de halo ou les stéréotypes) et ceux résultant des séries de copies à corriger (par exemple, si une copie témoignant d'une performance moyenne est corrigée juste après une copie de haute qualité, le correcteur risque de sous-évaluer cette copie moyenne).

- **Le manque de validité.** Contrairement aux QCM, les évaluations basées sur des QRO comprennent généralement moins de questions, car leur correction nécessite plus de temps. Les QRO peuvent donc être restrictives pour vérifier des connaissances dans des domaines plus vastes, là où des QCM pourraient être posées en nombre et plus appropriées pour évaluer un large contenu.
- **Le manque de sensibilité des mesures qui ignorent les états de connaissances partielles.** Comme pour les QCM, les QRO présentent également des limites en ce qui concerne l'évaluation de certains niveaux cognitifs. Les connaissances métacognitives, par exemple, sont également difficiles à évaluer systématiquement au travers de ce format. Cependant, dans la section suivante de ce travail (section 3), nous explorerons comment la technique des Degrés de Certitude, bien qu'initialement plus adaptés aux QCM, pourraient constituer une solution pour combler cette lacune.
- **Le manque d'équité des épreuves traditionnelles, en particulier les oraux.** Lors des épreuves traditionnelles, notamment les épreuves orales, les élèves sont souvent amenés à tirer au hasard une QRO. Cependant, cette approche peut engendrer plusieurs injustices. Tout d'abord, les niveaux de difficulté des questions ne sont pas nécessairement uniformes d'une épreuve à l'autre. De plus, un élève qui aurait étudié intensivement un seul des sujets d'examen et qui, par chance, tirerait une question portant sur ce même sujet, serait injustement récompensé pour l'ensemble du cours.

2.5 L'évaluation en mathématique

Les évaluations en mathématiques font fréquemment usage des QRO. En effet, l'un des objectifs fondamentaux des cours de mathématiques, notamment dans le contexte de l'école de culture générale, est d'« *apprendre à expliciter sa démarche en utilisant un vocabulaire adéquat* »[2]. Par conséquent, les élèves sont encouragés à présenter leurs calculs de manière correcte et précise. C'est pourquoi il est courant que les élèves détaillent leurs étapes de résolution.

Dans ce mémoire de Master of Advanced Studies pour le degré secondaire II, les évaluations ont été conçues exclusivement avec des QRO dont principalement des QROM (cf. section 5). En effet, la plupart des questions posées nécessitaient plusieurs étapes de résolution, incluse dans la réponse at-

[2] Plan d'étude DGEP, École de Culture Générale, https://www.vd.ch/themes/formation/formations-gymnasiales/ecole-de-culture-generale

tendue. Par conséquent, une réponse finale correcte ne garantit pas automatiquement l'obtention de tous les points pour l'exercice, et de même, une réponse finale incorrecte ne se traduit pas nécessairement par un score nul. Cela dépendra de la manière dont l'exercice est résolu et des détails inclus dans la réponse, qui sont indirectement liés à sa longueur.

3 Évaluation avec des Degrés de certitudes

L'idée de la technique des Degrés de certitude (DC) est d'amener l'élève à fournir une indication de la certitude associée à une réponse donnée. Les premières utilisations des DC remontent au début du siècle dernier, lorsqu'ils ont été employés pour accompagner les prévisions météorologiques (Cooke, 1906). Dans le domaine de l'éducation, DeFinetti (1965) est un précurseur dans l'utilisation des DC en évaluation, bien que leur adoption généralisée n'ait véritablement commencé qu'à partir des années 80, avec la démocratisation des outils informatiques qui a facilité leur intégration dans l'enseignement (Leclercq, 1982, 1987 ; Gilles, 1996 ; Gardner-Medwin et Gahan, 2003 ; Gardner-Medwin, 2006).

L'utilisation des DC dans l'évaluation n'est pas répandue, mais elle n'est pas non plus marginale. L'University College de Londres (UCL) utilise cette méthode pour certains examens depuis 2001 (Gardner-Medwin, 2006). L'université de Hasselt (UHasselt) en Belgique a systématisé l'usage des degrés de certitude depuis plus d'une décennie dans sa filière de Bachelor of Medicine et exporte cette technique dans d'autres facultés et universités dans le cadre de projets de coopération en enseignement supérieur (Rigo, El Jaafari et Gilles, 2020). Les DC sont aussi utilisés dans des épreuves de divers domaines à l'université de Liège (Gilles, 1996), en médecine d'urgence (Houzé-Cerfon et al., 2016), en didactique des mathématiques (Deruaz et Buenzli, 2015), en mathématiques (Foster, 2015 ; Caspari-Sadeghi et al., 2022) et en physique (Clark, 2020).

L'ajout d'une certitude à une réponse permet de faire des distinctions plus précises entre les niveaux de connaissance. Par exemple, une réponse correcte accompagnée d'un degré de certitude faible traduira une connaissance partielle tandis qu'on parlera de connaissance assurée lorsque cette réponse correcte est accompagnée d'une certitude élevée. Les mathématiques, grâce à leur nature vérifiable, semblent particulièrement appropriées pour l'utilisation des DC (section 3.1).

Shufford, Albert et Massengil (1966) ont montré que l'intégration des DC dans l'évaluation requiert une méthodologie spécifique : l'évaluation de la certitude se fait en fonction d'une échelle préétablie ; l'attribution des points suit un barème défini conforme à la théorie des décisions; les élèves ont besoin d'entraînement pour maîtriser cette approche et et des indices de performance d'autoévaluation tels que le réalisme doivent être fournis lors des feedbacks (section 3.2).

La Centration du sujet (Cs) et le Réalisme du sujet (Rs) sont des indices de performance métacognitive couramment utilisés lors d'évaluations avec la technique des DC. Les mesures de la Cs et du Rs

proposées par Gilles (2002), ainsi que la mesure probabiliste du réalisme développée par Prosperi (2015), sont présentées. Cependant, il est important de noter que la mesure de Prosperi tend à surestimer le réalisme pour les tests courts. Pour pallier cette surestimation, une version modifiée de la mesure est proposée. Enfin, en considérant le DC comme une estimation (auto-évaluation) du nombre de points obtenus, la mesure du réalisme est adaptée aux QRO (section 3.3).

3.1 Les principes

Le principe des évaluations utilisant la technique des Degrés de certitude (DC) repose sur l'évaluation préalable de la justesse de la réponse par l'élève lui-même. Cela permet de différencier plus finement l'acquisition des connaissances de l'élève. En considérant les DC, l'évaluation des connaissances se structure autour de deux axes, comme illustré dans le Tableau 2. L'axe vertical représente l'acquisition de l'objectif (la connaissance ou l'ignorance), tandis que l'axe horizontal indique l'évaluation de la connaissance subjective (reconnue ou ignorée) par rapport à la norme d'apprentissage.

Tableau 2

Les types de connaissances et ignorances subjectives.

	Reconnue	**Ignorée**
Connaissance	CR	CI
Ignorance	IR	II

Dans les évaluations traditionnelles, il n'y a pas de distinction claire entre les Connaissances reconnues (CR) et les Connaissances ignorées (CI). Par exemple, une CI pourrait résulter d'une réponse correcte mais donnée au hasard par l'élève à une QCM. De la même manière, l'Ignorance reconnue (IR), où l'élève sait qu'il ne sait pas, par exemple lorsqu'il accompagne une réponse incorrecte d'un degré de certitude faible, est souvent confondue avec l'Ignorance ignorée (II), où l'apprenant pense savoir mais se trompe (la réponse qu'il a accompagné d'un degré de certitude élevé s'avère incorrecte). Les DC permettent de faire ces distinctions et d'apporter des nuances plus précises lors des rétroactions issues des évaluations des connaissances des élèves.

Jans et Leclercq (1999) introduisent encore plus de nuances de ce type qu'ils situent sur un spectre partant à une extrémité de la pire des situations – *une réponse incorrecte accompagnée de la certitude la plus élevé*e – jusqu'à arriver à l'autre extrémité à la performance la meilleure – *une réponse correcte accompagnée de la certitude la plus élevée*. Ces auteurs parlent à ce propos de « connaissances spectrales ».

Selon Schraw et al. (2013), le Tableau 2 pourrait être analogue à un tableau de probabilités dans le contexte d'un test, où les CR correspondraient à de vrais positifs et les II à de faux positifs. Si l'on assimile la réponse à une question au résultat d'un test, il serait possible de décrire les connaissances de

l'élève en utilisant les concepts de *sensibilité* et de *spécificité*, similaires à ceux utilisés dans les tests médicaux.

3.1.1 L'importance de l'ignorance reconnue

Offrir aux apprenants la possibilité de pratiquer et d'améliorer leurs connaissances métacognitives est extrêmement bénéfique. En effet, dans certaines situations, une Ignorance ignorée (II) peut avoir des conséquences graves. Par exemple, l'affaire d'Amanda Knox[3] en novembre 2007, où elle a été injustement condamnée pour meurtre et violences sexuelles en raison d'une évaluation erronée d'un juge. Ce dernier n'a pas jugé nécessaire de procéder à un deuxième test ADN. L'Ignorance ignorée (II) du juge envers les éléments essentiels dans cette affaire a eu des conséquences dramatiques pour Amanda Knox, qui a finalement été acquittée en octobre 2011.

En revanche, l'ignorance reconnue (IR) n'est pas inutile. Par exemple, un médecin incertain de la posologie d'un traitement cherchera à vérifier les informations auprès d'un collègue expert avant d'administrer le traitement. Hunt (1977) a mis en évidence l'importance d'introduire la technique des degrés de certitude pour la maîtrise des connaissances métacognitives et le caractère crucial de ces dernières sur les plans de la réussite professionnelle des élèves et leur fonctionnement en société.

3.1.2 Évaluation plus fine et retour plus constructif

L'évaluation des connaissances subjectives offre une précision accrue en ce qui concerne l'acquisition des normes d'apprentissage. Elle permet ainsi d'établir une hiérarchie dans les connaissances et de porter une attention particulière aux connaissances ignorées (CI). Cette approche permet de faire la distinction entre un élève qui excelle grâce à une compréhension approfondie des sujets et un autre qui aurait réussi par chance. En effet, avec l'utilisation des Degrés de certitude (DC), ce dernier aurait probablement indiqué son incertitude sur certaines questions. Cette finesse dans l'évaluation est précieuse pour les enseignants, car elle leur permet de mieux répondre aux besoins variés de leurs élèves.

3.1.3 La certitude en mathématique

Nous avons vu que les mathématiques ont une préférence pour les Questions à réponse ouverte (QRO). Néanmoins, ces dernières sont moins adaptées à l'évaluation avec des Degrés de certitude (DC). En effet, les DC sont définis pour une interprétation binaire de la réponse (correcte ou incorrecte). Dans le cas des QRO, en particulier des QROM et QROL, les réponses pourraient être partiellement justes. Néanmoins, par leur essence, les mathématiques se prêtent bien pour évaluer la certitude d'une réponse. En effet, les réponses en mathématiques sont vérifiables (Deruaz et Buenzli, 2015). Par exemple, après avoir résolu une équation, il est d'usage de vérifier sa réponse en la substituant dans l'équation initiale. Il s'agit, par ailleurs, d'une étape importante en mathématique où un mathématicien

[3] "*Amanda Knox and bad maths in court*", article de BBC, https://www.bbc.com/news/magazine-22310186

est régulièrement amené à vérifier ses résultats. Cela se fait, naturellement, en vérifiant les calculs, mais également en utilisant d'autres méthodes de résolution. Cette dernière façon de vérifier est très instructive en mathématique.

3.2 Procédures

L'utilisation des DC dans l'évaluation nécessite une méthodologie spécifique, comprenant la définition d'une échelle de certitude, l'établissement d'un barème de notation et la mise en place d'un entraînement avec un retour d'information, dont le réalisme.

3.2.1 Échelle de certitude

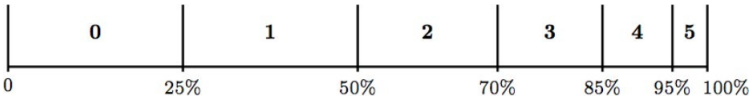
Il existe plusieurs études ayant conduit à la création de différents types d'échelles permettant à l'élève d'exprimer sa certitude (Gilles, 2002). Certaines échelles se basent sur des expressions reflétant la certitude (ou l'incertitude) de la réponse (Van Naerssen et Van Beaumonz, 1965 ; Jacobs, 1971 ; Leclercq, 1963 ; Gardner-Medwin et Gahan, 2003) par exemple:

- Sûr, Pas sûr.
- Sûr, Moyennement sûr, Pas sûr.

Cependant, selon les travaux de Leclercq (2016) et Gilles (1996), il est conseillé d'éviter les échelles utilisant des énoncés vagues et subjectifs. Ainsi, l'utilisation d'une échelle probabiliste est recommandée. Cette échelle peut être uniforme (Leclercq, 1963), non-uniforme (Edwards, 1967 ; Gilles, 2002) ou continue (DeFinetti, 1965). Dans le cadre de notre recherche, l'échelle probabiliste non-uniforme proposée par Gilles (2002) sera utilisée. La Figure 1 illustre les intervalles de pourcentages de certitude que représente chaque DC.

Figure 1

Échelle de certitude



Note. Pourcentage de certitude par DC (Gilles, 2002).

L'échelle des DC présentée dans la Figure 1 n'est pas uniforme. Elle devient plus fine à mesure que le DC augmente. Cette particularité découle du fait qu'on s'intéresse ici plus à la façon dont les élèves peuvent nuancer leurs certitudes que leurs incertitudes. Par ailleurs, étant donné que l'évaluation intervient après une séquence d'enseignement, on peut supposer que les élèves ont une meilleure certitude en général, ce qui justifie le besoin de plus de précision pour les pourcentages de certitude élevés.

D'un point de vue plus mathématique, soit NC_i (respectivement NU_i) le nombre de réponses correctes (respectivement d'utilisation) pour le DC de valeur $i=0,\ldots,5$. Alors, dans le cas d'une utilisation idéale des DC, NC_i suit une distribution binomiale $NC_i \sim B(NU_i, p_i)$ où p_i est une probabilité comprise dans l'intervalle de certitude correspondant au DC de valeur i (cf. Figure 1).

3.2.2 Barème des DC

La Figure 2 montre le barème proposé par Leclercq (1982) et Gilles (2002), qui est en accord avec la théorie des décisions (Leclercq, 1982 ; Gardner-Medwin et Gahan, 2003). En effet, pour maximiser ses points, un élève a intérêt à répondre de manière honnête, c'est à dire en faisant preuve de réalisme, sans surestimation ni sous-estimation. Pour être totalement réaliste, le taux de réponses correctes pour un certain DC doit correspondre à la valeur centrale de l'intervalle des pourcentages liés à la probabilité accordée par l'élève de donner la réponse correcte (cf. Figure 1).

Figure 2

Barème des évaluations avec DC.

