



Conseil ontarien
de la qualité de
l'enseignement
supérieur

Un organisme du gouvernement de l'Ontario

Guide du COQES pour l'élaboration de grilles critériées valides et fiables

Jess McKeown et Danielle Lenarcic Biss



Publié par le

Conseil ontarien de la qualité de l'enseignement supérieur

1, rue Yonge, bureau 2402
Toronto (Ontario) Canada M5E 1E5

Téléphone : 416 212-3893
Télécopieur : 416 212-3899
Site Web : www.heqco.ca
Courriel : info@heqco

Citer ce document comme suit :

MCKEOWN, J., et D. LENARCIC BISS. (2018). *Guide du COQES pour l'élaboration de grilles critériées valides et fiables*, Toronto, Conseil ontarien de la qualité de l'enseignement supérieur.



Les opinions exprimées dans le présent rapport de recherche sont celles des auteurs et ne reflètent pas nécessairement le point de vue ni les politiques officielles du Conseil ontarien de la qualité de l'enseignement supérieur ou des autres organismes ou organisations ayant offert leur soutien, financier ou autre, dans le cadre de ce projet. © Imprimeur de la Reine pour l'Ontario, 2018

Processus d'élaboration des grilles critériées

1

EXAMINER DES EXEMPLES DE GRILLES CRITÉRIÉES

- Déterminer le type de grille critériée nécessaire
- Fonder la conception sur des outils existants pour accroître la validité

2

CRÉER UNE ÉBAUCHE DE GRILLE CRITÉRIÉE

- Conceptualiser le but de la grille critériée et ses éléments structurels, et ébaucher des descriptions
- Envisager de demander l'avis des spécialistes pertinents

3

VÉRIFIER LA VALIDITÉ DE LA GRILLE CRITÉRIÉE

- Vérifier les types de validité pertinents
- Faire participer des spécialistes et des utilisateurs de première ligne des grilles critériées à des séances récursives de mise à l'essai de la grille pour obtenir de la rétroaction

4

VÉRIFIER LA FIABILITÉ DE LA GRILLE CRITÉRIÉE VALIDÉE

- Tester les types de fiabilité pertinents
- Faire participer les utilisateurs de première ligne à la mise à l'essai de la grille pour calculer les cotes de fiabilité

5

METTRE EN APPLICATION

- Évaluer la possibilité de former les utilisateurs de première ligne de la grille
- Fournir des tâches d'ancrage
- Supprimer les identificateurs des étudiants pour préserver la confidentialité, la validité et la fiabilité

Introduction

Quand il s'agit d'évaluation scolaire, il existe deux catégories générales de tests qui servent à mesurer l'apprentissage des étudiants : le test objectif, dans lequel il n'y a qu'une seule bonne réponse pour chaque élément du test (p. ex. choix multiples, jumelage, remplir les espaces vides) et le test subjectif, où il n'y a pas de bonne réponse unique, mais plutôt des réponses « composées » (p. ex. réponse courte, dissertation, présentation orale). Bien que les deux formes d'évaluation soient utiles pour mesurer l'apprentissage des étudiants, les préjugés et les incohérences des évaluateurs rendent les tests subjectifs beaucoup plus difficiles à évaluer, ce qui en diminue la fiabilité et parfois la validité.

Les grilles critériées pourraient atténuer ce problème; on estime qu'elles apportent un certain degré d'objectivité à la notation des tests subjectifs. Une grille critériée est un outil qui vise à orienter et à évaluer le travail des étudiants en articulant clairement les critères à mesurer (c.-à-d. les composantes ou les dimensions), ce qui peut comprendre les aptitudes, les connaissances, les attitudes et les comportements. Ces critères sont ensuite décrits de façon plus détaillée afin de s'harmoniser avec les niveaux croissants de réussite (c.-à-d. les niveaux de rendement), auxquels peut être attribuée une certaine valeur numérique pour calculer la note totale d'un étudiant pour un travail donné. Les enseignants qui n'ont pas reçu de formation particulière sur l'élaboration des grilles critériées peuvent éprouver des difficultés à concevoir des grilles qui saisissent adéquatement les critères qu'ils tentent de mesurer. Heureusement, l'Association des collèges et universités des États-Unis a créé 16 grilles critériées « VALUE » génériques pour évaluer les résultats d'apprentissage essentiels dans l'enseignement supérieur. Ces grilles critériées sont maintenant considérées comme une norme et se sont révélées extrêmement utiles pour les personnes qui ne savent pas par où commencer quand elles élaborent des grilles critériées en fonction de leur contexte particulier.

