RITPU | IJTHE
Revue internationale des technologies en pédagogie universitaire
International Journal of Technologies in Higher Education
RITPU.CA
IJTHE.CA
ISSN 1708-7570 Volume 22, n° 1, article 3 2025

Analyse comparative entre la génération automatique de questionnaires à choix multiples par ChatGPT et le corps enseignant pour l'évaluation de l'apprentissage

Comparative Analysis of ChatGPT Generated and Instructor-Designed Multiple-Choice Questions for Student Evaluation

https://doi.org/10.18162/ritpu-2025-v22n1-03

Nurgul MEIRMANOVA [a] **Université de Lille, France**

Mis en ligne : 4 avril 2025

Résumé

Cette étude, menée auprès de personnes apprenantes de l'Université, examine l'intégration de ChatGPT-4 dans la création de questionnaires à choix multiples (QCM). Nous comparons les QCM générés par l'outil d'intelligence artificielle (IA) à ceux conçus par des enseignantes et nous analysons la qualité des distracteurs ainsi que les indices de difficulté. L'étude révèle que les QCM générés par l'IA obtiennent des taux de réussite élevés, mais présentent des distracteurs moins efficaces, limitant leur capacité à évaluer des compétences de réflexion analytique. Cette recherche souligne l'importance de l'intervention humaine pour garantir des QCM de qualité, tout en reconnaissant le potentiel de l'IA comme outil complémentaire.

Mots-clés

QCM, qualité d'évaluation, distracteurs plausibles, ChatGPT, intelligence artificielle

Abstract

This study, conducted among university students, examines the use of ChatGPT-4 in the creation of multiple-choice questions (MCQs) by comparing AI-generated MCQs to those designed by instructors and analyzing the quality of distractors and the difficulty indices. The study reveals that AI-generated MCQs achieve high success rates but feature less effective distractors, limiting the possibility of assessing analytical thinking skills. This research highlights the importance of human intervention to ensure high-quality MCQs while acknowledging the potential of AI as a complementary tool.

Keywords

MCQ, assessment quality, plausible distractors, ChatGPT, artificial intelligence

(a) Faculté de psychologie, des sciences de l'éducation et de la formation.

Introduction

Le questionnaire à choix multiples (QCM) est un ensemble de questions utilisées dans le contexte universitaire pour évaluer rapidement et objectivement les connaissances des personnes apprenantes sur divers sujets, tout en simplifiant la correction. Parmi les modalités d'évaluation, les QCM sont couramment valorisés en pédagogie et sont particulièrement privilégiés par rapport aux examens en présentiel pour leur objectivité et leur standardisation (Bachman, 1990). Leur conception repose sur l'expertise des enseignants et enseignantes, un processus complexe nécessitant des compétences spécialisées pour garantir la pertinence et l'équilibre des questions, chaque réponse étant jugée correcte ou incorrecte par consensus (Leclercq, 1986). Malgré certaines limites, notamment leur tendance à évaluer la mémorisation au détriment de compétences cognitives plus complexes telles que l'esprit critique et l'analyse, les QCM présentent néanmoins des avantages notables. Grâce à leur structure prédéfinie, ils minimisent les biais de correction, assurent une comparaison cohérente des performances des personnes apprenantes et renforcent à la fois la fiabilité et la validité des résultats (Geisinger et Carlson, 202, p. 467-468).

Avec l'essor des technologies numériques et de l'intelligence artificielle générative (IAg), les QCM ont retrouvé une place importante, notamment dans les cours hybrides à l'Université, où leurs effets sont jugés bénéfiques (Spanjers *et al.*, 2015). Depuis 2022, la démocratisation de l'IA avec des modèles comme ChatGPT-4 a transformé le rapport au numérique (Alexandre, 2023) et ouvre une nouvelle ère dans la création de QCM automatisés à partir de vastes bases de données. Dans ce contexte, cette recherche compare les QCM générés par ChatGPT-4 à ceux créés par des enseignantes en matière de qualité des distracteurs, d'indice de difficulté et de pertinence globale. La qualité des distracteurs correspond à l'efficacité des mauvaises réponses, qui doivent être plausibles mais clairement incorrectes. L'indice de difficulté mesure le pourcentage de personnes apprenantes répondant correctement à une question; une question trop facile ou trop difficile peut ne pas bien différencier les niveaux de compétence (Lord, 1952; Sharma, 2021). À travers une analyse comparative menée auprès de deux groupes de personnes apprenantes, nous explorons l'efficacité des deux méthodes et leurs impacts sur la qualité de l'évaluation et la réussite des personnes apprenantes.

Cadre théorique

Conception des QCM par l'équipe pédagogique

Les QCM offrent des avantages significatifs par rapport à d'autres méthodes d'évaluation, notamment en assurant l'objectivité grâce à des conditions égales pour toutes les personnes apprenantes et à une échelle de notation standardisée (Gilles et Charlier, 2020). Ils sont adaptés aux évaluations formatives, qui permettent aux enseignants et enseignantes d'ajuster leurs contenus pédagogiques en fonction des résultats des personnes apprenantes. Cela aide les personnes apprenantes à s'autoévaluer et à ajuster leurs stratégies d'apprentissage en conséquence (Laoufi et Elkachradi, 2017; Leclercq, 1986). Ce mode d'évaluation favorise un apprentissage continu et développe les compétences nécessaires à la réussite des évaluations sommatives, tandis que l'évaluation sommative offre une efficacité pour évaluer rapidement et uniformément de grandes cohortes de personnes apprenantes (Leclercq, 1986). Les résultats des QCM servent à déterminer les notes, certifier les acquis et comparer les performances entre classes et établissements, offrant aux personnes apprenantes la possibilité d'ajuster leurs stratégies d'apprentissage (Rey et Feyfant, 2014).

