Dispositifs d'évaluation à distance à correction automatisée versus non automatisée : analyse comparative de deux formes emblématiques

Jean-Luc Gilles – jean-luc.gilles@hepl.ch
Haute école pédagogique du canton de Vaud, Suisse

Bernadette Charlier – bernadette.charlier@unifr.ch
Université de Fribourg, Suisse

Pour citer cet article : Gilles, J.-L., & 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,* Numéro Hors-série, *1*, 143-154.

Résumé
La crise sanitaire Covid-19 conduit chaque enseignant à envisager de concevoir des évaluations en ligne. Dès lors, la sensibilisation de ces derniers à la qualité de leurs dispositifs est cruciale, d'autant plus que la régulation efficace des enseignements-apprentissages à distance requiert des dispositifs d'évaluation fiables. Partant de ce constat, nous rappelons neuf critères de qualité tirés de la littérature et devant s'appliquer à tout type d'évaluation. Nous avons ensuite comparé deux formes emblématiques d'évaluations utilisées dans le contexte de la pandémie : les Questions à choix multiple (QCM) en ligne et les Questions à réponses ouvertes longues (QROL) gérées à distance. L'une autorise une correction entièrement automatisée et l'autre nécessite l'intervention de l'enseignant lors de la correction. L'examen de ces dispositifs met en évidence des avantages et des inconvénients pour chacune des deux formes. Il apparaît que les avantages de l'une pourraient compenser les désavantages de l'autre et, dans le contexte actuel de pandémie où des dispositifs d'évaluation à distance se mettent en place dans l'urgence, nous conseillons aux enseignants de ne pas choisir l'une au détriment de l'autre. Une utilisation combinée est préférable de manière à tirer parti des avantages des deux formes d'évaluation tout en limitant les désavantages de chacune. Nous concluons en soulignant que le gain de temps permis par la correction automatisée des QCM pourrait être avantageusement mis à profit par les enseignants pour améliorer la fiabilité de la correction des QROL et peaufiner des feedbacks dans une perspective de soutien aux apprentissages.

Mots-clés
Evaluation, évaluation à distance, critères qualité, docimologie, correction automatisée, questions à réponses ouvertes, questions à choix multiple, évaluation combinée.

Abstract

The Covid-19 health crisis is leading every teacher to consider designing online assessments. Consequently, it is crucial to make them aware of the quality of their systems, especially as the effective regulation of distance learning requires reliable evaluation systems. On the basis of this observation, we recall nine quality criteria drawn from the literature that should be applied to any type of evaluation. We then compared two emblematic forms of evaluation used in the context of the pandemic: online Multiple-Choice Questions (MCQs) and Open-ended Questions (OQs) managed via the Internet. One allows for fully automated scoring and the other requires teacher intervention during scoring. A review of these types of assessments reveals advantages and disadvantages for both. It appears that the advantages of one may outweigh the disadvantages of the other, and in the current pandemic context where remote assessment systems are being implemented in an emergency, teachers are advised not to choose one at the expense of the other. A combined use is preferable so as to take advantage of the benefits of both while limiting the disadvantages of each. We conclude by emphasizing that the time saved by the automated correction of MCQs could be advantageously used by teachers to improve the reliability of the correction of the QROLs and to refine feedback in a Assessment for Learning (AfL) perspective.

Keywords

Assessment, remote assessment, quality criteria, docimology, automated correction, open-ended questions, multiple choice questions, blended assessment.

1. Introduction

La crise sanitaire Covid-19 affectait au 12 avril 2020 91,3% des étudiants de la planète, soit à cette date près de 1,6 milliard d'apprenants dont l'établissement scolaire était fermé dans 188 pays (UNESCO, 2020). Dans ce contexte, avec plus de la moitié de la population mondiale confinée, la plupart des systèmes éducatifs se lancent dans des expériences sans précédent d'enseignement à distance et tentent ainsi d'éviter les ruptures dans les parcours de formation.

> Comme l'indique l'équipe spéciale internationale sur les enseignants pour l'éducation 2030 (PME, 2020) :
>
> *Les fermetures d'écoles dues à la crise du COVID-19 ont été soudaines. Peu d'attention a été accordée à la formation adéquate des enseignants sur la manière de garantir la poursuite de l'apprentissage ou sur la manière d'élaborer des programmes d'enseignement à distance pertinents et de qualité. Les enseignants, dont les compétences en matière de technologie numérique varient, doivent maintenant s'adapter aux nouvelles plateformes d'apprentissage et élaborer de nouvelles stratégies pour faire participer les enfants, tout en maintenant des normes élevées d'enseignement et d'apprentissage. Pour relever ces défis, les gouvernements et les autres parties prenantes doivent agir rapidement pour s'assurer que les enseignants reçoivent la formation nécessaire.* (p. 2).