Le présent document vise à donner un aperçu des considérations importantes et à proposer un processus aux personnes qui cherchent à élaborer ou à valider des grilles de notation pour l'évaluation subjective. Le processus proposé n'est pas la seule méthode valable pour l'élaboration de grilles critériées. Le présent guide a pour but de stimuler la réflexion et d'encourager l'adoption de méthodes d'évaluation rigoureuses.

Tableau 1 : Échantillon d'une grille critériée classique utilisant des niveaux de rendement génériques

		Niveaux de rendement (Notation avec pourcentage, nombre ou note alphabétique pour chaque niveau.)					Coefficient de pondération des critères (exemples)	Note de l'étudiant
		Avancé	Maîtrise	En développement	Débutant	Inadéquat		
Critères	Concept 1						30%	Note x 0,3
	Concept 2						20%	Note x 0,2
	Concept 3						10%	Note x 0,1
	Concept 4						40%	Note x 0,4
		↑ Descriptions des concepts à chaque niveau de rendement ↑					100%	Note globale

Processus d'élaboration des grilles critériées

ÉTAPE 1 : EXAMINER DES EXEMPLES DE GRILLES CRITÉRIÉES

Idéalement, la première étape consiste à commencer par un examen approfondi des grilles critériées validées propres à la compétence ou au domaine. Si les grilles critériées évaluant la compétence ou le domaine d'intérêt sont rares dans la documentation, ou si la vision de la structure de la grille est relativement unique, il est plus probable que le processus d'élaboration commence à zéro. En revanche, les grilles critériées préalablement validées ou génériques telles que la grille « VALUE » peuvent constituer une référence idéale pour les personnes qui cherchent à évaluer des compétences ou des sujets fréquemment évalués. En fonction des ressources disponibles pour le processus d'élaboration des grilles et du contexte d'utilisation, les grilles existantes peuvent être utilisées soit a) en les mettant en œuvre directement, dans leur forme originale, soit b) en les modifiant pour favoriser une meilleure harmonisation avec une tâche ou un travail en particulier. Bien qu'à première vue la mise en œuvre directe d'une grille existante puisse sembler moins exigeante en ce qui concerne les ressources, les concepteurs de grilles devraient tenir compte de la quantité de formation nécessaire pour les évaluateurs, car les grilles génériques tendent à être moins fiables que les grilles propres au contexte quand la formation est insuffisante.

ÉTAPE 2 : CRÉER UNE GRILLE CRITÉRIÉE PROVISOIRE

A) Définir le but de la grille

- Préciser la partie du travail de l'étudiant qui est évaluée : processus ou produit (ou les deux).¹
- L'objet de l'évaluation déterminera le type de grille à créer : individuelle ou collective. Les grilles conçues pour fournir de la rétroaction individuelle à des étudiants seront vraisemblablement plus détaillées et risquent de présenter des problèmes techniques en ce qui concerne une utilisation à grande échelle, tandis que les grilles conçues pour tirer des conclusions au sujet d'une population risquent de ne pas tenir compte du rendement individuel.

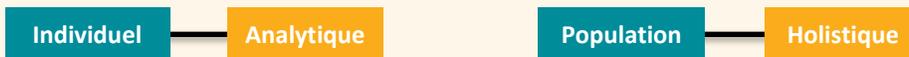
B) Examiner divers éléments pour la conception de la grille

- Choisir une structure : analytique ou holistique. Les grilles holistiques nécessitent que les évaluateurs portent un jugement global sur la qualité du travail d'un étudiant, tandis que les grilles analytiques comportent plusieurs dimensions pour l'évaluation du travail des étudiants.
- Créer des niveaux de notation. Déterminer si les niveaux de notation se rapporteront à la qualité du travail ou au degré de développement. Sélectionner le nombre de niveaux (le nombre le plus courant se situe entre 3 et 5) et décider s'ils seront jumelés à des valeurs numériques (se référer aux tableaux 1 et 2).²

¹ Les évaluations des processus portent sur les étapes à suivre pour accomplir une tâche; les évaluations des produits portent sur le résultat d'une tâche.