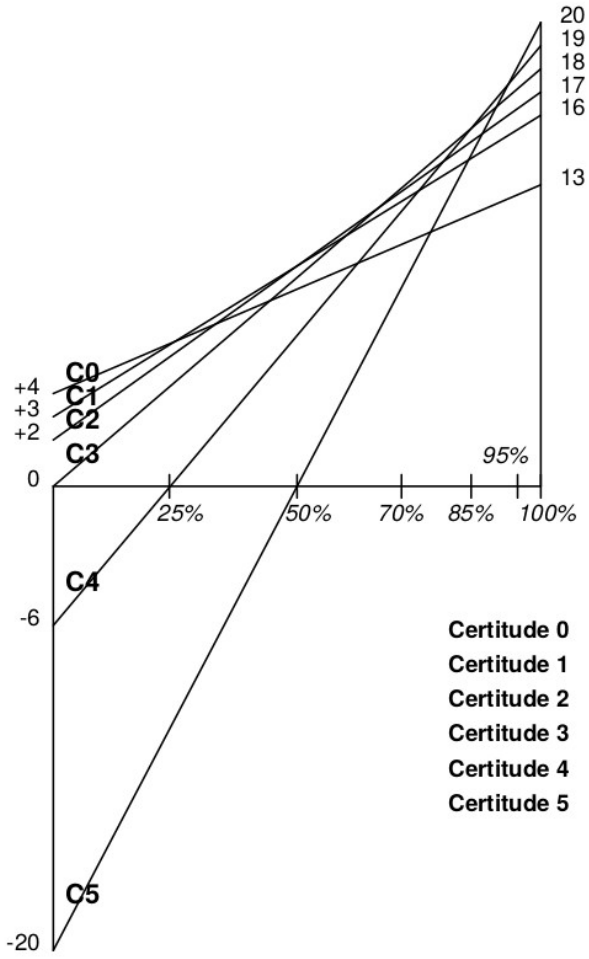
		Degrés de certitude					
		0	1	2	3	4	5
Réponse	Correcte	+13	+16	+17	+18	+19	+20
	Incorrecte	+4	+3	+2	0	-6	-20

Source. (Gilles, 2002)

La Figure 3 représente le barème des points. On remarque que ce barême est effectivement conforme à la théorie des décisions. Entre autres, la droite représentant le DC de valeur i est supérieure aux autres sur l'intervalle correspondant au degré de certitude i. Par exemple, la droite C3 est supérieure aux autres dans l'intervalle [50%, 70%].

Figure 3

Droites de certitude et conformité du barème à la théorie des décisions



Source. (Leclercqs et Gilles, 1994)

3.2.3 Entraînement

Pour garantir des résultats cohérents, il est nécessaire d'organiser des sessions d'entraînement à l'utilisation des DC. Les recherches menées par Leclercq et Gilles (1994) à l'aide d'une application web ont révélé que le réalisme (une mesure de la capacité de l'élève à s'autoévaluer correctement, qui sera abordée dans la section 3.3.4) varie avant de se stabiliser. De plus, les études de Miller et Geraci (2011) ainsi que de Callender et al. (2015) démontrent que les élèves ayant des performances moins élevées sont davantage susceptibles d'améliorer leur auto-évaluation grâce à des retours constructifs par rapport à ceux qui n'ont pas reçu de retours.

3.3 Analyse spectrale

Cette section débute en définissant le concept du Seuil d'acceptation d'une réponse (S_a) pour le calcul du score dans un test utilisant des DC avec des QRO. L'objectif est de définir un seuil au-delà duquel une réponse sera considérée comme correcte et en deçà duquel elle sera considérée comme incorrecte. De plus, une nouvelle méthode de calcul du score est proposée, ne nécessitant pas de seuil d'acceptation. Les notions de jugement, d'erreur de certitude et de centration sont également introduites. Les mesures du Réalisme (Rs) de Gilles (2002) et Prosperi (2015) sont présentées, et la mesure de Prosperi est ajustée pour s'adapter aux données limitées des tests courts. En fin de compte, une mesure du réalisme utilisant la centration est exposée. Ces différentes mesures permettent une meilleure interprétation des résultats d'un test avec des DC, tant pour les élèves que pour les enseignants.

Afin de présenter ces mesures, un langage mathématique semble être le plus approprié. Voici une liste de définitions et de notations pour certaines variables utilisées dans ce travail. Ces définitions ne seront pas nécessairement répétées dans le reste du texte.

- N_q: le nombre total de questions.
- P_n: le nombre de points qu'il est possible d'obtenir au maximum pour la question n.
- $\overline{P}_n$: le nombre de points obtenus pour la question n.
- P_{tot}: le nombre de points au total ($\sum_{n=1}^{N_q} P_n$).
- $\overline{P}_{tot}$: le nombre de points obtenus au total ($\sum_{n=1}^{N_q} \overline{P}_n$).
- P_i^+: le nombre de points obtenus si la réponse est correcte pour le DC de valeur i.
- P_i^-: le nombre de points obtenus si la réponse n'est pas correcte pour le DC de valeur i.
- $I_i = \left[I_i^1, I_i^2\right]$: l'intervalle du DC de valeur i (par exemple $I_3 = [0.7, 0.85]$).
- VC_i: la valeur centrale des intervalles de certitude I_i ($VC_i = \left(I_i^1 + I_i^2\right)/2$).
- DC_n: le DC choisi pour la question n.
- NU_i: le nombre d'utilisations du DC de valeur i.
- NC_i: le nombre de réponses correctes lors de l'utilisation du DC de valeur i.
- TE_i: le taux d'exactitude pour le DC de valeur i (NC_i / NU_i).
- TEQ_n: le taux d'exactitude à la question n ($\overline{P}_n / P_n$).
- TEG: le taux d'exactitude général ou le score ($\overline{P}_{tot} / P_{tot}$).

3.3.1 Adaptation aux QRO

L'utilisation des DC avec des QRO nécessite des réflexions et des adaptations particulières. En effet, dans de nombreux cas, les QRO attribuent plusieurs points aux réponses, ce qui rend difficile la catégorisation binaire (correcte ou incorrecte). Pour permettre malgré tout cette catégorisation binaire, un seuil d'acceptation de la réponse peut être établi. Par exemple, si un élève obtient 3.5 points sur 4 pour sa réponse, faut-il la considérer comme correcte ou incorrecte ? Un seuil d'acceptation de 0.8 indiquerait que la réponse est correcte car le taux d'exactitude (87.5 %) est supérieur au seuil d'acceptation de 80 %. Les points attribués à la question seraient ensuite calculés en utilisant le barème de la Figure 2, multiplié par le nombre total de points possibles pour la question (4 dans ce cas). La formalisation de ce seuil d'acceptation est présentée dans l'équation (2). Les implications de ce seuil sur les différentes mesures des évaluations avec DC sont explorées dans les sections 5.3.1 et 5.3.2.

Cependant, le choix de ce seuil d'acceptation peut sembler arbitraire. Une approche alternative consiste à considérer que, étant donné la nature non binaire de l'évaluation de l'exactitude d'une réponse à une QRO, l'interprétation des DC est modifiée. De manière intuitive, les DC peuvent être perçus comme représentant le taux d'exactitude qu'un élève estime avoir (auto-évaluation). Cette perspective permet d'éviter la nécessité d'utiliser un seuil d'acceptation. En effet, le taux d'exactitude d'une réponse se superpose à l'axe des abscisses de la Figure 3, qui représente la certitude. Cette approche est formalisée dans l'équation (4).

3.3.2 Score et gradation

Le score à un test est une valeur entre 0 et 1 qui mesure l'exactitude des réponses. En général, le score standard vaut

$$Score_{std} = TEG \tag{1}$$

mais ce dernier ne prend pas en compte les DC. En considérant un seuil d'acceptation, S_a, le nombre de points obtenus par question est donné par

$$\bar{P}_n^{S_a} = \begin{cases} P_{DC_n}^- & \text{si } TEQ_n < S_a, \\ P_{DC_n}^+ & \text{sinon.} \end{cases} \tag{2}$$

Ainsi, le score obtenu avec DC et avec un seuil d'acceptation est donné par

$$Score_{DC}^{S_a} = \frac{1}{P_{tot}} \sum_{n=1}^{N_q} \frac{\bar{P}_n^{S_a}}{20} \cdot P_n. \tag{3}$$

Dans les cas de QRO, les DC sont intuitivement perçus comme le taux d'exactitude de la réponse. Dans ce cas, nous pourrions considérer l'axe horizontal du graphique de la Figure 3 comme étant le taux d'exactitude à une QRO. Alors, le nombre de points attribués par question est défini par la valeur de la

droite de certitude choisie. Cette méthode à l'avantage de ne pas utiliser un seuil d'acceptation. Considérons $C_i(x)$ la fonction représentant la droite de certitude i (cf. Figure 3). On a

$$C_i(x)=\left(P_i^+ - P_i^-\right)x+P_i^- \tag{4}$$

Alors le nombre de points obtenus est donné par

$$\overline{P}_{DC}=\sum_{n=1}^{N_q}\frac{C_{DC_n}\left(TEQ_n\right)}{20}\cdot P_n.$$

Ainsi, le score avec DC et sans seuil d'acceptation est donné par

$$Score_{DC}=\frac{\overline{P}_{DC}}{P_{tot}} \tag{5}$$

La gradation, ou note, attribuée à un score est donnée par une transformation affine qui fait correspondre la valeur du score à une échelle entre 1 et 6 arrondie à le demie la plus proche. C'est-à-dire

$$Note=Arrondie(5\cdot Score+1,0.5)$$

3.3.3 Centration des sujets (Cs)

Seulement trois DC (bas, moyen, haut) sont utilisés par Gardner-Medwin et Gahan (2003), et cette approche est reprise dans la plupart des recherches anglophones (Miller et Geraci, 2011 ; Callender et al., 2015 ; Clark, 2020 ; Davies, 2002 ; Caspari-Sadeghi et al., 2022). Ces auteurs définissent le *jugement* comme l'estimation que l'élève fait de son Taux d'exactitude général (TEG). En d'autres termes, le jugement, $\widehat{TEG}$, représente une auto-évaluation du taux d'exactitude général. Gardner-Medwin et Gahan (2003), ainsi que Gardner-Medwin (2006), comparent le TEG avec le score obtenu en utilisant la technique des DC. Ils utilisent un barème conforme à la théorie des décisions pour les intervalles de certitude ([0, 67 %] ; [67 %, 80 %] ; [80 %, 100 %]). Miller et Geraci (2011) ainsi que Callender et al. (2015) demandent également une auto-évaluation générale. Callender et al. (2015) demandent une estimation du taux d'exactitude général à la fin de l'épreuve, tandis que Miller et Geraci (2011) utilisent une note. Ils comparent ensuite la différence entre cette auto-évaluation et le score obtenu au test. Dans le contexte de ce mémoire, le jugement de l'élève est estimé par rapport à l'échelle des certitudes présentée dans la Figure 1. Ainsi, le jugement de l'élève peut être défini de différentes manières

$$\widehat{TEG}_b=\frac{1}{P_{tot}}\sum_{n=1}^{N_q}P_n\cdot b\left(I_{DC_n}\right) où\, b(\cdot)=\sup(\cdot), moy(\cdot), inf(\cdot). \tag{6}$$

La Centration du sujet (Cs), de l'élève est alors donnée par

$$Cs_b=\widehat{TEG}_b - TEG$$

En utilisant $\widehat{TEG_b}=\widehat{TEG_{moy}}$, on obtient la définition de la Centration du sujet (Cs) donnée par Gilles (2002, p. 277). Les estimateurs proposés par l'équation (6) permettent de définir un intervalle du jugement $\left[\widehat{TEG_{inf}}, \widehat{TEG_{\sup}}\right]$de l'élève. La centration (Cs) est alors donnée par la distance relative de l'intervalle du jugement et le TEG. On a

$$Cs_{DC}=dist_{rel}\left(\left[\widehat{TEG_{inf}}, \widehat{TEG_{\sup}}\right], TEG\right). \tag{7}$$

La centration est une mesure de la surestimation, ou sous-estimation, d'un apprenant quant à l'auto-évaluation de ses connaissances. Un élève qui aurait tendance à répondre juste, mais de façon peu sûre aura une centration plus basse qu'un élève qui répondrait en général faux avec un DC élevé. La valeur de la Centration (Cs) varie entre -1 et 1 avec une valeur idéale en zéro. Ainsi, les auto estimations de l'élève comportent peu d'erreurs si la valeur de la Centration du sujet (Cs) est proche de zéro.

3.3.4 Réalisme des sujets (Rs)

Le Réalisme des sujets (Rs) est une mesure de la connaissance métacognitive de l'élève concernant l'auto- évaluation. Sa définition est proche des erreurs de certitude abordée plus haut. La mesure du Réalisme des sujets (Rs) de Gilles (2002, p. 272) se base sur le calcul de la Moyenne des Erreurs de Certitude (*MEC*)

$$MEC=\frac{1}{N_q}\sum_{i=0}^{5}\left|TE_i-VC_i\right|\cdot NU_i$$

qui consiste au calcul de la moyenne pondérée de l'erreur de certitude absolue par DC. En effet, sans les valeurs absolues, nous obtiendrions la Centration des sujet (CS). La moyenne des erreurs de certitude permet donc d'observer la variation de l'erreur de certitude par DC. En effet, si un élève a tendance à se sous-estimer lorsqu'il connaît la réponse et se surestimer s'il ne la connaît pas ("hard-easy effect" (Juslin et al., 2000)), cette différence de comportement pourrait ne pas être observable lors du calcul de la Centration (Cs) contrairement à la *MEC*. La mesure du réalisme des sujets est alors donnée par

$$Rs=\left(\left(1-MEC\right)-\beta\right)\cdot\alpha \tag{8}$$

avec $\alpha=\frac{20}{19}$ et $\beta=0.025$. Ces deux facteurs sont utiles pour que la valeur du réalisme soit comprise en 0 et 1. La différence avec 1 permet d'avoir un réalisme proche de 1 si l'élève s'autoévalue bien. Néanmoins, cette mesure comporte certains désavantages discutés par Prosperi (2015). Entre autres, si un élève à un taux d'exactitude contenu dans le bon intervalle (cf. Figure 1), mais différent de la valeur centrale correspondante, alors il ne devrait pas y avoir une erreur d'appréciation. Prosperi (2015) propose une approche probabiliste. Il considère le nombre de réponses correctes pour un DC comme une variable aléatoire suivant une distribution binomiale $NC_i \sim B\left(NU_i, p_i\right)$ où p_i représente la probabilité de ré-

pondre juste avec un DC de i. Dans l'idéal cette valeur s'approche de VC_i. Néanmoins, puisque les élèves ne sont pas idéaux, nous cherchons à savoir si l'estimation de leur probabilité de réussite correspond au choix de l'intervalle de certitude. Il est proposé d'utiliser la méthode de Wilson (1927) pour le calcul d'un intervalle de confiance à 90% pour la probabilité p_i noté $\hat{I}_i = \left[\hat{I}_i^1, \hat{I}_i^2\right]$. En effet, il s'agit d'un bon estimateur pour des données peu nombreuses. L'erreur du réalisme pour le DC de valeur i, err_i, est donnée par la plus courte distance entre l'intervalle $\hat{I}_i$ et l'intervalle du DC correspondant I_i. Ainsi, le réalisme d'un élève est donné par

$$Rs = 0.95 - \frac{1}{N_q}\sum_{i=0}^{5} err_i \cdot NU_i. \tag{9}$$

avec

$$err_i = dist\left(\hat{I}_i, I_i\right)$$

où dist(*A, B*) est la distance absolue entre les intervalles *A* et *B*. La soustraction de 0.95 permet d'avoir un réalisme compris entre 0 et 1 où un réalisme de 1 signifie que l'élève évalue bien sa certitude. L'erreur est également pondérée par le nombre d'utilisations du DC.

