Modération statistique et modération sociale des résultats scolaires : approches opposées ou complémentaires ?

Dany Laveault
Université d'Ottawa
Gonzague Yerly
Université de Fribourg et Haute École pédagogique de Fribourg

Mots clés : modération, modération sociale, modération statistique, évaluation certificative, évaluation sommative, comparabilité, équité, développement professionnel de l'enseignant

La modération des résultats d'évaluation est une pratique devenue courante dans de nombreux systèmes éducatifs. Ses buts sont de renforcer l'équité des résultats entre les élèves et/ou leur comparabilité lorsque l'évaluation sert à certifier ou à sélectionner les élèves. Selon les contextes, différentes méthodes sont appliquées. Deux approches sont présentées : la modération statistique et la modération sociale. La première consiste de manière générale en une transformation statistique des résultats des élèves à l'aide d'une épreuve de calibration. La seconde est une démarche de travail collaboratif et de développement professionnel qui amène les enseignants à construire collectivement des pratiques communes d'évaluation à l'aide de standards. Cet article propose une réflexion théorique qui permet une description de ces deux approches (différentes méthodes, apports et limites) afin de les comparer. Son apport est d'actualiser cette comparaison et de dégager quels en sont les enjeux pour les différents acteurs du système scolaire. Un certain nombre de recommandations liées à leurs conditions d'utilisation sont envisagées en conclusion.

Key words: moderation, social moderation, statistical moderation, certification, summative evaluation, comparability, equity, teacher professional development

The moderation of evaluation results is a common practice in many educational systems. Its goals are to reinforce equity between students' results and/or their

comparability when assessment is used to certify or select students. Depending on the context, different methods are applied. Two approaches are identified: statistical moderation and social moderation. The first is generally a statistical transformation of students' results using a calibration test. The second is a professional development and collaborative work approach that leads teachers to collectively build common evaluation practices using standards. This paper proposes a theoretical reflection that allows a description of these two approaches (their different methods, contributions and limits) in order to compare them and to identify what are the stakes for the different actors of the school system. A number of recommendations related to their conditions of use are considered in conclusion.

Palavras-chave: moderação, moderação social, moderação estatística, avaliação certificativa, avaliação sumativa, comparabilidade, equidade, desenvolvimento profissional de professores

A moderação dos resultados da avaliação é uma prática comum em muitos sistemas educativos. As suas metas são as de reforçar a equidade dos resultados entre os alunos e / ou a sua comparabilidade quando a avaliação é usada para certificar ou selecionar alunos. Dependendo dos contextos, são aplicados diferentes métodos. São apresentadas duas abordagens: a moderação estatística e a moderação social. A primeira consiste geralmente numa transformação estatística dos resultados dos alunos usando um teste de calibração. A segunda é uma abordagem de trabalho colaborativo e de desenvolvimento profissional que leva os professores a construir coletivamente práticas comuns de avaliação usando standards. Este artigo propõe uma reflexão teórica que permite uma descrição dessas duas abordagens (diferentes métodos, contribuições e limites) para compará-las. A sua contribuição é a de atualizar essa comparação e identificar quais são os desafios para os diferentes atores no sistema educativo. Conclui-se com uma série de recomendações relacionadas com as suas condições de utilização.

Note des auteurs : La correspondance liée à cet article peut être adressée aux adresses courriel suivantes : [dany.laveault@uottawa.ca] et [gonzague.yerly@unifr.ch].

La rédaction de cet article a pu compter sur le soutien financier du Fonds national suisse de la recherche scientifique (FNS), bourse postdoctorale de Gonzague Yerly (P2FRP1_161705).

Introduction

La modération des résultats scolaires pour en assurer la comparabilité

Le besoin de comparer les résultats scolaires des élèves est un enjeu important dans les situations où l'évaluation comporte des conséquences importantes, par exemple lorsqu'il est question de certifier les apprentissages à la fin d'un cycle d'études (primaire ou secondaire) et de sélectionner les élèves (c.-à-d. admission dans un établissement scolaire). La comparabilité des résultats des élèves est alors cruciale pour des raisons d'équité. Essentiellement, la comparabilité signifie que « les standards sont appliqués de manière cohérente dans tous les sites (écoles, régions) et par tous les juges (enseignants, évaluateurs) de sorte que les performances des élèves de niveau scolaire équivalent sont reconnues en tant que telles et se voient attribuer la même note » (Matters, 2006, p. 6, traduction libre). Pour y parvenir, plusieurs juridictions scolaires utilisent une forme ou l'autre de « modération » des résultats scolaires pour en assurer la comparabilité : la modération statistique et la modération sociale.

