UNIVERSITÉ DE LILLE

FACULTÉ DE MÉDECINE HENRI WAREMBOURG

Année : 2021

THÈSE POUR LE DIPLÔME D'ÉTAT

DE DOCTEUR EN MÉDECINE

Docimologie dans les formations pour les professions de santé

Présentée et soutenue publiquement le 14/06/2021 à 16h
au Pôle Formation
par Raphaël BENTEGEAC

_____

JURY

Président :
Monsieur le Professeur Philippe AMOUYEL
Assesseurs :
Monsieur le Professeur Grégoire FICHEUR
Monsieur le Docteur Pierre BALAYE
Directeur de thèse :
Monsieur le Professeur Emmanuel CHAZARD

_____

Avertissement

La Faculté n’entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs.

Sigles

FO	Fiche optique, synonyme de copie d'étudiant
UE	Unité d'enseignement

Sommaire

Introduction

1 Introduction générale

1.1 Professions de santé

En France et en Europe certaines professions de Santé sont réglementées(1). Leur accès est conditionné à la validation d'un diplôme d'études supérieures placées sous la responsabilité conjointe du Ministère de l'Enseignement Supérieur et de la Recherche (MESR) et du Ministère de la Santé. La plupart des filières suivent des programmes nationaux et délivrent des « diplômes d'État »(2).

Par ailleurs, ces filières sont en général sélectives, l'orientation des étudiants se jouant avec des concours locaux ou nationaux, comme par exemple en médecine avec la PASS (anciennement PACES) en première année et les ECNi en sixième année.

Dans ce contexte, l'organisation du contrôle des connaissances et des épreuves de sélection et d'orientation est au centre de la formation des professionnels de santé. Ces épreuves se trouvent donc naturellement dans l'accès aux études, la poursuite des études, l'orientation dans les spécialités et la validation finale.

1.2 Docimologie

La docimologie, l'étude des épreuves, est la discipline scientifique consacrée à l'étude du déroulement des évaluations en pédagogie.(3)

Dans le cas idéal, le système académique sélectionnerait et orienterait "la bonne personne au bon endroit", tout en prévenant le manque de compétence dans les métiers visés. Pour parvenir à un tel système, il est logiquement nécessaire d'évaluer les compétences attendues des étudiants dans leur futur métier. Cet objectif d'évaluation des compétences est particulièrement souhaitable dans les études de santé, au vu de la criticité de ces métiers(4).

Cependant si cet objectif est comme nous l'avons dit extrêmement souhaitable, sa réalisation est tout sauf triviale(5). En effet, pour évaluer strictement ces compétences, il faudrait pouvoir évaluer les décisions des professionnels en fonction de leurs conséquences réelles probables sur leurs patients et les comparer soit entre eux, soit à un niveau minimum requis(6).

Une solution alternative est d'utiliser des évaluations par variable proxy(7), en trouvant des proxy très prédictifs du niveau réel de compétences, comme par exemple une évaluation de connaissances pertinentes ou, mieux(8) des simulations. L'introduction d'un tel proxy représente un risque d'erreur s'il n'est pas correctement choisi.

Effectivement, un proxy peut être non pertinent (connaissance inutile pour le champ de compétences évalué), voire, pire, représenter un contre-sens (connaissance contre-productive pour le champ de compétences évalué). On pensera notamment à des connaissances pseudo-scientifiques(9) potentiellement vectrices de mauvaises décisions.

1.3 Méthodes d'évaluations fréquemment utilisées

Nous allons ici détailler quelques méthodes d'évaluation et décrire leur capacité à s'approcher de l'objectif souhaité d'évaluation des compétences.

Mais premièrement, insistons une dernière fois sur la différence entre évaluation des compétences et évaluation des connaissances.

L'évaluation des compétences est l'évaluation des décisions prises par les professionnels de santé dans le cadre de leur métier. L'optimisation de ces décisions est l'objectif principal de la formation et du processus d'orientation. Nous allons voir par la suite qu'évaluer directement les compétences est particulièrement difficile.

L'évaluation des connaissances est l'évaluation de l'acquisition de données et de raisonnements. L'objectif principal de la formation n'étant pas de simplement acquérir des connaissances, cette évaluation est en réalité un proxy plus ou moins pertinent selon les cas. L'avantage principal de l'évaluation des connaissances sur l'évaluation directe des compétences est qu'elle est plus simple à mesurer de manière objective.

1.3.1 Mise en situation réelle

L'évaluation par mise en situation réelle est probablement la méthode s'approchant le plus de l'évaluation des compétences.

En effet, il s'agit de la méthode directe n'utilisant aucun proxy. Si cette approche retire effectivement le risque de proxy, elle comportera toujours le risque résiduel de mauvaise évaluation des décisions en ne jugeant pas correctement ses conséquences probables, par exemple en jugeant uniquement les résultats réels tout en négligeant les fluctuations dues au hasard (bonne décision mais malchance ou l'inverse).

Par ailleurs, cette méthode a une capacité extrêmement faible à être étendue à un grand nombre d'étudiants (*scalability*[1]) et il serait quasiment impossible d'évaluer chaque étudiant de cette manière dans des promotions de PACES avec plus de mille étudiants.

Une autre limite très importante est la difficulté à mesurer la qualité des décisions prises de manière objective et reproductible.

On notera cependant que l'évaluation par les pairs en stage est débutée dès la deuxième année de médecine.

1.3.2 Mise en situation simulée

L'évaluation par mise en situation simulée(8) est quasiment identique à la méthode décrite précédemment, à la différence que la situation n'est pas réelle mais simulée.

Elle a pour avantage d'offrir une meilleure *scalability* que la méthode précédente, avec la possibilité d'évaluer plusieurs situations rapidement en série. Elle permet également d'évaluer des situations rares, ou posant un problème éthique ou légal.

Elle introduit cependant un risque de proxy, la situation étant souvent un "cas d'école" et paradoxalement potentiellement atypique de la "vraie vie".

Par ailleurs, la baisse en pression due à l'absence de conséquences réelles peut ajouter à ce risque de proxy, la personne évaluée pouvant prendre des décisions différentes sous pression.

1 Ce terme n'ayant pas de synonyme strict en Français, afin d'éviter d'utiliser systématiquement la périphrase synonyme, nous emploierons le terme anglais "*scablability*" dans le présent document.

Néanmoins on pourra noter que cette méthode reste très proche de l'évaluation réelle et limite grandement le risque de proxy par rapport aux méthodes suivantes.

1.3.3 Questions rédactionnelles

Ces variantes reposent sur des questions ouvertes ou fermées auxquelles les étudiants répondent parfois de longs textes argumentatifs et nécessitant de part ce fait une correction par un enseignant compétant et capable de discerner une copie suffisante d'une copie insuffisante.

En médecine, il s'agit dans la majorité des cas de questions fermées à réponses courtes, mais il est possible également d'y avoir des questions ouvertes ou à réponse longue.

Cette modalité a quelques avantages :

- elle permet une évaluation très fine des connaissances réelles de l'étudiant
- elle est très peu coûteuse en temps et en effort avant l'épreuve : le sujet d'épreuve peut être déterminé très rapidement, et la grille de correction peut être établie a posteriori. Il est même possible de tenir compte d'éventuelles erreurs faites au moment de la définition de l'épreuve.

Cependant, cette modalité pose plusieurs problèmes :

- le problème de la mise à l'échelle, en termes de nombres d'individus
- le problème de la mise à l'échelle, en terme de couverture du programme : mécaniquement, si chaque question doit être traitée en 20 minutes, une épreuve d'une heure ne permet de traiter que 3 questions. A contrario, une Unité d'Enseignement couvre volontiers plusieurs centaines de connaissances unitaires.
- le problème de reproductibilité de la mesure (lorsque le même évaluateur corrige deux fois la même copie, il peut donner deux notes différentes)
- le problème de reproductibilité inter-opérateur (lorsque deux évaluateurs corrigent la même copie, ils peuvent donner deux notes différentes)
- le problème de l'harmonisation des notes, dans le cas fréquent où plusieurs correcteurs se partagent un ensemble de copies, tandis que chaque copie n'est notée qu'une seule fois.

1.3.4 QCM

Les modalités plus récentes de questions à choix multiples(10), si elles ne permettent pas la finesse de correction des questions rédactionnelles, permettent de corriger un grand nombre de copies, en respectant la neutralité de la correction et ce en étant en général très corrélé au niveau de connaissances réel de l'étudiant(11).

Dans la prochaine partie nous détaillerons les différentes modalités des QCM ainsi que leurs propriétés.

Les QCM présentent plusieurs inconvénients par rapport aux questions rédactionnelles :

- Ils ne permettent pas une évaluation très fine
- C'est une méthode coûteuse en temps et en effort avant l'épreuve, en effet préparer des QCM à la fois suffisamment difficiles et sans ambiguïté est un exercice difficile à maîtriser

Cependant les QCM présentent également des avantages par rapport aux questions rédactionnelles :

- C'est la technique qui présente la meilleure *scablability* étant donné la capacité à automatiser complètement le processus de correction.
- Grâce à cette grande *scablability* il est possible de poser plus de questions et d'avoir une meilleure couverture du programme, pouvant aller jusqu'à l'exhaustivité
- C'est également la technique qui expose le moins à des problèmes de reproductibilité inter-opérateur et d'harmonisation, étant donné le caractère parfaitement déterministe de la correction : une fois l'épreuve rendue, à barème constant, les notes sont déjà déterminées.

Encore une fois le risque de proxy est important, en effet on pourra tout au plus évaluer des connaissances ou encore une partie des compétences face à un problème théorique. De ce fait, certaines compétences comme l'évaluation spontanée de l'attitude d'un patient ou bien des gestes techniques ne pourront pas être évaluées.

2 L’évaluation de connaissances par QCM

2.1 QCM “classiques”

Les QCM « classiques » sont la modalité la plus courante d’évaluation dans le groupe des QCM.

Une question comporte plusieurs réponses possibles, en général 5. Ces réponses sont appelées items. Un ou plusieurs de ces items sont vrais, les autres sont faux.

L’étudiant est évalué sur sa capacité à reconnaître les items vrais des items faux.

Une sous-catégorie souvent utilisée est la « Question à choix unique » ou QCU. C’est tout simplement le cas d’une question avec un seul item vrai et tous les autres faux, dans le cas où cette information étant donnée dans l’intitulé de la question.

- QCU : « Le traitement de première ligne de la cystite est : ABCDE»
- QCM : « Les traitements de seconde ligne de la cystite sont : ABCDE »

Le nombre de réponses possibles varie entre les deux méthodes : pour k cases cochables, et si on ne compte pas l’abstention :

- QCU : k réponses possibles, par exemple 5
- QCM : 2^k-1 réponses possibles, par exemple 31

2.2 Tablettes tactiles

Une tendance à l’informatisation des examens QCM est observable, notamment avec le développement des examens sur tablette tactile dont l’exemple le plus marquant est le passage des ECN aux ECNi en 2016(12).

Ces techniques ont l’avantage d’économiser du papier et de permettre des examens progressifs dans lesquelles les étudiants ne peuvent plus modifier leurs réponses précédentes.

Cependant cette solution présente également certains problèmes.

2.2.1 Coût d’accès et de maintenance

On notera tout d’abord le coût initial d’accès à la méthode d’évaluation avec la nécessité d’acheter des centaines de tablettes tactiles dédiées uniquement à la réalisation d’examens.

De plus, ces tablettes nécessitent une maintenance matérielle et logicielle, avec des risques de failles de sécurité ou d’impossibilité de composer si ces maintenances ne sont pas effectuées correctement.

2.2.2 Triches et failles

Par ailleurs la réalisation d’examens en ligne sur tablette augmentent les risques de triche avec plusieurs vecteurs d’attaques possibles.

Si l’examen est réalisé en local, des failles dans les tablettes pourraient permettre de revenir sur les questions précédentes.

Si l'examen est réalisé sur internet, ce qui est le cas des ECNi, une attaque possible serait par exemple d'attaquer le réseau avec un déni de service en cas d'examen défavorable. Le tricheur pourrait par exemple utiliser un matériel millimétrique caché dans son pantalon pour déclencher des attaques à la demande et forcer l'annulation de l'épreuve devant l'incapacité à synchroniser les facultés.

Une autre attaque serait rendue possible par une faille de sécurité sur le serveur d'examen, particulièrement si celle-ci permet de l'exécution de code à distance, ce scénario serait catastrophique, car il donnerait un accès privilégié à l'attaquant lui permettant de modifier les réponses ou les notes de ses concurrents et ce potentiellement de manière indétectable. Par ailleurs la probabilité d'un tel scénario paraît tout sauf négligeable(13).

De manière générale, dès lors que la complexité d'un système d'information augmente, les failles exploitables par des individus malveillants augmentent de même.(14)

2.2.3 Bugs

Même en l'absence d'attaque malveillante, tout système d'information aura une augmentation de ses échecs avec l'augmentation de son niveau de complexité, chaque chaînon dans le traitement de l'information ajoutant une possibilité d'échec(15).

Ce problème s'est notamment produit dans la deuxième instance des ECNi en 2017 avec la nécessité de reprogrammer certaines épreuves(16).

Ce risque est désormais contourné en organisant des épreuves de test en début de journée d'examen, permettant de détecter d'éventuelles erreurs dans le système d'information, mais cela demande une convocation de tous les candidats pour des épreuves ne comptant pas du tout dans la note finale. Le coût financier et humain d'une telle stratégie est considérable.

2.3 QCM papier

Les QCM papiers, s'ils sont certainement plus anciens, présentent tout de même quelques avantages par rapport à la méthode informatique.

En effet, le système d'information étant beaucoup plus simple il est donc plus résilient. Le risque de bug ou d'attaque malveillante est ainsi nettement plus faible.

La question de la maintenance des tablettes tactiles et de toutes l'infrastructure autour est donc éliminée.

Il est possible d'archiver les fiches optiques ainsi que leur numérisation si nécessaire dans un objectif de traçabilité.