Baturin et Melnikova (2009) ont mené une analyse psycho-didactique sur la création de QCM par les enseignants et enseignantes, décrivant un processus en huit étapes clés, chronophage et nécessitant des compétences professionnelles spécifiques, souvent assurées par des équipes pédagogiques : 1) formulation d'un plan thématique sur les concepts essentiels, 2) établissement du modèle du QCM (nombre de questions, ratio de réponses correctes/incorrectes), 3) conception des questions, 4) vérification pour éliminer les biais et reformulation, 5) révision corrective et évaluation empirique après pré-test, 6) évaluation préliminaire de la difficulté, 7) correction orthographique, et 8) relecture. Dans ces étapes, la relecture, la réalisation de pré-tests et la révision sont essentielles pour garantir la validité et la fiabilité des questions, incluant la vérification de la cohérence interne et l'ajustement de la difficulté (Baturin et Melnikova, 2009). Zhilin (2023) suggère également une approche exigeante en matière de temps pour la conception de QCM, nécessitant un équilibre entre questions simples et complexes ainsi qu'une variété de modalités (choix unique ou multiple). Les questions doivent être exemptes de biais et démontrer un niveau de difficulté bien ajusté pour garantir une évaluation juste. Ce processus, bien que chronophage, est crucial pour élaborer des QCM de qualité. Mais il reste difficile de concevoir des questions valorisant les compétences analytiques ou de pensée critique, les QCM étant généralement plus adaptés à l'évaluation de connaissances factuelles et de compétences de mémorisation. De plus, leur format majoritairement textuel peut désavantager certains profils de personnes apprenantes, notamment les visuelles, qui assimilent mieux l'information à travers des schémas ou des illustrations, ainsi que les auditives, qui retiennent plus efficacement grâce aux explications orales et aux discussions. Cette approche d'évaluation favorise ainsi les personnes apprenantes qui sont à l'aise avec la reconnaissance de réponses, au détriment de celles qui apprennent mieux en structurant et en expliquant leurs idées (Zhilin, 2023).

L'émergence de ChatGPT dans l'éducation

L'intégration de l'IA dans les systèmes éducatifs est un enjeu majeur. Alvarez (2023) insiste sur la nécessité d'adapter les programmes scolaires pour permettre aux personnes apprenantes de développer des compétences numériques spécifiques telles que la programmation, afin de tirer pleinement parti de ces outils. Il s'agit également de maintenir un équilibre avec les méthodes d'enseignement traditionnelles pour garantir une approche pédagogique innovante. Les innovations technologiques ont permis l'intégration de l'IAg et de l'analyse de données dans les évaluations éducatives, menant à la création de systèmes adaptatifs personnalisés en fonction du niveau de chaque personne apprenante (Jabraoui et Vandapuye, 2024).

Depuis son lancement en novembre 2022, ChatGPT incarne un tournant dans l'utilisation des technologies numériques. Il peut traiter efficacement des volumes massifs de données et générer des textes pertinents et adaptés. Alimenté par des modèles linguistiques avancés et entraîné sur un large éventail de données accessibles en ligne (Anctil, 2023; Lelepary *et al.*, 2023), ChatGPT est capable de répondre aux requêtes des utilisateurs et utilisatrices en produisant des réponses textuelles de haute qualité qui imitent souvent le style d'écriture humain. Cette technologie a ainsi profondément modifié la manière dont nous comprenons et générons du texte (Alexandre, 2023). D'après Brown *et al.* (2020), la capacité des modèles de ChatGPT à générer du contenu pertinent et adapté grâce à l'apprentissage profond a des applications directes en éducation. Celui-ci peut accomplir diverses tâches telles que la synthèse de documents, la rédaction de contenu technique, l'analyse de supports, la création de questions ainsi que les calculs statistiques, entre autres (Belkaim, 2023). Cela rend ChatGPT utile pour la création de tests éducatifs et aide à évaluer les connaissances des personnes apprenantes dans divers domaines.

Génération de QCM avec ChatGPT

L'utilisation de ChatGPT pour la création de QCM implique une approche avancée qui intègre des techniques et algorithmes divers issus du traitement du langage naturel et de l'apprentissage automatique (Gefen, 2023). D'après Du *et al.* (2017), générer automatiquement des questions à partir de textes nécessite non seulement de comprendre le contenu, mais aussi de reformuler les informations clés sous forme de questions, une compétence que les grands modèles de langage ont nettement améliorée. Cependant, la création de distracteurs à la fois plausibles et incorrects demeure un défi. Comme l'indique Gierl *et al*. (2017), cela souligne l'importance d'une finesse linguistique afin d'éviter de produire des réponses trop facilement identifiables comme fausses. Selon Petrov *et al*. (2011), l'analyse syntaxique joue un rôle crucial dans la compréhension des structures linguistiques, permettant une génération de contenu plus nuancée et adaptée.