La modération, dans son sens le plus large, consiste en une série de processus conçus pour assurer que les résultats scolaires sont comparables, c'est-à-dire qu'ils ne sont pas affectés par le degré de sévérité ou de clémence de l'évaluateur, peu importe l'établissement scolaire ou l'enseignant (Matters, 2006). Il existe une grande variété de procédures de modération à l'intérieur de chacun des deux « types d'approches » ou grandes « familles » (Linn, 1993) :

1) Procédures de modération statistique

Elles reposent essentiellement sur différentes méthodes de transformation[1] des résultats institutionnels de l'enseignant ou de l'établissement scolaire au moyen d'une épreuve externe unique, aussi appelée « épreuve de calibration ». La modération statistique intervient essentiellement une fois que les résultats ont été obtenus, tant à l'échelle de l'institution que de l'épreuve de calibration.

2) Procédures de modération sociale

Elles font appel au jugement d'évaluation de l'enseignant. Elles relèvent de processus d'assurance de la qualité intervenant avant et après l'obtention des résultats. Elles ne se contentent pas de vérifier ou de corriger la comparabilité des résultats a posteriori (Maxwell, 2007), mais font intervenir des procédures de vérification, d'inspection ou de coordination à toutes les étapes conduisant à la notation finale. Elles visent à assurer une plus grande conformité de l'évaluation et, par ricochet, une plus grande uniformité dans la notation des résultats des élèves, quel que soit l'établissement scolaire ou l'enseignant.

Bien que ces deux approches de modération aient des buts identiques (comparabilité et équité des résultats), elles ne mobilisent pas les mêmes outils et mécanismes. Elles s'inscrivent dans des paradigmes bien différents : la modération statistique s'ancre de manière forte dans la « mesure », ce qui n'est pas le cas de la modération sociale, qui fait plutôt appel au « développement professionnel » et au « jugement professionnel » des enseignants. Chronologiquement, la modération sociale est un phénomène plus récent que la modération statistique, ce qui se traduit par des références plus récentes dans le domaine de la recherche par rapport à la modération statistique. Par ailleurs, il n'existe pas de bilan récent sur l'utilisation de chaque type de modération dans différentes juridictions à travers le monde. Or, alors que la modération statistique des résultats, sous ses différentes formes, est encore utilisée, il nous paraît important de remettre en question sa place dans le monde éducatif contemporain, où les courants de l'évaluation-soutien d'apprentissage (*assessment for learning*) et de l'approche par compétences sont de plus en plus au centre des préoccupations concernant l'évaluation des apprentissages.

Les objectifs de cet article sont donc d'examiner et de différencier ces deux approches et leurs différentes méthodes afin de mieux comparer leurs apports et leurs limites. Nos conclusions soulèvent les enjeux que représentent ces deux approches de modération sur les différents acteurs de la scolarité (élèves et leurs parents, enseignants, gestionnaires) et envisagent un certain nombre de recommandations liées à leurs conditions d'utilisation.

Cadre théorique

Le besoin de modération des résultats d'évaluation

Les résultats scolaires produits par les enseignants et les écoles sont une source précieuse de renseignements quant aux apprentissages des élèves. Ils ont l'avantage de se fonder sur une grande diversité d'observations quantitatives et qualitatives échelonnées sur de longues périodes. Les deux approches de la modération partent du principe que l'évaluation réalisée par les enseignants mérite d'être prise en considération, à des degrés différents toutefois. Plusieurs recherches confirment qu'il existe une concordance élevée entre les enseignants lorsqu'il s'agit de mettre en rang les résultats des élèves (Maxwell, 2007 ; Elley & Livingston, 1972), mais que les résultats peuvent être plus ou moins affectés par le degré de sévérité au moment de la correction. C'est ainsi que, même si plusieurs enseignants peuvent situer les résultats des élèves dans le même ordre, un plus grand nombre d'élèves sont à risque de rater le seuil de réussite lorsque certains des enseignants agissent plus sévèrement (faux négatifs). Inversement, lorsque les correcteurs agissent de façon indulgente, les élèves peuvent être promus alors qu'ils ne sont pas prêts (faux positifs). Le même raisonnement peut s'appliquer aux différences se produisant entre établissements. C'est la principale raison invoquée par la Hong Kong Examinations and Assessment Authority (HKEAA) pour justifier la modération statistique des résultats (HKEAA, 2010).