² Les formateurs peuvent avoir l'habitude de la méthode de notation traditionnelle à 5 lettres (A, B, C, D et F) et peuvent préférer les grilles à 5 niveaux (Kapelus, Miyagi et Scovill, 2017).

Figure 1 : Lien entre le niveau d'interprétation et la structure d'une grille



Les grilles analytiques sont souvent plus exactes quand les résultats sont interprétés à l'échelle individuelle, par exemple en classe, pour discerner les forces et les besoins d'apprentissage des étudiants. On estime que les grilles holistiques sont plus faciles et moins coûteuses à mettre en œuvre pour les évaluations à grande échelle qui nécessitent des interprétations à l'échelle de la population.

Source : Jonsson et Svingby (2007); Brophy (2012)

- S'il s'agit de notes traditionnelles, attribuer une pondération aux critères de notation. Le fait d'avoir un plus grand nombre de niveaux permet une plus grande variabilité des notes des étudiants, ce qui peut être préférable.
- Tenir compte de la note ou du niveau minimal exigé pour que le travail d'un étudiant soit jugé satisfaisant. Autrement, les utilisateurs des grilles pourraient attribuer une note de passage aux étudiants qui n'ont même pas tenté de faire l'exercice.

Tableau 2 : Exemple de grille analytique contenant des descripteurs génériques de la qualité du travail

Échec		En-deçà des attentes	Attentes comblées	Au-delà des attentes
Non démontré	Méconnaissance			
L'indicateur n'est pas démontré parce qu'il n'y a pas suffisamment de travail à évaluer.	Absence totale de qualité ou preuve d'une incompréhension fondamentale du concept.	Manque de qualité; le travail doit être largement revu pour être acceptable.	La définition de la qualité. Le travail est acceptable et présente un certain degré de maîtrise.	L'étudiant dépasse les attentes normales et produit un travail de qualité supérieure.

Source : Lesmond, McCahan et Beach (2017)

C) Rédiger des descriptions claires et exactes des concepts

- Utiliser une terminologie claire et s'assurer que tous les concepts sont mesurables.
- Créer des niveaux unidimensionnels (c.-à-d. que chaque rangée ne devrait servir à évaluer qu'un seul concept et que de nouveaux concepts ne devraient pas être introduits à mesure que les niveaux augmentent).
- Envisager d'harmoniser les niveaux avec des taxonomies des résultats d'apprentissage comme celles de Bloom, comme ce fut le cas pour les grilles « VALUE », ou avec les différents niveaux d'exactitude, d'originalité, de couverture du contenu, d'effort ou d'erreur (Association des collèges et universités des États-Unis, 2017).

- Harmoniser le langage avec le but de la grille, en veillant à ce que le langage convienne aux évaluateurs et aux étudiants. Certains chercheurs recommandent d'utiliser un langage subjectif et qualitatif (p. ex. « sophistication », « crédibilité », « uniformité » et « pertinence du travail »). Toutefois, un langage quantitatif objectif (p. ex. « tout », « la plupart », « aucun ») peut être utile pour les grilles destinées à une interprétation à l'échelle de la population.