Lors de la rédaction, il est important de formuler les requêtes (*prompts*) qui sont adressées à une intelligence artificielle, de manière claire et précise, pour obtenir des réponses pertinentes. Une requête trop vague risque de donner des réponses imprécises, tandis qu'un excès de détails peut détourner l'attention du modèle des points importants. De plus, une requête formulée avec un biais peut influencer la réponse de manière incorrecte. Les spécialistes qui entraînent les modèles de langage et conçoivent les requêtes poursuivent un double objectif à travers ChatGPT : d'une part, favoriser une réflexion critique éclairée chez les utilisateurs et utilisatrices; d'autre part, encourager l'émergence de nouveaux usages en facilitant leur appropriation de l'outil (Gefen, 2023). Ce processus facilite une meilleure compréhension et une maîtrise accrue de ChatGPT, notamment en perfectionnant « l'art de la requête », c'est-à-dire l'élaboration de requêtes optimales. Nous considérons que les requêtes sont la clé pour utiliser efficacement ChatGPT. Leur bonne formulation permet d'obtenir des réponses précises et utiles, ce qui simplifie grandement l'interaction avec l'IA.

Pour créer le QCM, nous avons utilisé la méthode de génération augmentée par récupération (RAG, *retrieval-augmented generation*) à l'aide de ChatGPT. C'est une méthode qui consiste à utiliser un modèle préentraîné pour une tâche spécifique à partir de ressources de données supplémentaires. Nous exploitons ainsi les connaissances qu'il a déjà acquises sur de vastes ensembles de données sans le réentraîner (*cf.* figure 1). Les modèles de langage les plus performants commencent généralement par être entraînés sur un large éventail de sujets, ce qui leur confère une base généraliste. Par la suite, ils sont affinés et spécialisés pour répondre à des domaines précis. Ce processus est comparable au parcours d'un étudiant en sciences humaines qui, après avoir acquis une formation générale, se spécialise progressivement dans des disciplines comme la science, la technologie, l'ingénierie ou les mathématiques au cours de ses études supérieures (Latif et Zhai, 2024).

Objectifs de la présente étude

L'objectif de cette recherche est de comparer la qualité des QCM générés par ChatGPT et ceux créés par des enseignantes, en se basant sur deux indicateurs principaux : l'indice de facilité, qui mesure le pourcentage de bonnes réponses pour évaluer la difficulté, et la qualité des distracteurs plausibles, ou options incorrectes mais crédibles qui testent la compréhension des personnes apprenantes. En comparant la formulation et la pertinence de ces distracteurs, la recherche examine la capacité de chaque type de QCM à encourager la réflexion critique et à éviter des réponses de simple élimination.

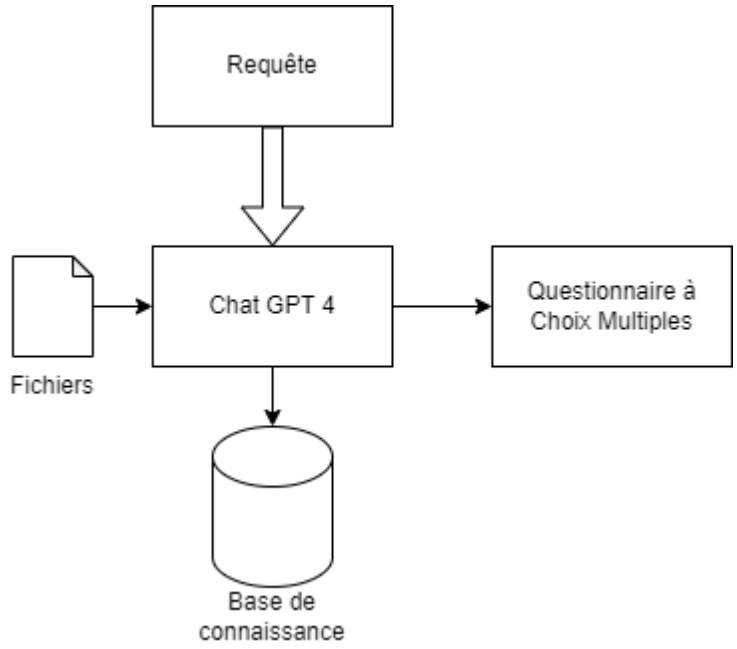

Figure 1
Méthode de génération de questionnaires à choix multiples (QCM) augmentée par récupération (RAG, Retrieval-augmented generation)

Méthodologie

Personnes apprenantes

Cette étude a été réalisée dans le cadre du cours *Gouvernance en éducation : les aspects numériques* à l'Université de Lille, organisé au début d'avril 2024. Elle a été menée auprès de 152 personnes apprenantes inscrites en deuxième année de licence *Sciences de l'éducation et de la formation*, réparties en deux groupes égaux : le groupe ChatGPT et le groupe traditionnel, avec 76 personnes apprenantes dans chaque groupe. L'âge des personnes apprenantes dans chaque groupe est compris entre 19 et 22 ans. Par groupe, 73 femmes et 3 hommes ainsi que trois enseignantes âgées de 39 à 50 ans en sciences de l'éducation et de la formation ont participé à la conception des QCM.