Ce qui est en cause ici, ce n'est pas l'erreur relative du score de l'élève, mais l'erreur absolue entachant son résultat scolaire et pouvant donner lieu à des décisions différentes selon que ce résultat situe l'élève au-dessus ou en dessous d'une note de passage ou d'une condition d'acceptation de la performance (Laveault & Grégoire, 2014). La théorie de la généralisabilité prédit que l'erreur absolue sera toujours plus grande que l'erreur relative, car elle fait intervenir un plus grand nombre de sources d'erreurs affectant la valeur du résultat scolaire (Cardinet, Johnson & Pini, 2010). Conséquemment, la généralisabilité absolue sera toujours plus faible. Devant ce problème d'erreur absolue, la modération statistique consiste à ajuster les résultats de tous les élèves en fonction d'une épreuve de calibration. Wilmut et Tuson (2005) ainsi que Johnson (2011) déplorent le peu de recherches dans ce domaine, tant sur les examens à grande échelle que sur les évaluations se produisant en salle de classe.

La modération sociale vise, quant à elle, à intervenir à la source même du problème en tâchant de réduire dès le départ les variations entre enseignants dans l'interprétation des standards ou niveaux d'exigence des objectifs d'apprentissage. Plusieurs sources de variation entraînent des différences entre enseignants dans le jugement d'évaluation. Or, ces variations ne sont pas le seul fait des enseignants : elles dépendent aussi de la clarté et de la précision des standards qui leur sont communiqués. À cet égard, Harlen (2007) fait valoir que, lorsqu'il est question d'avoir recours au jugement de l'enseignant, il est plus profitable de lui transmettre les critères qui sont de bons indicateurs de la progression de l'apprentissage et de lui laisser le choix des tâches d'évaluation que de lui spécifier directement quelles tâches utiliser. La publication de copies-types ou copies d'ancrage peut servir à illustrer des niveaux différents de qualité d'une production d'élève et aider à mieux comprendre les standards. Bref, une meilleure transmission des standards vaut mieux qu'une transmission de tâches spécifiques.

Enfin, il est important de souligner que l'une ou l'autre des deux approches de modération n'est pas toujours possible et qu'elles ne sont pas une panacée pour rectifier des manquements importants dans la livraison du programme d'études, que ceux-ci soient le fait des enseignants ou encore de carences dans le curriculum ou dans la supervision du système éducatif. Par contre, lorsque toutes les conditions sont réunies pour que les deux formes de modération soient envisageables, il y a lieu de se demander si l'une est préférable à l'autre selon le contexte, ou encore si elles sont complémentaires. Dans ce dernier cas, leur utilisation de concert peut-elle être pertinente, que ce soit dans l'amélioration de la comparabilité des résultats ou du système éducatif dans son ensemble ?

La modération statistique

Définition et objectifs

La modération statistique consiste à ajuster les résultats obtenus par les élèves de différents groupes-classe ou groupes-école en fonction de leur rendement à une épreuve de calibration passée à tous les élèves de la même juridiction scolaire, locale ou nationale. Le but de la modération est d'assurer que les résultats issus du processus d'évaluation des apprentissages de tous les élèves d'une même juridiction sont comparables et ainsi d'introduire plus de justice et d'équité dans ce processus.

La modération statistique intervient surtout dans les situations d'évaluation à enjeux élevés et permet d'assurer la comparabilité des résultats en l'absence de standards précis ou d'échelons décrivant la progression attendue des apprentissages. Elle se retrouve surtout dans les situations de promotion ou de sélection des élèves dans des cycles d'études supérieurs ou encore dans les juridictions scolaires (p. ex., un État ou une province). Il s'agit ainsi de s'assurer que les variations observées dans la notation par des enseignants qui œuvrent dans le cadre du même programme et sous la même juridiction scolaire n'affectent pas indûment les résultats des élèves, peu importe le groupe d'élèves ou l'établissement.

Cet ajustement des notes accordées par l'enseignant en fonction des résultats de l'élève et de son groupe à l'épreuve de calibration repose sur la série de conditions suivantes :

- Le classement de l'élève par l'enseignant est essentiellement le même que le classement de l'élève obtenu par l'épreuve de calibration ;
- Seule change la valeur absolue du résultat (en fonction du niveau d'exigence ou de sévérité de l'enseignant). La valeur relative (rang de l'élève) par rapport à son groupe ne change pas ;
- Pendant une certaine période de temps, l'enseignant évalue l'élève essentiellement sur les mêmes connaissances et compétences que celles de l'épreuve de calibration (Burton & Linn, 1994) ;
- Les résultats classe/école portent sur les mêmes notions et contenus que ceux de l'épreuve de calibration.