Ce changement abrupt lié à la crise sanitaire entraîne également une nouvelle nécessité : la mise en place de dispositifs d'évaluation à distance soutenant ou certifiant les apprentissages et permettant la régulation des enseignements. Des pratiques réelles (et non idéales) émergent dans le cadre de scénarios d'usages adoptés par les enseignants. Dans un tel contexte « Il s'agit d'inventer des pratiques, minimisant le coût de la tâche, adaptées à celle-ci et tirant parti des apports spécifiques des TIC » (Charlier, 2010, p. 147).

L'accès aux solutions technologiques visant à assurer la continuité des apprentissages nécessite des ressources telles qu'une (bonne) connexion à internet, le matériel hardware et les softwares *ad hoc* pour les enseignants, les apprenants et les familles dépendant du niveau scolaire considéré. Dans plusieurs milieux, cet accès aux ressources peut poser problème. C'est encore plus vrai dans les contextes du Sud où même l'approvisionnement en électricité n'est pas garanti. Par ailleurs, que ce soit au Nord ou au Sud, même lorsque les ressources sont là, encore faut-il que le lieu de confinement réunisse les conditions de quiétude, d'environnement et de climat favorables aux apprentissages. Comment éviter que les apprenants issus de milieux défavorisés soient encore plus laissés pour compte lors de la mise en place d'évaluations à distance dans ce contexte Covid-19 ?

En plus des questions d'accès aux ressources, Hettiarachchi et Huertas (2013) mettent en évidence dans le domaine des évaluations à distance les préoccupations des praticiens concernant la détection du plagiat et la surveillance pour éviter la tricherie. Audet (2011) considère qu'il s'agit d'un véritable défi en évaluation à distance, ajoutant à la tricherie et au plagiat la fabrication et la contrefaçon ainsi que le sabotage.

Parmi les enjeux préoccupants dans ce contexte de confinement généralisé des apprenants, la qualité des évaluations proposées à distance devrait nous préoccuper au premier plan. Prendre des décisions qui permettent de certifier ou de réguler efficacement les apprentissages et les enseignements à distance, nécessite des informations fiables récoltées avec des dispositifs dont on s'est assuré qu'ils répondent à des critères de qualité sur le plan docimologique et technique (Blais, Gilles & Tristan-Lopez, 2015). Mais comment assurer la

qualité des évaluations à distance si les enseignants ne se sont pas sensibilisés à ces aspects dès la mise en œuvre de leurs scénarios ?

2. Critères de qualité des dispositifs d'évaluation à distance

Il importe de clarifier les attentes quant à la qualité des évaluations à distance, qu'il s'agisse de prendre des décisions de régulation des apprentissages et des enseignements ou des décisions centrées sur la certification. Reprenons ici les propos de Brown (2019) à propos des qualités que devraient démontrer les évaluations des apprentissages :

> *There needs to be empirical and theoretical evidence supporting the interpretations and decisions being made from the test* (Messick, 1989). *That evidence must cover the processes used to ensure the validity of the test design, administration, and interpretation—that is:*
>
> • *Can you show that the test itself validly represents the domain?*
>
> • *Was it administered fairly and properly? and*
>
> • *Were appropriate procedures used to evaluate the data arising from the assessment?*
>
> *Secondly, the evidence must show that the scoring processes were reliable, accurate, and credible. In other words, questions such as these need to be addressed:*
>
> • *Were the right and wrong answers given the right score?*
>
> • *Would another judge give, following the same scoring protocol, the same or nearly the same score? or*
>
> • *Was the scoring free from biases?*
>
> *Without evidence that the design, implementation, and scoring were done in a robust way, the testing process fails to meet fundamental requirements, and should not be used as the basis of decision-making. These are the standards and expectations the psychometric industry places on tests* (AERA, APA, and NCME, 2014). *My view of assessment is, notwithstanding its non-standardized or non-systematic procedures, if it is to be the basis for decisions about students* (see Newton, 2007 for 17 different purposes or functions to which assessments can be put)*, that it needs to be judged against the criteria by which standardized tests are evaluated. Otherwise, assessment practices do not merit the term assessment.* (p. 2).