On notera tout de même l'impossibilité de faire des "dossiers progressifs" avec cette méthode. Cependant, il s'agit d'un faux problème : même en QCM, il est toujours possible d'évaluer n'importe quelle étape de la prise en charge, y compris une étape terminale. Pour ce faire, il suffit de présenter un énoncé court, incluant le début de la prise en charge (" vous avez suspecté telle pathologie, et demandé tel examen qui revient avec tel résultat ") et poser une seule question, qui se situe en aval de la prise en charge initiale.

Un autre point négatif par rapport aux tablettes est l'iconographie limitée. On pensera notamment à l'impossibilité de zoomer et de naviguer dans un résultat de scanner comme c'est le cas sur un écran.

2.4 Limites des QCM classiques

Les problèmes soulevés par les QCM classiques sont nombreux et ont été soulevés de maintes fois(17).

Nous ferons ici une liste non-exhaustive de ces problèmes.

2.4.1 Suggestion de réponse

Le fait de suggérer des réponses représente un problème important, particulièrement dans des questions de raisonnement clinique(17).

En effet, il est facile dans la vie réelle d'oublier de penser à une embolie pulmonaire devant un patient asymptomatique présentant un S1Q3 léger à l'électrocardiogramme.

Alternativement il est bien difficile de ne pas y penser quand "suspicion d'embolie pulmonaire" fait partie des cinq réponses proposées, à moins de recourir exagérément à la proposition " aucune des réponses ci-dessus n'est exacte " ou " autre réponse ", ce qui ne semble pas être une bonne solution.

Nous décrirons dans les parties suivantes quelques pistes pour limiter cette propriété indésirable de l'évaluation par QCM.

2.4.2 Exagération du niveau de certitude

Un des problèmes soulevés par les QCM classiques est que l'étudiant n'a absolument pas intérêt à répondre honnêtement(18) s'il ne connaît pas la réponse.

En effet, il vaut bien mieux répondre au hasard et avoir une chance même réduite de gagner un point plutôt que de ne pas répondre du tout.

Cette propriété peut être éliminée par la présence de points négatifs, mais cela poussera l'étudiant à ne pas répondre s'il n'est pas complètement sûr, biaisant encore une fois ses réponses.

On notera par la suite quelques pistes pour corriger ce problème de manière plus efficace.

2.5 Pistes d'amélioration, nouvelles modalités

Il existe de nombreuses possibilités de modalités d'examens réalisables avec des fiches optiques au-delà des QCM classiques. Nous en donnons ici quelques exemples.

2.5.1 QCM Bayésiens

Le principe des QCM Bayésiens(19) (20) est de permettre à l'étudiant d'exprimer ses réponses en niveau de certitude plutôt qu'en un avis tranché binaire.

L'étudiant cochera donc la case correspondant à son niveau de certitude de la question.

- 0% signifiant « je suis sûr que la réponse est non »
- 100 % signifiant « je suis sûr que la réponse est oui »
- 50 % signifiant « je ne sais pas »

La principale problématique de cette modalité d'examen est le risque d'exagération de certitude de l'étudiant, préférant cacher son ignorance et artificiellement tendre vers 0 % ou 100 %.

Ceci implique l'absolue nécessité de points négatifs pour les réponses fausses exprimées avec un grand niveau de confiance, donnant ainsi un intérrêt à exprimer de manière honnête son incertitude(18).

Les deux principaux algorithmes de notation donnant l'intérêt à l'étudiant d'être parfaitement honnête pour obtenir une note maximale sont :

1. *La notation logarithmique*

Notons *p* la certitude de l'étudiant que l'item proposé soit vrai :

Ses poins seront alors (*S* désignant le score obtenu pour la question) :

- $S = ln(1-p) - ln(p)$ si l'item est en réalité faux
- $S = ln(p) - ln(1-p)$ si l'item est en réalité vrai

2. *La notation quadratique*

Notons p la certitude de l'étudiant que l'item proposé soit vrai :

Ses poins seront alors :

- $S = 1 - 2p^2$ si l'item est en réalité faux
- $S = 1 - 2(1-p)^2$ si l'item est en réalité vrai

Décrivons ici un exemple de cette dernière notation :

Un étudiant ne connaissant pas la réponse et ayant répondu 50 % obtiendra ainsi 0.5 points que la réponse soit vrai ou fausse, ce qui équivaut à une espérance de 0.5 points.

Un étudiant ne connaissant pas la réponse et ayant répondu 100 % obtiendra ainsi -1 point si la réponse est fausse et +1 point si la réponse est vraie, ce qui équivaut à une espérance de 0 point.

L'honnêteté est dans ce cas la stratégie optimale, et cette propriété est démontrable dans le cas général(20).

A noter que des pondérations et transformations de notes peuvent toujours être appliquées pour sélectionner le niveau requis des étudiants.

2.5.2 Très nombreuses propositions

Avec la liberté de définir des fiches optiques personnalisées vient la capacité de faire des questions avec de très nombreuses propositions.

En effet, une des limites précédemment citée des QCM est de restreindre les possibilités de réponses et offre donc des indices non représentatifs(17) (21) des vraies situations cliniques.

Cette propriété peut-être fortement limitée en proposant un grand nombre de réponses possibles pour une question. Dans les faits, proposer plus de 40 réponses pour une question, limite l'effet « indice », particulièrement dans le cas d'un choix multiple.

D'une certaine manière, de tels QCM pourraient mieux refléter des situations de vies réelles. Ainsi par exemple, un interne confronté à un patient symptomatique aux urgences, aura une liste très importante de possibilités, notamment en ce qui concerne les examens complémentaires qu'il est possible de prescrire.

2.5.3 Schéma anatomique à compléter

Il est également possible de poser des questions directement sur la fiche optique, voire de compléter un schéma anatomique pré-imprimé.

Par exemple on pourra demander de cocher la case proche de l'os hyoïde et l'étudiant devra trouver l'os sur un schéma dessiné sur la fiche. On pourrait même par exemple, sur une coupe transversale de membre, demander à l'étudiant de figurer une artère à l'aide d'un petit cercle.

Cette perspective n'a pas été retrouvée dans la littérature scientifique.

2.6 Méthodes d'évaluation de la qualité d'un examen

Les examens par QCM étant très fréquents, la question de l'évaluation de la qualité d'un examen a été soulevée de nombreuses fois, et il existe donc de nombreuses méthodes d'évaluation de cette qualité.

On pourra noter deux approches possibles à ces évaluations :

2.6.1 Mesure de la capacité à mesurer le niveau absolu d'un étudiant

Cette capacité de mesurer le niveau absolu, si elle est souhaitable pour sélectionner les étudiants qui « ont le niveau » pour la validation, reste en pratique assez difficile(21) à mettre en place, est en général non testée de manière formelle et laissé à la discrétion des professeurs qui jugent du niveau requis pour la validation.

2.6.2 Mesure de la capacité à sélectionner les meilleurs étudiant

Cette capacité est plus facile à mesurer en pratique de manière formelle, et de nombreux indices de difficulté et de discrimination ont été développés au cours du temps, allant de la simple analyse des courbes de distribution, à des techniques d'analyse question par question.

Nous détaillerons dans la partie « Matériel et méthodes » différentes méthodes de validation(22) docimologie que nous avons utilisées dans notre étude.

3 Contexte à la Faculté de Médecine de Lille

A la faculté médecine de Lille, la question à choix multiple est la modalité d'examen la plus fréquente au cours des études de médecine.

3.1 PASS Santé

L'enjeu d'une correction la plus rapide et neutre possible se situe au sein des examens du PASS Santé (ancienne PACES, ancienne PCEM1)(23).

Figure 1: Photo d'un examen de PACES ayant lieu

En effet il s'agit d'une année de concours pendant laquelle plusieurs milliers d'étudiants sortant de l'enseignement secondaire sont en concurrence pour intégrer plusieurs branches des études de santé (Maïeutique, Odontologie, Médecine, Kinésithérapie).

Le volume de copies corrigées lors des deux sessions d'examens correspondant à 7 unités d'enseignement (UE) est de l'ordre de 22000 copies par an, ce qui serait extrêmement difficile, voire impossible à corriger avec modalité de questions rédactionnelles.

Par ailleurs, l'université ne possède pas assez de tablettes tactiles et utilise donc la modalité des QCM sur papier pour cette année.

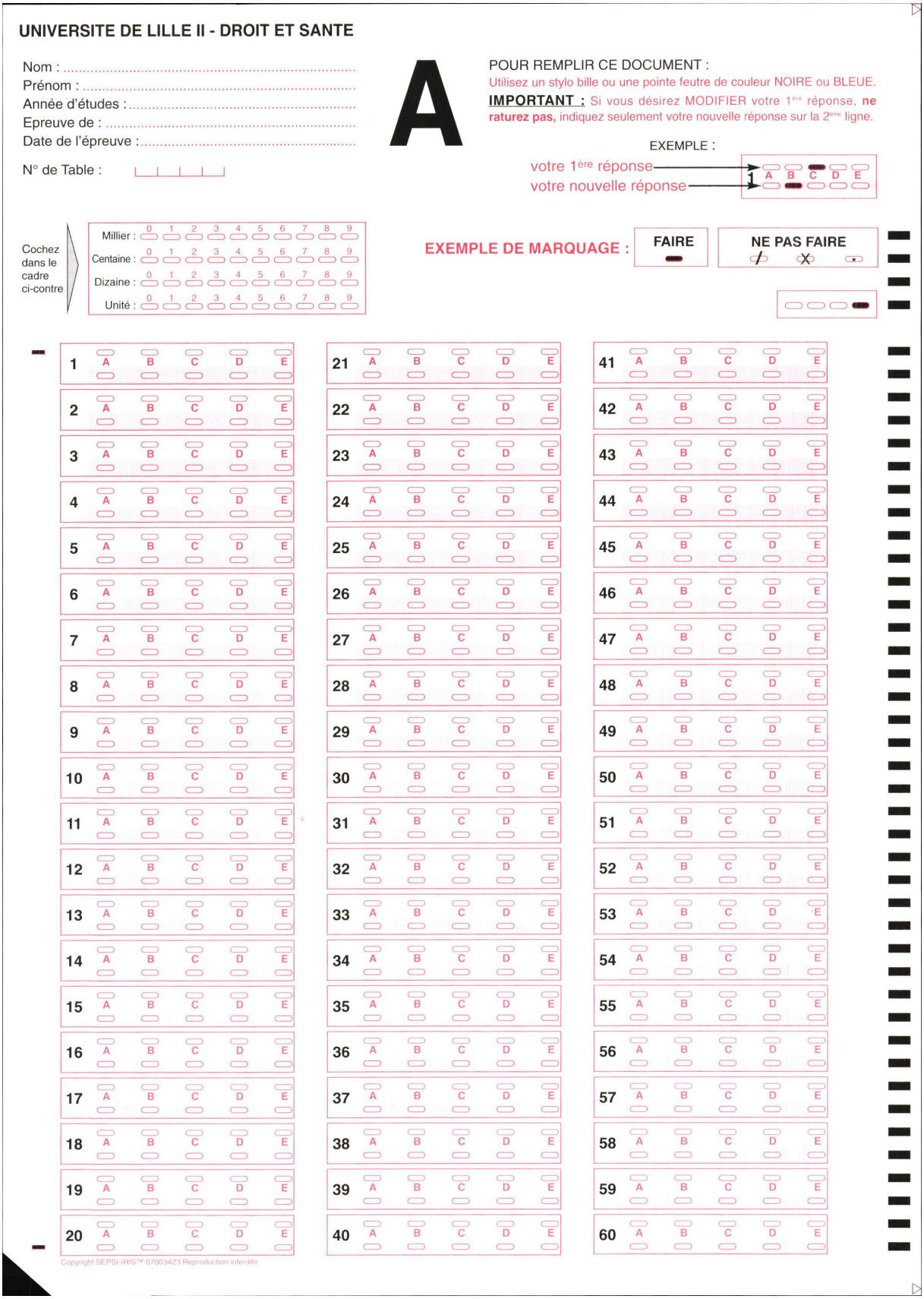

Figure 2: Exemple de fiche optique utilisée par la faculté de médecine

Figure 3: Zone de coche

3.2 Années de Med-2 à Med-6

Entre la deuxième et la sixième année de médecine, les QCM restent la principale modalité d'évaluation. Cependant un bon nombre d'examens ont lieu sur tablette. Ceci est permis par le relativement faible nombre d'étudiants dans chaque année d'étude (de l'ordre de 500, redoublants inclus).

Nous noterons tout de même la présence de quelques épreuves de QCM sur papier.

3.3 Sous-traitance

De nombreuses autres formations de l'Université sous-traitent actuellement la correction de QCM papier à la faculté de Médecine. Cette correction est alors, comme pour la formation médicale, automatisée.

3.4 Correction manuelle

Certaines formations de l'Université de Lille, comme l'école des sages-femme, corrigent leurs QCM entièrement à la main. La réponse à ces QCM utilise des supports de réponse conçus pour chaque épreuve : il n'est alors pas nécessaire de disposer de fiches optiques standardisées.

En outre, et de manière qu'il est difficile de quantifier, il arrive souvent que des épreuves dites rédactionnelles, présentées avec un sujet et une copie de réponse libre, comportent des QCM : il revient alors à l'étudiant de saisir sur sa copie le numéro de question, et les items choisis.

3.5 Correction automatisée actuelle

Il existe déjà dans la Faculté de Médecine de Lille un matériel permettant la correction automatisée des examens à QCM.

Le système actuellement en place requiert que les étudiants noircissent des cases sur une grille en papier au format A4, appelée « fiche optique » (FO). Ces fiches optiques sont rassemblées et passées dans une machine de correction, qui grâce à une lumière rouge et un capteur optique, détecte les cases cochées et transmet la liste des réponses à un ordinateur branché.