La valeur des résultats modérés statistiquement dépend de l'algorithme de calcul employé. Quelques auteurs ont recensé les écrits dans ce domaine et ont comparé les différents algorithmes (Mislevy, 1992 ; Linn, 1993 ; Wilmut & Tuson, 2005). Dans toutes ces méthodes, le rang du résultat de chaque élève est préservé à l'intérieur de son groupe respectif. Il est possible de les regrouper en trois grandes catégories :

1) Méthodes reposant sur l'échelonnage linéaire (*linear regression scaling*)

Cette famille de méthodes consiste à ajuster la moyenne et la variance de la note institutionnelle (celle qui provient de l'enseignant ou de l'établissement d'enseignement) en fonction du résultat à l'épreuve de calibration.

2) Méthodes reposant sur l'échelonnage équipercentile (*equipercentile scaling*)

Les méthodes regroupées dans cette catégorie sont particulièrement utiles dans le cas où la relation entre les résultats à l'épreuve de calibration et la note institutionnelle est curvilinéaire. Essentiellement, la note institutionnelle et le résultat à l'épreuve de calibration sont considérés comme comparables lorsqu'ils sont atteints par le même pourcentage d'élèves (percentile) d'un groupe donné. Les méthodes de ce type sont particulièrement sensibles aux effets de plafond ou de plancher.

3) Méthode reposant sur l'échelonnage par rang (*rank order scaling*)

Cette méthode consiste à remplacer les scores attribués par l'enseignant par les scores à l'épreuve de calibration, l'élève de plus haut rang recevant le score le plus élevé et ainsi de suite. Cette méthode est particulièrement utilisée lorsque la modération porte sur plusieurs matières à la fois – pas toujours les mêmes – et qu'il faut situer tous les résultats à ces différentes matières sur une seule échelle calibrée. Ce genre de pratique se retrouve dans les politiques d'admission aux études postsecondaires, où les résultats des élèves du secondaire se retrouvent dans autant de « bouquets » de matières scolaires variées et doivent être comparables.

Exemple pratique et conditions d'utilisation

Calcul d'un score modéré

La modération statistique, lorsqu'elle satisfait aux conditions précédentes, parvient à assurer la comparabilité des résultats institutionnels. Prenons un exemple de calcul fondé sur la technique de transformation linéaire, couramment utilisée à Hong Kong (HKEAA, 2010), au Québec (MELS, 2016) et dans l'État australien de Nouvelle-Galles du Sud (Board of Studies, Teaching & Educational Standards NSW, 2011).

Notre exemple de calcul illustre l'ajustement à la baisse de l'élève Éric, qui pourrait très bien être représentatif d'un élève québécois. Il a obtenu une note globale de 58 % pour son cours de physique de 5e secondaire, soit la dernière année du secondaire ou 12e année. La moyenne de son groupe-classe a été de 72 % et les résultats des 24 élèves de sa classe se sont distribués avec un écart-type de 7. À l'épreuve uniforme obligatoire de fin

d'études qui sert d'épreuve de calibration, le groupe d'Éric a obtenu une moyenne bien inférieure à celle assignée par l'enseignant, soit 66 %, avec un écart-type légèrement moindre de 6. La moyenne des élèves à l'épreuve de calibration tend à indiquer que les élèves sont surévalués par leur enseignant, car leurs résultats y sont inférieurs et moins dispersés. Le tableau 1 liste les données de notre exemple et leur représentation symbolique :

Tableau 1

Données sur le cas d'Éric et représentations symboliques

Représentation symbolique	**Définition**	**Valeur**
N_i	Note institutionnelle	58
$\overline{N}_i$	Moyenne des notes institutionnelles du groupe-classe	72
S_i	Écart-type des notes institutionnelles du groupe-classe	7
$\overline{N}_u$	Moyenne des notes du groupe-classe à l'examen uniforme national	66
S_u	Écart-type des notes du groupe-classe à l'examen uniforme national	6
N_m	Note institutionnelle calculée après modération statistique ou note modérée	?

Il existe plusieurs méthodes d'échelonnage. Aux fins de notre exemple, nous emploierons une procédure d'échelonnage utilisant une simple transformation linéaire pour effectuer le calcul de la note modérée. Dans un premier temps, il s'agira de calculer le score *z* d'Éric en rapport avec sa note institutionnelle :

$$Z_i = \frac{N_i - \overline{N}_i}{s_i} = \frac{58-72}{7} = -2$$

Dans un second temps, il s'agira de calculer à quel résultat à l'épreuve uniforme correspond le score *z* institutionnel en effectuant la transformation linéaire suivante :

$$N_m = (S_u \times Z_i) + \overline{N}_u = (6 \times -2) + 66 = 54$$

La modération statistique a donc contribué à ajuster à la baisse la note d'Éric attribuée par cet enseignant en fonction de sa note institutionnelle et des résultats de son groupe-classe à l'épreuve uniforme. La note institutionnelle modérée d'Éric passe donc de 58 % à 54 % après modération. Dans la province de Québec, où une procédure de modération statistique de ce genre est couramment employée (MELS, 2011, section 7.5.2), la note finale d'Éric versée à son dossier de fin d'études secondaires serait composée à parts égales de la note institutionnelle modérée et de la note à l'épreuve de calibration. Dans le cas d'Éric, ce résultat serait la moyenne de 58 % et 54 %, soit 56 %.