En vue de prises de décisions éducatives sur base d'informations fiables, les chercheurs en docimologie tentent de clarifier depuis des décennies les processus d'élaboration des évaluations (p. ex. Ebel, 1972 ; De Landsheere, 1980 ; Cardinet, 1986 ; Millman & Greene, 1989 ; Mehrens & Lehmann, 1991 ; Gilles & Leclercq, 1995 ; Boulet, McKinley, Whelan & Hambleton, 2003 ; Downing & Haladyna, 2006 ; Tillema, Leenknecht & Segers, 2011). Les critères d'évaluation de la qualité des dispositifs d'évaluation font également l'objet de discussions, notamment des débats à propos de l'évolution des conceptions principalement en matière de validité et de fidélité (p. ex. Kuder & Richardson, 1937 ; Cronbach, 1951 ; Ebel, 1969 ; Messick, 1988, 1995 ; AERA, APA et NCME, 1985, 1999, 2014 ; Angoff, 1988 ; Downing & Haladyna, 1997 ; Popham, 1997 ; Phelps, 2005 ; Osterlind, 2006 ; Leclercq, 2006 ; Birenbaum, 2007 ; Newton, 2007 ; André, Loye & Laurencelle, 2015). Par ailleurs, l'approche *Assessment for Learning* (Assessment Reform Group, 1999 ; Black & Wiliam, 2006 ; Laveault & Allal, 2016) apporte une nouvelle dynamique dans le champ des évaluations formatives. L'accent y est mis sur l'engagement actif des apprenants et sur des dispositifs qui soutiennent les apprentissages et fournissent des feedbacks informatifs en vue de favoriser l'autorégulation et l'apprentissage tout au long de la vie (CCSSO, 2009 ; McMillan, 2007 ; Tillema, Leenknecht & Segers, 2011 ; Reinholz, 2016). Les discussions dans le domaine sont

loin d'être figées et les propos de Galton (2019) cités plus haut, issus de son article intitulé "*Is Assessment for Learning Really Assessment?*" illustrent la nature d'une partie des débats en cours dans les milieux de la recherche en évaluation.

En référence au contexte des travaux énoncés plus haut, nous focaliserons la suite de notre réflexion sur une série de critères de qualité en évaluation qu'il nous semble important de présenter aux enseignants et que nous souhaitons aisément appréhendables par ces derniers. *A priori* ces critères concernent tout type d'évaluation, qu'elles soient en ligne ou non. Notre visée est pragmatique et tend à fournir aux praticiens des critères qu'ils peuvent utiliser en vue de concevoir leurs évaluations à distance.

Tableau 1. Critères de qualité pour des dispositifs d'évaluation

Critères	Explicitations	Références
Validité	Les informations résultant des évaluations doivent représenter ce que l'enseignant veut mesurer, permettre des inférences solides, couvrir les aspects importants qui étaient à évaluer, et ce, en lien avec les objectifs et le contenu enseigné.	p. ex. Cronbach et Meehl (1955) ; Linn, Baker et Dunbar (1991) ; Anderson (2002)
Fidélité	Les traitements des résultats doivent fournir des garanties d'objectivité. La subjectivité de l'enseignant doit être contrôlée lors des corrections des évaluations (concordance intra-évaluateur, mais aussi inter-évaluateurs si plusieurs enseignants interviennent). Ceci peut être fait au moyen de grilles d'évaluation critériées.	p. ex. Cronbach (1951) ; Pieron (1963) ; Ebel (1969) ; Brown (2019)
Sensibilité	Les mesures des apprentissages réalisés doivent être précises, refléter des phénomènes subtils.	p. ex. Abernot (1996) ; Leclercq (2003)
Diagnosticité	Les feedbacks renvoyés aux apprenants après correction doivent permettre le diagnostic précis des points forts et des points à améliorer et faciliter la régulation des apprentissages et des enseignements. Le feedback est d'autant plus bénéfique qu'il aide les apprenants à corriger leurs erreurs et fournit des indications sur la marche à suivre pour améliorer les apprentissages.	p. ex. Alderson (2005) ; Hattie et Timperley (2007) ; Jang (2008) ; Jang et Wagner (2013)
Équité	Les apprenants doivent être traités de façon juste, sans discrimination, en principe de la même façon.	p. ex. Messick (1981)
Praticabilité	La réalisation des évaluations doit être faisable endéans des délais raisonnables et à l'aide de ressources humaines et matérielles disponibles.	p. ex. Gilles et Leclercq (1995)
Transparence	Les informations non confidentielles relatives aux processus et aux enjeux de l'évaluation doivent être communiquées et comprises par tous les acteurs de l'évaluation. La sensibilisation aux objectifs visés par l'évaluation doit permettre aux évalués de comprendre ce qui est attendu, et ce, dans une	p. ex. Birenbaum (2007) ; Reinholz (2016)