Cette machine propriétaire effectue à la fois la lecture optique et la correction du lot de copies. Elle est reliée par port LPT ou USB à un ordinateur dédié, lequel doit impérativement exécuter un logiciel privé. La machine est directement pilotée par ce logiciel, et lui retourne directement un message propriétaire que le logiciel exploite pour fournir *in fine* un compte-rendu de correction au format CSV. Il n'existe aucune forme de transmission ni d'enregistrement d'image. Cette machine ne peut être utilisée de manière déconnectée, ni avec un autre logiciel. Cette machine ne peut être utilisée avec d'autres

copies vierges que celle commercialisée par l'éditeur, et réciproquement ces copies vierges ne peuvent être corrigées avec une autre machine.

Figure 4: Vue globale de la machine permettant les corrections optiques

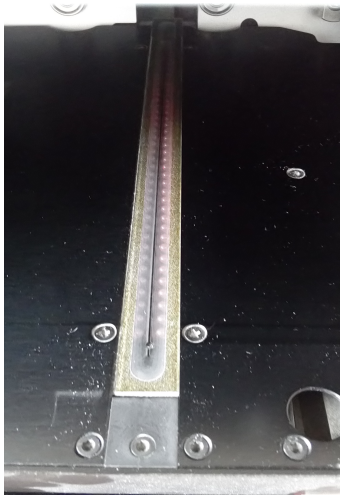

Figure 5: Vue rapprochée de la barrette de lecture optique

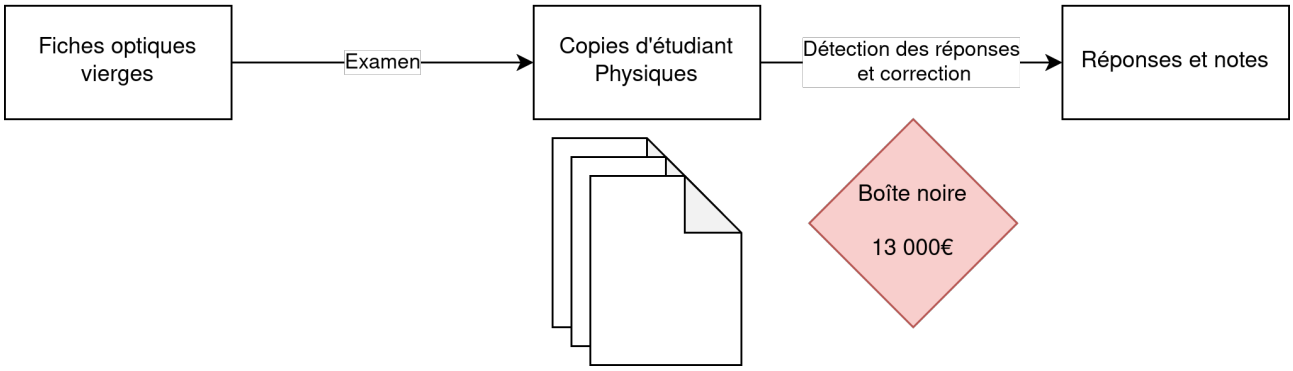

Figure 6: Processus de correction actuel

Cette machine coûte près de 13 000€, mais en mettant de côté la question du prix, les principaux problèmes soulevées par l'utilisation de ce système sont les suivants :

3.5.1 Rupture de continuité

Le premier problème soulevé par l'utilisation de ce système est le risque de rupture de la continuité du service.

En effet ces machines nécessitent une maintenance, et il existe un risque d'appauvrissement progressif de l'offre de ces machines sur les marchés industriels dans un contexte d'informatisation des examens. Par ailleurs la pérennité de l'entreprise qui commercialise cette solution n'est pas garantie.

La possibilité d'utilisation de n'importe quel scanner du marché résoudrait ce problème.

3.5.2 Asservissement logiciel

Le second problème soulevé peut être décrit comme le pendant logiciel du problème précédent. En effet le système utilise des logiciels à sources fermées uniquement maintenables par le producteur des machines de corrections.

Cela crée une ces logiciels potentiellement non maintenus dans le futur ou incompatibles avec les nouveaux OS.

L'utilisation d'un logiciel libre et open source résoudrait ce problème.

3.5.3 Bourrage papier

Malgré la grande qualité de la mécanique d'entraînement, qu'on peut qualifier de supérieure aux scanners du marché, la machine actuelle subit parfois des bourrages de papier. La copie de l'étudiant se froisse et devient alors illisible, car il n'est pas possible de numériser la copie à plat : la machine ne fonctionne qu'avec un entraînement mécanique des feuilles, à grande vitesse, dans un collimateur très étroit et spécifique. La seule solution est alors de recopier la FO à la main sur une copie vierge, et la présenter de nouveau à la machine. Cela pose un problème de risque d'erreur, et de contestation d'un

étudiant qui pourrait ne pas reconnaître son écriture sur la partie d'identification de la copie.

Une possibilité de numérisation de la copie à plat sur la vitre d'un scanner classique résoudrait ce problème.

4 Objectifs

Ce travail poursuivra trois objectifs détaillés ci-après :

1. Le développement d'un logiciel de correction optique pour les copies existantes : " logiciel de correction "
2. La preuve de concept d'utilisation de ce logiciel pour des modalités d'examen innovantes : " nouvelles modalités d'examen "
3. La mise en place d'un rapport d'analyse statistique automatisé : " validation docimologique "

4.1 Logiciel de correction

Le premier objectif de ce travail est de développer une procédure informatique de correction automatisée ne dépendant ni d'une machine propriétaire, ni d'un modèle de fiche optique particulier.

La lecture optique *stricto sensu* sera déléguée à un photocopieur-scanner quelconque du marché, permettant une potentielle baisse à la fois du coût de l'accès à la correction automatisée ainsi que de la maintenance en cas de dommage matériel. Les images des copies seront stockées numériquement, et notre produit aura pour fonction d'analyser ces images.

Par ailleurs, la correction pourra être réalisée de manière asynchrone à la numérisation des fiches optiques et les copies pourront être recorrigées avec différents paramètres autant de fois que nécessaire.

Les fiches optiques numérisées pourront être archivées à la fois sous forme papier et numérique, permettant d'effectuer des sauvegardes par réplication et permettant une meilleure traçabilité des différentes actions.

4.2 Nouvelles modalités d'examen

Cette nouvelle technique de correction, puisqu'elle ne dépend ni d'une machine spécifique, ni d'un modèle de fiche optique spécifique peut permettre l'émergence de nouvelles modalités d'examen décrites précédemment.

Dans cette partie, nous proposerons et évaluerons le logiciel précédemment développé pour des modalités d'examen différentes et innovantes.

4.3 Validation docimologique

La modalité des QCM permet également d'effectuer une validation de la qualité de chaque question à départager les étudiants, et ce de manière automatisée. Pour ce faire, nous développerons et testerons un rapport d'analyse statistique automatisée. Cela permettra de mettre en évidence des erreurs dans la correction ou dans le sujet.

Matériel et méthodes

1 Logiciel de correction

1.1 Déroulement d'une correction dans le nouveau système

Le nouveau mode permet l'utilisation de n'importe quel scanner du marché, ouvrant l'utilisation de QCM à des structures ayant des budgets plus faibles, et permettant aux structures plus grosses de s'affranchir de la dépendance d'un fabriquant.

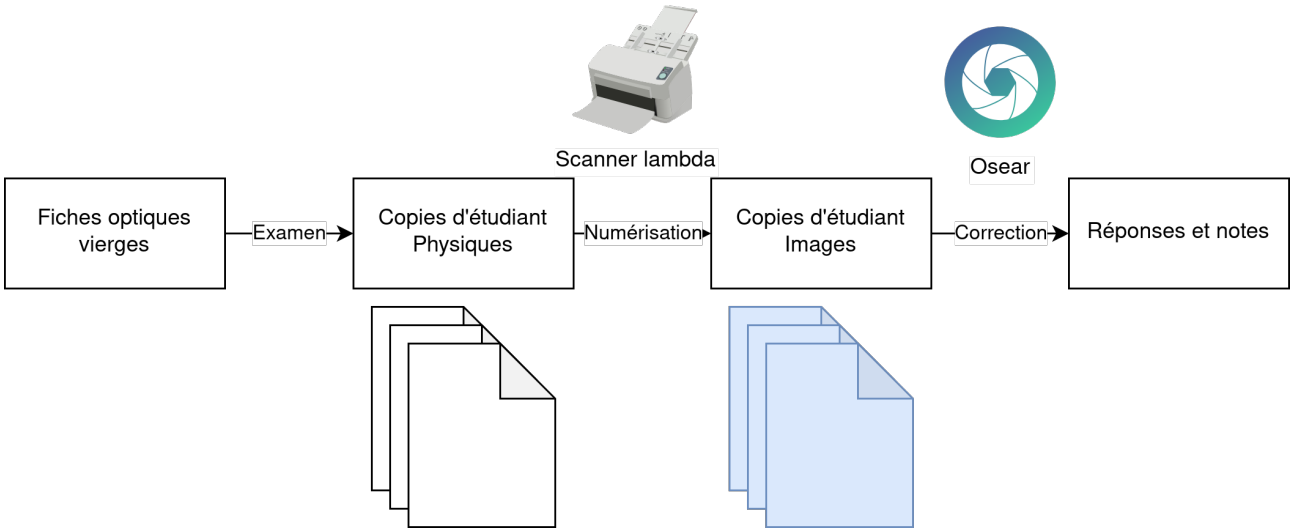

Figure 7: Nouveau processus de correction

Ceci est accompli grâce à une séparation entre l'étape de numérisation et l'étape de correction.

1.2 Osear

Le programme Osear, permet d'effectuer l'analyse des images obtenues par n'importe quel scanneur du marché.

Il s'agit d'un programme libre et open-source, composé en deux parties écrites en C++17 :

libosear : La bibliothèque de correction de copies

osear : L'interface graphique de correction

La bibliothèque libosear dépend d'autres bibliothèques pour fonctionner :

- OpenImageIO
- libboost
- libmagic

- openmp

L'interface osear dépend bien évidemment de libosear mais aussi de :

- wxwidgets 3.1.4 pour l'interface graphique

Le code source du programme Osear fait environ 3122 lignes de C++, et quelques dizaines de lignes d'autres langages.

1.3 Descripteur de document

Le descripteur de document est un fichier xml décrivant la liste des positions théoriques des cases sur une fiche optique, les différents repères pour réorienter une fiche mal scannée, et des zones noires et blanches pour échantillonner les couleurs.

Ce descripteur décrit également le contexte d'une case :

- chiffre pour le numéro identifiant l'étudiant
- réponse de l'étudiant à une question

Les cases cochables sont décrites par des balises <zone> comprises dans deux des trois types de balises principales :

1.3.1 <detection_system>

Cette balise décrit tout le système d'orientation, permettant la réorientation d'une copie penchée et la vérification de la réorientation.

Elle ne contient aucune balise <zone>.

1.3.2 <id_field>

Cette balise décrit les cases cochable permettant à un étudiant de s'identifier,

Ce champ comprend des balises <digit order="n"> décrivants le chiffre du numéro étudiant pour la position n, par exemple <digit order="3"> correspondant au chiffre des milliers.

Par exemple un étudiant dont le numéro est « 4523 », cochera la <zone> 4 du digit <digit order="3">, la <zone> 5 du <digit order="3">, la <zone> 2 du <digit order="1"> et la <zone> 3 du <digit order="0">

1.3.3 <response_field num='n'>

Cette balise décrit la liste des cases cochables pour la question « n ».

1.4 Flux des données

Les images sont analysées les unes après les autres avec l'aide du descripteur de fiche optique.

Les réponses des étudiants peuvent être exportées au format csv, et/ou comparées par rapport à une correction qui établit les réponses vraies et fausses ainsi que leur pondération.

Ces étapes mises bout-à-bout permettent de corriger une épreuve sans intervention humaine.

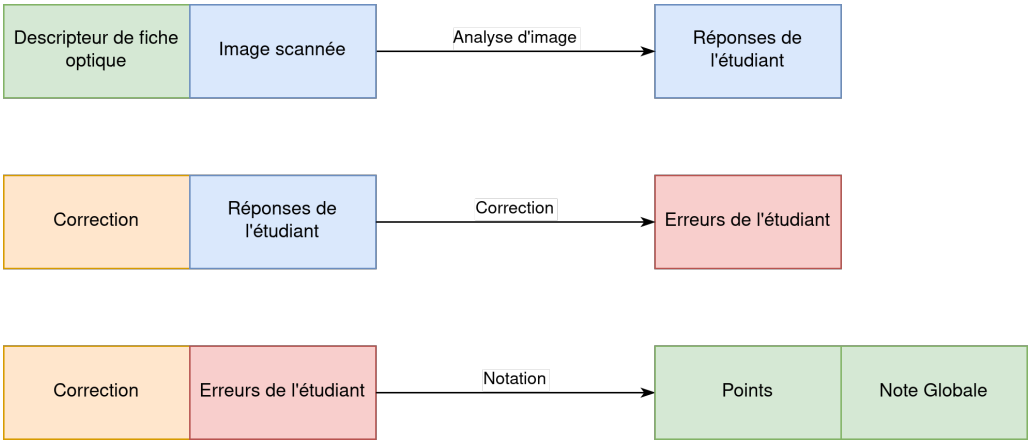

Figure 8: Flux des données

1.5 Erreurs de correction

Une mauvaise reconnaissance des éléments d'une fiche peuvent amener à des erreurs dans les réponses de l'étudiant et donc de sa note et de son classement.

Plusieurs parties du programme Osear ont été créées pour minimiser le risque d'erreur et de le mesurer pour demander une correction manuelle dans les cas ou le risque minimal d'erreur n'est pas acceptable.

L'évaluation de ce risque est faite avec une analyse statistique interne.

1.6 Repères

Il existe sur chaque copie des repères permettant l'orientation de la fiche optique, ce sont des rectangles noirs entourés d'une zone blanche.

D'abord une recherche de ces motifs est effectuée dans des zones décrites dans le fichier descripteur du modèle de fiche dans sa section <*detection_system*>.