Conditions d'utilisation

Le calcul des scores modérés ne présente pas de difficulté particulière. Dans la pratique, cependant, la modération statistique peut nécessiter un certain nombre de précautions quant à sa mise en œuvre administrative afin de tenir compte des limites des algorithmes statistiques utilisés. Tout document officiel devrait donc comporter une série de directives claires sur les conditions suivantes de la modération statistique :

1) Taille du groupe de modération

Les pratiques varient à cet égard. Dans la plupart des cas, le groupe de modération est constitué par un groupe-classe, mais il peut s'agir d'un groupe-école selon les juridictions et la procédure de notation des élèves (p. ex., si les élèves sont notés par l'ensemble des enseignants d'une même année de la même école). Cela a naturellement un impact sur la taille des groupes de modération. Les groupes comportant un petit effectif peuvent limiter l'usage de certaines méthodes statistiques et nécessiter d'autres formes de regroupements permettant de mieux estimer la moyenne et la variance.

2) Composition du groupe de modération

Une procédure claire doit être établie pour définir quels élèves font partie du groupe de modération. Voici des conditions qu'il faut prendre en compte parce que la composition du groupe a un impact sur la modération – à la hausse ou à la baisse – des résultats des autres élèves de ce groupe :

a) Depuis combien de temps l'élève fait-il partie du groupe de modération ? Les élèves qui sont de nouveaux arrivants ou qui viennent de changer d'école pourraient être exclus du groupe de modération.

b) Qu'est-ce qui constitue un résultat individuel acceptable ? Comment sont traités les cahiers blancs ou les copies comportant un nombre élevé de réponses omises à l'épreuve de calibration ? La question se pose aussi pour les élèves qui ont besoin d'accommodation (aide visuelle ou auditive, support d'un ordinateur, temps supplémentaire, etc.).

3) Nettoyage des données (*data cleaning*)

Tous les résultats à l'épreuve de calibration doivent-ils être utilisés ? Si non, quelles procédures sont prévues dans le nettoyage des données pour réduire l'incidence des cas extrêmes (*outliers*) sur la modération de la note institutionnelle ? Cette incidence est d'autant plus grande lorsque le groupe de modération est de petite taille.

L'absence de directives précises sur ces questions ouvre la porte à toutes sortes de possibilités de manipulation de données pour mieux faire paraître les résultats d'une classe ou d'une école ou pour toutes autres raisons. Ces questions peuvent être résolues assez facilement en rendant publique une série de directives qui clarifient le traitement des cas extrêmes et en s'assurant qu'un protocole strict est suivi dans la formation des groupes de modération.

Principales juridictions qui emploient la modération statistique

À notre connaissance, il n'existe pas à proprement parler de registre ni de référence récente faisant le bilan de l'utilisation de la modération statistique dans différentes juridictions à travers le monde. De plus, la situation est appelée à évoluer à cause des changements se produisant dans les systèmes d'éducation. Ainsi, une juridiction peut avoir abandonné cette pratique, alors que d'autres peuvent l'avoir intégrée récemment.

Wilmut et Tuson (2005) fournissent plusieurs exemples de pays où la modération statistique était employée au moment de rédiger leur rapport. Selon eux, elle avait lieu dans plusieurs États australiens, à Hong Kong et en Afrique du Sud. La modération statistique a été abandonnée par la Suède et par la Nouvelle-Zélande. En Nouvelle-Zélande, elle a été remplacée par des dispositifs de contrôle de la qualité des évaluations réalisées par les enseignants. Ainsi, le *Education Review Office* évalue périodiquement les établissements scolaires à cet effet (Nusche, Laveault, MacBeath & Santiago, 2012). Au Royaume-Uni, la modération statistique n'est pas employée pour assurer la comparabilité des résultats à différents

examens à enjeux élevés, mais des dispositifs de surveillance et d'inspection sont en place. Aux États-Unis, elle n'est pas employée, du moins pas pour les examens de fin d'études secondaires, car de nombreux examens d'admission aux études supérieures sont couramment utilisés dans la sélection des étudiants. Enfin, à la liste précédente des pays qui utilisent présentement la modération statistique, il faut ajouter la province de Québec, qui ne figure pas dans la liste de Wilmut et Tuson (2005). La province de Québec utilise la conversion et la modération statistique depuis une quarantaine d'années[2].