D) Envisager de demander l'avis de spécialistes pertinents

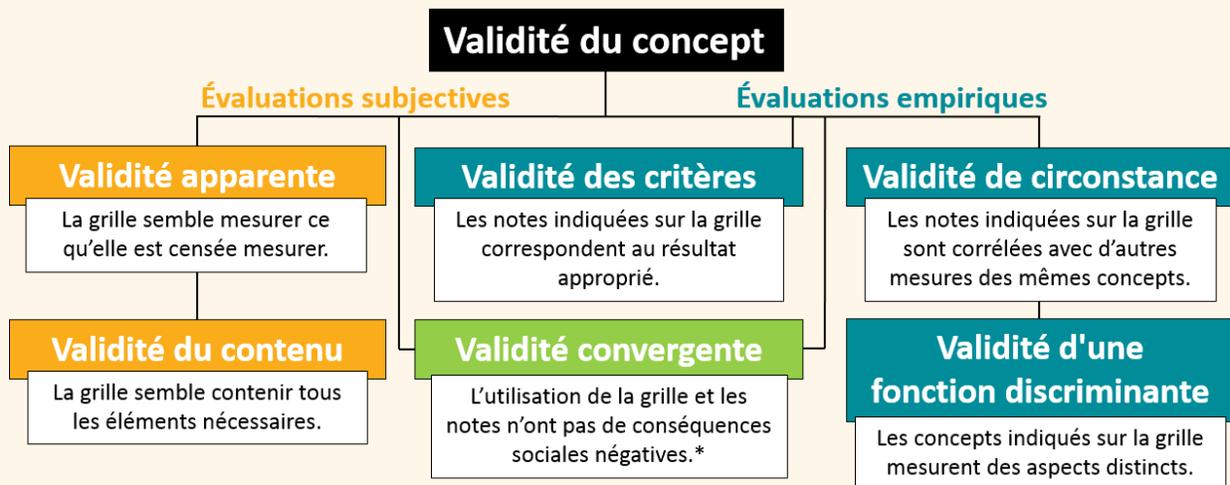
- Si le temps et le budget du projet le permettent, consulter des experts en la matière ou en évaluation.

ÉTAPE 3 : TESTER LA VALIDITÉ DE LA GRILLE

A) Comprendre les types de validité pertinents

- S'assurer que la grille mesure ce qu'elle est censée mesurer (se référer à la figure 2 et au tableau 3).

Figure 2 : Types de validité



La définition globale de la validité d'un concept est la mesure dans laquelle la grille d'évaluation mesure le ou les concepts visés, et leur concordance exacte avec le cadre théorique approprié. La validité peut être évaluée soit subjectivement, soit par consensus entre spécialistes, soit empiriquement en analysant les notes. Pendant l'élaboration des grilles, les jugements subjectifs seront les plus importants. Il convient de noter qu'il y a chevauchement entre ces types de validité et qu'il existe d'autres types de validité qui ne sont pas mentionnés ici.

Source : Adapté de Morling (2014, p. 138)

* Slomp, Corrigan et Sugimoto (2014) proposent plusieurs façons de mesurer la validité corrélative, y compris des évaluations subjectives et empiriques.

Tableau 3 : Questions pour l'évaluation de divers types de validité

Type de validité	Questions à envisager
Validité apparente	<ul style="list-style-type: none"> • Est-ce que toutes les facettes importantes du concept visé sont évaluées à l'aide des critères de notation? • Y a-t-il des critères d'évaluation qui ne sont pas pertinents pour le concept visé?
Validité du contenu	<ul style="list-style-type: none"> • Les critères d'évaluation de la grille portent-ils sur tous les aspects du contenu prévu? • Les critères d'évaluation portent-ils sur du contenu superflu?
Validité de circonstance	<ul style="list-style-type: none"> • Quel est le but de la grille, et comment les notes seront-elles utilisées? • Quels intervenants sont importants pour aider à comprendre les conséquences de la grille? • Quelles sont les conséquences prévues et non prévues, en fonction de l'objet et de l'utilisation de la grille?
Validité des critères	<ul style="list-style-type: none"> • Comment les critères de notation reflètent-ils les compétences qui laissent entrevoir la réussite par rapport au rendement futur ou connexe? • Y a-t-il des aspects du rendement futur ou connexe qui ne sont pas pris en compte dans les critères de notation?
Validité convergente	<ul style="list-style-type: none"> • Y a-t-il d'autres mesures bien établies de ce concept auxquelles les notes peuvent être comparées?
Validité d'une fonction discriminante	<ul style="list-style-type: none"> • Y a-t-il un chevauchement dans ce qui est mesuré dans les divers concepts? • Les notes associées à un concept sont-elles corrélées avec un concept non connexe?

Source : Adaptation de Moskal et Leydens (2000); Jonsson & Svingby (2007)

B) Tester la validité de la grille en obtenant plusieurs séries de rétroaction récursive

- Valider la grille par consensus subjectif d'intervenants tels que des concepteurs de grilles, des experts en la matière, des enseignants, des auxiliaires d'enseignement et des étudiants.