Matériels

QCM généré par ChatGPT-4

Pour générer un QCM avec l'IA, nous avons utilisé une requête spécifique afin de produire 20 questions, chacune ayant quatre options de réponse. Ces questions portaient sur les thématiques abordées dans le fichier que nous avions chargé en parallèle avec la requête dans ChatGPT-4. La première requête utilisée était : « Créez un QCM de 20 questions, avec trois réponses incorrectes et une correcte, en utilisant les options A, B, C et D. » Après avoir constaté que les questions générées par ChatGPT étaient trop simples, nous avons reformulé notre demande avec une seconde requête : « Créez un QCM de 20 questions, avec trois réponses incorrectes et une correcte, en utilisant les options A, B, C et D, en vous basant sur le document fourni, avec indice de difficulté élevé et des distracteurs plausibles de qualité. » Les résultats issus de cette deuxième requête ont été utilisés pour le groupe ChatGPT. Pour générer les questions de QCM, la méthode à **RAG** a été employée. Cette méthode combine une étape de recherche d'informations précises à partir d'un

document source avec la génération de contenu par l'IA, permettant ainsi de s'assurer que les questions formulées s'alignent étroitement avec le contenu fourni (*cf.* figure 1).

QCM traditionnel élaboré par des enseignantes

Les enseignantes ont également conçu un QCM de 20 questions en se basant sur le même contenu de cours. Ce nombre de questions a été déterminé en fonction de la pertinence des thèmes abordés dans le document. Les trois enseignantes ont participé à toutes les étapes de la conception du QCM traditionnel, en commençant par l'élaboration d'un plan thématique et la sélection des réponses correctes et incorrectes. Ces questions ont ensuite été réparties entre les enseignantes. Le modèle de QCM utilisé était identique à celui du QCM ChatGPT, avec des réponses au format A, B, C et D. Après la rédaction initiale des questions, les enseignantes ont procédé à une vérification minutieuse des items pour éliminer d'éventuels biais et reformuler certains éléments si nécessaire. L'objectif était de créer un QCM couvrant différents niveaux de difficulté, allant de facile à difficile. Enfin, chaque enseignante a relu les questions pour assurer la cohérence et la qualité du questionnaire final.

Protocole

Respect des principes éthiques en recherche

Cette expérimentation suit les principes éthiques de la recherche en éducation. Les personnes apprenantes ont été informées des objectifs de l'étude et leur consentement éclairé a été obtenu. La confidentialité des données a été assurée, garantissant leur anonymat et la protection de leurs résultats. L'étude a respecté les procédures institutionnelles et administratives incluant l'information et l'invitation à participer à ce QCM. Une approbation éthique a été obtenue auprès de la direction du Département. Enfin, aucune influence n'a été exercée sur la réussite universitaire, assurant une participation volontaire.

Les deux groupes ont été soumis aux mêmes conditions d'évaluation. Les questions ont été distribuées aux personnes apprenantes des deux groupes. Le groupe ChatGPT a reçu le QCM généré par ChatGPT, tandis que le groupe traditionnel a reçu celui élaboré par des enseignantes. Les deux groupes se sont installés dans le même amphithéâtre; ils avaient 30 minutes pour répondre aux 20 questions. Les résultats ont ensuite été collectés pour une analyse complète.

Critères d'évaluation

Les réponses incorrectes dans ces deux types de QCM doivent agir comme des distracteurs plausibles, c'est-à-dire qu'elles doivent sembler réalistes pour tester véritablement la compréhension des personnes apprenantes. L'importance de la plausibilité des distracteurs réside dans leur capacité à évaluer plus efficacement les connaissances. La qualité des questions dépend donc non seulement de la précision de la réponse correcte, mais aussi de la pertinence des choix incorrects, qui doivent être suffisamment crédibles. Un indice de difficulté élevé suggère que la question est facile, alors qu'un indice faible suggère qu'elle est difficile. Ces indices servent à évaluer la difficulté perçue de chaque question.

Résultats

Résultats globaux des performances

En analysant les résultats globaux, nous avons constaté que les personnes apprenantes du groupe ChatGPT ont montré des performances supérieures pour la majorité des questions par rapport à ceux du groupe traditionnel. La note moyenne obtenue avec le QCM de ChatGPT est de 16,6 sur 20 (écart-type 1,4), contre 10,8 sur 20 (écart-type 1,8) pour le QCM traditionnel (*cf.* figure 2a).