Appréciation de la modération statistique

Avantages

La modération statistique peut s'avérer utile lorsque les hypothèses de départ sont respectées et que les résultats de la modération ne se traduisent pas par des changements extrêmes. C'est un moyen économique et pratique d'assurer la comparabilité des résultats lorsque ceux-ci sont numériques. Elle peut aussi s'avérer la seule option lorsque le programme d'études est peu documenté et que la description des standards est si sommaire que la modération sociale ne saurait pas constituer une autre option.

Par ailleurs, l'analyse des résultats de la modération statistique peut contribuer à déclencher un processus de réflexion conduisant à des améliorations du système d'enseignement et d'évaluation. Des modérations trop importantes à la hausse ou à la baisse peuvent être le signal que quelque chose ne va pas dans l'alignement[3] de l'enseignement et de l'évaluation en salle de classe sur le programme d'études prescrit.

La modération statistique ne peut réussir à corriger les problèmes qui se produisent lorsqu'un enseignant donne une formation qui n'est pas conforme aux objectifs du programme d'études ou n'évalue pas correctement les apprentissages visés par la formation (Wilmut & Tuson, 2005). C'est pourquoi, dans quelques juridictions où la modération statistique est employée, celle-ci s'accompagne de suivi au moyen de rétroaction ou d'activités d'inspection afin d'aider les enseignants et les écoles à améliorer l'alignement de l'enseignement et de l'évaluation sur les objectifs du curriculum. Par exemple, la Hong Kong Examinations and Assessment Authority (HKEAA, 2010) fournit un rapport de modération qui informe si l'ajustement de la moyenne et de la dispersion des scores se situe à l'intérieur d'une marge plus ou moins adéquate. Un tel rapport peut servir

d'alerte aux enseignants et aux établissements afin qu'ils puissent corriger l'alignement pour les prochaines cohortes d'élèves, ce qui peut à long terme contribuer à l'amélioration du système éducatif.

Limites

Wilmut et Tuson (2005) affirment que la transformation des scores des enseignants à l'intérieur de la « boîte noire » que constitue la modération statistique n'est pas de nature à encourager les enseignants à améliorer leurs pratiques d'évaluation ni à les rendre plus redevables auprès du public. Il s'agit là d'un effet non souhaitable de la modération statistique, surtout lorsque l'évaluation des apprentissages des élèves se fait non seulement dans une perspective de notation, mais aussi d'amélioration du système éducatif.

La modération statistique repose sur des modèles mathématiques qui se fondent sur une série d'hypothèses quant à la nature de la population, sa distribution et ses paramètres. Lorsque ces hypothèses ne tiennent pas, les modèles utilisés peuvent rendre la modération statistique risquée ou carrément inappropriée. Voici deux exemples relevés par Wilmut et Tuson (2005) :

– **Effet de l'homogénéité/hétérogénéité des groupes (*the "company you keep" factor*)**

  Alors que les résultats individuels à l'épreuve de calibration ne sont pas affectés par ceux des autres élèves du groupe, ce n'est pas le cas pour la valeur des scores institutionnels modérés, qui dépendent de quels autres élèves font partie du groupe de modération.

– **Effet d'une très faible corrélation entre les résultats institutionnels et ceux à l'épreuve de calibration ou d'une forte variation des corrélations selon les groupes**

  Ce type de condition peut avoir des effets non désirables et faire en sorte que la comparabilité souhaitée des résultats soit plus ou moins fiable.

En plus des cas mentionnés ci-dessus, il faut également se préoccuper des situations de distribution bimodale des résultats, qui se produisent – bien que très rarement – lorsque le groupe est constitué de deux populations différentes d'élèves en ce qui concerne la réussite scolaire. Il y a également lieu de porter attention au fonctionnement différentiel des items (FDI), qui se produit lorsque, à habiletés égales, la probabilité de réussir

un item particulier est différente selon le groupe d'appartenance de l'élève. Par exemple, ce serait le cas si la difficulté d'un item n'était pas la même pour les garçons et pour les filles ayant obtenu des scores équivalents à l'épreuve de calibration.

La modération sociale

Définition et objectifs

Parmi les cinq procédures de modération relevées par Linn (1993), la modération sociale est la seule qui n'est pas une approche statistique, mais une approche axée sur le jugement professionnel des enseignants. La modération sociale (*social moderation* ou *standard-based moderation*[4]) consiste en une procédure collective (donc sociale) de confrontation de « jugements professionnels » sur l'évaluation de travaux d'élèves (Mottier Lopez, Tessaro, Dechamboux & Villabona, 2012). Elle vise la construction d'un consensus professionnel parmi un groupe d'enseignants. Elle est aussi appelée *consensus moderation* (Klenowski & Wyatt-Smith, 2010 ; Wilson, 2004), *group moderation* (Harlen, 2007) ou encore *auditing* et *verification* (Matters, 2006).