Première étape

- Tester la grille d'évaluation auprès d'utilisateurs de première ligne (p. ex. enseignants ou auxiliaires d'enseignement) en leur demandant de noter de nombreux échantillons de travaux préexistants des étudiants afin³ de comparer les différences entre les notes.⁴

FORMER OU NE PAS FORMER?

Former les utilisateurs de la grille avant les tests de validité *uniquement si* vous avez l'intention de former l'ensemble des utilisateurs avant la mise en application.

³ Le nombre idéal d'évaluateurs et d'échantillons de travaux d'étudiants dépend fortement du contexte. Habituellement, plus le nombre d'utilisateurs de première ligne qui participent aux essais est élevé, moins il faut d'échantillons de travaux d'étudiants.

⁴ Les travaux d'étudiants peuvent être sélectionnés afin de disposer d'un échantillon varié pour le plus grand nombre possible de concepts et de tenir compte d'une gamme de réalisations. Vérifier si les travaux étaient facultatifs ou notés, étant donné que cela peut influencer le degré d'effort investi dans la réalisation du travail et ainsi compromettre la validité des notes (Timmerman, Strickland, Johnson et Payne, 2011)

- Organiser une séance de rétroaction pour discuter des notes qui diffèrent dans deux dimensions ou plus afin de clarifier les concepts, de repérer un langage flou ou compliqué et d'évaluer la pondération des critères.
- Susciter les commentaires des participants à l'aide de questions comme celles du tableau 3.

Deuxième étape

- Remanier la grille en fonction de la première séance de rétroaction.
- Organiser une autre séance de rétroaction avec des étudiants pour discuter de la terminologie, de la clarté des critères et de l'utilité globale perçue de la grille quand il s'agit d'orienter et d'évaluer leur travail.
- Communiquer avec d'autres intervenants pertinents (p. ex. des employeurs) pour obtenir leurs commentaires sur la grille.

Troisième étape

- Réviser la grille d'évaluation en fonction des suggestions obtenues lors des séances de rétroaction supplémentaires.
- Tester la grille une deuxième fois auprès d'un groupe élargi d'utilisateurs de première ligne, en leur demandant de noter de nombreux échantillons de travaux préexistants d'étudiants.
- Veiller à ce que la rétroaction des intervenants soit adéquatement intégrée aux grilles et à ce que les étapes 1 à 3 soient répétées jusqu'à ce qu'aucune autre révision ne soit nécessaire.

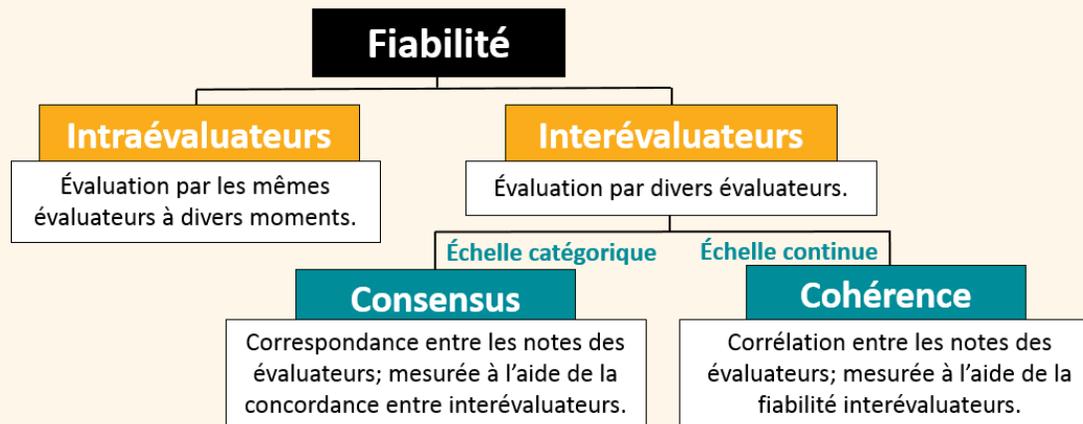
ÉTAPE 4: TESTER LA FIABILITÉ DE LA GRILLE VALIDÉE

A) Comprendre les types de fiabilité pertinents

- Veiller à ce que la grille soit fiable quand il s'agit d'évaluer avec exactitude et uniformité les travaux des étudiants (se référer au graphique 3).⁵

⁵ Rezaei et Lovorn (2010) soulignent l'importance d'utiliser uniquement des grilles fiables; ils estiment que le fait d'utiliser de façon inappropriée un outil qui n'est pas fiable est pire que le fait de ne pas l'utiliser.