L'analyse est basée sur un modèle de réduction d'un coût, en l'occurrence la différence entre le motif théorique attendu et le motif détecté.

Le coût minimal est ensuite passé dans une régression logistique entraînée préalablement sur des données de correction officielles et labellisées, permettant l'évaluation de la probabilité d'un motif reconnu à tort comme valide.

Si le risque d'erreur dépasse 1 % la fiche est marquée comme à risque, permettant sa lecture manuelle.

Une fois l'assurance de la validité des repères obtenue, une transformation affine est calculée, permettant de traduire des coordonnées théoriques en coordonnées réelles sur l'image scannée, « orientant » ainsi la feuille.

1.7 Cases

Une fois la copie orientée, les cases sont repérées grâce à leur position théorique décrite dans le fichier descripteur.

Les pixels des cases sont analysés en fonction de leur luminance dans l'axe ou les axes sélectionnés dans les paramètres de correction (ex : axe rouge uniquement, permettant d'ignorer les délimitations pré-imprimées en rouge, ou tous les axes, soit une correction en niveaux de gris).

Une distribution approximativement bimodale des luminances est attendue, en effet les pixels doivent être soit sombres (zone cochée de la case si la case est cochée, luminance faible proche de 0/255), soit clairs (case non-cochée, ou zone non-cochée autour d'une case cochée, luminance élevée proche de 255/255).

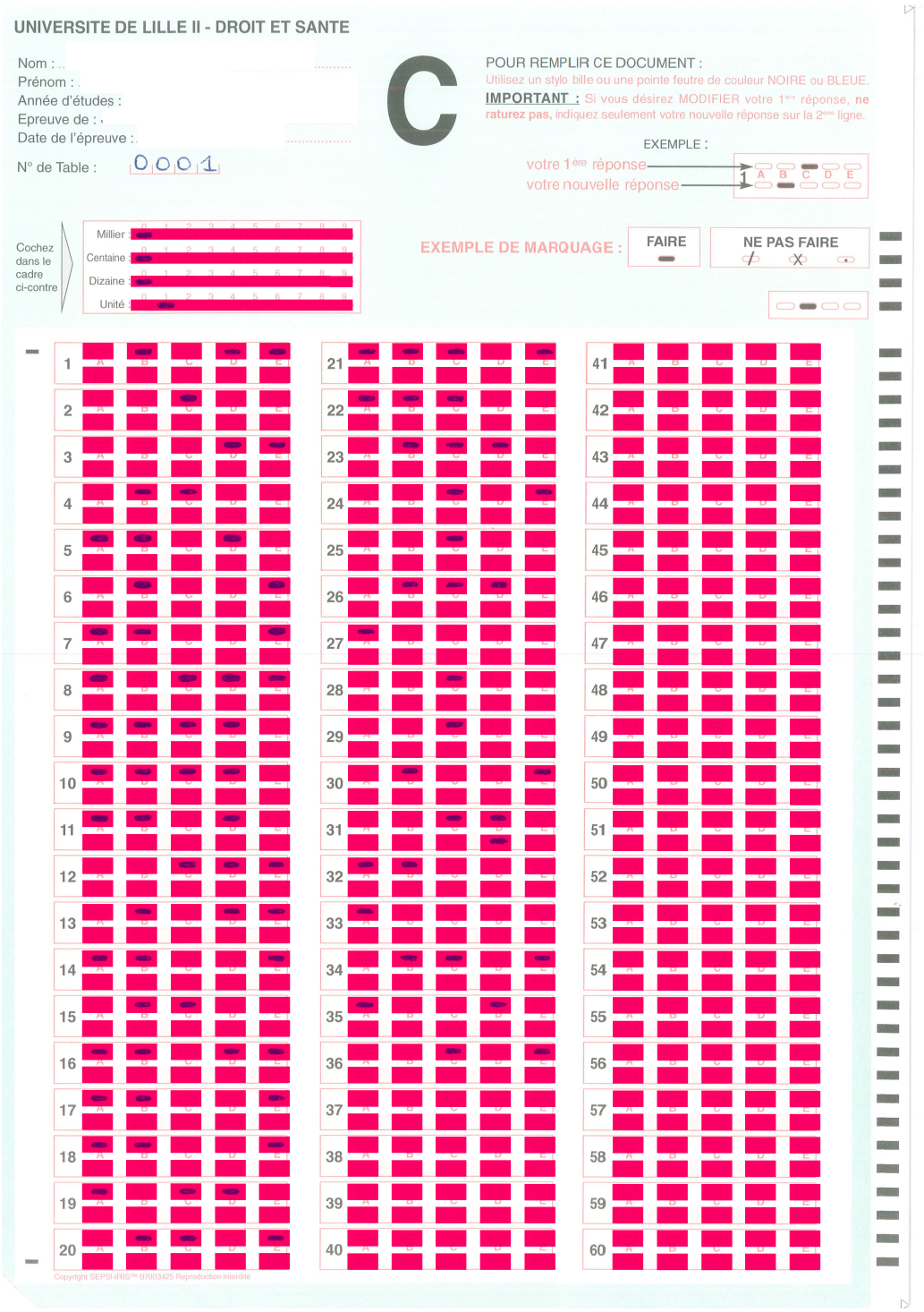

Figure 9: Exemple de surlignage automatique des zones de coche

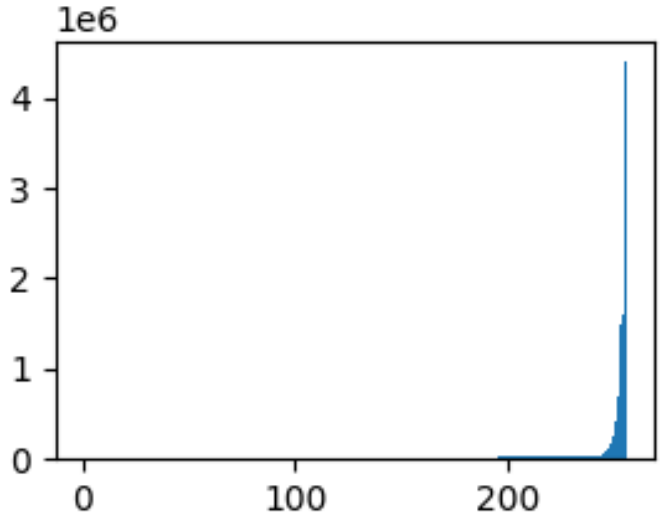

Figure 10: Distribution des luminances (échelle linéaire)

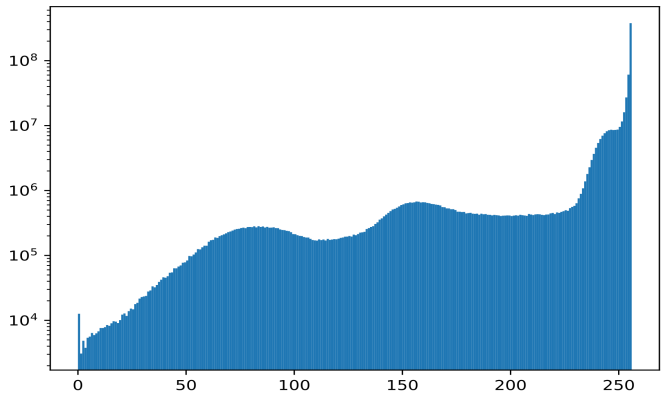

Figure 11: Distribution des luminances (échelle logarithmique)

Un test de cet bimodalité est effectué, s'il revient négatif au risque de première espèce de 1 %, la copie est marquée comme à risque d'erreur de correction.

Une fois cette bimodalité des couleurs établie, un seuil est calculé de manière dynamique permettant de corriger la copie.

Les cases sont analysées une à une par un algorithme de comptage des pixels noirs, avec un seuil établi encore une fois grâce à la distribution bimodale du nombre de pixels noirs dans les cases.

La case est considérée comme cochée si le nombre de pixel noirs est plus proche de la moyenne des cases cochées que de la moyenne des cases non-cochées.

A noter qu'un biais intrinsèque en faveur des cases cochées a été mis en place suite à de nombreuses corrections de copies : la présence de ce biais minimise les erreurs de correction.

1.8 Identifiant copie

L'identifiant de la copie est détecté en considérant la case la plus noircie comme l'unique case cochée.

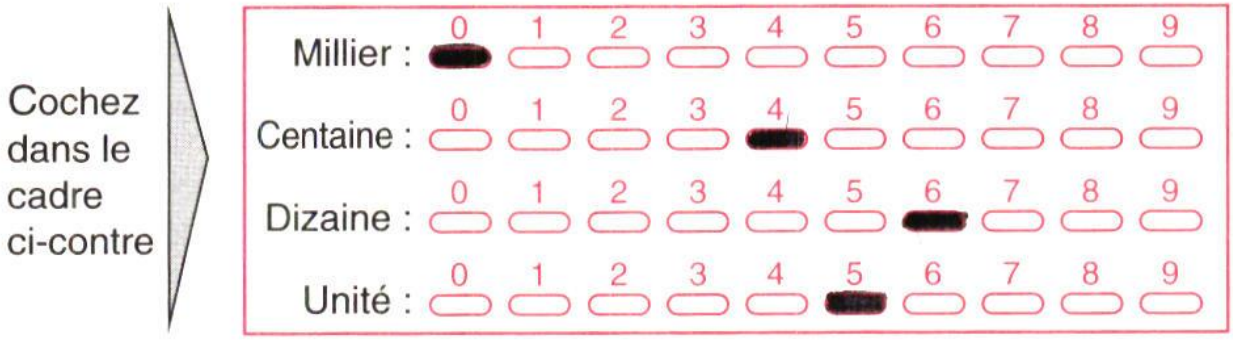

Figure 12: Identifiant de copie

Le numéro de copie est simplement la concaténation des chiffres cochés, ici 0465.

1.9 Correction de la liste de réponses

Une fois la liste de réponses obtenue, elle est corrigée grâce à la liste pondérée des réponses justes.

Cette liste peut-être décrite soit par un fichier XML soit par une “ copie de correction ” numérisée dans le même temps que les copies des élèves.

Dans le deuxième cas, aucune pondération n’est effectuée, chaque question comptant alors autant que n’importe quelle autre dans la note finale.

1.10 Fichier en sortie

Un fichier final csv est produit, contenant les identifiants, notes, réponses des étudiants, ainsi que le risque d’erreur.

En voici un exemple visuel :

Table 1: Exemple de fichier csv de sortie

Numéro de copie	Note	Réponse 1	...	Réponse 20	A risque d’erreur
0465	10.0	AB		BDE	0
0012	12.0	A		BDE	1
...					
1263	5.0	AC		BE	0

1.11 Évaluation de la correction

L'évaluation de la qualité de la correction est effectuée de manière comparative à la correction officielle produite par l'ancienne méthode.

Les données ont été analysées avec le logiciel R 4.0.3 (2020-10-10, Bunny-Wunnies Freak Out).

1.12 Tests de concordance

La liste des réponses de chaque étudiant fournie par la nouvelle méthode a été comparée avec la liste fournie par l'ancienne méthode grâce au test de concordance du Kappa de Cohen.

Trois tests ont été effectués, sur trois jeux de données :

1. Toutes les copies
2. Copies "non à risque d'erreur"
3. Copies "à risque d'erreur"

Il s'agit d'une concordance par copie, sur la liste complète des réponses (ex : 2^{100} réponses possibles pour une épreuve de 20 questions avec 5 cases ABCDE).

Le Kappa de Cohen avait donc une concordance attendue par hasard d'environ 0 étant donné un nombre de motifs possibles de 2^{100} pour une copie comportant 20 questions, le Kappa représente donc directement le niveau de concordance observée.

Il s'agit d'une évaluation statistique très sévère : une copie est discordante si une seule coche diffère entre l'analyse de la copie par l'ancienne méthode et l'analyse par Osear.

2 Nouvelles modalités d'examen

2.1 Introduction

S'affranchir d'un modèle particulier de fiche optique permet de créer de nouvelles modalités d'examen avec des fiches optiques spécifiquement créées pour celui-ci.

Nous avons ici simplement testé si des fiches optiques peuvent être corrigées sans erreurs pour les modalités d'examen suivante sans effectuer de validation docimologique.

Un kappa de Cohen a été effectué pour tester la concordance.

2.2 QCM Bayésiens

Voici un exemple visuel de fiche optique pour un examen par QCM bayésiens.

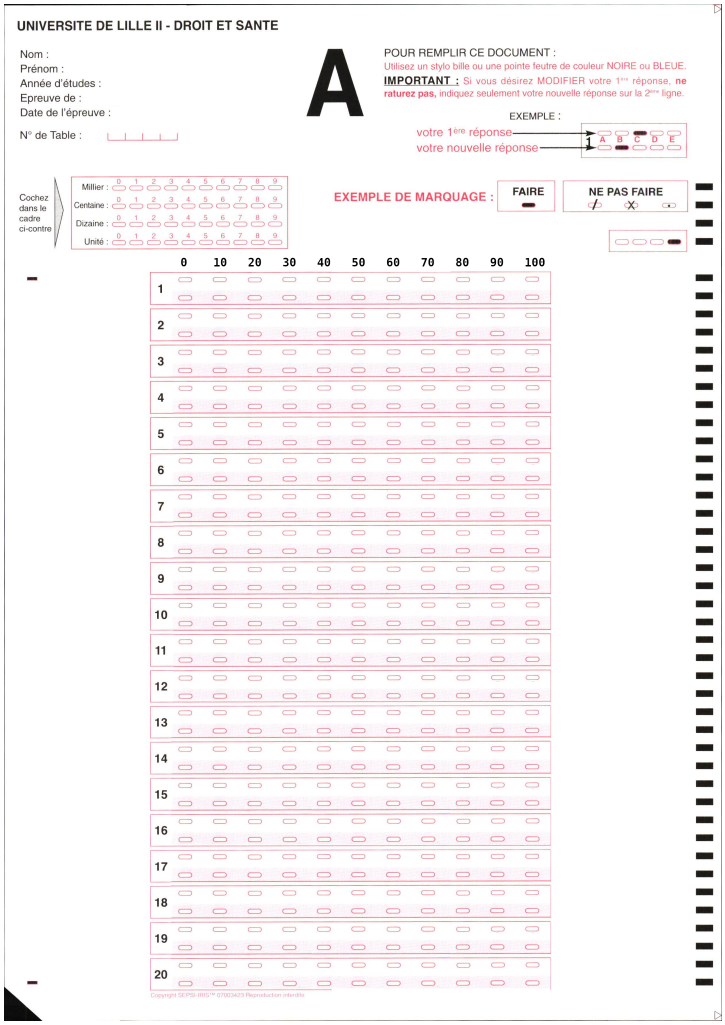

Figure 13: Fiche optique pour 20 QCM bayésiens

Le logiciel de correction détectera les réponses et appliquera l'algorithme de correction spécifié dans le fichier de correction.