Le terme « harmonisation » est parfois utilisé en français au lieu de « modération sociale ». C'est le cas en Ontario dans le document ministériel *Faire croître le succès* (MEO, 2007, 2010) ou au Québec (MELS, 2015, 2016). Pour Linn (1993), la modération sociale est surtout nécessaire lorsque les tâches demandées aux élèves sont plus complexes (p. ex., production textuelle, réponses ouvertes) que d'autres tâches plus basiques (réponses à choix multiples, réponses courtes). Or, elle est également un moyen de contrôle et de développement en cas de non-recours à des évaluations standardisées, nécessaires pour la modération statistique (Colbert, Wyatt-Smith & Klenowski, 2012).

Tout comme la modération statistique, le but de la modération sociale est de s'assurer que les résultats issus du processus d'évaluation sommative/certificative/pronostique des élèves en classe sont de qualité, équitables et comparables (Harlen, 2007 ; Matters, 2006 ; Mottier Lopez et al., 2012). La modération sociale doit soutenir le développement des pratiques d'évaluation et de jugement professionnel des enseignants sous forme de communauté de pratique (Adie, Klenowski & Wyatt-Smith, 2012). Elle permet également un « alignement » des pratiques évaluatives des enseignants entre eux et sur les « standards » locaux et/ou nationaux (Laveault,

2009a ; Wilson, 2004). En effet, cette procédure doit permettre une compréhension commune de « standards » qui mènera à terme à une meilleure interprétation des résultats d'évaluation par les enseignants et par les établissements (Adie et al., 2012 ; Matters, 2006).

Les « standards » sont généralement apparents dans différents cadres de référence externes. Parfois, ils ne sont pas véritablement opérationnalisés ni commentés (programme d'études, standards de performance, évaluations standardisées, directives institutionnelles, politiques d'évaluation nationale, etc.). C'est cet espace d'interprétation que la modération sociale cherche à combler. Parfois, au contraire, ces standards sont illustrés au moyen de copies-types ou copies d'ancrage pouvant servir de modèles et limitant par le fait même l'étendue des interprétations possibles par les enseignants (p. ex., en Ontario). En s'inspirant de la définition de Maxwell (2001), les « standards » formulés peuvent indiquer plusieurs éléments aux enseignants :

1) les impératifs d'apprentissage : ce que devraient faire les élèves ;
2) les exigences institutionnelles : ce que doivent faire les élèves ;
3) les indicateurs comparables : ce que devraient maîtriser les élèves ;
4) les critères de réussite : à quel(s) niveau(x) d'atteinte devraient arriver les élèves ;
5) les étapes successives dans le cursus de l'élève : à quel(s) moment(s) les élèves devraient maîtriser les contenus attendus.

Il existe donc des standards de contenu (*content standards*) et des standards de performance (*achievement standards*) (Klenowski & Wyatt-Smith, 2010).

Ainsi, la modération sociale est un moyen pour les enseignants de comprendre de façon collective ces différents éléments afin de les intégrer dans leurs pratiques, surtout lors de l'implantation d'un nouveau curriculum (Adie et al., 2012 ; Wyatt & Smith, Klenowski & Gunn, 2010). Toutefois, Maxwell (2007) nomme deux finalités que les autorités peuvent donner à la modération sociale. D'un côté, la modération sociale est un instrument d'*accountability* ; elle sert à la « reddition de comptes », au contrôle et à la validation des pratiques enseignantes par l'administration. D'un autre côté, elle peut servir d'instrument de développement professionnel et de perfectionnement continu en matière d'évaluation qui, à terme, devrait permettre d'améliorer le système scolaire.