	démarche de soutien aux apprentissages.	
Authenticité	Les questions et les tâches proposées lors des évaluations doivent avoir du sens pour les apprenants interrogés, être pertinentes par rapport à leur contexte.	p. ex. Kerka (1995)
Auto-évaluation	Favoriser l'auto-évaluation et l'explicitation chez les apprenants évalués permet une prise de conscience des points à améliorer, ce qui peut contribuer à soutenir leurs apprentissages.	p. ex. Leclercq (1982) ; Chi, de Leeuw, Chiu et Lavancher (1994) ; Reinholz (2016)

3. Formes d'évaluations à distance

Whitelock et Brasher (2019) définissent l'évaluation numérique au sens large de la façon suivante :

> *e-Assessment is defined in its broadest sense, where information technology is used for any assessment-related activity. e-Assessment can be used to assess cognitive and practical abilities. Cognitive abilities are assessed using e-testing software, while practical abilities are assessed using e-portfolios or simulation software.* (p. 3).

Dans ce domaine, Audet (2011) distingue trois types d'évaluations à distance. Le premier concerne les "*évaluations entièrement en ligne*", le plus fréquemment sous la forme de questionnaires (souvent des QCM) formatifs et sommatifs avec correction automatique et feedbacks personnalisés, mais cette catégorie contient aussi les jeux-questionnaires, les tests adaptatifs, les simulations et les mondes virtuels. Le deuxième type d'évaluation à distance est en rapport avec les "*activités d'évaluation en ligne*" « dont la correction (...) n'est toutefois pas automatisée »" (Audet, 2011, p. 36). Ces activités, comprennent par exemple les forums, les cyber portfolios, les fonctionnalités de dépôt de travaux dans les environnements numériques d'apprentissage, les blogues, les wikis, l'audio et la vidéoconférence, les cartes heuristiques. Parmi les activités réalisables en ligne répertoriées par Audet signalons l'accès aux travaux des pairs et l'évaluation par les pairs, des approches préconisées par l'*Assessment for Learning (AfL)* (Assessment Reform Group, 1999). Le troisième type d'évaluations à distance concerne les formes de "*soutien à l'évaluation*" où l'évaluation ne se déroule pas sur le web, mais est facilitée par les outils en ligne. Il s'agit par exemple d'outils de dépôt et de transmission des informations relatives aux évaluations avec par exemple l'envoi par courriel aux apprenants de rétroactions visuelles ou multimédias en guise de feedback ou encore l'utilisation par l'évaluateur de logiciels d'aide à la correction. L'évaluation à distance couvre donc un large éventail de possibilités allant de l'utilisation du courrier électronique aux questionnaires automatisés (souvent des QCM) en passant par des instruments plus sophistiqués comme les tests adaptatifs qui s'adaptent aux performances des étudiants comme certains tests en langue ou encore les simulations en ligne.

Deux catégories peuvent être distinguées au sein de ces multiples formes d'évaluation à distance, d'une part les évaluations complètement automatisées depuis l'envoi des questions jusqu'à la transmission des feedbacks après correction automatique et d'autre part les évaluations à distance qui nécessitent l'intervention d'un correcteur humain dans la phase de traitement des réponses. Le QCM en ligne utilisé dans une perspective formative ou sommative est un exemple emblématique de la première catégorie. Concernant la seconde

catégorie, voici un exemple typique : (1) l'enseignant envoie par email aux apprenants un fichier (devoir) contenant une ou plusieurs questions ouvertes longue (QROL) ; (2) les apprenants y répondent à distance ; (3) ils renvoient par email le fichier complété ; (4) l'enseignant corrige et élabore les feedbacks personnalisés (éventuellement directement dans les fichiers des élèves) et (5) transmet les QROL corrigées et son feedback par email aux apprenants.