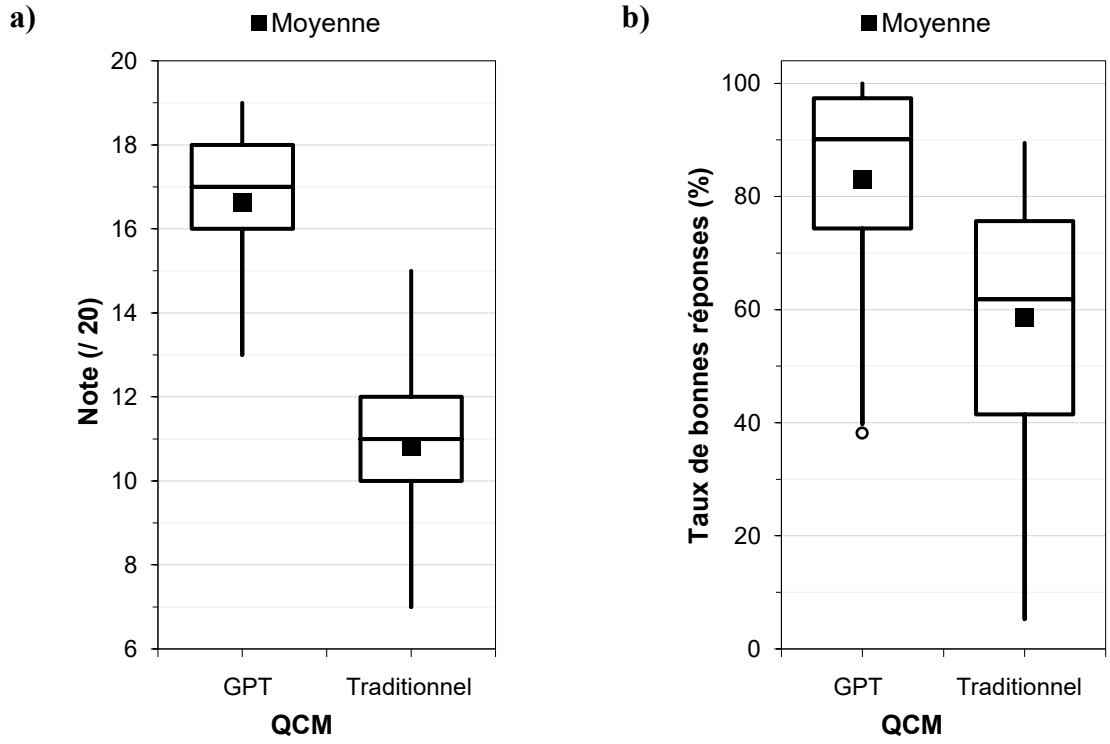

Figure 2
Répartition pour le QCM ChatGPT et le QCM traditionnel : a) des notes (N = 76); b) des taux de bonnes réponses par question (N = 20)

Les moyennes de bonnes réponses pour chaque question sont de 83 % (écart-type 17 %), pour le groupe ChatGPT, et de 58 % (écart-type 23 %) pour le groupe traditionnel (figure 2b). Les taux de bonnes réponses par question, pour le QCM généré par ChatGPT, varient de 38 % à 100 % avec une majorité de questions où plus de 90 % des personnes apprenantes ont répondu correctement. Dans le QCM traditionnel, les taux de bonnes réponses sont plus dispersés et souvent inférieurs, plusieurs questions affichant moins de 60 % de réponses correctes.

Résultats sur l'indice de facilité

L'indice de facilité est une mesure qui indique le niveau de difficulté d'une question. Il permet de déterminer si une question est facile ou difficile pour les personnes apprenantes. Plus l'indice de facilité est élevé, plus la question est jugée facile, car un grand nombre de personnes apprenantes y ont répondu correctement.

Nous avons analysé les indices de facilité de deux séries de questions à choix multiples, le QCM traditionnel et le QCM ChatGPT, en calculant pour chaque question le rapport entre le nombre de réponses correctes et le nombre total de réponses. Les valeurs des indices de facilité par question

dans le QCM traditionnel varient de 5 % à 89 %, montrant une distribution étendue des niveaux de difficulté. En revanche, le QCM ChatGPT présente des indices de facilité plus élevés et homogènes, oscillant entre 38 % et 100 %. Nous avons comparé les résultats question par question et constaté que, dans le QCM ChatGPT, les questions atteignent des indices de facilité plus élevés. Certaines questions affichent même un indice de facilité de 100 %, indiquant une réponse correcte de l'ensemble des personnes apprenantes, alors que le QCM traditionnel ne présente aucun cas de cette nature. Cette comparaison montre une tendance générale vers une accessibilité accrue dans le QCM ChatGPT, avec des valeurs d'indice de facilité qui restent plus proches du maximum possible pour la plupart des questions.

Résultats sur les distracteurs

Tel que nous l'avons démontré, dans le QCM ChatGPT, les résultats révèlent que les questions atteignent souvent des indices de facilité élevés, certaines obtenant même un score de 100 %, ce qui signifie que toutes les personnes apprenantes ont répondu correctement. Cette tendance peut être liée à la formulation des distracteurs, souvent orientés négativement, tandis que les bonnes réponses sont formulées positivement. Par exemple, une question affiche un indice de facilité élevé. Dans la question « Quelle est l'importance de la personnalisation de l'apprentissage? », les distracteurs sont formulés de manière négative : b) *« **inefficace** et coûteuse » (0 %),* c) ***« seulement** importante pour les jeunes élèves » (0 %),* et d) *« **réduit** l'autonomie des personnes apprenantes » (0 %).* Par contre, la réponse correcte, positive et alignée avec les avantages reconnus, a) *« **améliore** l'engagement et l'efficacité de l'apprentissage »*, choisie par 100 % des personnes répondantes, devient évidente. Ce choix de distracteurs négatifs dans le QCM ChatGPT oriente implicitement les personnes apprenantes vers la bonne réponse, rendant la réponse facile à repérer. Les distracteurs de ce QCM montrent une attractivité moindre, avec une moyenne de 17 % de réponses incorrectes. Dans plusieurs questions, certains distracteurs n'ont été choisis par aucune personne apprenante (0 % de sélection), ce qui signifie qu'il était trop évident qu'ils étaient incorrects. Cela remet en question leur plausibilité, car un bon distracteur doit sembler suffisamment crédible pour induire une hésitation chez certaines personnes apprenantes. Le maximum atteint par les distracteurs est de 62 %, observé dans des questions où l'option incorrecte est relativement crédible. Le taux de réussite moyen pour le QCM ChatGPT est plus élevé (83 %), suggérant que les distracteurs jouent un rôle limité dans la complexité du test et que les personnes apprenantes repèrent souvent la bonne réponse sans grande hésitation. De plus, aucun élément sans réponse n'est enregistré dans le QCM ChatGPT, ce qui laisse penser que chaque question est perçue comme claire et les options incorrectes comme moins engageantes.