3.3.5 Adaptation de la mesure du réalisme

L'échantillon utilisé pour ce travail est composé de tests relativement courts. Certains comportent seulement cinq questions. De ce fait, il y a peu d'utilisations des différents DC (trop souvent $NU_i = 1$ pour certains i). Ainsi l'intervalle de confiance de Wilson est très large et donc le réalisme est très élevé. On observe, alors, peu de variations dans les résultats. Dans ce mémoire, il est donc proposé de considérer comme intervalle de confiance celle des pourcentages de certitude correspondant au choix du DC et centré autour de l'estimateur $\widehat{p_i^{S_a}} = NC_i / NU_i$. C'est-à-dire

$$\widehat{I_i^{S_a}} = \left[\widehat{p_i^{S_a}} - \Delta_i, \widehat{p_i^{S_a}} + \Delta_i\right] avec\ \Delta_i = \frac{1}{2}\left(I_i^2 - I_i^1\right). \tag{10}$$

En remplaçant $\hat{I}_i$ par $\widehat{I_i^{S_a}}$, une nouvelle mesure est définie. La mesure du réalisme utilisant l'intervalle donné par (10) est nommée *mesure de Prosperi modifié discrète,* car il se base sur un estimateur discret de p_i. On remarque (cf. Figure 1) que lors d'un choix de DC bas l'intervalle de confiance est plus grand que pour un DC élevé. Ainsi, les erreurs d'appréciation pour un DC bas sont moins pénalisées. En effet, cette mesure suppose que la variance de la certitude est équivalente au DC.

L'estimateur, $\widehat{p_i^{S_a}}$, utilisé dans la mesure de Prosperi modifiée discrète dépend du seuil d'acceptation et sa précision dépendra alors de NU_i. Il peut alors être pertinent de considérer le score obtenu par degré de certitude comme la valeur de l'estimateur de p_i. C'est-à-dire

$$\widehat{p_i^{DC}} = \frac{1}{NU_i} \sum_{n=1}^{N_q} TEQ_n \cdot Id_i\left(DC_n\right) \tag{11}$$

où

$$Id_i(x) = \begin{cases} 1 & \text{si } x = i, \\ 0 & \text{sinon.} \end{cases}$$

En substituant l'estimateur défini par (11) dans la mesure précédente une nouvelle *mesure de Prosperi modifiée continue* est définie. Cette mesure du réalisme à l'avantage de n'utiliser aucun seuil d'acceptation et être moins sensible aux données peu nombreuses.

Finalement, une mesure du réalisme est donnée en utilisant la centration définie par l'équation (7). Ainsi, le R*éalisme (Rs) utilisant la Centration (Rs)* est donné par

$$Rs = 1 - \left|Cs_{DC}\right|. \tag{12}$$

Cette mesure du réalisme à l'avantage d'être facile à calculer. Par ailleurs, le réalisme défini par l'équation (12) n'utilise pas de seuil d'acceptation. Néanmoins, cette mesure a le désavantage de ne pas capturer la variation du jugement à travers les différents niveaux de certitude.

4 Questions de recherche

Dans ce mémoire, mon objectif est d'explorer l'application des DC dans l'évaluation en mathématique de Questions à réponse ouverte (QRO). Je vais comparer diverses mesures du réalisme, en particulier leurs adaptations pour les QRO (section 4.1). J'analyserai également la manière dont le réalisme évolue avec le temps (section 4.2).

4.1 Mesures et seuil d'acceptation

Le retour principal aux élèves concernant leur utilisation des DC consistait à fournir la note standard accompagnée de la note attribuée avec les DC, en appliquant un seuil d'acceptation de 0.8. Une analyse qualitative des diverses méthodes de gradation pourrait aider à déterminer celle qui est la mieux adaptée, à la fois pour fournir des retours utiles concernant les connaissances subjectives et pour évaluer de manière adéquate les connaissances générales de l'élève. Dans la section 3.3.2 sont présentées trois méthodes de gradation utilisant :

- le score standard $Score_{std}$ (cf. équation (1)) ;
- le score avec DC et avec un seuil d'acceptation $Score_{DC}^{S_a}$ (cf. équation (3)) ;
- le score avec DC et sans un seuil d'acceptation $Score_{DC}$ (cf. équation (5)).

La méthode n'utilisant pas de seuil d'acceptation semble être un candidat pratique, car il ne nécessite pas un choix arbitraire.

Dans le cas où la méthode de gradation se base déjà sur les DC. Il est essentiel de fournir d'autres informations que la note concernant les connaissances subjectives de l'élève. Ces informations supplémentaires peuvent être données par la Centration (Cs) et le Réalisme (Rs). Plusieurs mesures du réalisme ont été présentées dans les sections 3.3.4 et 3.3.5 :

- La mesure du Réalisme des sujets (Rs) de Gilles (2002) (cf. équation (8)).
- La mesure de Prosperi (2015) (cf. équation (9)).
- La mesure de Prosperi modifiée discrète (cf. équation (10)).
- La mesure de Prosperi modifiée continue (cf. équation (11)).
- La mesure utilisant la Centration (Cs) (cf. équation (12)).

La mesure utilisant la Centration (Cs) de certitude semble être une mesure pratique. En effet, son calcul est plus simple et elle peut facilement être implémentée à l'aide d'un software tout public comme E*xcel*. Par ailleurs, elle ne dépend pas d'un choix arbitraire du seuil d'acceptation. Je voudrais déterminer si cette dernière reflète correctement le réalisme.

4.2 Le réalisme des élèves s'améliore-t-il avec le temps ?

Plusieurs études ont déjà établi que le réalisme s'améliore avec la pratique (Leclercq et Gilles, 1994). On a observé que le jugement des élèves s'améliore, en particulier ceux ayant obtenu des notes basses, et que ces améliorations du jugement et de la performance sont plus significatives lorsqu'un retour constructif est donné (Callender et al., 2015 ; Miller et Geraci, 2011). Dans cette étude, nous évaluerons la progression du réalisme à travers plusieurs tests. Cependant, il est important de noter que le réalisme dépend également d'autres facteurs, tels que :

- **La difficulté de la question.** La certitude d'une personne varie selon la difficulté de la question. Pour des questions faciles, on a tendance à se sous-estimer (en pensant qu'il y un piège) tandis que pour des questions difficiles la tendance est inversée. Ce phénomène s'appelle l'*effet difficile-facile*, en anglais, *hard-easy effect* (Juslin et al., 2000). Juger de la difficulté d'une question ou d'un test n'est pas évident. La moyenne générale pourrait être un indicateur pour un test. Pour une QCM on se basera sur le pourcentage de réponses correctes et pour une QRO sur le score moyen obtenu pour une question. Par ailleurs, estimer en tant qu'élève la difficulté d'un test est une notion subjective. Elle peut donc être différente pour chaque élève selon son niveau de réalisme. De plus, si l'effet *hard-easy* existe au sein des données, elle pourrait être observée par les mesures de Gilles (2002) et Prosperi (2015). Néanmoins, je n'entreprendrai pas cette analyse après avoir étudié son impact sur les données récoltées (section 5.3.2).

- **Les compétences de l'élève.** Les travaux de Callender et al. (2015) et Miller et Geraci (2011) ont montré que le réalisme s'améliore plus chez les élèves avec des compétences basses que les autres. Ils ont observé que le réalisme ne s'améliore pas chez les élèves avec des performances hautes. Par ailleurs, Gilles et Melon (2000) montrent que le réalisme diminue puis augmente à nouveau pour se stabiliser lors d'entraînements. En catégorisant les élèves selon leur performance, je vérifierai si une progression significative du réalisme existe dans trois catégories de performance (basse, moyenne et haute).
- **La qualité des retours.** Il est montré par Callender et al. (2015) et Miller et Geraci (2011) qu'un retour constructif et détaillé est nécessaire pour permettre aux élèves d'améliorer leur jugement. Néanmoins, ils ont également noté une progression du jugement pour les élèves avec des performances basses n'ayant reçu aucun retour hormis leur score avec les DC. Les données présentées dans ce travail se situent entre ces deux catégories (feedback détaillé et constructif lors d'une correction collective en classe *versus* feedback uniquement basé sur la transmission d'un score DC). En effet, une correction détaillée pour chaque test a été faite en classe. Les élèves ont également reçu une note avec DC. La section 5.2 traitera plus en détail de la méthodologie expérimentale. Nous verrons (section 5.3) que les retours effectués ont été suffisants pour observer une progression du réalisme.
- **Le temps à disposition.** Davies (2002) et Ahlgren (1969) ont montré qu'un test avec des DC requiert un peu moins d'une fois et demie de temps en plus pour la réflexion métacognitive nécessaire pour répondre à une question avec des DC. Dans le cadre de ce travail, ce facteur n'a pas été pris en compte *a priori*. Le temps à disposition pour les tests courts est en général suffisant. Ceci est moins le cas pour les tests longs. Cette différence de temps à disposition pourrait avoir un impact entre le réalisme des tests courts et longs.
- **La nature de l'évaluation sommative ou non.** Il a été noté par Gardner-Medwin et Gahan (2003) que le comportement des élèves change lors des tests sommatifs où les résultats du jugement et de la performance sont meilleurs. Pour ce mémoire, les données ont été récoltées pour trois types de tests : formatif court, semi-sommatif court, sommatif long (section 5.2). Il est donc important d'être vigilant à cet aspect lors de l'analyse des résultats.

Dès lors, dans le cadre de ce travail, je voudrais savoir si les élèves améliorent leur réalisme par rapport au premier test effectué.

5 Partie expérimentale

Des tests avec la technique des Degrés de certitude (DC) ont été effectués dans deux classes (A et B). Dans la section 5.1, les différents types de tests sont exposés ainsi que la procédure de récolte des don-

nées et leur structure. Les différentes méthodes d'analyse des données utiles pour répondre aux questions de recherches sont présentées dans la section 5.2. Finalement, les résultats seront analysés et discutés dans la section 5.3.

5.1 Récolte des données

Des tests avec des DC ont été effectués dans deux classes (nous les intitulerons pour les besoins de cette recherche : classe A et classe B) de première année à l'école de culture générale au gymnase de Burier d'août à décembre 2021. Pour chaque question, les élèves doivent accompagner leur réponse d'un DC en se référant à l'échelle de la Figure 1. De plus, il y a trois types de tests :

- Le test formatif (TF) qui consiste en la résolution de questions nécessitant des calculs de plusieurs étapes. Le temps consacré à ce test est de 15 minutes. La note est informative pour l'élève et n'est pas comptabilisée dans sa moyenne annuelle.
- Le test assimilé ou semi-sommatif (TA) consiste en la résolution de questions nécessitant des calculs ou d'autres procédures mathématiques de plusieurs étapes à résoudre en quinze minutes. À la fin de l'année, la moyenne des quatre meilleures notes aux TA forme une note unique utilisée dans le calcul de la moyenne annuelle.
- Le test sommatif (TS) qui consiste en la résolution de questions nécessitant des calculs ou d'autres procédures mathématiques. Le temps consacré à ce test est de quarante-cinq minutes. Il se compose de quelques questions nécessitant une plus longue résolution. La note est sommative et compte pour la moyenne annuelle.

Il y a un total de sept tests par classe (cf. Annexe B) :

- Un test formatif (TF) ayant comme sujet le calcul numérique.
- Trois tests assimilés (TA1, TA2, TA3) ayant comme sujet les puissances, le calcul littéral et la lecture de graphe.
- Trois tests significatifs (TS1, TS2, TS3) ayant comme sujet le calcul numérique, les puissances et le calcul littéral, et la lecture de graphe de fonction.

Les tests sont exclusivement composés de QRO. Les consignes des questions ne précisent pas explicitement que le détail du raisonnement est demandé, à l'exception du TS3, ce qui pourrait les faire ressembler à des QROC. Le choix de ne pas signaler la nécessité de détailler les calculs, par exemple, relève d'une décision didactique de ma part. En effet, l'objectif est de faire comprendre aux élèves l'utilité de la notation mathématique comme outil de résolution. Prenons l'exemple de la question 1(a) du TS1 :

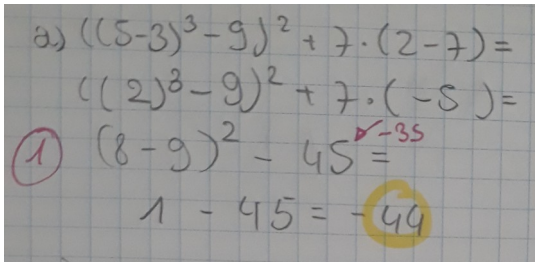
Calculer la valeur de c[ette] expression. Donner la réponse sous forme de fractions irréductibles.

(a) $\left((5-3)^3-9\right)^2 7\cdot(2-7)$

Il serait possible de répondre de manière similaire à une QROC en donnant la bonne réponse de -34. Cependant, cette approche présente un risque. En cas de réponse incorrecte, l'élève ne reçoit aucun point. À titre d'exemple, la Figure 4 illustre la réponse d'une élève à cette question. On peut observer qu'elle fournit une réponse finale incorrecte. Cependant, elle obtient tout de même 1 point sur 1.5, car elle a commis une erreur de calcul. Il incombe donc à l'élève de fournir suffisamment de détails pour démontrer sa compréhension du sujet et ainsi éviter de perdre des points.

Figure 4

Réponse d'une élève à la question 1(a) du TS1



Note. Les corrections de l'enseignant sont notées en rouge.

Un autre exemple est illustré par l'exercice 4 du TS1 :

Exprimer ces nombres sous forme de fractions irréductibles.