Les deux algorithmes de correction implémentés sont « quadratic » et « logarithmic » et sont décrits dans l'introduction.

2.3 Nombreuses propositions

Voici un exemple de fiche optique avec 20 questions ayant chacune 16 réponses possibles. Les algorithmes classiques de corrections sont utilisables sur ce type de fiches.

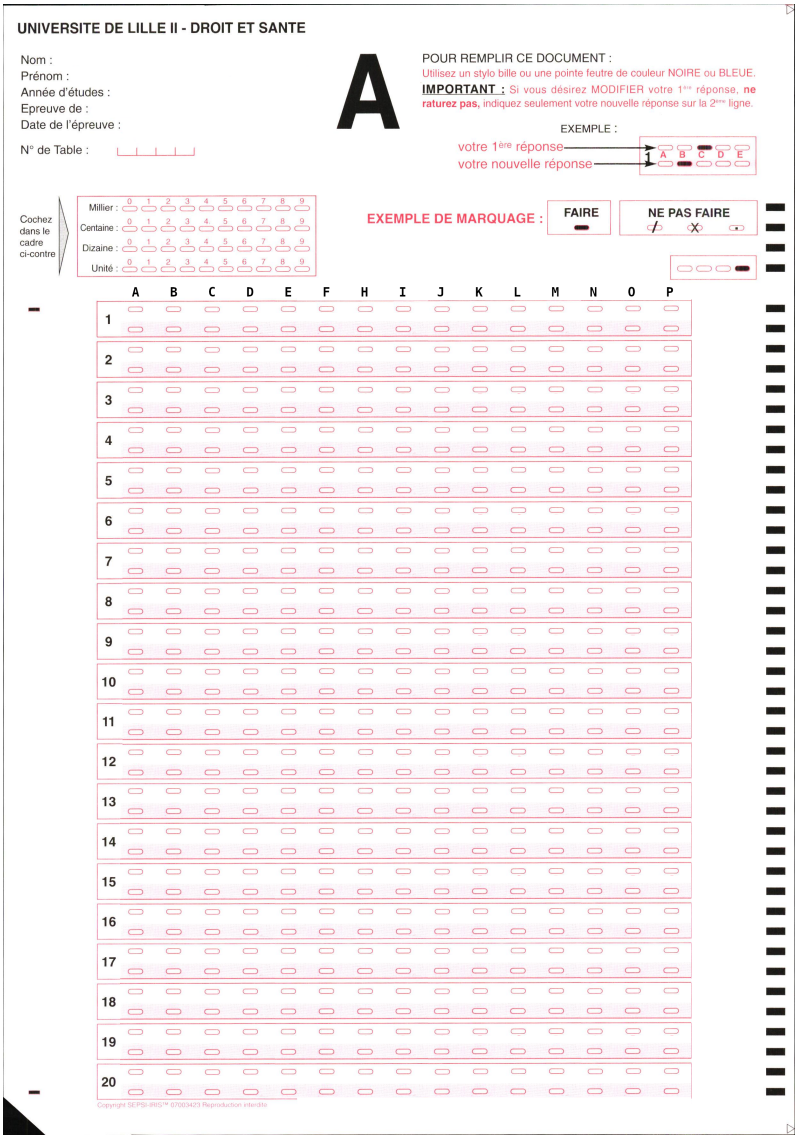

Figure 14: Fiche optique avec de nombreuses propositions par question

2.4 Schéma d'anatomie à compléter

Voici un exemple de fiche optique avec un schéma anatomique à compléter.

A noter qu'il est possible de demander le coloriage en plusieurs couleurs pour poser de multiples questions.

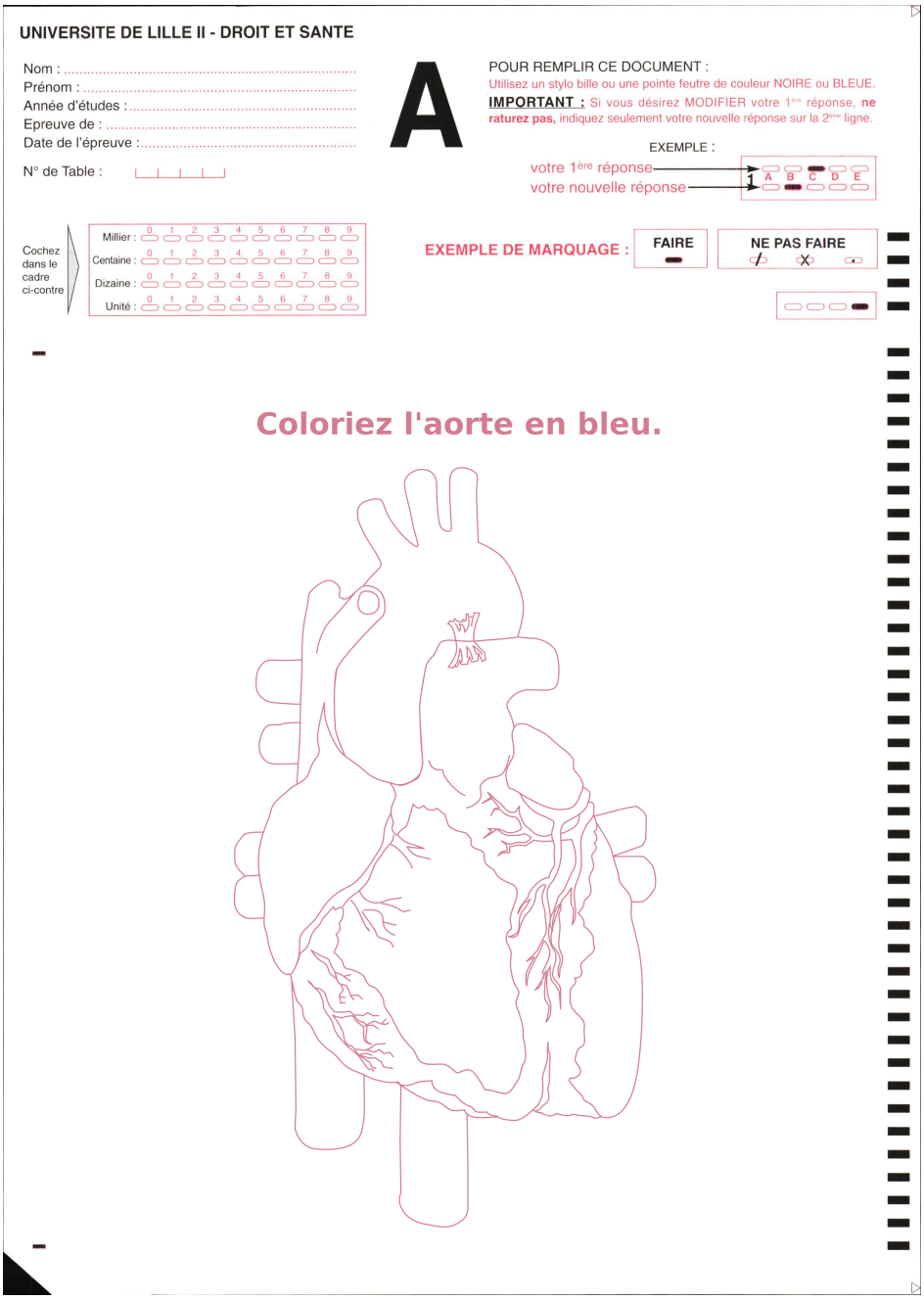

Figure 15: Fiche optique avec un schéma anatomique à compléter

3 Validation docimologique

3.1 Introduction

L'évaluation d'une instance d'examen peut également être effectuée après correction complète d'un examen en faisant une analyse statistique.

En effet, il est possible de mesurer la crédibilité d'une erreur dans le sujet même de l'examen en repérant des distributions de réussites aux questions atypiques.

Toute la validation docimologique a été effectuée avec le logiciel R 4.0.3(2020-10-10, Bunny-Wunnies Freak Out).

Devant le très faible nombre de copies dans les examens tests des nouvelles modalités, l'analyse docimologique de ces derniers n'a pas pu être effectuée.

3.2 Distribution des notes

Classiquement la réussite ou non d'un item est un facteur prédictif de la note globale à l'épreuve, les étudiants ayant bien répondu à l'item ayant des meilleures notes que les personnes n'ayant pas bien répondu.

Les figures 16, 17 et 18 illustrent le comportement attendu d'une « bonne » question : elles représentent l'estimation graphique de la densité de probabilité de la note finale à l'épreuve, en superposant en rouge cette courbe pour les étudiants qui ont échoué à la question, et en bleu cette courbe pour les étudiants qui ont réussi à la question.

Si la question est discriminante, les étudiants qui la réussissent devraient plutôt être de bons étudiants, donc avec une distribution de la note finale décalée vers la droite, et inversement pour ceux qui y échouent. Si la question n'est pas discriminante, la réponse est aléatoire, et les deux distributions devraient se confondre. A l'extrême, si la question est discriminante mais qu'il existe une erreur dans la grille de correction, les deux distributions apparaissent inversées.

Une question comportant une erreur, soit dans son intitulé, soit dans sa correction officielle ne séparera pas bien les “bons” et les “mauvais étudiants”, soit en ayant des distributions identiques soit, cas plus extrême, une inversion18 de la distribution avec les “mauvais” étudiants répondant mieux que les “bons”.

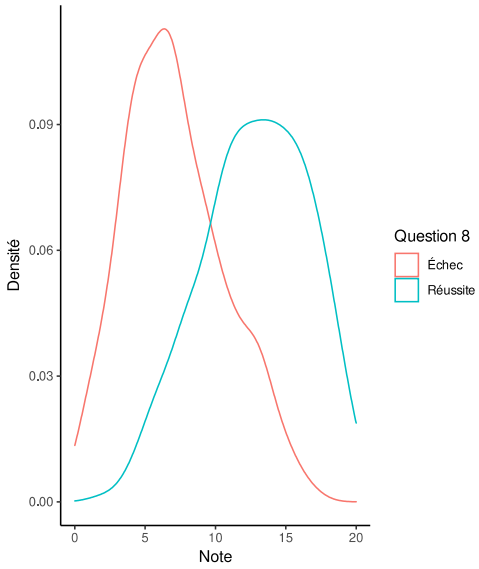

Figure 16: Question discriminante

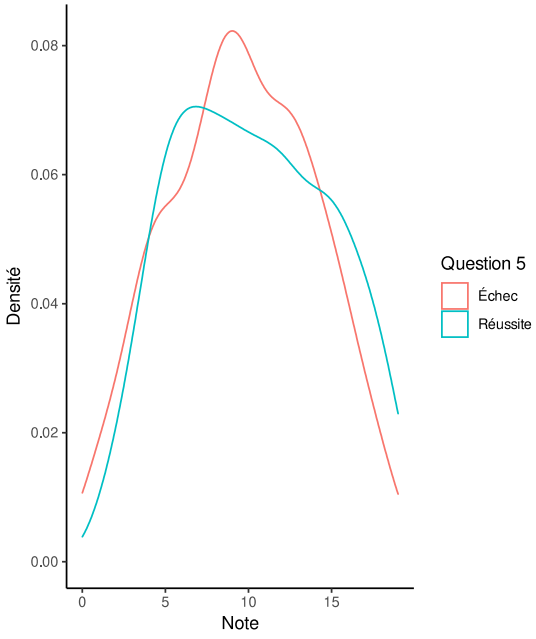

Figure 17: Question non-discriminante

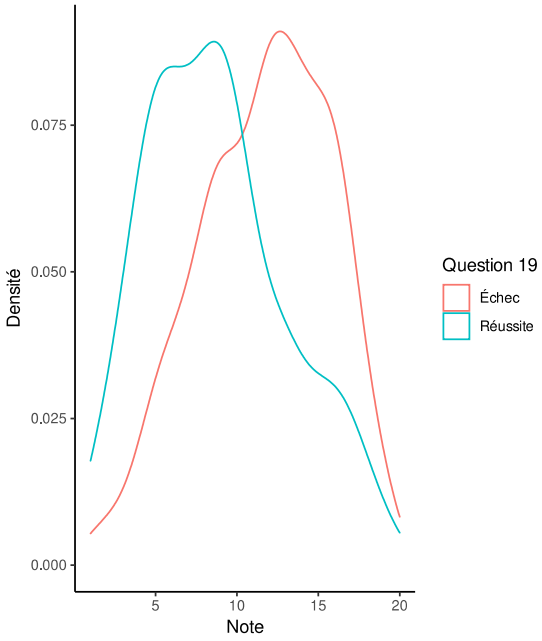

Figure 18: Question "contre-sens"

3.3 Indices de discriminations

Partant de ce principe, il est possible d'imaginer plusieurs approches statistiques ou informatiques pour formaliser et quantifier le décalage attendu des deux distributions. On citera par exemple :

- un test de Student
- un test non-paramétrique des rangs de Wilcoxon
- un test de comparaison de deux moyennes à l'aide d'une loi normale (qui est alors en rapport avec le cas particulier d'une régression linéaire simple avec une variable explicative binaire)
- une régression logistique à un facteur, en inversant X et Y
- une ANOVA, qui permet de calculer le coefficient de détermination R^2, qui est dans ce cas simplement le coefficient de corrélation empirique de Pearson r^2
- des indices de discrimination documentés dans la littérature :
    - L'Indice de discrimination, qui est en réalité encore une fois un simple coefficient de corrélation de Pearson
    - Indice *B* de Brennan(24)
    - Indice de Findley (d-index)(25)

Nous avons choisi d'utiliser un test de Wilcoxon devant son caractère non-paramétrique et de distributions de notes en générales décentrées(26) dans les classes d'étudiants.