À l'inverse, dans le QCM traditionnel, les distracteurs sont conçus par les enseignantes de manière plus variée et crédible, attirant davantage de personnes apprenantes vers les mauvaises réponses. Pour garantir leur plausibilité, ils ont été construits selon différentes approches. Certains sont dérivés de bonnes réponses à d'autres questions du QCM, induisant une confusion naturelle; d'autres intègrent des idées courantes mais non étudiées; enfin, certains s'inspirent de notions déjà abordées, rendant les réponses trompeusement crédibles. Cela rend le QCM plus exigeant et équilibré. Les distracteurs montrent une attractivité variée, avec une moyenne de 31 % de réponses incorrectes. Les distracteurs les moins attractifs captent un minimum de 5 % des réponses (notamment dans certaines questions où des réponses sont facilement écartées), tandis que les plus attractifs peuvent atteindre jusqu'à 84 %. Ce niveau élevé de sélection montre que certains distracteurs sont perçus comme fortement crédibles, incitant de nombreuses personnes apprenantes à les choisir, ce qui signifie qu'ils jouent bien leur rôle en détournant l'attention de la bonne réponse. Le taux de bonnes réponses moyen pour le QCM traditionnel est de 58 %, ce qui suggère

que les personnes apprenantes ont été modérément défiées par les distracteurs. Les éléments sans réponse sont également présents, avec une moyenne de 11 %, montrant une certaine indécision face aux questions, ce qui renforce l'efficacité des distracteurs.

Discussion

Dans cette étude, nous avons examiné les performances des personnes apprenantes confrontées à deux types de QCM. Celles qui ont répondu au QCM ChatGPT ont des résultats globaux plus élevés que celles qui ont répondu au QCM traditionnel.

Le fait qu'aucune question du QCM ChatGPT n'ait été laissée sans réponse suggère que les personnes apprenantes se sentaient plus assurées ou moins intimidées par les questions générées par l'IA. À l'inverse, celles qui ont utilisé le QCM traditionnel ont laissé plusieurs questions sans réponse, ce qui peut refléter une difficulté perçue plus élevée ou un certain degré d'incertitude face à certaines questions. Les absences de réponse indiquent ici une complexité accrue dans la formulation des items, ainsi qu'un niveau de confiance plus variable parmi les personnes apprenantes, ce qui souligne le potentiel du QCM traditionnel à explorer plus finement les nuances dans la compréhension des personnes apprenantes (Malcourant, 2020).

L'analyse des indices de facilité soutient ces observations. Les questions du QCM ChatGPT montrent une homogénéité des indices de facilité, ce qui peut restreindre leur capacité à discriminer entre différents niveaux de compétences. En revanche, le QCM traditionnel propose une plus grande variété de difficultés, offrant une évaluation plus nuancée des connaissances. Une diversité dans les degrés de difficulté est en effet cruciale pour mettre en lumière les besoins spécifiques d'apprentissage et pour cibler plus précisément les domaines nécessitant un approfondissement (Régnier, 2013).

Les distracteurs dans le QCM ChatGPT étaient souvent formulés de manière négative, tandis que les réponses correctes étaient formulées positivement, créant parfois un biais qui pouvait orienter les personnes apprenantes vers la bonne réponse de manière inconsciente. En revanche, le QCM traditionnel se distinguait par des distracteurs crédibles et diversifiés, conçus de manière à générer une répartition plus équilibrée des réponses. Cette qualité de conception a permis d'identifier plus finement les lacunes dans la compréhension des personnes apprenantes et d'évaluer plus précisément leurs compétences en matière de réflexion critique et d'analyse (Boch et Sorba, 2020; Grühn et Cheng, 2015). Cette rigueur dans la conception des distracteurs met en lumière le rôle essentiel des enseignants et enseignantes dans l'élaboration de tests diagnostiques approfondis, qui non seulement évaluent les acquis, mais aussi permettent de cibler précisément les besoins d'apprentissage. L'intervention humaine est essentielle pour garantir que les outils d'évaluation conservent une pertinence pédagogique élevée, en intégrant des niveaux de difficulté variés et des distracteurs mieux adaptés aux objectifs d'apprentissage.

Nous considérons que, pour optimiser les QCM générés par des machines, l'intégration de principes pédagogiques éprouvés, comme l'ajustement des niveaux de difficulté et la création de distracteurs crédibles, ainsi qu'une collaboration étroite entre concepteurs numériques et enseignants, renforceraient la pertinence de ces outils tout en préservant leur valeur pédagogique (Amadieu et Tricot, 2014).