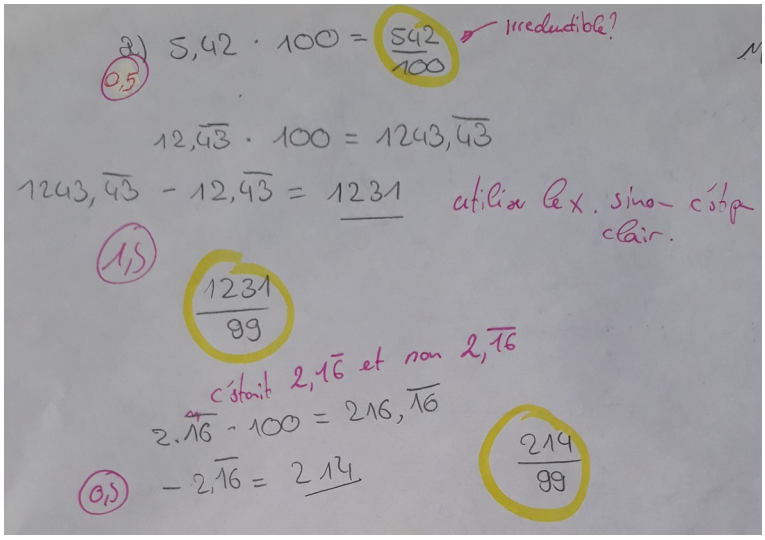
(a) 5.42 ; (b) $12.\overline{43}$; (c) $2.1\overline{6}$

La question (a) pourrait être considérée comme une QROC (car il est possible d'y répondre de manière immédiate). Cependant, il est à noter que l'évaluation de la réponse n'est pas strictement binaire, même dans ce cas, car l'élève reçoit la moitié des points en raison de sa réponse non réduite à sa forme irréductible (cf. Figure 5). Néanmoins, il est appréciable que l'élève ait pris le temps de recopier la donnée. Cette compétence s'avérera utile pour la question (c), où elle commet une erreur lors de la copie de la question. Dans cette situation, elle obtient tout de même 0.5 point sur 1.5 (avec une perte de 0.5 point due à l'erreur de copie et une autre perte de 0.5 point en raison de la modification de la complexité de la question).

Enfin, dans la réponse à la question (b), on peut observer que les erreurs de notation mathématique sont identifiées mais non pénalisées. Cependant, il est important de noter que ce type d'erreurs de notation pourrait avoir des conséquences défavorables pour l'élève lors de la résolution de problèmes plus complexes.

Figure 5

Réponse d'une élève à la question 4 du TS1



Note. Les corrections de l'enseignant sont notées en rouge.

Le nombre d'élèves pour le TF et les TA ne correspondent pas forcément au nombre d'élèves dans chaque classe. En effet, dû aux fréquentes absences des élèves en raison de la Covid, tous les TA n'ont pas été rattrapés. Les élèves n'ayant pas l'habitude d'évaluations avec les DC et pour les encourager dans leur auto-évaluation, nous avons retenu la meilleure note entre celle obtenue avec le score standard et celle utilisant le score avec DC et un seuil d'acceptation de 0.8. Nous verrons plus loin que ce choix n'était pas le plus adapté, car il désavantage beaucoup les élèves. Néanmoins, ils reçoivent ainsi deux notes et ont un retour implicite sur leurs connaissances subjectives.

Pour chaque élève, test et question, j'ai relevé dans un tableau E*xcel* le nombre de points, le nombre de points obtenus, le DC utilisé. Les données ont ensuite été analysées à l'aide du langage de programmation P*ython* et de la librairie d'outils statistiques S*cipy*.

5.2 Méthodologie

Dans cette section sont présentées les méthodes d'analyse des données utilisées pour répondre aux questions de recherches (section 4). La catégorisation des élèves selon leur niveau de compétence est exposée ainsi que les différents tests statistiques utilisés pour l'évaluation des différences significatives du réalisme. Les statistiques qui ne seront pas discutées se trouvent dans l'annexe A.

5.2.1 Biais entre les classes

Afin de vérifier qu'il n'y a pas de biais entre les classes A et B (dus à leur niveau, à mon enseignement, aux biais intra-correcteurs, etc.), un test de normalité Shapiro-Wilk pour la Centration (Cs) et le Réalisme (Rs) est réalisé. Si le test normalité est réussi à un seuil de 5% ($p < 0.05$) un test U de Mann-Whitney est effectué pour voir s'il y a une différence entre les distributions de la centration et du réa-

lisme des élèves des deux classes. Si le test de Shapiro-Wilk n'est pas rejeté, un test de Student pour deux variables indépendantes est utilisé.

5.2.2 Mesures et seuil d'acceptation

La mesure du Réalisme (Rs) utilisant la Centration (Cs) a l'avantage d'être facile à calculer et n'a pas besoin d'un seuil d'acceptation de la réponse. C'est-à-dire qu'il n'est pas nécessaire de catégoriser la réponse comme correcte ou incorrecte. Je voudrais vérifier si cette mesure du réalisme est suffisamment proche des mesures du Réalisme (Rs) de Gilles (2002) et Prosperi (2015) (cf. section 3.3.4), mais également aux mesures de Prosperi modifiée discrète et continue (cf. section 3.3.5) pour être cohérente et prétendre capturer des effets semblables à ces dernières. La corrélation de Pearson entre la mesure du Réalisme (Rs) utilisant la Centration (Cs) et les autres mesures pour des seuils d'acceptation compris dans l'intervalle [0.3, 1] est mesurée. J'espère observer une bonne corrélation entre la mesure utilisant la centration (cf. équation (12)) et les autres en particulier avec celle de Prosperi modifiée continue. J'espère également pouvoir définir un seuil d'acceptation cohérent.

Je calcule la corrélation de Pearson entre le score standard et les différentes mesures du Réalisme (Rs) pour voir si ces dernières reflètent, en effet, d'autres aspects de l'apprentissage. Une forte corrélation avec le score indiquerait que les variations observées dans le réalisme sont principalement liées au score. Cela nécessiterait de reconsidérer l'interprétation des résultats des tests statistiques pour la progression du réalisme présentée plus loin. En effet, en cas de forte corrélation, deux interprétations sont possibles. Soit le DC utilisé par l'élève varie peu, ce qui rendrait son jugement (auto-évaluation) constant et donc sa centration dépendrait principalement du score au test. Soit le réalisme des élèves est intrinsèquement lié à leurs performances globales. Dans le premier scénario, les données pourraient être biaisées. Dans le second scénario, cela remettrait en question l'indépendance entre les performances et les connaissance subjectives.

5.2.3 Progression du réalisme

Puisque la progression du réalisme semble différente pour les élèves avec des performances basses, les élèves sont divisés en trois catégories d'après leur score standard au premier test (TF). Les trois catégories sont définies comme suit (pour $Score_{std}^{TF}$ le score standard au TF) :

- Basse : $0 \leq Score_{std}^{TF} < 0.5$;
- Moyenne : $0.5 \leq Score_{std}^{TF} < 0.75$;
- Haute : $0.75 \leq Score_{std}^{TF} \leq 1$;
- Tous : tous les élèves sont considérés.

Un élève étant absent au TF, il ne fait pas partie de l'analyse par niveau de performance. Comme pour le calcul du biais entre les classes, un test de normalité de Shapiro-Wilk est effectué. En ce qui concerne la Centration (Cs), la différence significative de la moyenne avec zéro sera vérifiée à l'aide d'un test de Student ou d'un test de Wilcoxon selon le résultats du test de normalité. Pour le réalisme, les différences entre le premier test (TF) et les suivants seront comparées. Si le test de Shapiro-Wilk est rejeté pour les deux réalismes, un test de Student pour deux variables dépendantes est appliqué. Sinon, un test des rangs signés de Wilcoxon est utilisé.

5.3 Résultats

Les deux classes ont été combinées en un seul échantillon. En effet, le test de normalité de Shapiro-Wilk a été rejeté pour la Centration (Cs) et le Réalisme (Rs) des classes A et B ($p = 0.000$ dans les quatre cas). Il peut sembler surprenant que la centration ne suive pas une distribution normale. Cela entre en contradiction avec les résultats de D'Antò et Rosset (2019), ainsi qu'avec l'hypothèse de normalité de la centration dans d'autres études (Gardner-Medwin et Gahan, 2003 ; Callender et al., 2015). L'une des raisons de cette divergence pourrait être liée à une légère variation de la définition de la Centration (Cs) (cf. équation (7)). En effet, le jugement est considéré comme un intervalle, ce qui entraîne une plus grande fréquence de centrations nulles.

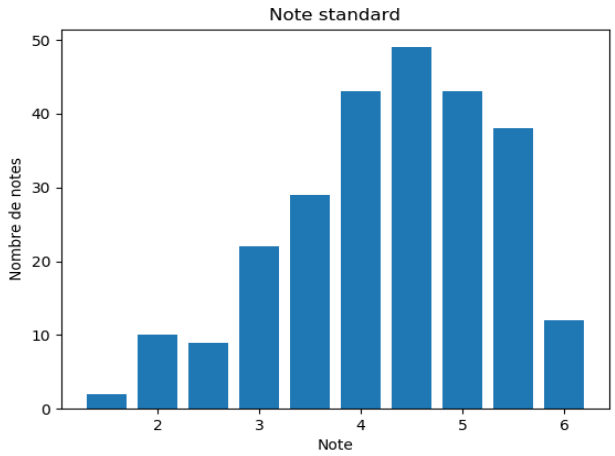
Le test U de Mann-Whitney ne révèle aucune différence significative, que ce soit pour la centration ($p = 0.413$) ou pour le réalisme ($p = 0.084$). Par conséquent, il semble justifiable de considérer les deux classes comme un seul et unique échantillon.

5.3.1 Gradation

La Figure 6 illustre la répartition des notes des élèves en utilisant la gradation standard. On observe que la courbe n'est pas totalement symétrique et que le mode se situe au-dessus du seuil de réussite (note de 4). Une courbe en forme de J, comme celle-ci, est généralement indicative d'un effet d'apprentissage. Pour confirmer cette hypothèse, il serait nécessaire de comparer la distribution des notes entre un post-test et un pré-test.

Figure 6

Répartition des notes standards



Note. Ne prend pas en compte les DC

Les notes ont été calculées pour trois seuils d'acceptation de la réponse (0.55, 0.75 et 0.9), ainsi que sans l'utilisation d'un seuil. Le Tableau 3 présente les différences entre les méthodes de gradation. Lorsqu'on utilise la méthode de gradation avec ou sans un seuil d'acceptation, les notes sont généralement plus basses. On remarque qu'un seuil trop élevé a un impact négatif significatif sur les élèves. La méthode de gradation avec l'utilisation des DC, mais sans seuil d'acceptation, offre le plus de résultats identiques à la méthode standard, tout en offrant un avantage à certains élèves. En revanche, la méthode de gradation avec un seuil de 0.55 favorise davantage d'élèves, mais conduit à moins de notes identiques. Du point de vue qualitatif, il semble donc que la méthode de gradation sans seuil d'acceptation (cf. équation (5)) soit mieux adaptée. Je remarque, *a posteriori*, qu'un seuil d'acceptation de 0.8 est trop élevé pour encourager au mieux les élèves à améliorer leurs connaissances subjectives.

Tableau 3

Comparaison des différentes méthodes de gradation avec la méthode standard

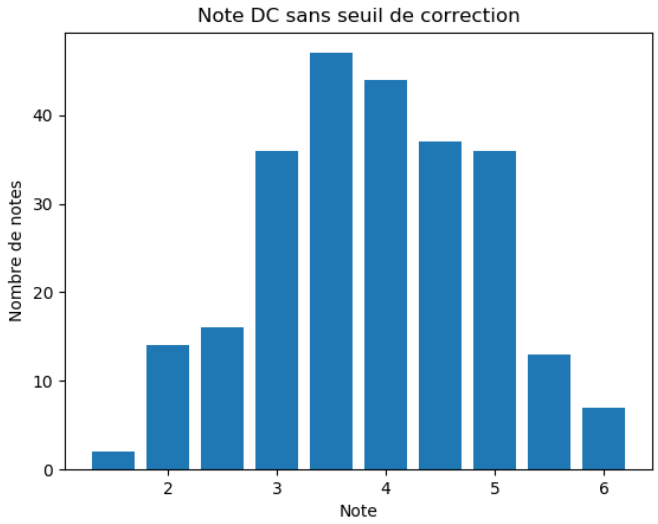
Gradation	Identique	Supérieur	Inférieur	Moyenne
Pas de seuil	77	7	173	3.83
$S_a = 0.55$	55	10	192	3.49
$S_a = 0.75$	46	6	205	3.35
$S_a = 0.9$	28	4	225	2.93

Note. La moyenne des notes standards est de 4.29. Le seuil d'acceptation de la réponse correspond au niveau d'exactitude requis pour qu'une réponse soit qualifiée de correcte.

La Figure 7 illustre la distribution des notes avec l'utilisation des DC, mais sans l'application d'un seuil d'acceptation (cf. équation (5)). On observe que le mode de la distribution (à 3.5) se rapproche du seuil de réussite. De plus, la forme de la distribution en cloche semble cohérente avec la nature des notes (cf. Figure 6). Néanmoins, cette distribution semble supposer un apprentissage faible quant à l'utilisation des DC par les élèves.

Figure 7

Répartition des notes standards

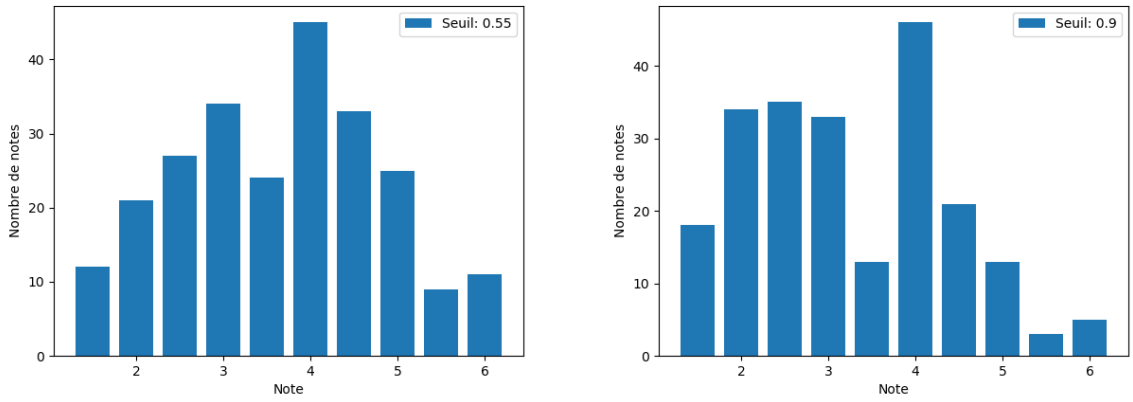


Note. Ne prend pas en compte les DC

Finalement, la Figure 8 illustre la distribution des notes en utilisant un seuil d'acceptation de 0.55 et de 0.9. On remarque que cette méthode de gradation génère un "creux" au niveau de la note 3.5, en particulier avec S_a = 0.9. Cette observation pourrait partiellement découler de la sévère pénalité en cas de réponse incorrecte. En effet, lorsque S_a = 0.9 et qu'un élève obtient trois quarts des points pour une question en utilisant le DC 4, il se voit infliger une pénalité négative. Cette transition le fait passer de +75 % des points dans le scénario standard à -20 % des points. Cette pénalité semble excessivement punitive. Plus loin, nous constaterons que le seuil d'acceptation d'une réponse optimal semble se situer autour de 0.5.