Figure 3 : Types de fiabilité



L'élément le plus pertinent de la fiabilité, dans ce contexte, est la cohérence des notes d'évaluation découlant de la variation du jugement des évaluateurs. Cela peut se produire pour un évaluateur individuel au fil du temps (c.-à-d. fiabilité intraévaluateurs) ou entre évaluateurs (c.-à-d. fiabilité interévaluateurs). Quand on compare les jugements de divers évaluateurs, les mesures diffèrent en fonction de l'échelle de notation (c.-à-d. catégorique ou continue), ce qui détermine s'il convient d'examiner le consensus des évaluateurs ou la cohérence des évaluateurs.

B) Vérifier et interpréter la fiabilité de la grille

Première étape

- Tester la grille d'évaluation auprès d'utilisateurs de première ligne en leur demandant d'examiner de nombreux échantillons de travaux préexistants d'étudiants. Les travaux d'étudiants devraient être différents de ceux qui ont été utilisés pour les tests de validité si le groupe des utilisateurs de première ligne grille demeure le même.

Deuxième étape

- Calculer les résultats en matière de fiabilité (se référer à la figure 3). La mesure statistique la plus courante est la fiabilité interévaluateurs (FI), qui calcule la corrélation entre les notes attribuées par différents évaluateurs; autrement, la concordance interévaluateurs (CI) calcule la concordance exacte des notes attribuées par les évaluateurs.⁶

Troisième étape

- Analyser les mesures de fiabilité en fonction de leur contexte particulier, en tenant compte de l'objet et de la portée de la grille.⁷

⁶ La fiabilité sera inévitablement plus élevée pour les tests qui mesurent le rendement des étudiants pour un même examen ou une même tâche que pour les travaux individuels que les étudiants peuvent orienter en fonction de leurs intérêts (Jonsson et Svingby, 2007).

⁷ Les grilles utilisées uniquement pour des interprétations individuelles peuvent être considérées comme fiables même quand les notes sont inférieures à 0,7, mais les grilles utilisées pour des interprétations à l'échelle de la population nécessitent habituellement des scores de fiabilité supérieurs à 0,7 (Jonsson et Svingby, 2007).

- Si les cotes de fiabilité sont jugées appropriées et qu'aucun autre changement n'est proposé par les utilisateurs de première ligne de la grille, celle-ci est prête à être mise en application.
- Si les notes de fiabilité inférieures à 0,7 ne sont pas acceptables dans les circonstances, répéter les tests de validité pour comprendre les incohérences dans la notation ou répéter l'étape 1 des tests de fiabilité à l'aide d'échantillons plus importants de travaux d'étudiants ou d'évaluateurs supplémentaires jusqu'à ce que les coefficients de fiabilité soient supérieurs à 0,7.

Interprétation des coefficients de fiabilité

- > 0,9 = très élevé
- 0,8 à 0,89 = élevé
- 0,7 à 0,79 = acceptable
- < 0,7 = faible

Source : Morrison, Ross, Kemp et Kalman (2010)

ÉTAPE 5: METTRE LA GRILLE EN APPLICATION

A) Former les utilisateurs de la grille si la mise en application est à grande échelle

- Expliquer clairement aux évaluateurs le but de la grille, le nombre d'examens qui comporteront une note pour chaque élément des travaux d'étudiants, la façon dont les différences entre les évaluateurs seront évaluées et ce qu'il convient de faire avec les notes.
- Idéalement, accompagner la grille de travaux d'ancrage.⁸

B) Maintenir la validité et la fiabilité de la grille

- Préserver l'anonymat des étudiants en dépouillant leurs travaux de tous les identificateurs, en assignant à chaque document un code d'identification et en normalisant le format (p. ex. police, marges, espacement des lignes).
- Calculer le total des notes des étudiants de manière appropriée.
 - Faire preuve de diligence en additionnant les notes attribuées à des concepts distincts (rangées) dans une grille analytique pour obtenir une note globale, car certains concepts peuvent être plus importants que d'autres pour les résultats visés par le cours.
 - Dans les grilles analytiques, ne pas faire la moyenne des notes des divers concepts, étant donné que chaque note s'applique uniquement à la dimension qui lui est assignée.