Notons qu'il existe des dispositifs d'évaluation à distance qui combinent QCM et QROL et facilitent leur organisation en banques de questions à l'aide de systèmes de gestion ou "*Assessment Management Systems (AMS)*". Certains AMS proposent des contrôles qualité et des approches en cycles d'évaluation (Piette *et al.* ; Gilles & Tinnirello, 2017), malheureusement ces systèmes sont encore peu répandus dans le monde enseignant.

4. Comparaison "correction automatisée" versus "correction ayant recours à un correcteur " à la lumière des critères qualité

Dans cette partie, dans le tableau 2, nous comparons les QCM et les QROL à la lumière des critères de qualité pour des dispositifs d'évaluation.

Tableau 2. Comparaison de deux formes d'évaluation : QCM en ligne avec correction automatique versus QROL gérée à distance avec correction par l'enseignant

Forme →	*QCM en ligne*	*QROL gérée à distance*
Correction →	*Entièrement automatisée*	*Recours au correcteur humain*
Critères qualité ↓		
Validité	La correction automatisée permet beaucoup de questions (fermées) et donc une large couverture des contenus, mais les niveaux taxonomiques sont peu élevés et ne dépassent pas l'analyse.	Le temps de correction entraîne peu de questions (ouvertes) d'où une faible couverture des contenus, mais le format des QROL permet potentiellement d'évaluer des niveaux taxonomiques plus élevés.
Fidélité	La correction automatisée permet une évaluation objective. On évite les biais liés à la subjectivité dans l'interprétation des réponses.	L'intervention d'un correcteur humain introduit des problèmes de concordance intra et inter-correcteurs. L'utilisation d'échelles descriptives est conseillée.
Sensibilité	Les QCM ne permettent pas de recueillir des informations subtiles sauf en y associant des techniques telles que les degrés de certitude ou les justifications de réponses.	La formulation de réponses longues et argumentées permet un riche recueil d'informations sur l'état des apprentissages.
Diagnosticité	Les QCM permettent potentiellement des feedbacks rapides, précis, personnalisés indiquant les apprentissages acquis par l'apprenant et ceux qui ne le sont pas encore ainsi que des remédiations. Le diagnostic porte	Lors de la correction des QROL, l'enseignant a la possibilité d'analyser en profondeur les explications des apprenants. Son expertise professionnelle permet de remonter aux causes d'éventuels problèmes d'apprentissage. Il peut

	sur des niveaux taxonomiques qui ne vont pas au-delà de l'analyse.	proposer des remédiations individuelles, mais ce type de démarche est chronophage.
Équité	La possibilité de standardisation de l'évaluation par QCM permet de garantir une certaine neutralité lors de la correction lorsque celle-ci est entièrement automatisée. Cependant, le format QCM peut aussi potentiellement faciliter la triche.	L'influence +/- inconsciente sur l'évaluateur de certaines caractéristiques de l'évalué comme son origine sociale sont bien établies dans la littérature. Par ailleurs, les devoirs avec QROL gérés à distance n'excluent pas les éventuels plagiats et la malhonnêteté scolaire.
Praticabilité	Rédiger de "bonnes" QCM avec feedbacks diagnostiques et remédiations en fonction des types d'erreurs est chronophage, mais la démarche peut être considérablement allégée lorsque des banques de questions existent et qu'un travail collaboratif à plusieurs enseignants est organisé. De plus, l'automatisation de la correction et des feedbacks permet un gain de temps considérable.	La correction des QROL portant sur des niveaux taxonomiques élevés et la rédaction de feedbacks personnalisés après analyse des erreurs des apprenants est chronophage, mais la mise en œuvre d'échelles descriptives peut faire gagner du temps. A cela peut s'ajouter le temps nécessaire pour élaborer des stratégies individualisées de soutien aux apprentissages.
Transparence	Outre les précautions habituelles d'information préalable sur les caractéristiques de l'évaluation (consigne, système de correction, temps alloué, etc.), on peut aussi transmettre une version de la table de spécifications. Par ailleurs, les QCM offrent potentiellement cette particularité de permettre à l'évalué de savoir exactement ce qu'il faut faire et ce qui est attendu-	Les précautions habituelles d'information préalable doivent être prises. Dans le contexte d'une démarche de soutien aux apprentissages, un effort particulier est requis pour permettre aux évalués de comprendre précisément les objectifs visés et ce qui est attendu.
Authenticité	Une série de problèmes inhérents aux QCM, comme la centration sur des détails (souvent lorsque des connaissances factuelles sont en jeu) et la contraction du champ cognitif lié au format, peuvent entraîner une perte de sens chez les apprenants qui ne perçoivent pas de liens avec leur contexte.	Le format des QROL autorise des tâches évaluatives qui potentiellement peuvent avoir beaucoup de sens pour les évalués. En outre, il permet d'évaluer des objectifs d'expression qui, combinés à des niveaux taxonomiques élevés, permettent aux évalués d'élaborer des réponses complexes.
Auto-évaluation	L'auto-évaluation systématique est possible avec des QCM si on y associe la technique des degrés de certitude. Correctement utilisée, cette technique permet des feedbacks métacognitifs et renseigne les apprenants sur leurs états de connaissances partielles.	L'auto-évaluation est possible avec les QROL et est recommandée dans une approche évaluative qui vise à soutenir les apprentissages en ce sens qu'elle aide les apprenants à prendre conscience des progrès réalisés et des prochaines étapes dans leur parcours de formation.