Figure 8

Répartition des notes avec DC et avec deux seuils d'acceptation de la réponse (0.55 et 0.9)



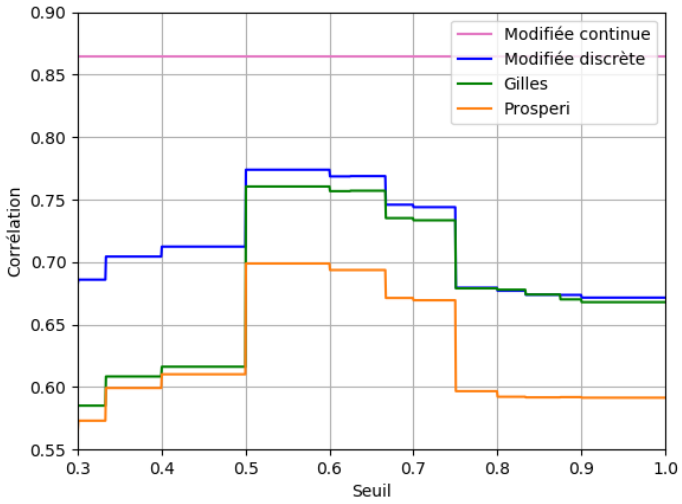
Note. À gauche : seuil d'acceptation = 0.55. À droite : seuil d'acceptation = 0.9

5.3.2 Mesures et seuil d'acceptation

La Figure 9 illustre la corrélation de Pearson entre le Réalisme (Rs) utilisant la Centration (Cs) et les autres mesures. Il est très encourageant d'observer une corrélation élevée avec la mesure de Prosperi modifiée continue (r = 0.853). On remarque que la corrélation maximale se situe dans l'intervalle [0.5, 0.6] pour le seuil d'acceptation, atteignant une valeur proche de r = 0.8 pour la mesure de Gilles (2002) ainsi que pour celle de Prosperi modifiée discrète. Cette valeur de 50 % pour le seuil d'acceptation semble cohérente, car elle correspond à l'arrondissement aux entiers les plus proches.

Figure 9

Corrélation de Pearson entre le Réalisme (Rs) utilisant la Centration (Cs) et les autres mesures



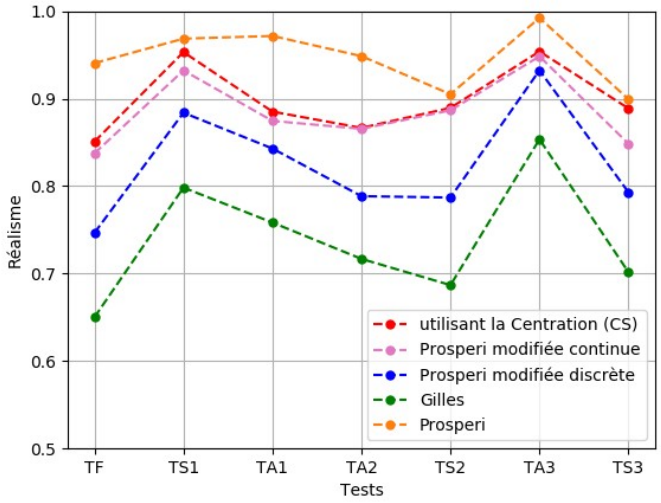
Note. Le seuil d'acceptation varie dans l'intervalle [0.3, 1]

La corrélation est moindre avec la mesure de Prosperi (2015), car celle-ci a tendance à surestimer le réalisme lorsqu'il y a peu d'utilisation d'un DC. Cette situation survient fréquemment dans des tests comprenant peu de questions. En effet, il est possible d'observer sur la Figure 10 que la moyenne des réalismes par test est plus élevée avec la mesure de Prosperi (2015). Il est à noter particulièrement que le réalisme diminue lors des TS. L'augmentation du nombre de questions dans ces tests pourrait améliorer la précision de la mesure de Prosperi (2015) et ainsi l'éloigner de la valeur 1.

À l'exception de la mesure de Prosperi (2015), le réalisme semble progresser en général. En se limitant aux TF et aux TA, on remarque que cette progression semble plus marquée, tandis qu'elle diminue pour les TS. Je n'étudierai pas la différence de réalisme entre les TA et les TS, mais les données suggèrent qu'il pourrait y en avoir une. Avant d'aborder l'analyse quantitative de cette progression, la corrélation entre le score et les différentes mesures du réalisme est étudiée.

Figure 10

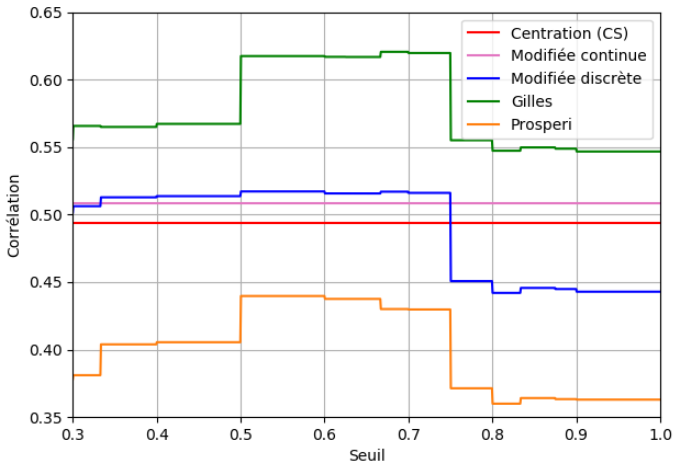
Réalisme moyen par tests pour les différentes mesures (Sc = 0.55)



La Figure 11 indique que le modèle utilisant la Centration (Cs) présente une corrélation faible (r = 0.501) avec le score. Les autres modèles présentent également des corrélations peu élevées avec le score. Par conséquent, il semble raisonnable d'ignorer les effets du score des élèves sur le réalisme dans la suite de notre analyse.

Figure 11

Corrélation de Pearson entre le score standard et les mesures du réalisme



Note. Le seuil d'acceptation varie dans l'intervalle [0.3, 1]

En plus des avantages pratiques de la mesure du Réalisme (Rs) utilisant la Centration (Cs) (équation (12)), cette mesure semble être cohérente avec le concept de réalisme. En effet, elle présente une bonne corrélation avec les autres mesures, en particulier pour un seuil d'acceptation de 50 %, ce qui pa-

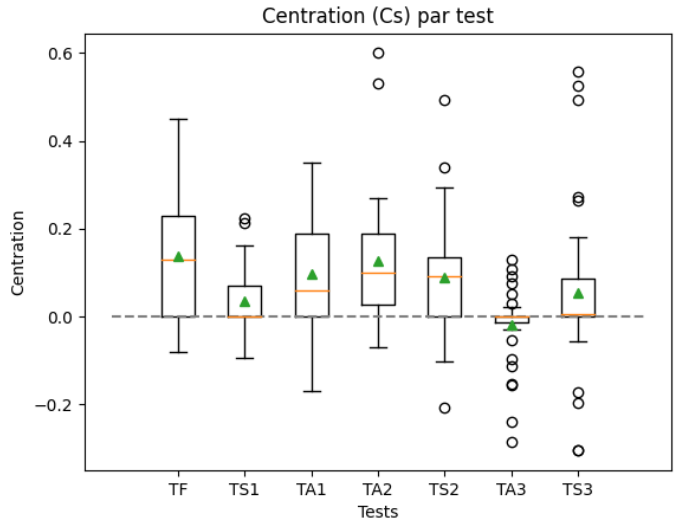
raît cohérent. De plus, sa corrélation avec la performance (score) de l'élève est faible. C'est pourquoi la mesure basée sur la Centration (Cs) sera utilisée pour les tests statistiques à venir.

5.3.3 Centration (Cs)

Sur la Figure 12 est représentée la répartition de la Centration (Cs) moyenne par test. On observe que les élèves ont tendance à se surestimer.

Figure 12

Diagramme en boîte à moustache de la Centration (Cs) par test (ordonné chronologiquement).



Note. Le rectangle représente le premier et troisième quartile. L'intervalle de confiance est de 99% en supposant une loi normale. Le segment vertical est la médiane et le triangle la moyenne.

En effet, il est notable dans le Tableau 4 que la Centration (Cs) est fréquemment positivement significative, notamment chez les élèves ayant des notes basses. Ce constat, illustrant une tendance à la surestimation des connaissances subjectives, est en accord avec les conclusions de Juslin et al. (2000). Toutefois, cette tendance est moins marquée chez les élèves performants, chez qui une centration plus précise semble être observée (centrée autour de zéro).

Tableau 4

Moyenne de la Centration (Cs).

Groupe	Nombre d'individus	Moyenne de la Centration (Cs)						
		TF	TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	**0.221**	**0.057**	**0.134**	**0.109**	**0.160**	-0.010	0.102
Moyenne	16	**0.118**	0.029	**0.064**	**0.125**	0.037	-0.025	0.028
Haute	6	-0.013	0.006	**0.108**	0.174	0.065	-0.018	0.019
Tous	38	**0.138**	**0.036**	**0.098**	**0.127**	**0.089**	-0.017	**0.055**

Note. Les différences significatives avec zéro sont notées en gras.

Finalement, les résultats du test de normalité diffèrent lorsque l'on prend en compte toutes les catégories de performances. Comme indiqué dans le Tableau 5, nous pouvons constater que la centration tend à présenter une distribution gaussienne au sein des groupes de performance, mais pas dans l'ensemble de la population d'individus. Il est envisageable que la distribution de la Centration (Cs) pour l'ensemble de l'échantillon soit multimodale et donc considérablement distincte d'une distribution gaussienne.

Tableau 5

p-valeur du test de Shapiro-Wilk pour la centration.

Groupe	Nombre d'individus	*p*-valeur du test de Shapiro-Wilk						
		TF	TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	**0.369**	0.009	0.019	**0.296**	**0.089**	**0.150**	0.028
Moyenne	16	**0.842**	**0.088**	**0.497**	0.004	**0.366**	0.000	**0.142**
Haute	6	0.007	0.000	**0.172**	**0.701**	**0.065**	0.038	**0.353**
Tous	38	0.019	0.000	**0.154**	0.000	0.032	0.000	0.000

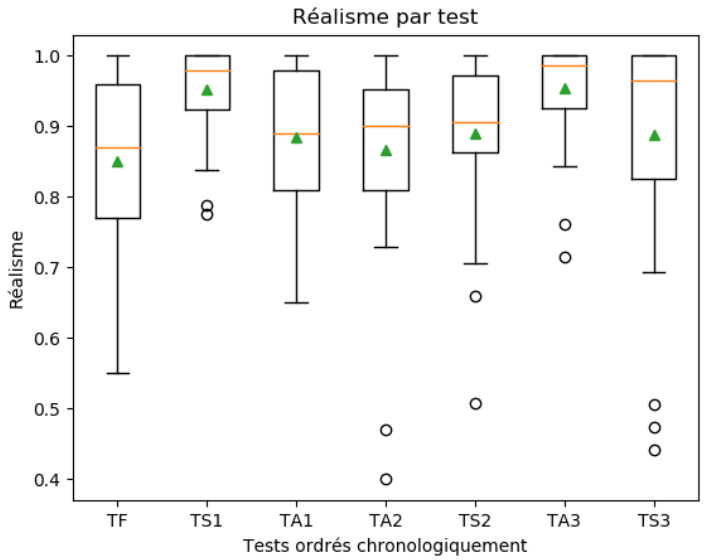
Note. Les valeurs en gras correspondent à p > 0.05 (on ne peut pas rejeter l'hypothèse d'une distribution gaussienne).

5.3.4 Réalisme (Rs)

Les distributions du Réalisme des sujets (Rs) pour les différents tests ne semblent pas être normales à première vue. En effet, elles montrent une accumulation de valeurs en dessous et près de 1. Cette tendance est illustrée dans la Figure 13, où le rectangle représentant les quartiles ne se situe pas au centre de l'intervalle de confiance et où les moyennes se trouvent systématiquement en dessous de la médiane.

Figure 13

Diagramme en boîte à moustache du réalisme par test (ordonné chronologiquement).



Note. Le rectangle représente le premier et troisième quartile. L'intervalle de confiance est de 99 % en supposant une loi normale. Le segment vertical est la médiane et le triangle la moyenne.

Les *p*-valeurs des tests de Shapiro-Wilk sont répertoriées dans le Tableau 6. Les distributions normales semblent plus fréquentes dans la catégorie des performances élevées. Cela pourrait s'expliquer par la taille relativement réduite de l'échantillon de cette catégorie, qui peut être insuffisante pour rejeter l'hypothèse nulle (c'est-à-dire, que la distribution est normale).

Tableau 6

p-valeur du test de Shapiro-Wilk pour le réalisme.

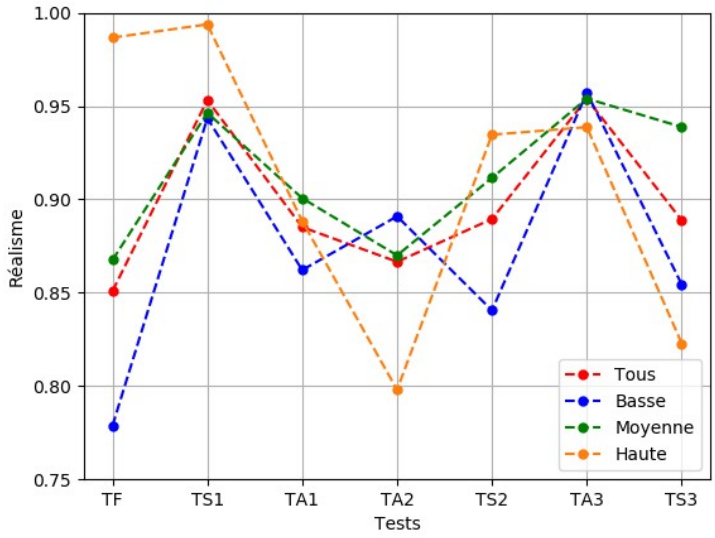
Groupe	Nombre d'individus	*p*-valeur du test de Shapiro-Wilk						
		TF	TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	0.002	0.007	0.010	**0.296**	**0.089**	0.004	0.002
Moyenne	16	0.013	0.006	**0.082**	0.001	0.011	0.000	0.013
Haute	6	**0.257**	0.000	**0.200**	**0.132**	**0.065**	0.016	**0.257**
Tous	38	0.000	0.000	0.003	0.000	0.000	0.000	0.000

Note. Les valeurs en gras correspondent à p > 0.05 (on ne peut pas rejeter l'hypothèse d'une distribution normale).

La Figure 14 illustre la moyenne des réalismes par tests pour les différentes catégories de performance. On constate que chez le groupe ayant des performances faibles et moyennes, le réalisme tend à augmenter par rapport au premier test. En revanche, pour le groupe affichant des performances élevées, cette tendance semble être inverse.

Figure 14

Réalisme moyen par test pour les différentes catégories de performance.