Cependant devant l'absence de notion de force de l'effet discriminant d'une p valeur, nous avons également effectué un indice d de Findley et un « indice de discrimination »pour chaque question.

L'indice de Findley varie entre -1 et 1, il est considéré comme acceptable au dessus de 0.5, comme discutable entre 0 et 0.5, et inacceptable (contre-sens) en dessous de 0.

On remarquera une interprétation du d-index très similaire au coefficient de corrélation de Pearson, nous effectuerons un coefficient de corrélation intraclasse pour évaluer leur concordance.

3.4 Analyse par composante principale

Une analyse par composante principale a également été effectuée sur la variable Note et sur toutes les variables “question_numéro_x_réussie”.

3.4.1 Composantes principales

Une épreuve bien conçue et bien comprise par les étudiants aura classiquement une première composante principale expliquant une partie conséquente de la variance totale et qui correspondra dans les grandes lignes au niveau global de l’étudiant. Les composantes suivantes devraient être moins contributives.

Dans un épreuve problématique, on peut s’attendre à ce qu’au moins une deuxième composante, par définition indépendante de la première, porte une part importante de la variance totale.

3.4.2 Note totale

La note totale sera attendue comme un vecteur quasiment uniquement orienté dans la première composante principale, et donc son amplitude dans les autres composantes sera attendue très faible.

3.4.3 Questions individuelles

Les questions départageant bien les étudiants seront attendues comme des vecteurs dont l’amplitude dans la première composante sera maximale et dans le sens du vecteur “Note totale”.

Une question ne départageant pas les étudiants aura une amplitude faible dans la première composante.

Une question dont la correction est un contre-sens aura une forte amplitude dans la première composante mais dans le sens contraire de la note globale.

Résultats

1 Correction d'une épreuve classique

1.1 Introduction

Pour effectuer une analyse de la qualité de la correction à grande échelle, une épreuve d'UE5 Anatomie de la PACES 2019 nous a été fournie :

- 3255 copies physiques
- Résultats de la correction officielle au format csv

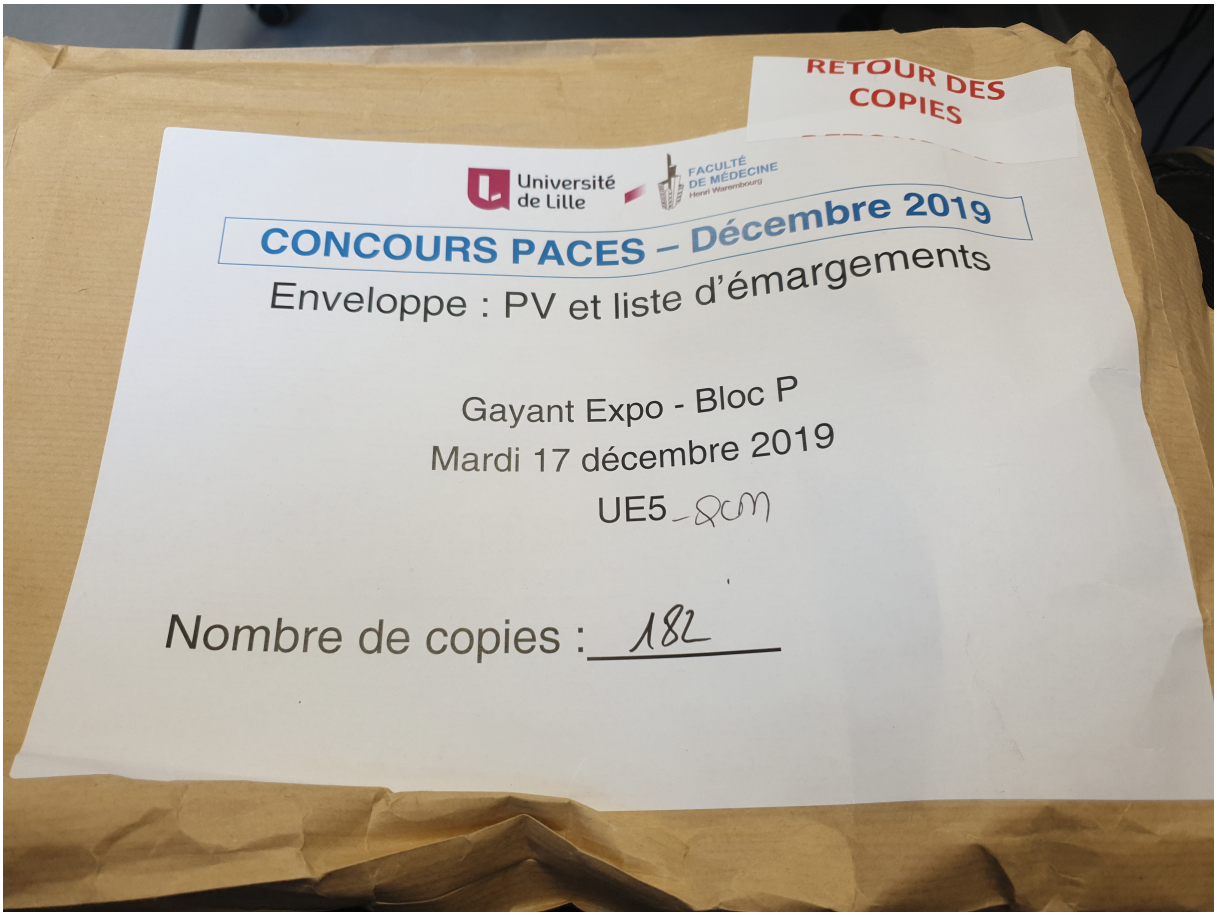

Figure 19: Un des cartons de copies de l'épreuve

1.2 Numérisation

Les 3255 fiches optiques ont été numérisées en 10 minutes sur un photocopieur du CERIM.

Les images JPEG produites ont été transférés avec une clé USB 2.0 de 8Go vers l'ordinateur effectuant l'analyse optique, la totalité des fichiers numérisés pesait 800Mo.

1.3 Analyse optique

L'ordinateur utilisé pour effectuer la correction était un ordinateur portable du CERIM.

- CPU : Intel(R) Core(TM) i7-8650U @ 1.90GHz
- SSD : NVMe Samsung PM981 256GB
- RAM : 16GB

L'ensemble des copies a été corrigé par Osear en 50s, en utilisant 100 % de la puissance du processeur et 300Mo de mémoire.

Le facteur limitant lors de l'exécution a été la puissance du CPU, mais le facteur limitant global lors de la correction dans son ensemble a été l'étape de numérisation.

L'analyse optique du jeu de copies a donc conduit, grâce au programme décrit précédemment à la production du fichier csv contenant les réponses de chaque étudiant.

1.4 Concordance

Aucune copie n'a été marquée comme « à risque d'erreur ».

Aucune discordance entre Osear et la machine officielle n'a été observée.

Le kappa de Cohen calculé était donc de 1, IC95 % = [1;1], devant l'absence complète de variance du kappa, la p-value était incalculable.

1.5 Données descriptives

210 étudiants attendus ont été absents et leurs données ont été retirées de l'analyse, 3045 copies ont donc été analysées.

La médiane des notes dans l'épreuve était de 7.5/20 avec un intervalle inter-quartile de [4.9; 11.9].

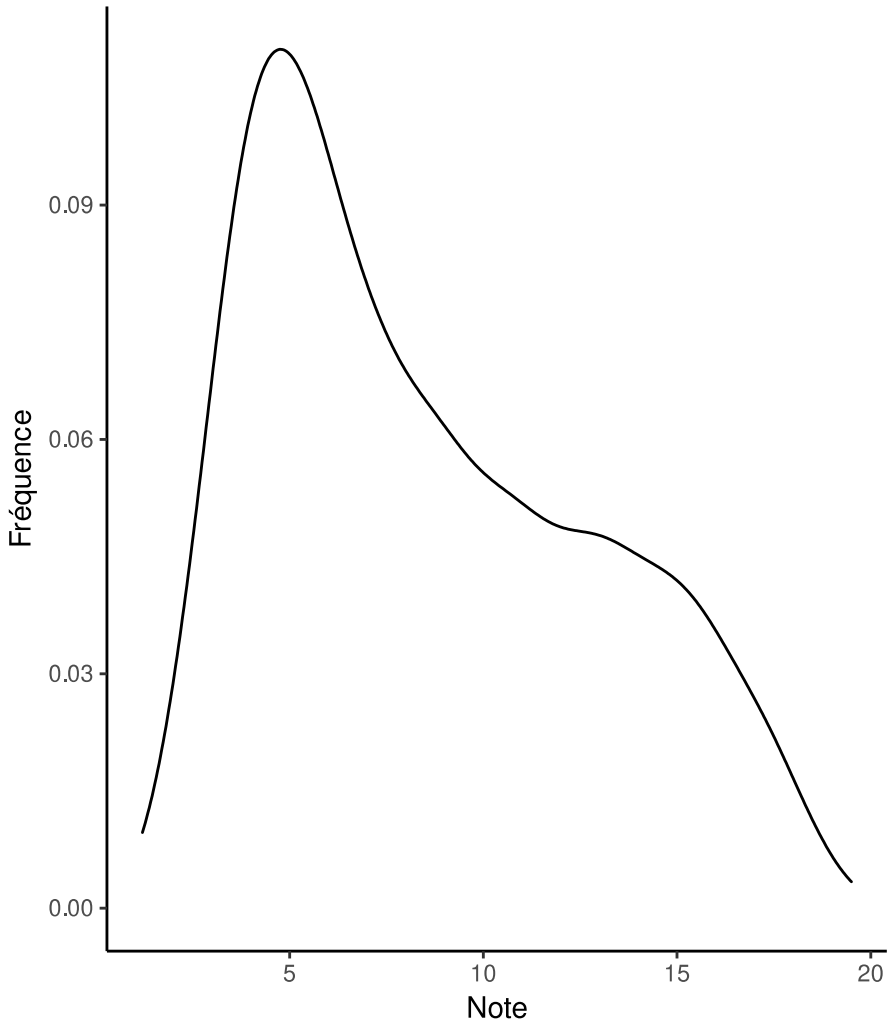

Figure 20: Distribution des notes des étudiants.

2 Nouvelles modalités de correction – Capacité à automatiser

2.1 Correction d'une épreuve à QCM bayésiens

Une épreuve de test à été effectuée avec la méthode des QCM bayésiens.

30 copies ont été produites et corrigées d'une part avec Osear et d'autre part à la main.

La concordance entre les deux méthodes de correction a été parfaite avec un kappa de Cohen de 1, IC95 % = [1;1] et devant l'absence complète de variance du kappa, la p-value était incalculable.

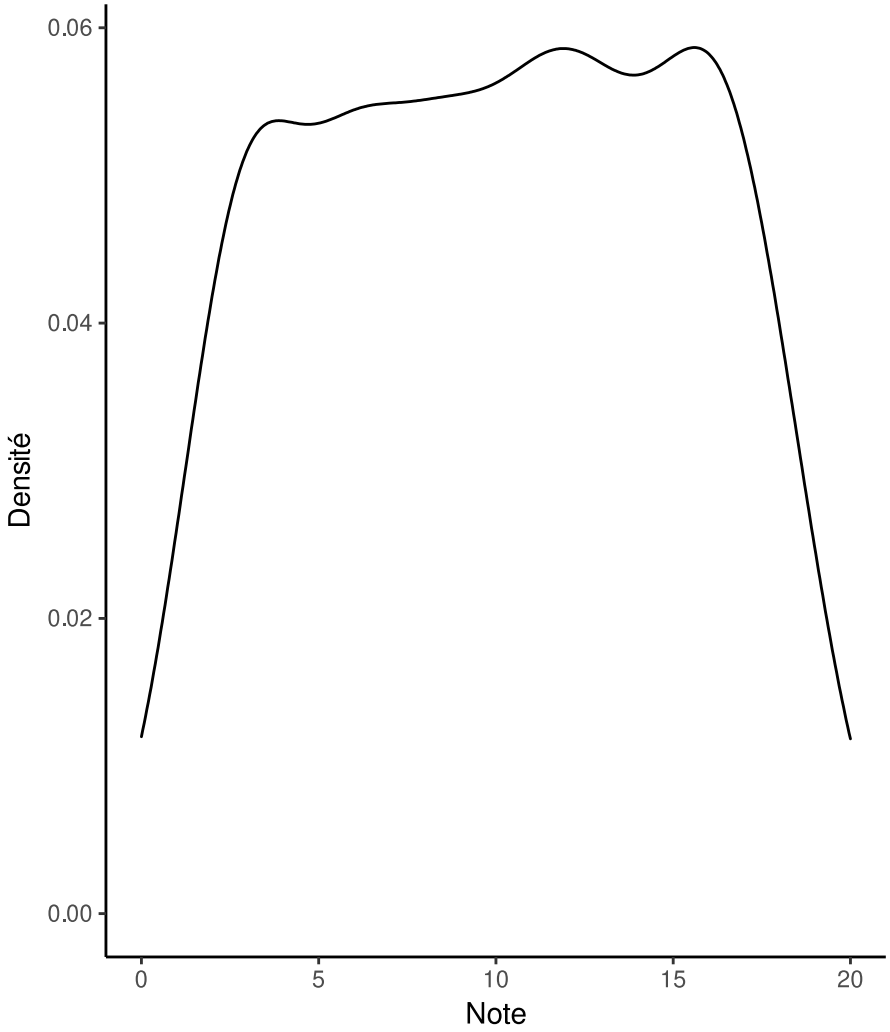

Figure 21: Distribution des notes de l'examen test à QCM bayésiens

2.2 Correction d'une épreuve à nombreuses réponses

Une épreuve de test à été effectuée avec la méthode des QCM à nombreuses réponses possibles.