5. Conclusions

Les dispositifs d'évaluation à distance à correction automatisée et non-automatisée présentent des avantages et des inconvénients mis en lumière à l'aide de neuf critères de qualité élaborés à partir de la littérature en docimologie. Nous avons comparé deux formes emblématiques d'évaluation utilisables dans le contexte de la crise sanitaire Covid-19 : les QCM en ligne entièrement automatisés et les QROL gérées à distance qui nécessitent l'intervention d'un correcteur humain.

A l'aide des critères proposés, nous pointons des aspects qui peuvent se révéler décisifs lors des choix des enseignants dans la mise en place de leurs dispositifs d'évaluation à distance. Par exemple, les QCM en ligne offrent des avantages sur le plan de la couverture des contenus, mais ne permettent pas d'évaluer des niveaux taxonomiques élevés contrairement aux QROL gérées à distance. Par ailleurs, les QCM en ligne présentent des avantages indéniables en termes de fidélité tandis que sur le plan du critère de sensibilité les QROL gérées à distance sont potentiellement plus performantes. Concernant la diagnosticité, les deux formes d'évaluations à distance présentent des caractéristiques intéressantes, mais dans le cas des QCM en ligne cette diagnosticité n'opèrera que dans le cadre de questionnements qui ne franchiront pas le niveau taxonomique d'analyse, alors que pour les QROL elle opèrera pour les niveaux taxonomiques les plus élevés. En ce qui concerne l'équité, l'automatisation et la standardisation permettent aux QCM en ligne de se démarquer des QROL gérées à distance. De même, en ce qui concerne la praticabilité, les QCM en ligne offrent en effet des avantages en termes de gain de temps de correction. En ce qui concerne le critère de transparence, les mêmes précautions en termes d'informations préalables à fournir doivent être appliquées aux deux formes d'évaluation à distance, mais ce qui est attendu des évalués est plus intuitif et évident pour la forme QCM en ligne. Pour ce qui est de l'authenticité, c'est l'inverse, les QROL gérées à distance permettent potentiellement des réponses complexes plus en lien avec la réalité que le format de sélection de réponses offert par les QCM en ligne. Concernant l'auto-évaluation, un critère particulièrement prégnant dans une approche évaluative en soutien aux apprentissages, elle n'est possible dans le cadre des QCM en ligne qu'à la condition d'utiliser la technique des degrés de certitude.

Dans le contexte Covid-19 de mise en place en urgence des dispositifs d'évaluation à distance, nous conseillons aux enseignants de ne pas choisir l'une de ces deux formes au détriment de l'autre. En effet, une utilisation combinée des QCM en ligne et des QROL gérées à distance permettrait des dispositifs d'évaluation où les avantages de l'une compenseraient les désavantages de l'autre. Dans une telle approche, il convient de souligner que le gain de temps permis par la correction automatisée offerte par les QCM en ligne pourrait être avantageusement mis à profit par les enseignants pour améliorer la fiabilité de la correction des QROL et peaufiner des feedbacks combinant les informations issues des deux formes dans une perspective de soutien aux apprentissages.

La période que nous vivons nous demande beaucoup de créativité et d'adaptation et nous conduit au retour à l'essentiel, les critères de qualité énoncés en font partie pour améliorer de manière générale la qualité de l'évaluation.