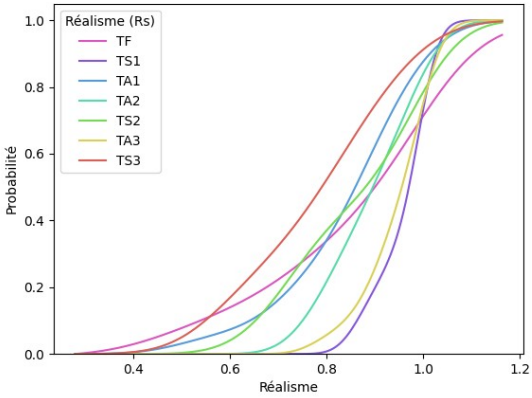


La Figure 15 représente la fonction de répartition du réalisme pour chaque test et niveau de performance, ainsi que pour l'ensemble des deux classes. On observe que, chez les élèves ayant des performances basses, leur réalisme s'améliore avec le temps. En effet, la répartition du réalisme semble se res-

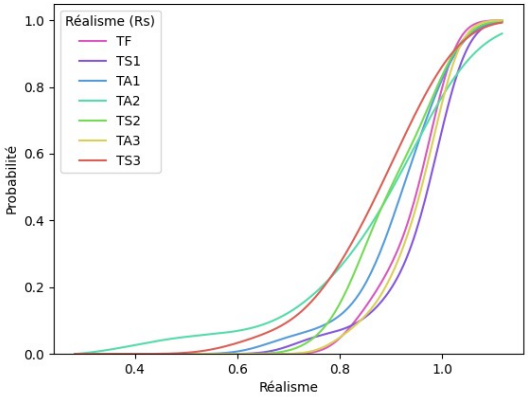
serrer vers la valeur de 1. Cette tendance est également constatée chez les élèves ayant des performances élevées. Néanmoins, cet effet est moins prononcé chez ceux ayant une performance moyenne. Enfin, il est à noter qu'en prenant en compte l'ensemble des élèves, la répartition du réalisme semble également se rapproche de 1, à l'exception du groupe TS3.

Figure 15

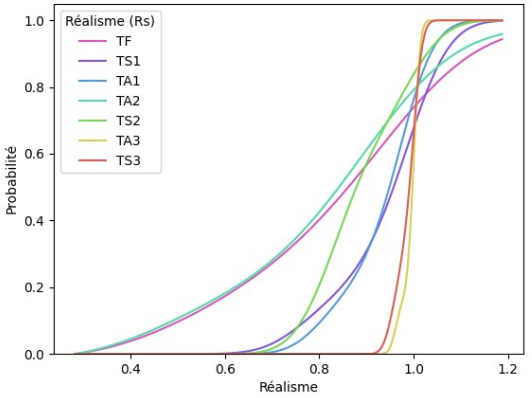
Fonction de répartition du Réalisme (Rs) par test et par niveau de performance



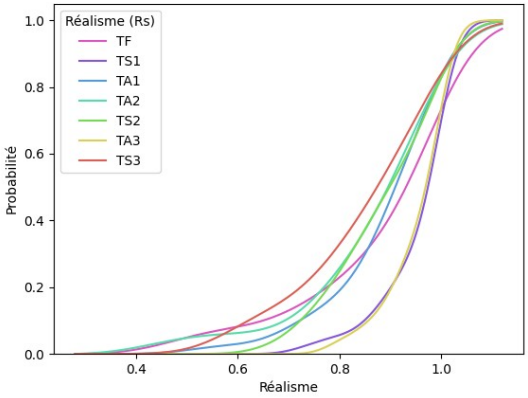
Note. Réalisme pour les performances basses



Note. Réalisme pour les performances moyennes



Note. Réalisme pour les performances hautes



Note. Réalisme pour l'ensemble des élèves

Dans le Tableau 7, il est noté que les élèves ayant obtenu des notes basses ont amélioré leur réalisme pour trois tests (le premier, troisième et cinquième). Les autres valeurs sont positives, mais non significatives. Concernant les élèves de la catégorie moyenne, une augmentation du réalisme est également observable, mais avec une amplitude et une fréquence moins marquée.

Tableau 7

Différence de la moyenne des réalismes par rapport au premier test (TF).

Groupe	Nombre d'individus	Moyenne des différences de réalisme					
		TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	**0.164**	0.084	**0.112**	0.062	**0.179**	0.076
Moyenne	16	**0.079**	0.033	0.002	0.044	**0.087**	0.071
Haute	6	0.007	**-0.098**	-0.189	-0.052	-0.048	-0.164
Tous	37	0.102	0.034	0.016	0.039	0.103	0.038

Note. Une valeur positive signifie une progression du réalisme. Les valeurs significatives sont notée en gras (test de Wilcoxon).

Enfin, pour la catégorie de performances hautes, une seule différence significative se dénote. En effet, le réalisme de ces élèves est en baisse au troisième test (TA1). Ce résultat peut sembler en contradiction, mais plusieurs éléments pourraient l'expliquer. Tout d'abord, la taille de l'échantillon est réduite, ce qui peut engendrer une incertitude dans les résultats. Deuxièmement, Leclercq et Gilles (1994) ont observé que le réalisme diminue puis s'améliore pour se stabiliser au fil de l'entraînement. Cependant, les échelles de temps de cette étude diffèrent. Troisièmement, en se référant au Tableau 4, il apparaît que les élèves de la catégorie élevée se sont surestimés au TA1. Il est envisageable que ces élèves, ayant obtenu de bons résultats lors du premier test, aient développé une confiance excessive en leurs performances.

Les résultats obtenus sont en concordance avec ceux de Callender et al. (2015), Miller et Geraci (2011), ainsi que Gilles et Melon (2000). En effet, nous observons une amélioration du réalisme dans le groupe affichant des performances basses, tandis que cette tendance s'estompe dans le groupe présentant des performances élevées.

6 Conclusion

L'un des objectifs fondamentaux de ce mémoire était d'adapter la méthode d'évaluation avec les DC, principalement utilisée pour les QCM, aux épreuves constituées de QRO. Cette adaptation a nécessité des ajustements au niveau des mesures de Centration (Cs) et de Réalisme (Rs). Ces nouvelles mesures s'harmonisent bien avec celles employées par Gilles (2002) et Prosperi (2015). Notamment, la mesure du Réalisme (Rs) utilisant la Centration (Cs) présente l'avantage d'être aisément calculable, ce qui en fait une méthode pratique à mettre en œuvre. De plus, elle ne dépend pas d'un seuil d'acceptation de la réponse, évitant ainsi le choix arbitraire inhérent aux méthodes avec seuil. Néanmoins, il est important de noter que la méthode de gradation mise en place conduit à des résultats inférieurs par rapport à la note standard. Ainsi, la définition de nouvelles échelles de notation ou méthodes de gradation qui refléteraient plus fidèlement le jugement et la performance des élèves pourrait s'avérer intéressante. La méthode de gradation sans seuil d'acceptation (équation (5)) n'assure pas systématiquement l'obtention de la note maximale même en cas d'utilisation adéquate des DC. Cette limitation est due au fait que les courbes de certitude de la Figure 2 ne se superposent pas parfaitement à l'intervalle de certitude (cf. sec-

tion 3.2), notamment pour les niveaux élevés de certitude. En conséquence, cette méthode de gradation pourrait encore bénéficier d'ajustements supplémentaires.

Un autre objectif de ce travail était l'application pratique de la méthode d'évaluation avec des DC dans un contexte réel. Les données collectées révèlent une progression significative du réalisme chez les élèves ayant des performances basses par rapport au premier test. Cette observation corrobore les résultats obtenus par Callender et al. (2015), Leclercq et Gilles (1994) ainsi que Miller et Geraci (2011). Toutefois, il est essentiel de considérer que le premier test était formatif, ce qui pourrait potentiellement introduire un biais. En effet, les élèves tendent à présenter de meilleurs jugements et performances lors des évaluations sommatives. Par conséquent, il est possible que les résultats de cette étude ne confirment que partiellement les conclusions de Gardner-Medwin et Gahan (2003). Cependant, la progression du réalisme s'est avérée significative pour deux TA par rapport à un seul TS. Il convient également de noter que les TA possèdent certaines caractéristiques des test formatifs, étant donné que les élèves ne sont pas nécessairement pénalisés en cas d'échec à certains (cf. section 5). Dans ce contexte, les TA peuvent être assimilés à des tests formatifs.

L'absence de recueil du ressenti des élèves concernant l'utilisation des DC à travers un questionnaire constitue une lacune dans cette étude. Cependant, des échanges informels ont eu lieu au sujet de l'utilisation des DC. Certains élèves ont signalé que l'évaluation de leur certitude leur prenait du temps et qu'ils se sentaient, à juste titre, peu récompensés pour une utilisation attentive des DC. La méthode de gradation employée, avec un seuil d'acceptation de 0.8 (cf. équation (3)), pourrait expliquer en partie ces réticences. L'utilisation d'une méthode de gradation sans seuil d'acceptation (cf. équation (5)) aurait potentiellement encouragé les élèves à une utilisation plus fréquente des DC, tout en fournissant des retours plus favorables. En effet, les notes obtenues avec cette méthode sont moins sujettes à des valeurs extrêmes contrairement aux mesures utilisant un seuil d'acceptation (cf. Figure 8). Néanmoins, il est important de noter qu'un élève avait exprimé des préoccupations concernant les notes attribuées par les DC, les trouvant parfois sévères. Nous avions pris l'exemple d'un de ses tests et avions remarqué qu'il avait tendance à surestimer sa performance. Il a réussi à obtenir une meilleure note avec DC lors d'un futur test. Finalement, une décision délicate reste celle d'utiliser exclusivement les notes avec des DC, car les aspects des connaissances subjectives ne sont pas explicitement mentionnés dans le programme d'études. Par conséquent, l'adoption d'un tel modèle nécessiterait une justification solide auprès de la direction de l'établissement. Par ailleurs, l'introduction d'un questionnaire à la fin de la session aurait permis d'exclure les élèves qui n'ont pas utilisé les DC de manière attentive.

L'utilisation des DC dans le processus d'évaluation était une nouveauté pour moi. Cependant, j'ai souhaité éviter de recourir aux QCM pour évaluer les connaissances des élèves. Les tests ont en effet été effectués par des élèves de première année du gymnase, où les attentes diffèrent généralement entre le secondaire I et II. Les concepts mathématiques deviennent plus complexes, en particulier avec l'intro-

duction intensive du calcul littéral. Dans ce contexte, la capacité à exprimer correctement les idées mathématiques devient un élément crucial de l'apprentissage des élèves. D'autant plus que ces derniers risquent de rencontrer cette technique des DC dans l'enseignement supérieur (section 3). Il est donc impératif de vérifier leur compréhension et leur raisonnement. Les DC se sont avérés utiles à plusieurs égards. Ils m'ont permis d'obtenir une meilleure vision des connaissances des élèves. D'un point de vue qualitatif, j'ai pu constater une tendance à la surestimation. Certains élèves ont tenté de négocier des points pour des questions où ils avaient répondu avec une grande certitude. Ces situations étaient souvent liées à de petites erreurs de calcul ou de notation. En conséquence, j'ai eu l'occasion de confronter les élèves à leur certitude en leur demandant pourquoi ils étaient si certains de leur réponse et s'ils avaient vérifié leurs calculs. Un élève particulièrement brillant perdait rarement des points et terminait les tests deux fois plus rapidement que les autres. Il répondait systématiquement avec la certitude maximale. Cela m'a convaincu que le niveau de difficulté du cours n'était pas adapté à ses compétences. J'ai donc ajusté les exercices pour le stimuler et l'encourager à progresser. Il est également devenu une ressource précieuse pour expliquer certains concepts aux autres élèves. D'autres élèves doués existaient, mais leur utilisation des DC était plus hésitante.

J'espère que les adaptations proposées pour l'évaluation des connaissances subjectives à l'aide des DC dans le contexte des QRO et des tests courts encourageront d'autres enseignants à adopter cet outil d'évaluation, notamment en mathématiques, où la vérification des réponses est une partie intégrante du processus d'apprentissage. Malgré l'inconvénient de devoir consigner les points et les DC pour chaque question, la mesure du Réalisme (Rs) utilisant la Centration (Cs) développée peut être facilement mise en œuvre dans un tableau *Excel*. Par ailleurs, toutes les mesures présentées dans ce travail ont été implémentées en *Python* à l'aide de la bibliothèque statistique *Scipy*. Ce logiciel est mis à disposition en ligne[4] pour d'éventuelles études en docimologie ou pour une utilisation dans le domaine de l'enseignement.

[4] Répertoire GitHub : https://github.com/baminian2/DC

7 Références

Agazzi, A. (1967). Les aspects pédagogiques des examens.

Ahlgren, A. (1969). Reliability, predictive validity, and personality bias of confidennce- weighted scores. Confidence on Achievement Tests – Theory, Applications.

Anderson, L. W. (2002). Curricular Alignment: A Re-Examination. *Theory to into Practice*, Vol. 41, n°4, pp.255-260

Anderson, L. W. (2004). Accroître l'efficacité des enseignants.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., and Wittrock, M. C. (2000). A taxonomy for learning, teaching, and assessing : A revision of Bloom's taxonomy of educational objectives, abridged edition.

Callender, A. A., Franco-Watkins, A. M., and Roberts, A. S. (2015). Improving metacognition in the classroom through instruction, training, and feedback. Metacognition and Learning.

Caspari-Sadeghi, S., Mille, E., Epperlein, H., and Forster-Heinlein, B. (2022). Stimulating reflection through self-assessment : Certainty-based marking (cbm) in online mathematics learning. Mathematics Teaching Research Journal, 14 :145–156.

Clark, C. (2020). The impact of confidence-based marking on unit exam achievement in a high school physical science course.

Cooke, E. (1906). Forecast and verifications in western australia. Monthly Weather Review, 34.

Davies, P. (2002). There's no confidence in multiple-choice testing, ... Proceedings of 6th CAA Conference, pages 118–130.

DeFinetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item.

Deruaz, M. and Buenzli, L.-O. (2015). L'utilisation des degrés de certitude comme outil de professionnalisation en formation des maîtres du premier degré. Pluralités culturelles et universalité des mathématiques : enjeux et perspective pour leur enseignement et leur apprentissage.

D'Antò, K. and Rosset, D. (2019). L'usage des degrés de certitude (dc) améliore-t-il le réalisme des étudiants ? analyse de l'évolution du réalisme dans deux classes du secondaire post-obligatoire soumises à des évaluations ayant recours à la technique des dc. HEP Lausanne.

Edwards, W. (1967). Probabilistic information processing by men and man-machine systems. La simulation du comportement humain.

Foster, C. (2015). Confidence and competence with mathematical procedures. Educ Stud Math.

Gardner-Medwin, A. (2006). Confidence-based marking - towards deeper learning and better exams. Innovative Assessment in Higher Education.

Gardner-Medwin, A. and Gahan, M. (2003). Formative and summative confidence- based assessment.

Gilles, J.-L. (1996). Utilisation des degrés de certitude et normes de réalisme en situation d'examen et d'auto-estimation à fa.p.s.e. - ulg.

Gilles, J.-L. (2002). Qualité spectrale des tests standardisés universitaires. PhD thesis, Université de Liège.