⁸ Les travaux d'ancrage sont de véritables échantillons de travaux qui présentent les divers niveaux de réussite sur la grille (Jonsson et Svingby, 2007; Brophy, 2012).

Conclusion

Le présent document donne un aperçu des considérations importantes que doivent examiner les personnes qui cherchent à élaborer ou à valider des grilles critériées. Au lieu de présenter ce processus comme une science exacte, nous espérons encourager tout processus entrepris pour créer des outils d'évaluation valides et fiables à des fins d'évaluation subjective. Outre ses avantages manifestes à titre d'outil d'évaluation pour les enseignants, la grille critériée devrait être considérée comme avantageuse en raison de sa capacité potentielle de contribuer à la création de meilleurs messages-guides ou travaux, à l'orientation du travail des étudiants et à la production d'une rétroaction axée sur l'amélioration pour ceux-ci.

Il importe de se rappeler que la validité et la fiabilité ne sont pas des points fixes à atteindre aux dernières étapes de l'élaboration des grilles. Les contextes éducatifs dynamiques dans lesquels de telles évaluations sont mises en œuvre devraient évoluer en fonction des changements des besoins des étudiants et de la société. Par conséquent, la vérification de la validité et de la fiabilité de tels outils d'évaluation et la formation des évaluateurs qui les utilisent devraient constituer un processus continu.

Bibliographie

- ASSOCIATION OF AMERICAN COLLEGES AND UNIVERSITIES (2017). *On solid ground: Value report 2017*. [<https://www.luminafoundation.org/files/resources/on-solid-ground.pdf>]
- BROPHY, T. S. (2012). *Writing effective rubrics*. [https://assessment.aa.ufl.edu/media/assessment_aaufledu/academic-assessment/writing_effective_rubrics_guide_v2.pdf]
- JONSSON, A., et G. SVINGBY (2007). "The use of scoring rubrics: Reliability, validity and educational consequences", *Educational Research Review*, 2(2), 130–144.
- KAPELUS, G., N. MIYAGI et V. SCOVILL (2017). « Renforcer la capacité de mesurer les compétences essentielles à l'employabilité : cibler la pensée critique », Toronto, Conseil ontarien de la qualité de l'enseignement supérieur.
- LESMOND, G., S. MCCAHAN ET D. BEACH (2017). « Élaboration de rubriques d'analyse pour évaluer les compétences », Toronto, Conseil ontarien de la qualité de l'enseignement supérieur.
- MORLING, B. (2014). *Research methods in psychology: Evaluating a world of information*, New York, WW Norton & Company.
- MORRISON, G. R., S. M. ROSS, J. E. KEMP et H. KALMAN (2010). *Designing effective instruction*, Hoboken, John Wiley & Sons.
- MOSKAL, B. M., et J. A. LEYDENS (2000). "Scoring rubric development: Validity and reliability", *Practical Assessment, Research & Evaluation*, 7(10), 71–81.
- REZAEI, A. R., et M. LOVORN (2010). "Reliability and validity of rubrics for assessment through writing", *Assessing Writing*, 15(1), 18–39.
- SLOMP, D. H., J. A. CORRIGAN et T. SUGIMOTO (2014). "A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study", *Research in the Teaching of English*, 48(3), 276–302.
- TIMMERMAN, B. E. C., D. C. STRICKLAND, R. L. JOHNSON et J. R. PAYNE (2011). "Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing", *Assessment & Evaluation in Higher Education*, 36(5), 509–547.



Conseil ontarien
de la qualité de
l'enseignement
supérieur

Un organisme du gouvernement de l'Ontario