6. Bibliographie

Abernot, Y. (1996). *Les méthodes d'évaluation scolaires*. Paris : Dunod.

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The Interface between Learning and Assessment*. London: Continuum.

American Educational Research Association [AERA], American Psychological Association [APA], and National Council for Measurement in Education [NCME]. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Anderson, L. (2002). Curricular Alignment. A Re-Examination. *Theory into Practice*, *Vol. 41*, *4*.

André, N., Loye, N., & Laurencelle, L. (2015). Regard actuel sur le concept centenaire de validité psychométrique, sa genèse et ses avatars. *Mesure et évaluation en éducation*, *37*(3), 125-148. https ://doi.org/10.7202/1036330ar

Angoff, W.H. (1988). Validity: An evolving concept. In Wainer, H. & Braun, I.H. (Eds.). *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 19-32.

Assessment Reform Group (ARG). (1999). *Assessment for learning: Beyond the black box*. Cambridge: University of Cambridge.

Audet, L. (2011). *Les pratiques et défis de l'évaluation en ligne*. Réseau d'enseignement francophone à distance du Canada (REFAD).

Birenbaum, M. (2007). Evaluating the Assessment: Sources of Evidence for Quality Assurance. *Studies in Educational Evaluation*, *Vol. 33*, Issue *1*, 29-49.

Black, P., & William, D. (2006). Developing a Theory of Formative Assessment. In J. Gardner (Ed.), *Assessment and Learning* (pp. 81-100). London: Sage.

Blais, J.-G., Gilles, J.-L., & Tristan-Lopez, A. (Eds.). (2015). *Evaluation des apprentissages et technologies de l'information et de la communication – Bienvenue au 21e siècle*. Berne, Suisse : Peter Lang. http://hdl.handle.net/20.500.12162/135

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). Quality assurance methods for performance-based assessments. *Advances in Health Sciences Education: Theory and Practice*, 2003;8(1):27-47. https ://doi.org/10.1023/a:1022639521218

Brown, G. (2019). Is Assessment for Learning Really Assessment? *Frontiers in Education*. *4*(64). https ://doi.org/10.3389/educ.2019.00064

Cardinet, J. (1986). *Evaluation scolaire et mesure.* Bruxelles : Éditions De Boeck-Wesmael, Pédagogie en développement.

CCSSO. (2009). Report of the Council of Chief State School Officers. *Vision for developing assessment systems that support high quality learning*. Washington, DC.

Charlier, B. (2010). Les TIC ont-elles transformé l'enseignement et la formation ? In B. Charlier et F. Henri. (Eds.). *Apprendre avec les technologies*. Paris : Presses universitaires de France.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, *Vol 18*, Issue *3*, 439-477.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334. https://doi.org/10.1007/BF02310555

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

De Landsheere, G. (1980). *Evaluation continue et examens – Précis de docimologie*. 6e édition. Bruxelles : Edition Labor, Education 2000.

Downing, S. M., & Haladyna, T. M. (1997). Test Item Development: Validity Evidence from Quality Assurance Procedures. *Applied Measurement in Education*, *10*(1), 61-82.

Downing, S. M., & Haladyna, T. M. (Eds). (2006) *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ebel, R. L. (1972). *Essentials of educational measurement* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Gilles, J.-L., & Tinnirello, S. (2017). *DOCIMO: an Online Platform Dedicated to the Construction and Quality Management of Learning and Impact Assessments in the Digital Age*. Poster presented at The #dariahTeach Open Resources Conference Lausanne, Suisse. Retrieved from http://hdl.handle.net/20.500.12162/107

Gilles, J.-L., & Leclercq, D. (1995). *Procédures d'évaluation adaptées à des grands groupes d'étudiants universitaires - Enjeux et solutions pratiquées à la FAPSE-ULG*. Communication présentée à Symposium International sur la Rénovation Didactique en Biologie, Tunis, Tunisie.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research. Vol 77*, Issue *1*, 81–112. https://doi:org/10.3102/003465430298487

Hettiarachchi, E., & Huertas, M.-A. (2013). *Skill and Knowledge E-Assessment: A Review of the State of the Art*. Universitat Oberta de Catalunya: Internet Interdisciplinary Institute, IN3 Working Paper Series.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y.R. Chung, & J. Xu (Eds.), *Towards adaptive CALL: Natural language processing for diagnostic language assessment* (pp. 117-131). Ames, IA: Iowa State University.