Gilles, J.-L. and Melon, S. (2000). Comparaison de trois modalités de "testing" des compé- tences en français chez les étudiants en médecine lors de leur première candidature à l'ulg. In J.-M. Defays, S. Melon, M. M., editor, La maîtrise du français du niveau secondaire au niveau supérieur, pages 161–178. De Boeck.

Houzé-Cerfon, C., Lauque, D., and Charpentier, S. (2016). Intégration du degré de certitude dans l'évaluation des connaissances des étudiants en médecine d'urgence. Ann. Fr. Med. Urgence.

Hunt, D. (1977). *The human self-assessment process. Study II: The effects of the number of self-assessment categories on acquisition*. Interim Report from U.S. Army Research Institute for the Behavioral and Social Sciences Grant #DAHC19-76-G-002, New Mexico State University, Las Cruces, NM.

Jacobs, S.-S. (1971). Correlates of unwarranted confidence in response to objective test items. 8.

Jans & Leclercq (1999). Mesurer l'effet de l'apprentissage à l'aide de l'analyse spectrale des performances, in C. Depover & B. Noel (Ed.), *Evaluation des compétences et des processus cognitifs*. Bruxelles : De Boeck, pp. 303-317.

Juslin, P., Winman, A., and Olsson, H. (2000). Naive empiricism and dogmatism in confi- dence research : A critical examination of the hard–easy effect. Psychological Review, 107 :384–396.

Leclercq, D. (1963). Critiques de méthodes d'application, de correction et de cotation des questions à choix multiple. X.

Leclercq, D. (1982). Confidence marking : Its use in testing. Evaluation in Education, 6(2) :161–287.

Leclercq, D. (1986). La conception des questions à choix multiple.

Leclercq, D. (1987). Qualité des questions et signification des scores avec application aux qcm.

Leclercq, D. (2016). Probable inference, the law of succession. Evaluer. Journal international de Recherche en Education et Formation, 2(3).

Leclercq, D. and Gilles, J.-L. (1994). Guess, un logiciel pour entraîner à l'auto-estimation de sa compétence cognitive.

Leclercq, D. and Gilles, J.-L. (1995). Le kaléidoscope des techniques de questionnement.

Miller, T. M. and Geraci, L. (2011). Training metacognition in the classroom : The influence of incentives and feedback on exam predictions. Metacognition and Learning, 6.

Prosperi, O. (2015). Le réalisme avec degrés de certitude. mesure et évaluation en éducation. 38.

Rigo, J.-M., Ef Jaafari, S., et Gilles, J.-L. (2020). CORETEV: Un réseau d'expertises Nord-Sud en ingénierie des évaluations. Communication présentée à Séminaire international de recherche SIR CORETEV, Lausanne, Suisse. http://hdl.handle.net/20.500.12162/3776

Shufford, A., Albert, A. & Massengil, N.E. (1966). Admissible probability measurement procedures, *Psychometrika*, 31, 125-145.

Schraw, G., Kuch, F., and Gutierrez, A. P. (2013). Measure for measure : Calibrating ten commonly used calibration scores. Learning and Instruction, 24.

Van Naerssen, R.-F. and Van Beaumonz, E. (1965). Ervaringen met een zekerheidsaanduiding bij objektieve tentamens. 20.

Wilson, E. B. (1927). Probable inference, the law of succession. Journal of the Americn Statistical Association, 22(158).

Annexes A : Statistique et tailles des échantillons

Centration (Cs)

Tableau 8

Centration (Cs). p-valeur du test de Student ou Wilcoxon pour l'hypothèse nulle d'une moyenne des Centration (Cs) nulles. Les valeurs significatives (p < 0.05) sont notées en gras.

Groupe	Nombre d'individus	*p*-valeur						
		TF	TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	**0.000**	**0.008**	**0.005**	**0.003**	**0.000**	0.592	0.124
Moyenne	16	**0.002**	0.172	**0.039**	**0.003**	0.241	0.400	0.209
Haute	6	0.180	0.317	**0.035**	0.166	0.089	1.000	0.879
Tous	38	**0.000**	**0.003**	**0.000**	**0.000**	**0.000**	0.313	**0.046**

Tableau 9

Centration (Cs). Statistique du test de Student ou Wilcoxon pour l'hypothèse nulle d'une moyenne des centrations nulles.

Groupe	Nombre d'individus	Statistique						
		TF	TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	5.83	0.000	2.50	3.97	4.69	-0.55	7.00
Moyenne	16	3.69	1.44	2.27	3.00	1.22	12.00	1.31
Haute	6	0.00	0.00	2.87	1.69	2.11	3.00	0.16
Tous	37	16.00	30.5	4.89	15.00	53.00	78.00	106.00

Réalisme (Rs)

Tableau 10

Réalisme. p-valeur du test de Student ou Wilcoxon pour l'hypothèse nulle d'une différence nulle des moyennes du réalisme par rapport au premier test (TF). Les valeur significative (p < 0.05) sont notée en gras.

Groupe	Nombre d'individus	*p*-valeur					
		TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	**0.002**	0.104	**0.042**	0.208	**0.000**	0.124
Moyenne	16	**0.047**	0.426	0.701	0.1	**0.023**	0.065
Haute	6	0.593	**0.035**	0.107	0.199	0.273	0.119
Tous	38	0.000	0.374	0.043	0.140	0.000	0.169

Tableau 11

Réalisme. Statistique du test de Student ou Wilcoxon pour l'hypothèse nulle d'une différence nulle des moyennes du réalisme par rapport au premier test (TF).

Groupe	Nombre d'individus	Statistique					
		TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	4.00	26.00	10.00	37.00	0.	28.00
Moyenne	16	25.00	46.00	40.00	31.00	20.00	32.00
Haute	6	2.00	2.88	2.07	1.48	2.00	1.88
Tous	38	69.00	245.50	181.00	225.00	76.00	231.00

Tableau 12

Réalisme. Taille des échantillons pour le test de Student ou Wilcoxon (élève présent aux deux tests).

Groupe	Nombre d'individus	Taile de l'échantillon					
		TS1	TA1	TA2	TS2	TA3	TS3
Basse	15	15	14	11	15	14	15
Moyenne	16	16	16	15	16	16	16
Haute	6	6	6	5	6	6	6
Tous	38	37	36	31	37	36	37

Annexe B : Tests

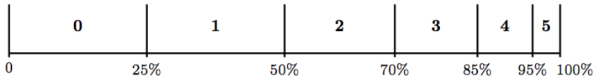
Mathématiques: 1C5 Test formatif: Calcul numérique 6 septembre 2021

Nom : Prénom :

Points : Note :

15 minutes.
Formulaire autorisé.

Degrés de certitude et points selon le degré :



		Degrés de certitude					
		0	1	2	3	4	5
Réponse	Correcte	+13	+16	+17	+18	+19	+20
	Incorrecte	+4	+3	+2	0	-6	-20

Exercice 1.

Calculer la valeur de ces expressions. Donner la réponse sous forme de fractions irréductibles.

(a) $-(-1)^2 - 3 \cdot 2 \cdot (3 - 3 \cdot (-1))$

Degré de certitude : ☐

(b) $-5^2 \cdot 3 - (5 - 20) \cdot 2$

Degré de certitude : ☐

(c) $\frac{2}{15} + \frac{5}{12} - \frac{1}{10} + 2$

Degré de certitude : ☐

(d) $\left(\frac{1}{2} - \frac{1}{3}\right) - \left(\frac{3}{8} - \left(\frac{1}{3} + \frac{5}{12}\right)\right)$

Degré de certitude : ☐

(e) $-(-2)^3 - 2^3$

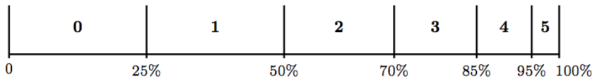
Degré de certitude : ☐

Nom : Prénom :

Points : Note :

15 minutes.
Formulaire autorisé.

Degrés de certitude et points selon le degré :



		Degrés de certitude					
		0	1	2	3	4	5
Réponse	Correcte	+13	+16	+17	+18	+19	+20
	Incorrecte	+4	+3	+2	0	-6	-20

Exercice 1.

Réduire ces expressions au maximum.

(a) $2^{-5} \cdot 2^{8}$

Degré de certitude : ☐

(b) $\dfrac{5^{-3}}{5^{-5}}$

Degré de certitude : ☐

(c) $\dfrac{3^{7}}{3^{-3} \cdot 3^{8}}$

Degré de certitude : ☐

(d) $\left(\dfrac{2^{2}a^{3}}{3b^{-2}}\right)^{2} \left(\dfrac{2b^{5}}{3a^{2}}\right)^{-2}$

Degré de certitude : ☐

(e) $\dfrac{8^{8}}{4^{11}}$

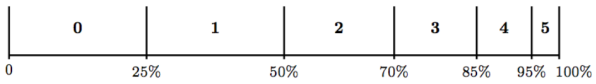
Degré de certitude : ☐

Nom : Prénom :

Points : Note :

45 minutes.
Formulaire et calculatrice autorisés.

Degrés de certitude et points selon le degré :



		Degrés de certitude					
		0	1	2	3	4	5
Réponse	Correcte	+13	+16	+17	+18	+19	+20
	Incorrecte	+4	+3	+2	0	-6	-20

Exercice 1. (5 points)

Calculer la valeur de ces expressions. Donner la réponse sous forme de fractions irréductibles.

(a) $((5-3)^3-9)^2+7\cdot(2-7)$

Degré de certitude : ☐

(b) $\left(\frac{1}{2}+\frac{5}{3}-\frac{1}{12}\right)^2 \div \frac{45}{42}$

Degré de certitude : ☐

(c) $\left(\frac{5}{7}+\frac{8}{14}\right)\cdot\left(2-\frac{1}{18}\right)$

Degré de certitude : ☐

(d) $\left(-\frac{3}{5}\right)^3-\left(\frac{5}{3}\right)^2$

Degré de certitude : ☐

Exercice 2. (3 points)

Substituer les valeurs numériques dans les expressions algébriques. Donner la réponse sous forme de fractions irréductibles.

(a) $-2x^2+5x-\frac{1}{2}$ pour $x=-3$.

Degré de certitude : ☐

(b) $\dfrac{(a-5)\cdot(b-7)}{a\cdot b}$ pour $a=3$ et $b=-5$.

Degré de certitude : ☐

Exercice 3. (3 points)

Exprimer ces fractions en code décimal à l'aide de la division euclidienne. Détailler vos calculs. Zéro point si pas de division euclidienne.

(a) $\dfrac{21}{25}$

Degré de certitude : ☐

(b) $\dfrac{227}{22}$

Degré de certitude : ☐

Exercice 4. (4 points)

Exprimer ces nombres sous forme de fractions irréductibles.

(a) 5.42

Degré de certitude : ☐

(b) $12.\overline{43}$

Degré de certitude : ☐

(c) $2.1\overline{6}$

Degré de certitude : ☐

Exercice 5. (5 points)

Une voiture fait le trajet de Berne (Suisse) à Bucharest (Roumanie). Le premier jour, elle fait un tiers du voyage. Le deuxième jour, elle parcoure trois cinquièmes du trajet restant. Sachant que le trajet total mesure 2'000 kilomètres.

(a) Quelle proportion du trajet a-t-elle parcouru les deux premier jours ?

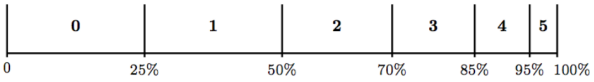
(b) Combien de kilomètre lui reste-t-elle à parcourir le troisième jours ?

Nom : Prénom :

Points : Note :

15 minutes.
Formulaire autorisé.

Degrés de certitude et points selon le degré :



		Degrés de certitude					
		0	1	2	3	4	5
Réponse	Correcte	+13	+16	+17	+18	+19	+20
	Incorrecte	+4	+3	+2	0	-6	-20

Exercice 1.

Réduire ces expressions au maximum.

(a) $(-2a)^3 \cdot (-3b)^2 + 2ab$

Degré de certitude : ☐

(b) $5xy(-xy^2)(3x^3y)2x$

Degré de certitude : ☐

(c) $x^2 + 2x + 3x^2 - 4x$

Degré de certitude : ☐

(d) $xy - yz - (3xy - 5yz)$

Degré de certitude : ☐

(e) $-2(3x - 2) + 6x$

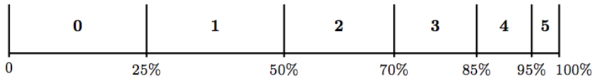
Degré de certitude : ☐

Nom : Prénom :

Points : Note :

45 minutes.
Formulaire autorisé.

Degrés de certitude et points selon le degré :



		Degrés de certitude					
		0	1	2	3	4	5
Réponse	Correcte	+13	+16	+17	+18	+19	+20
	Incorrecte	+4	+3	+2	0	-6	-20

Exercice 1. (8 points)

Réduire ces expressions au maximum. Détailler tous les calculs.

(a) $2x(-3xy^2)(-y)$

Degré de certitude : ☐

(b) $\dfrac{(2a)^{12}(3b)^{11}}{(3a)^9(2b)^{10}}$

Degré de certitude : ☐

(c) $2x^2 - 3 - (5 - 3x^2)$

Degré de certitude : ☐

(d) $(3x - y)(7x + 3y)$

Degré de certitude : ☐

(e) $5x(2x - 3)^2$

Degré de certitude : ☐

(f) $\dfrac{1}{2x} - \dfrac{1}{3x}$

Degré de certitude : ☐

(g) $\dfrac{3xy^2}{2} \div \dfrac{9x^3y}{10}$

Degré de certitude : ☐

(h) $\dfrac{3}{2x} - \dfrac{5}{x+3}$

Degré de certitude : ☐

Exercice 2. (6 points)

Écrire ces nombres en notation scientifique. Détailler tous les calculs.

(a) 1200300

Degré de certitude : ☐

(b) 0.0042

Degré de certitude : ☐

(c) $0.00014 \cdot 0.0000034$

Degré de certitude : ☐

(d) $(0.00005)^3$

Degré de certitude : ☐

(e) $\dfrac{60000}{0.00015}$

Degré de certitude : ☐

(f) $\dfrac{48 \cdot 10^{10} \cdot 7 \cdot 10^{5}}{(2 \cdot 10^{-3})^4}$

Degré de certitude : ☐

Exercice 3. (3 points)

La Terre existe depuis environ 4,5 milliards d'années. Combien de minutes cela représente-t-il ?

Degré de certitude : ☐

Exercice 4. (2 points)

L'énergie contenue dans une masse m (en kg) est donnée par la formule

$$E = m \cdot c^2 \,.$$

La lettre $c = 3 \cdot 10^8$ m/s est la vitesse de la lumière. Calculer l'énergie E que contient 0.2 kg de matière.

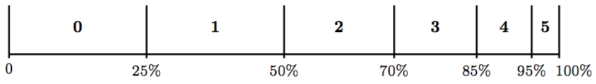
Degré de certitude : ☐

Nom : Prénom :

Points : Note :

15 minutes.
Formulaire et calculatrice autorisés.