35 copies ont été produites et corrigées d'une part avec Osear et d'autre part à la main.

La concordance entre les deux méthodes de correction a été parfaite avec un kappa de Cohen de 1, IC95 % = [1;1] et devant l'absence complète de variance du kappa, la p-value était incalculable.

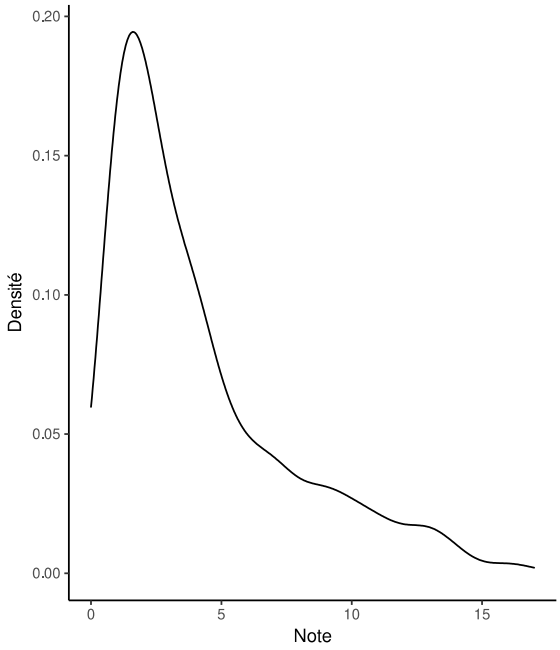

Figure 22: Distribution des notes de l'examen test à QCM avec nombreuses réponses possibles

2.3 Correction d'un schéma anatomique

Une épreuve de test à été effectuée avec la méthode du schéma anatomique.

Une seule question, « coloriez l'aorte en bleu », a été posée.

Un médecin a manuellement rempli 25 fiches optiques avec des réponses justes et fausses tout en les labellisant juste/faux dans un tableau. Nous avons ensuite scanné et corrigé les copies avec Osear.

La concordance observée a été nettement inférieure aux autres méthodes avec un kappa de 0.75, un intervalle de confiance à 95 % de [0.49 ; 1], avec tout de même une p-value inférieure au seuil 5 %.

3 Validation docimologique

3.1 Analyse question par question

D'après les tests de Wilcoxon, toutes les questions ont séparé significativement (au seuil de 5%) les meilleurs étudiants des moins bons, cependant une simple valeur p ne donnant pas de notion de la force de l'effet discriminant.

Pour autant tous les indices de Findley ont été entre 0 et 0.5, c'est à dire dans la tranche « discutable ».

Aucune question n'a eu d'index de Findley ou de coefficient de corrélation en dessous de 0, ce qui aurait signalé une question « contre-sens ».

Table 2: Indices docimologiques question par question

	P Valeur du Wilcoxon	Indice d de Findley	Indice de discrimination
Question 1	1.4e-205	0.36	0.57
Question 2	8.3e-199	0.28	0.59
Question 3	1.5e-185	0.30	0.55
Question 4	5.0e-173	0.33	0.52
Question 5	2.7e-57	0.32	0.26
Question 6	3.9e-130	0.18	0.48
Question 7	1.6e-142	0.32	0.47
Question 8	3.4e-27	0.05	0.22
Question 9	3.9e-230	0.31	0.63
Question 10	2.1e-234	0.46	0.59
Question 11	8.9e-236	0.38	0.62
Question 12	1.3e-87	0.13	0.41
Question 13	1.8e-187	0.24	0.57
Question 14	7.4e-104	0.16	0.44
Question 15	1.4e-209	0.32	0.59
Question 16	2.9e-135	0.18	0.50
Question 17	5.6e-206	0.25	0.61
Question 18	4.4e-214	0.34	0.59
Question 19	1.2e-94	0.20	0.40
Question 20	8.2e-151	0.27	0.51

3.1.1 Indice d de Findley vs comparaison visuelle

Voici également ci-dessous une comparaison graphique des distributions conditionnelles entre la question la plus et la moins discriminante selon l'index d de Findley.

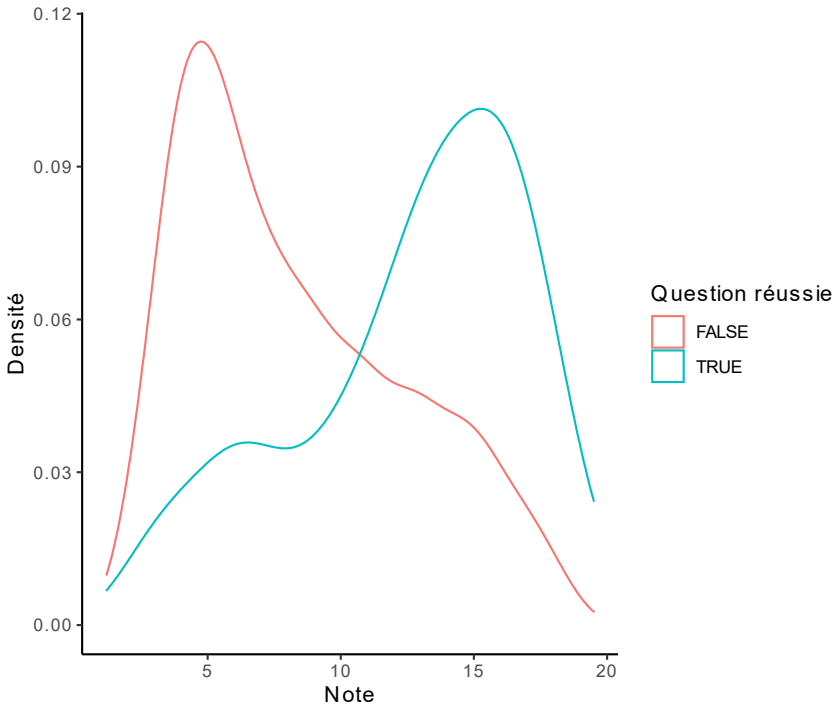

Figure 23: Question 8 : la moins discriminante selon l'index d de Findley valant ici 0.05

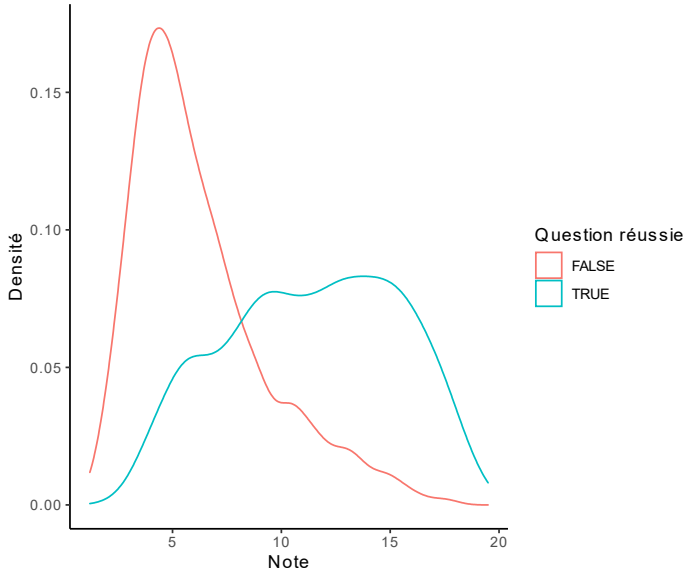

Figure 24: Question 10 : la plus discriminante selon l'index d de Findley valant ici 0.46

Visuellement la pertinence de l'index d de Findley peut sembler discutable.

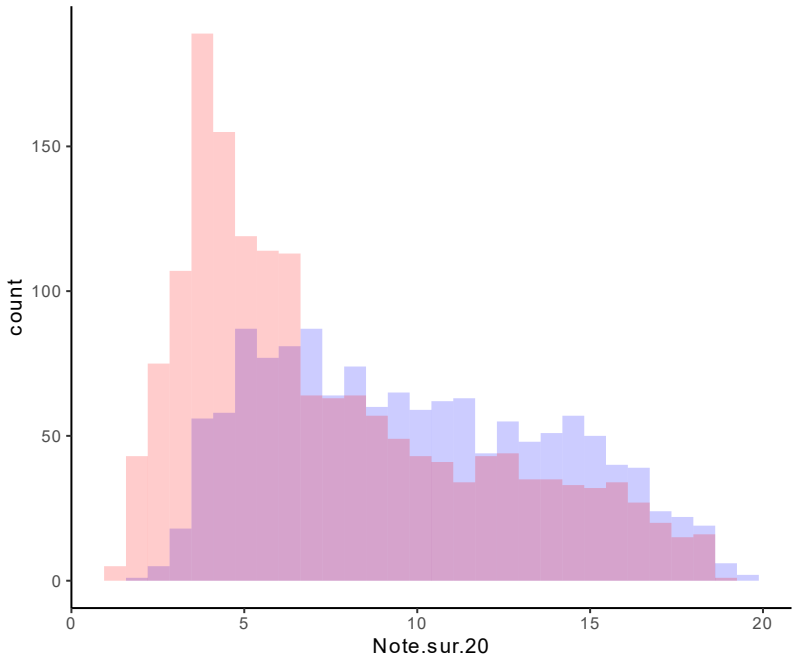

Figure 25: Histogramme de la question 5, dont l'index d est de 0.32

En regardant l'histogramme de la question 5 qui visuellement paraît assez peu discriminante, ayant un index d de Findley de 0.32, on peut également douter de la pertinence de l'index.

3.1.2 Indice d de Findley vs Pearson

L'index d de Findley s'est révélé peu concordant avec le coefficient de corrélation de Pearson, avec un indice de corrélation intra-classe faible de 0.173, avec un IC95 % de [0, 0.528] et une p-value de 0.188 sous l'hypothèse nulle d'absence de concordance.

3.2 Analyse par composantes principales

Une composante principale se dégage particulièrement, avec une contribution à la variance totale de 32.13 %.

Table 3: 5 premières composantes principales

Composante	Eigenvalue	Contribution à la variance
1	6.75	32.13
2	1.04	4.971
3	0.97	4.642
4	0.95	4.509
5	0.87	4.128

La note sur 20 est en très grande majorité orientée dans cette composante avec une coordonnée de 0.97 dans cette composante.

Lors de l'analyse graphique des composantes principales on peut remarquer que la majorité des questions sont également orientées dans le sens de la note sur 20, ce qui vient corroborer l'hypothèse d'un bon pouvoir discriminant de chaque question.

On remarquera cependant la question 5 ayant une orientation principalement dans la 2ème composante principale. Ceci est relativement attendu étant donné l'histogramme observé précédemment : avec des notes très similaires entre les étudiants ayant réussi et raté cette question.

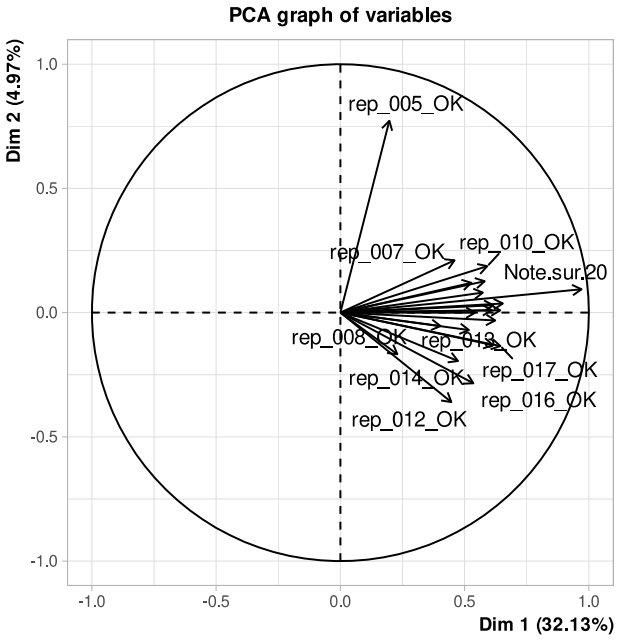

Figure 26: Projection des questions dans les 2 composantes principales

Discussion

1 Principaux résultats

L'objectif stratégique était de libérer l'accès aux QCM à grande échelle en baissant le coût d'accès à cette méthode, tout en ouvrant la possibilité à des techniques d'évaluation novatrices permettant de répondre à certains problèmes rencontrés par l'utilisation des QCM classiques.

Pour ce faire, notre objectif opérationnel était de développer et évaluer une méthode indépendante des périphériques permettant de corriger des fiches optiques de manière fiable et rapide.

Notre solution de correction automatisée a permis de corriger 3255 copies en un peu plus de 10 minutes tout en utilisant uniquement du matériel déjà présent dans le secrétariat du CERIM et ce avec une concordance observée de 100 % par rapport à la machine de correction traditionnelle.

2 Discussion des résultats

2.1 Correction de l'épreuve à QCM classiques

L'absence complète de discordance indiquée par la valeur de 1 du kappa sur le jeu de 3255 copies est encourageante, même si des tests à plus grande échelle encore serait souhaitable.

On peut noter la rapidité d'exécution de la nouvelle méthode de correction sur un matériel raisonnablement accessible. On peut acheter un scanner avec une mécanique d'entraînement de 25 pages par minute pour 230€ dans le commerce et n'importe quel ordinateur de moins de 10 ans corrigera les copies plus rapidement que le scanner pourra numériser. On pourra évidemment également utiliser n'importe quel photocopieur-scanner déjà présent dans la majorité des secrétariats des facultés pour corriger les copies, ces matériels ont généralement une mécanique d'entraînement bien plus rapide que les scanners bas de gamme(27).