Jang, E. E., & Wagner, M. (2013). Diagnostic feedback in language classroom. In A. Kunnan (Ed.), *Companion to language assessment*. Wiley-Blackwell.

Kerka, S. (1995). *Techniques for authentic assessment: Practice application brief* (ERIC Clearinghouse on Adult, Career, and Vocational Education, No. ED381688). Retrieved from https://files.eric.ed.gov/fulltext/ED381688.pdf

Kuder, G. F., & Richardson, M. W. (1937). The Theory of Estimation of Test Reliability. *Psychometrika*, 2, 151-160.

Laveault, D., & Allal L. (2016). Implementing Assessment for Learning: Theoretical and Practical Issues. In D. Laveault & L. Allal (Eds.), *Assessment for Learning: Meeting the Challenge of Implementation*. Bern: Springer International Publishing Switzerland.

Leclercq, D. (1982). Confidence Marking, its Use in Testing. In B. Choppin & N. Postlethwaite (Eds.), *Evaluation in Education: International Review Series*. Oxford: Pergamon, *vol. 6*, *2*, 161-287.

Leclercq, D. (2003). *Diagnostic cognitif et métacognitif au seuil de l'université. Le projet Mohican mené par les 9 universités de la Communauté Française Wallonie Bruxelles*. Liège : Éditions de l'Université de Liège.

Leclercq, D. (2006). L'évolution des QCM. In G. Figari & L. Mottier-Lopez. *Recherches sur l'évaluation en Education*. Paris : L'Harmattan, 139-146.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, *20*(8), *15*.

McMillan, J. H. (2007). *Formative classroom assessment: Research, theory and practice*. New York: Teachers College Press.

Mehrens, W. A., & Lehmann, I. J. (1991). *Measurement and evaluation in education and psychology* (2nd ed.). New York, NY: Houghton Mifflin Company.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational researcher*, *10*(9), 9-20.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In Wainer, H. & Braun, I.H. (Eds.). *Test validity* (pp. 33-45). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, *14*(4), 5–8.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). New York: American Council on Education and MacMillan.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*, 149-170. https://doi.org/10.1080/09695940701478321

Osterlind, S. J. (2006). *Modern Measurement: Theory, Principles, and Applications of Mental Appraisal*. Upper Saddle River, N.J: Pearson/Merrill Prentice Hall.

Partenariat mondial pour l'éducation (PME) (2020, 27 mars). *Réponse à l'épidémie de COVID-19 - Appel à l'action pour les enseignants*. Consulté sur https://www.globalpartnership.org/fr/news/reponse-lepidemie-de-covid-19-appel-laction-pour-les-enseignants.

Phelps, R. P. (Ed.) (2005). *Defending Standardized Testing*. Lawrence Erlbaum Associates Publishers.

Pieron, H. (1963). *Examen et docimologie*. Paris : PUF.

Piette, S.-A., Tinnirello, S., Bruyère, F., & Gilles, J.-L. (2012). D'ExAMS à DOCIMO, évolution d'une plateforme web-based soutenant la création de tests selon un modèle scientifique de création et gestion qualité de tests standardisés. In G. Baillat (Ed.), *Livre des résumés du 17e Congrès de l'Association mondiale des sciences de l'éducation* (pp. 104-105). Reims, France : Université de Reims. http://hdl.handle.net/20.500.12162/1646

Popham, W. J. (1997). Consequential validity: Right Concern-Wrong Concept. *Educational Measurement: Issues and Practice*, *Vol 16*, Issue *2*, 9-13. https://doi.org/10.1111/j.1745-3992.1997.tb00586.x

Reinholz, D. (2016). The assessment cycle: a model for learning through peer assessment. *Assessment & Evaluation in Higher Education*, *41*(2), 301-315. https://doi.org/10.1080/02602938.2015.1008982

Tillema, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning – A review of research studies. *Studies in Educational Evaluation, Vol 37*, 25-34. https://doi.org/10.1016/j.stueduc.2011.03.004

UNESCO (2020, 12 avril). *Impact du COVID-19 sur l'éducation.* Consulté sur https://fr.unesco.org/covid19/educationresponse

Whitelock, D., & Brasher, A. (2019). *Roadmap for e-assessment: Which Way Now?* Loughborough University. https://hdl.handle.net/2134/4451