Degrés de certitude et points selon le degré :



		Degrés de certitude					
		0	1	2	3	4	5
Réponse	Correcte	+13	+16	+17	+18	+19	+20
	Incorrecte	+4	+3	+2	0	-6	-20

Exercice 1.

Considérez le graphe de la figure 1.

(a) Trouver l'ordonnée à l'origine.

Degré de certitude : ☐

(b) Trouver les zéros de la fonction f.

Degré de certitude : ☐

(c) Trouver les coordonnées du minimum de f.

Degré de certitude : ☐

(d) Trouver l'image de 2.

Degré de certitude : ☐

(e) Trouver x tels que $f(x) = -10$.

Degré de certitude : ☐

Exercice 2.

Le réfrigérateur fonctionne à l'aide d'un gaz qui est refroidi en modifiant la pression. D'abord le gaz est relativement chaud, puis, en baissant la pression du gaz, celui-ci se refroidit. Il passe dans les tuyaux du réfrigérateur et se réchauffe en prenant la chaleur. Ensuite, le gaz est légèrement refroidi, passe par quelques tuyaux et le processus se répète. Sur la figure 1, nous avons représenté la température en degré Celsius du gaz par rapport au temps en minutes.

(a) Interpréter le résultat du point (a).

Degré de certitude : ☐

(b) Interpréter le résultat du point (b).

Degré de certitude : ☐

(c) Interpréter le résultat du point (c).

Degré de certitude : ☐

(d) Interpréter le résultat du point (d).

Degré de certitude : ☐

(e) Interpréter le résultat du point (e).

Degré de certitude : ☐

(f) Trouver la température maximale du gaz.

Degré de certitude : ☐

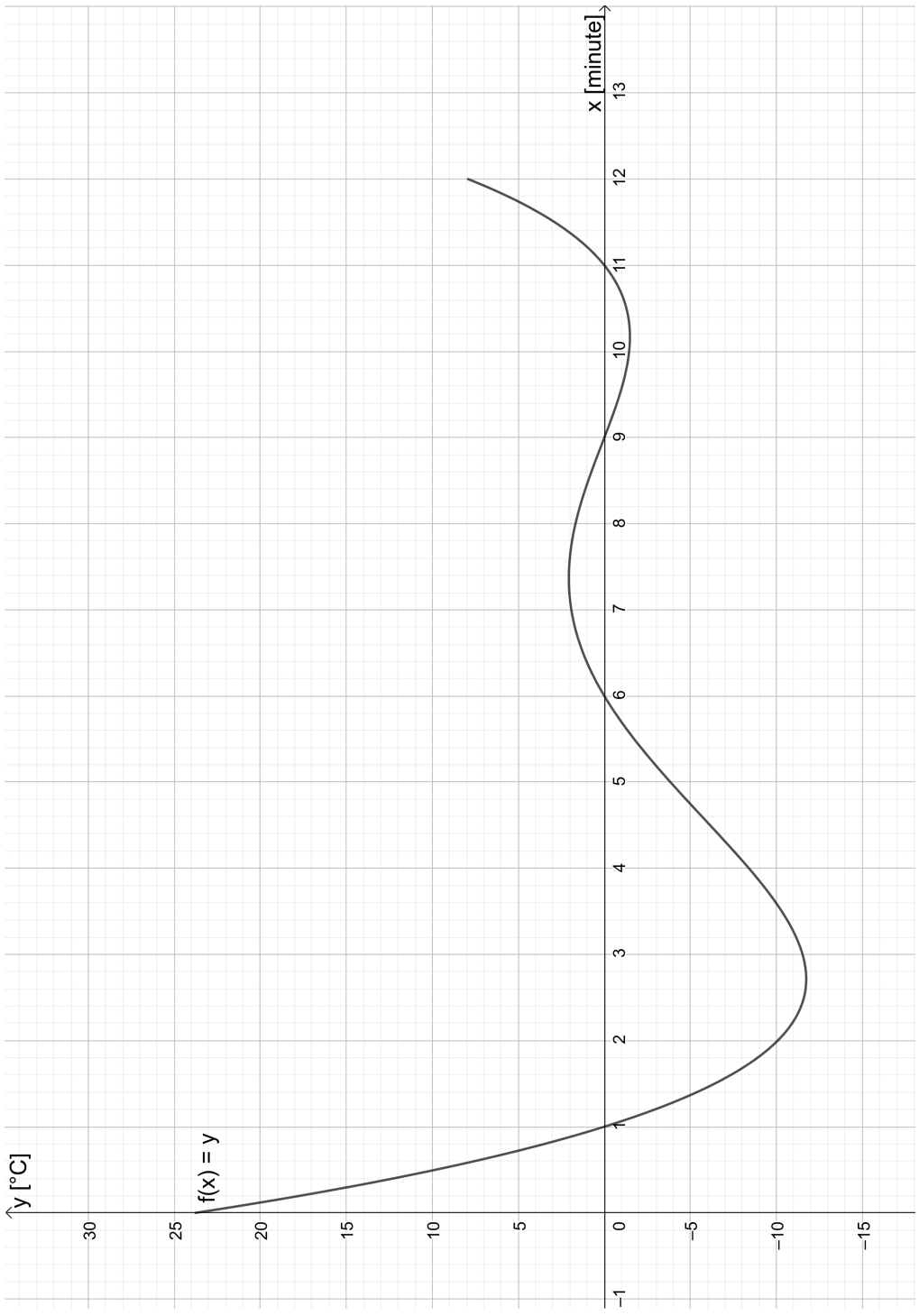


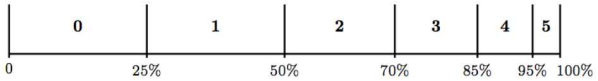
FIGURE 1 – Température du gaz d'un réfrigérateur selon le temps en minute

Nom : Prénom :

Points : Note :

45 minutes.
Formulaire et calculatrice autorisés.

Degrés de certitude et points selon le degré :



		Degrés de certitude					
		0	1	2	3	4	5
Réponse	Correcte	+13	+16	+17	+18	+19	+20
	Incorrecte	+4	+3	+2	0	-6	-20

Exercice 1. (Noter clairement les traits de construction sur le graphe de la figure 1.)

Sur la figure 1 est représenté le graphe de la fonction qui donne le pourcentage des élèves qui ont obtenu une note supérieur à la moyenne (réussite) à leur test de mathématique selon le nombre d'heures de travail qu'ils ont fournies lors de leur révision.

(a) Trouver l'échelle pour chaque axe en considérant le millimètre comme unité.

Degré de certitude : ☐

(b) Quel pourcentage des élèves ont réussi en ayant travailler 2.5 heures.

Degré de certitude : ☐

(c) Combien d'heures faudrait-il travailler pour que 50% des élèves réussissent leur test ?

Degré de certitude : ☐

(d) Quel pourcentage d'élèves réussissent leur test s'ils travaillent entre 2h et 3h ?

Degré de certitude : ☐

(e) Dans une classe, les élèves ont travaillé 1.5 heures pour le test de math. Quel pourcentage des élèves vont réussir le test ?

Degré de certitude : ☐

Exercice 2.

Dans un marché, le kilo de pommes coûte 2,30 CHF.
À partir de cette phrase :

(a) Faire un tableau des valeurs pour le prix de 1 ; 2 ; 3 ; 3.5 et 4 kg de pommes.

Degré de certitude : ☐

(b) Retrouver l'expression algébrique qui donne le prix en fonction du nombre de kilogrammes.

Degré de certitude : ☐

(c) Reporter les points du tableau des valeurs dans un système d'axes.

Degré de certitude : ☐

(d) Esquisser la fonction qui donne le prix en fonction du nombre de kilogrammes (entre 0 et 4 kg).

Degré de certitude : ☐

Exercice 3. (Noter clairement les traits de construction sur le graphe de la figure 2.)

Sur la figure 2 est représenté le graphe d'une fonction.

(a) Trouver l'ordonnée à l'origine de la fonction.

Degré de certitude : ☐

(b) Trouver les zéros de la fonction.

Degré de certitude : ☐

(c) Trouver le maximum de la fonction.

Degré de certitude : ☐

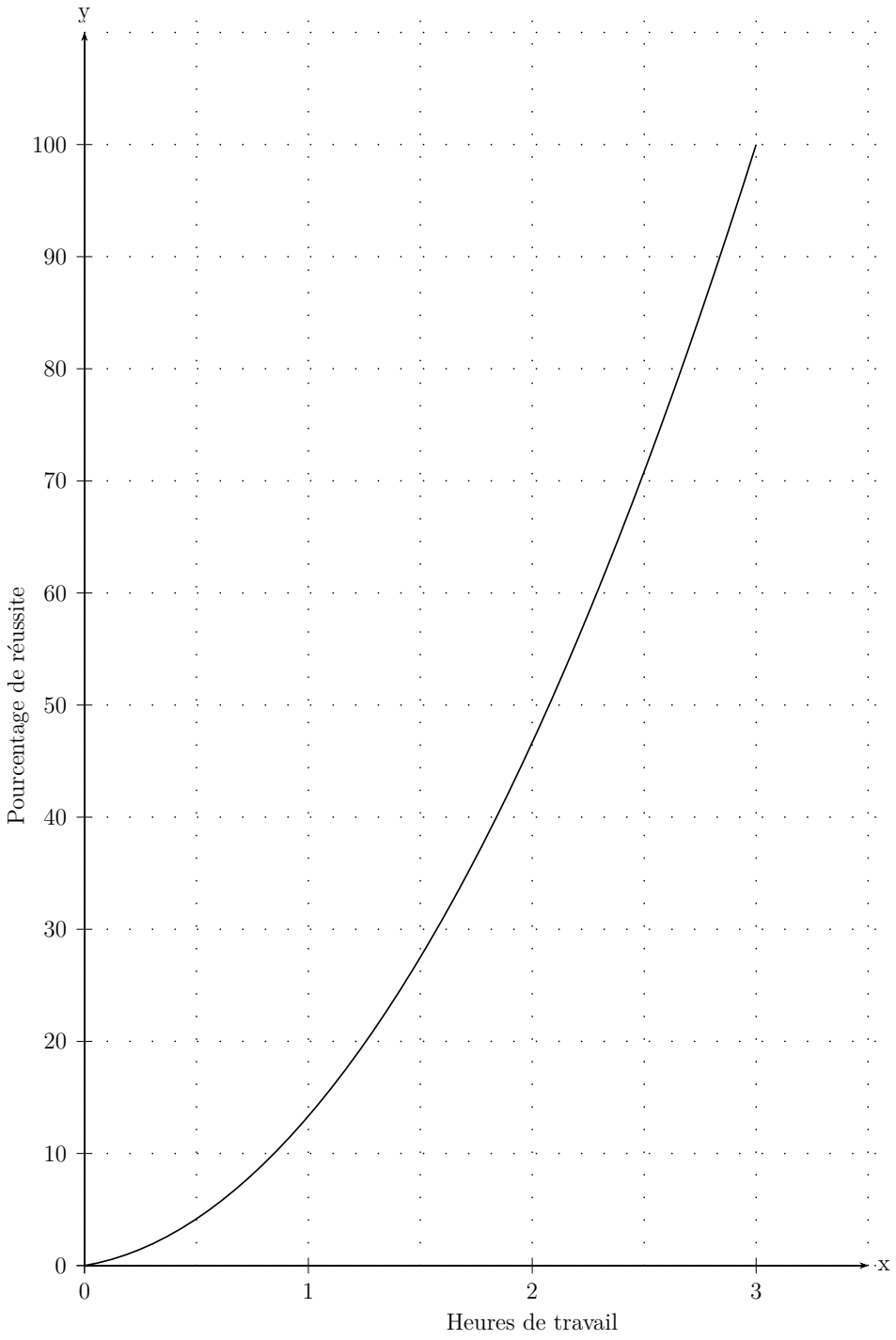


FIGURE 1 – Pourcentage de réussite pour un test de math selon le nombre d'heures de travail (exercice 1).

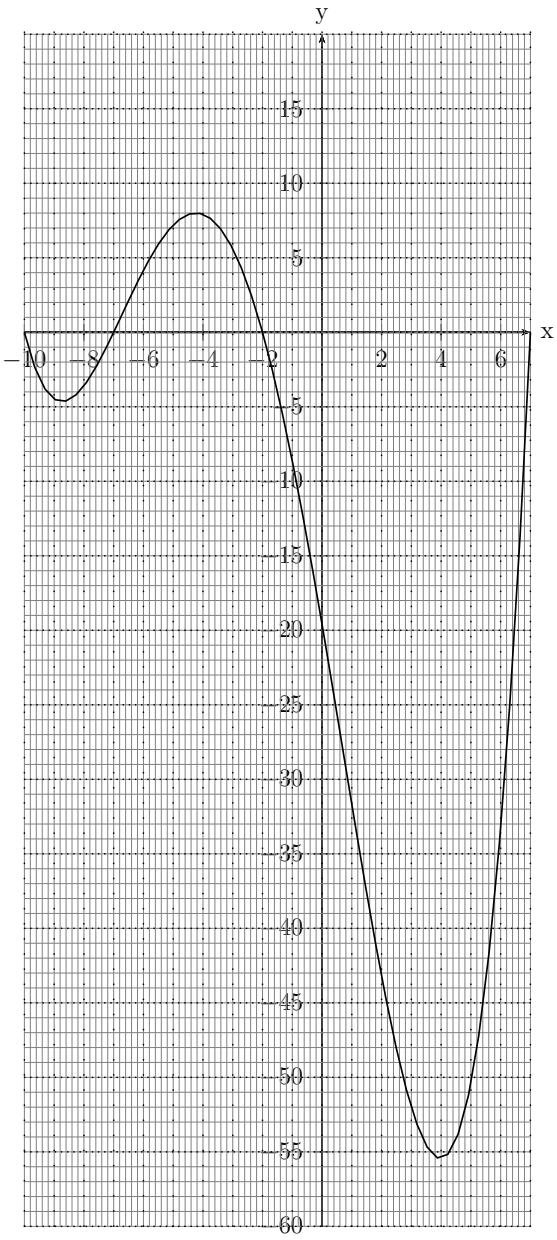


FIGURE 2 – Graphe d'une fonction (exercice 3).

4e de couverture

La méthode d'évaluation basée sur les Degrés de certitude (DC) est généralement employée pour évaluer les réponses à des Questions à choix multiple (QCM). Cependant, ce mémoire en docimologie propose une adaptation de cette méthode pour les Questions à réponse ouverte (QRO). Les mesures de la centration, du réalisme et du score utilisant les DC sont ainsi présentées et ajustées pour convenir aux QRO. De plus, les mesures probabilistes du réalisme sont également adaptées pour les tests courts. En appliquant cette méthode à sept tests de mathématiques effectués dans deux classes du secondaire II, une tendance positive de progression du réalisme est observée chez les élèves ayant des performances moins élevées.

Mots clefs : Docimologie, Degré de Certitude, Question à réponse ouverte (QRO), Réalisme, Mathématiques.