Ces résultats soulèvent des questions importantes quant à la conception des évaluations. Il est crucial de reconnaître que les QCM ne servent pas uniquement à mesurer les acquis, mais aussi à déterminer les domaines nécessitant une attention particulière et à enrichir l'apprentissage des

personnes apprenantes, notamment grâce au soutien du personnel enseignant, qui joue un rôle essentiel dans l'accompagnement et la remédiation pédagogique. Une approche équilibrée dans la conception des évaluations contribue à une meilleure compréhension et à un apprentissage plus profond.

Limites et perspectives

Les consignes (la mise en place des indices de difficulté élevés, la création de distracteurs plausibles, etc.) données aux enseignantes et à ChatGPT étaient identiques, excluant ainsi tout biais lié aux directives initiales. Toutefois, bien que cette recherche mette en évidence une différence significative entre les performances aux QCM générés par ChatGPT et ceux conçus par des enseignantes, certaines limites doivent être prises en compte. D'une part, la formulation des items peut influencer les résultats, notamment en fonction de la manière dont les enseignantes interprètent et appliquent les consignes, ou encore de la structuration automatique des questions par ChatGPT-4. De plus, la perception du niveau de difficulté des questions peut varier en fonction de leur formulation, même si les objectifs d'évaluation restent les mêmes.

D'autre part, l'expérience des enseignants et enseignantes dans la conception des QCM pourrait jouer un rôle. Certains d'entre eux, plus expérimentés, sont en mesure d'élaborer des items plus efficaces, tandis que ChatGPT génère des questions selon des algorithmes fondés sur des données, mais sans réflexion humaine ni esprit critique, ce qui ne lui permet pas toujours de prendre en compte les nuances pédagogiques. D'autres facteurs contextuels tels que le niveau d'engagement des personnes apprenantes, leur familiarité avec les QCM ou encore leur perception des tests pourraient également influencer les résultats. Ces aspects mériteraient d'être explorés dans de futures recherches afin d'affiner la comparaison entre la génération humaine et la génération automatisée des QCM.

Conclusion

L'intégration de ChatGPT pour la génération de QCM offre des avantages indéniables en matière de rapidité et d'efficacité. Cependant, cette étude, basée sur une analyse comparative, met en lumière plusieurs limites importantes. Le QCM généré par ChatGPT simplifie souvent les réponses en proposant des options incorrectes négatives, ce qui rend la réponse correcte plus évidente et élève les indices de facilité. Cependant, une autre limite est associée à l'utilisation de la méthode d'ajustement basée sur le RAG. En effet, ce dernier peut induire un surapprentissage des données spécifiques, réduisant la flexibilité du modèle face à de nouvelles informations. Si les données employées pour le RAG sont biaisées ou de qualité médiocre, le modèle ajusté risque d'hériter de ces défauts, entraînant des résultats inexacts ou, potentiellement, des biais discriminatoires.

Dans tous les cas, l'intervention du corps enseignant reste cruciale pour vérifier et ajuster les questions, assurant ainsi une évaluation plus nuancée et équitable des connaissances des personnes apprenantes, et bonifier l'évaluation en ajoutant les volets analytique et réflexif, amenant ainsi les personnes apprenantes à un niveau supérieur sur le plan des acquis et du raisonnement. Bien que l'automatisation puisse offrir des avantages considérables, il est impératif de corriger les biais potentiels introduits par l'IA pour maintenir l'intégrité et l'efficacité des évaluations éducatives. Cette étude souligne la nécessité d'une approche équilibrée, combinant l'efficacité de l'IA avec l'expertise humaine pour optimiser la création et l'utilisation des QCM dans le domaine éducatif.

Le corps enseignant, fort de son expertise et de sa compréhension nuancée des besoins pédagogiques, apporte une valeur ajoutée essentielle qui ne peut être entièrement reproduite par

une machine. Ce travail souligne ainsi le rôle central et irremplaçable des enseignants et enseignantes dans la conception d'outils d'évaluation de qualité. L'intelligence artificielle, bien qu'utile, doit être perçue comme un complément et non un substitut à l'intuition et au discernement humain. En conclusion, cette étude plaide pour une collaboration harmonieuse entre les technologies avancées et l'expertise des enseignants et enseignantes, garantissant ainsi des évaluations justes, équilibrées et véritablement adaptées aux personnes apprenantes.

Notes

Disponibilité des données

Les données collectées au cours de la présente recherche et sur lesquelles l'article s'appuie sont disponibles à **https://osf.io/hnc4b/files/osfstorage**

Références

Alexandre, L. (2023). *La guerre des intelligences à l'heure de ChatGPT*. JC Lattès.

Alvarez, L. (2023). L'IA à l'école ou l'école de l'IA. *Universitas*, *2023*(3), 25-27. **https://unifr.ch/universitas/...**

Amadieu, F. et Tricot, A. (2014). *Apprendre avec le numérique : mythes et réalités*. Retz.

Anctil, D. (2023). L'éducation supérieure à l'ère de l'IA générative. *Pédagogie collégiale*, *36*(3). **https://eduq.info/xmlui/handle/11515/38833**

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Baturin, N. A. et Melnikova, N. N. (2009). Tekhnologiya razrabotki testov: chast'I [The technology of test development: Part I]. *Bulletin of the South Ural State University. Series "Psychology"*, *30*(163), 4-14.