Selon nos observations, la qualité de la mécanique d'entraînement impacte essentiellement la rotation de la page numérisée. Si elle était un souci majeur au début du développement, est devenue secondaire. En effet, grâce aux systèmes de repérages de notre solution logicielle, toute anomalie de rotation légère de moins de 10° est parfaitement compensée par le logiciel28. Cependant on notera tout de même que l'utilisation d'un matériel haut de gamme baisse probablement le risque d'erreur de correction, même si le risque a été en pratique indétectable dans notre instance de correction.

Le jeu de copie pourra être stocké physiquement mais également dans sa version numérisée à des buts de traçabilité. En effet la gestion électronique de document reste d'un intérêt majeur à des fins d'archivage(28).

Les résultats de cette correction sont difficiles à comparer avec les données de la littérature étant donné l'absence de publication concernant ce sujet précis.

2.2 Nouvelles modalités de correction

Osear s'est montré capable de corriger des épreuves a base de QCM bayésiens et à nombreuses réponses possible sans problème et avec une concordance parfaite.

La concordance a été nettement moins bonne sur la méthode d'examen du schéma anatomique, mais tout de même supérieure à 0 au seuil alpha de 5 %. Cependant étant donné l'objectif d'une concordance parfaite, les fonctions logicielles de correction de ce type d'examen ont une importante marge de progression à réaliser avant de pouvoir être raisonnablement utilisées en pratique.

Une méthode de correction basée sur du deep-learning(29) entraîné sur des données labellisées(30) pourrait être une solution permettant d'atteindre une meilleure concordance sur les schémas d'anatomie à compléter.

2.3 Validation docimologique

Les différentes techniques de validation docimologique (analyse visuelle, Wilcoxon, d index et Pearson), ont toutes conclu à une certaine qualité docimologique de l'examen d'anatomie.

Le d-index de Findley s'est montré peu concordant avec les autres techniques de validation docimologique, on pensera notamment à la question 5 qui conjuguait d index relativement élevé et une distribution visuellement peu convaincante sur le plan de sa capacité à discriminer les étudiants. De manière globale, la concordance entre le d-index de Findley et le coefficient de corrélation de Pearson s'est révélée faible, ce qui est inquiétant devant des indices dont l'interprétation est si proche en pratique(25).

Devant l'absence de démonstration mathématique de la pertinence du d-index et la faible concordance avec des techniques statistiquement validées comme le test de Wilcoxon et le coefficient de corrélation de Pearson, on peut sincèrement douter de l'utilité de cette métrique.

3 Discussion de la méthode

3.1 Tests de concordance

Nous avons évalué la concordance entre notre méthode de correction et la méthode actuellement employée à la Faculté de Médecine, sur un jeu de 3255 copies issues d'un examen d'anatomie. Le principal point fort de cette évaluation est probablement la puissance statistique grâce à l'effectif très important de 3255 copies.

L'absence totale de valeurs manquantes et la présence de résultats officiels de référence permettent de s'assurer d'une qualité des données brutes élevée.

L'utilisation d'un coefficient kappa n'est pas forcément très adaptée lorsque l'objectif est une concordance absolument parfaite entre deux méthodes de correction, c'est à dire quand le kappa souhaité est strictement égal à 1(31).

3.2 Validation docimologique

Il y a plusieurs limites liées aux méthodes de validation docimologiques utilisées. La plus évidente est que juger la pertinence des questions uniquement sur leur capacité à discerner les étudiants entre eux ne mesure en aucun cas le niveau de connaissance absolue des étudiants.

Par ailleurs cette méthode ne permet pas directement la mesure du niveau de compétence de l'étudiant, qui constitue en réalité le but principal d'une formation professionnelle(5,6).

Il serait souhaitable de tenter de mesurer dans des études futures à quel point la mesure des connaissances en relatif, la mesure des connaissances en absolu et la mesure des compétences concordent en pratique.

Une autre limite de notre méthode est l'exploitation de peu d'examens facultaires différents pour juger de la qualité de la validation docimologique, conduisant à des conclusions peu généralisables. Il serait probablement utile de mener des travaux de recherche sur la concordance entre l'avis d'enseignants sur la qualité docimologie de l'examen et les méthodes automatisées utilisées dans notre étude.

On pourra aussi noter qu'il est dommage de ne pas avoir pu effectuer la validation docimologique des examens innovants, il serait utile de mener une étude à plus grande échelle pour comparer leur qualité docimologique avec les méthodes plus classiques.

4 Perspectives

Ce travail ne va pas en soi faire évoluer la formation des médecins, mais il lève un obstacle technique à des progrès pédagogiques et docimologiques. Il protège également la Faculté de Médecine d'une possible panne survenant à un moment critique, comme par exemple juste après la première session d'examens de PASS, alors que les résultats doivent être rendus suffisamment tôt pour permettre une réorientation d'une partie de la promotion.

La barrière d'accès posée par l'acquisition et de la maintenance soit de tablettes tactiles, soit de machines de corrections spécialisée peut ainsi être levée.

On pensera notamment à l'institut des sages-femmes corrigeant toujours les épreuves QCM à la main, et qui s'est annoncé intéressé par l'utilisation de notre solution gratuite.

Par ailleurs l'inclusion de la validation docimologique dans le processus de correction peut permettre de corriger des erreurs ou de proposer un feed-back aux enseignants concernant leurs examens.

Conclusion

Pour conclure, nous avons d'une part créé une solution qui permet l'utilisation de techniques d'évaluations QCM tant classiques que nouvelles, et ce pour un coût d'accès extrêmement faible.

D'autre part nous avons évalué des formes d'examens compatibles avec la technique des QCM proposant certains avantages par rapport à la méthode classique.

Liste des tables

Liste des figures

Références

1. Directive 2005/36/CE du Parlement européen et du Conseil du 7 septembre 2005 relative à la reconnaissance des qualifications professionnelles (Texte présentant de l'intérêt pour l'EEE) [Internet]. 255, 32005L0036 sept 30, 2005. Disponible sur: http://data.europa.eu/eli/dir/2005/36/oj/fra

2. Article D613-7 - Code de l'éducation - Légifrance [Internet]. [cité 26 mars 2021]. Disponible sur: https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000037219445

3. Docimologie. In: Wikipédia [Internet]. 2020 [cité 26 mars 2021]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Docimologie&oldid=171126706

4. N° 2475 - Rapport d'information de Mme Cécile Untermaier et M. Philippe Houillon déposé en application de l'article 145 du règlement, par la commission des lois constitutionnelles, de la législation et de l'administration générale de la République, en conclusion des travaux d'une mission d'information sur les professions juridiques réglementées [Internet]. [cité 26 mars 2021]. Disponible sur: https://www.assemblee-nationale.fr/14/rap-info/i2475.asp

5. Watson R, Stimpson A, Topping A, Porock D. Clinical competence assessment in nursing: a systematic review of the literature. J Adv Nurs. 2002;39(5):421-31.

6. Wass V, Van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. The Lancet. 24 mars 2001;357(9260):945-9.

7. Proxy (variable). In: Wikipédia [Internet]. 2020 [cité 30 mars 2021]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Proxy_(variable)&oldid=166163625

8. Wood RE, Beckmann JF, Birney DP. Simulations, learning and real world capabilities. Educ Train. 26 juin 2009;51(5-6):491-510.

9. Pseudoscience. In: Wikipedia [Internet]. 2021 [cité 30 mars 2021]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Pseudoscience&oldid=1014254497

10. Multiple choice. In: Wikipedia [Internet]. 2021 [cité 30 mars 2021]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Multiple_choice&oldid=1005382247

11. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Med Teach. 1 déc 2004;26(8):709-12.

12. ECN informatisées : ce qui va changer pour les étudiants en 2016 [Internet]. Le Quotidien du médecin. [cité 27 mars 2021]. Disponible sur: https://www.lequotidiendumedecin.fr/internes/etudes-medicales/ecn-informatisees-ce-qui-va-changer-pour-les-etudiants-en-2016

13. Sommestad T, Holm H, Ekstedt M. Estimates of success rates of remote arbitrary code execution attacks. Inf Manag Comput Secur. 1 janv 2012;20(2):107-22.

14. Shin Y, Meneely A, Williams L, Osborne JA. Evaluating Complexity, Code Churn, and Developer Activity Metrics as Indicators of Software Vulnerabilities. IEEE Trans Softw Eng. nov 2011;37(6):772-87.

15. Cotroneo D, Natella R, Pietrantuono R. Predicting aging-related bugs using software complexity metrics. Perform Eval. 1 mars 2013;70(3):163-78.

16. Médecine : et de deux épreuves d'ECN annulées, l'état de crise décrété [Internet]. L'Etudant. [cité 30 mars 2021]. Disponible sur: https://www.letudiant.fr/etudes/medecine-sante/medecine-deux-epreuves-des-ecn-annulees-etat-de-crise-decrete.html

17. Elstein AS. Beyond multiple-choice questions and essays: The need for a new way to assess clinical competence. Acad Med. 1993;68(4):244-9.

18. Scoring rule. In: Wikipedia [Internet]. 2021 [cité 26 mars 2021]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Scoring_rule&oldid=1006554244

19. Science4All. Les QCM bayésiens [Internet]. 2019 [cité 26 mars 2021]. Disponible sur: https://www.youtube.com/watch?v=1fuIG7rhIXo

20. Lê Nguyên Hoang L. Bayesian examination - LessWrong [Internet]. [cité 2 févr 2021]. Disponible sur: https://www.lesswrong.com/posts/8wzKawHmh4d3h2otw/bayesian-examination

21. Dufresne RJ, Leonard WJ, Gerace WJ. Marking sense of students' answers to multiple-choice questions. Phys Teach. mars 2002;40(3):174-80.

22. Considine J, Botti M, Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian. janv 2005;12(1):19-24.

23. Accès aux études de santé en France. In: Wikipédia [Internet]. 2021 [cité 27 mars 2021]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Acc%C3%A8s_aux_%C3%A9tudes_de_sant%C3%A9_en_France&oldid=180494977

24. Brennan RL. A Generalized Upper-Lower Item Discrimination Index. Educ Psychol Meas. 1 juill 1972;32(2):289-303.

25. Findley WG. A Rationale for Evaluation of Item Discrimination Statistics. Educ Psychol Meas. 1 juill 1956;16(2):175-80.

26. Arthurs N, Stenhaug B, Karayev S, Piech C. Grades Are Not Normal: Improving Exam Score Models Using the Logit-Normal Distribution [Internet]. International Educational Data Mining Society. International Educational Data Mining Society; 2019 [cité 6 avr 2021]. Disponible sur: https://eric.ed.gov/?id=ED599204

27. Scanner (informatique). In: Wikipédia [Internet]. 2020 [cité 27 mars 2021]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Scanner_(informatique)&oldid=172743388

28. Piwowar HA, Vision TJ, Whitlock MC. Data archiving is a good investment. Nature. mai 2011;473(7347):285-285.

29. Deep learning. In: Wikipedia [Internet]. 2021 [cité 30 mars 2021]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Deep_learning&oldid=1012107396

30. Supervised learning. In: Wikipedia [Internet]. 2021 [cité 30 mars 2021]. Disponible sur: https://en.wikipedia.org/w/index.php?title=Supervised_learning&oldid=1014088467

31. Sim J, Wright CC. The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. Phys Ther. 1 mars 2005;85(3):257-68.

AUTEUR : Nom : BENTEGEAC **Prénom :** RAPHAEL

Date de Soutenance : 14/06/2021

Titre de la Thèse : Docimologie dans les formations pour les professions de santé

Thèse - Médecine - Lille 2020

Cadre de classement : Médecine Générale

DES + spécialité : Santé Publique

Mots-clés : Docimologie, QCM, Indices de discrimination

Résumé :

Contexte :

La correction optique d'épreuves de QCM sur papier nécessite l'utilisation de solutions commerciales « packagées ». Ces solutions lient entièrement : des copies imprimées par l'éditeur, immuables et non paramétrables, une machine réalisant à la fois la lecture et la correction, des algorithmes de notation imposés, un logiciel propriétaire, et une machine dont le système d'exploitation est imposé. Cette rigidité fait obstacle à l'évolution docimologique, et à la résilience en cas de panne. En outre, les solutions du marché ne proposent pas d'analyse évoluée des résultats. Notre objectif est de spécifier et produire une solution logicielle paramétrable et indépendante du matériel, et de proposer des méthodes d'analyse statistique des résultats.

Matériel et Méthodes :

Nous concevons, développons et évaluons une méthode de correction ne reposant pas sur une machine spécialisée et nous comparons la concordance entre cette nouvelle méthode et la machine d'origine sur un examen de PACES. Nous évaluons la capacité de la nouvelle méthode de correction à corriger des examens par modalités QCM innovantes. Nous évaluons la qualité docimologique d'un examen à l'aide de plusieurs critères quantifiables issus de la littérature.

Résultats :

La nouvelle méthode de correction ne reposant pas sur une machine spécialisée a une concordance parfaite avec la machine spécialisée pour l'épreuve testée (n=3255, k=1). La nouvelle méthode de correction est capable de corriger les modalités innovantes testées (QCM avec niveau de certitude, QCM à réponses très nombreuses, schéma d'anatomie) avec une concordance parfaite sauf pour la modalité par complétion de schéma anatomique. L'examen de PACES possède des critères de qualité docimologique satisfaisant selon le d-index de Findley, l'indice de discrimination, le test de Wilcoxon et l'analyse par composante principale.

Conclusion :

La nouvelle méthode de correction permet à la fois un coût faible de correction, une résilience complète en dissociant les étapes matérielles et logicielles, et la mise en place de nouvelles modalités d'examen.

Composition du Jury :

Président : Monsieur le Professeur Philippe AMOUYEL

Assesseurs : Monsieur le Professeur Grégoire FICHEUR

Monsieur le Docteur Pierre BALAYE

Monsieur le Professeur Emmanuel CHAZARD