Belkaim, L. (2023). ChatGPT à l'université : ami ou ennemi? *Analele Universităţii din Craiova, seria Psihologie-Pedagogie*, *45*(2), 22-30. **https://aucpp.ro/...**

Boch, F. et Sorba, J. (2020). Tester la compétence lexicale des adultes francophones : réflexion sur le choix des distracteurs dans un test à choix multiples. *Lidil*, (62). **https://doi.org/10.4000/lidil.8023**

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Dans H. Larochelle, M. Ranzatom, R. Hadsell, M. F. Balcan et H. Lin (dir.), *Advances in Neural Information Processing Systems 33 – Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)* (p. 1877-1901). **https://proceedings.neurips.cc/...**

Du, X., Shao, J. et Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. Dans R. Barzilay et M.-Y. Kan (dir.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Volume 1: Long papers (1342-1351). **https://aclanthology.org/P17-1123**

Gefen, A. (2023). *Vivre avec ChatGPT : séduire, penser, créer, se cultiver, s'enrichir… L'intelligence artificielle aura-t-elle réponse à tout?* Éditions de l'Observatoire.

Geisinger, K. F. et Carlson, J. F. (dir.). (2021). *The twenty-first mental measurements yearbook*. University of Nebraska Press.

Gierl, M. J., Bulut, O., Guo, Q. et Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education. *Review of Educational Research*, *87*(6), 1082-1116. **https://doi.org/gdxrhj**

Gilles, J.-L. et Charlier, B. (2020). Dispositifs d'évaluation à distance à correction automatisée versus non automatisée : analyse comparative de deux formes emblématiques. *Évaluer – Journal international de recherche en éducation et formation*, (hors-série nº 1), 143-154. **http://journal.admee.org/...**

Grühn, D. et Cheng, Y. (2015, 15 mars). *EPP-APS. L'auto-correction des QCM* (L. Libeyre, trad.). Association for Psychological Science. **https://psychologicalscience.org/...** [Article orginal paru en 2014 dans *Teaching of Psychology*, *41*(4), 335-339. **https://doi.org/n7z4**]

Jabraoui, S. et Vandapuye, S. (2024). L'intelligence artificielle dans l'enseignement : histoire et présent, perspectives et défis. *Dossiers de recherches en économie et management des organisations*, *9*(1), 118-128. **https://doi.org/10.34874/PRSM.dremo-vol9iss1.1777**

Laoufi, A. et Elkachradi, R. (2017). Pratiques et défis de l'usage des technologies numériques pour l'évaluation pédagogique : cas des universités marocaines. *The Journal of Quality in Education*, *7*(9). **https://doi.org/10.37870/joqie.v7i9.8**

Latif, E. et Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, article 100210. **https://doi.org/10.1016/j.caeai.2024.100210**

Leclercq, D. (1986). *La conception des questions à choix multiples*. Labor.

Lelepary, H. L., Rachmawati, R., Zani, B. N. et Maharjan, K. (2023). ChatGPT: Opportunities and challenges in the learning process of Arabic language in higher education. *Journal International of Lingua and Technology*, *2*(1), 11-23. **https://doi.org/10.55849/jiltech.v2i1.439**

Lord, F. M. (1952). The relationship of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, *17*(2), 181-194. **https://doi.org/10.1007/BF02288781**

Malcourant, É. (dir.). (2020). *QCM or not QCM? Processus de conception d'une évaluation par QCM* (cahiers du LLL nº 10). Presses universitaires de Louvain. **https://hdl.handle.net/...**

Petrov, S., Das, D. et McDonald, R. (2011). A universal part-of-speech tagset. Dans N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk et S. Piperidis (dir.), *Proceedings of LREC 2012 – Eighth International Conference on Language Resources and Evaluation* (p. 2089-2096). **http://lrec-conf.org/proceedings/lrec2012/summaries/274.html**

Régnier, N. (2013, août). *Systèmes de réponse instantanée pour une pédagogie active* [communication]. *CFM 2013 – 21e Congrès français de mécanique*, Courbevoie, France. **https://hal.science/CFM2013/hal-03441139v1**

Rey, O. et Feyfant, A. (2014). Évaluer pour (mieux) faire apprendre. *Dossier de veille de l'IFÉ* (94). **https://ens-lyon.hal.science/ensl-01576226**

Sharma, L. R. (2021). Analysis of difficulty index, discrimination index and distractor efficiency of multiple choice questions of speech sounds of English. *International Research Journal of MMC*, *2*(1), 15-28. **https://doi.org/10.3126/irjmmc.v2i1.35126**

Spanjers, I. A. E., Könings, K. D., Leppink, J., Verstegen, D. M. L., de Jong, N., Czabanowska, K. et van Merriënboer, J. J. G. (2015). The promised land of blended learning: Quizzes as a moderator. *Educational Research Review, 15*, 59-74. **https://doi.org/10.1016/j.edurev.2015.05.001**

Zhilin, V. V. (2023). Prilozhenie dlya generatsii testovykh zadaniy s pomoshch'yu modeli ChatGPT [Application pour la génération de questions de test à l'aide du modèle ChatGPT]. Dans V. V. Zhilin (dir.), *Molodoy issledovatel': ot idei k proektu* [Le jeune chercheur: de l'idée au projet] (p.97-100). Université d'État Mari.