Exemple pratique et conditions d'utilisation

La modération sociale consiste en une confrontation de jugements d'évaluation entre enseignants afin de trouver un consensus sur les standards attendus. En principe, les enseignants évaluent individuellement les productions de leurs propres élèves. Puis, de manière collective, les enseignants se regroupent pour confronter leurs jugements sur ces travaux d'élèves et pour tenter de trouver un consensus sur leur évaluation (Klenowski & Wyatt-Smith, 2010 ; Linn, 1993). Les différences observées entre les évaluations sont donc la base des discussions au sein du groupe. Par cette démarche, les enseignants créent ainsi une représentation commune des attentes officielles, c'est-à-dire des « standards ». Pour atteindre ce consensus, il est donc nécessaire que les enseignants échangent sur une base commune de travaux d'élèves (méthode des signets ou copies d'ancrage ; en anglais, *benchmark or anchor papers*) issus d'une ou de plusieurs classes (Linn, 1993). Or, pour ce faire, il est également nécessaire qu'ils utilisent certains référents externes (objectifs d'apprentissage, standards de performance, copies-types, évaluations standardisées, etc.) (Colbert et al., 2012 ; Harlen, 2007 ; Mottier Lopez et al., 2012). Généralement, ces interactions sociales sont encadrées par un expert, c'est-à-dire une autorité locale capable d'observer les besoins de formation des enseignants (Harlen, 2007). Les enseignants et experts peuvent provenir d'un ou de différents établissements scolaires, voire de différentes juridictions scolaires. Enfin, la procédure et, en amont, les standards devraient être documentés au minimum au sujet du processus d'évaluation, de l'expertise des évaluateurs et du degré de consensus entre les juges (Linn, 1993).

Allal et Mottier Lopez (2014) distinguent quatre niveaux croissants d'exigence pour parvenir à une comparabilité accrue des résultats des élèves :

1) Transparence

Les enseignants sont d'accord que chacun doit pouvoir expliquer et illustrer ses pratiques d'évaluation sommative, tout en reconnaissant que celles-ci puissent varier.

2) Cohérence accrue

Les enseignants s'accordent sur un ensemble de lignes directrices qu'ils vont tous respecter dans leurs pratiques d'évaluation, mais leurs pratiques sont libres de varier sur d'autres aspects non mentionnés dans ces mêmes lignes directrices.

3) Qualité améliorée

Les enseignants développent ensemble des tâches d'évaluation bien alignées sur le curriculum ou encore s'accordent pour utiliser les mêmes tâches tirées d'une banque de tâches officielles. Une certaine variabilité peut toutefois subsister dans la manière de corriger les épreuves et de déterminer les notes.

4) Consistance et comparabilité des résultats d'évaluation

Les enseignants s'engagent dans des démarches de collaboration qui satisfont aux exigences de tous les niveaux précédents, avec en plus l'objectif d'assurer que la correction et les scores qui en résultent sont rigoureusement comparables.

Le quatrième niveau est le seul, selon ces auteures, à correspondre véritablement au but poursuivi par la « modération sociale ».

Principales juridictions qui ont institutionnalisé la modération sociale

Différentes procédures de modération sociale ont été développées dans certains contextes éducatifs, surtout dans les pays anglo-saxons. C'est notamment le cas au Pays de Galles, en Angleterre, en Écosse et en Australie (Klenowski & Wyatt-Smith, 2010). Néanmoins, les finalités (*accountability* et développement professionnel) et les résultats attendus (type et degré de consensus) sont variables selon les contextes.

Au Pays de Galles, les évaluations à grande échelle ont été graduellement abandonnées. Désormais, l'évaluation des apprentissages repose sur une plus grande confiance dans l'évaluation réalisée dans les établissements et par les enseignants. Dans ce contexte, la modération sociale joue un rôle important de contrôle et de développement de la qualité. Les écoles du primaire et du secondaire sont tenues de suivre les procédures de modération internes de l'établissement. Des rencontres entre les enseignants servent à confronter des évaluations de travaux d'élèves afin de construire une compréhension partagée et négociée des standards. Le niveau de compréhension des enseignants est contrôlé par des procédures de modération externes (*external moderation*) qui donnent aux établissements une accréditation. En Angleterre, de manière similaire, les évaluations à grande échelle perdent du terrain au profit de l'évaluation faite par les enseignants (Klenowski & Wyatt-Smith, 2010).

En Australie, l'État du Queensland a également choisi de faire surtout confiance à l'évaluation par les enseignants et par les écoles (Adie et al., 2012 ; Maxwell, 2001). Toutefois, ce modèle impose à cette juridiction de développer davantage les compétences évaluatives des enseignants et une compréhension commune des standards. Pour ce faire, la Queensland Studies Authority (QSA) a invité les enseignants du secondaire à expérimenter différentes formes de modération sociale (Adie et al., 2012). Les enseignants, en petits groupes de deux à six provenant de la même école ou d'écoles différentes, ont pour mission de discuter de productions d'élèves. Ces productions sont issues d'exercices de la banque de données des Queensland Comparable Assessment Tasks (QCAT). Trois procédures de modération sociale utilisant des modèles différents sont proposées aux enseignants (Queensland Studies Authority, 2007) :

1) Modèle de calibration (*Calibration model*)

Un « facilitateur » sélectionne un échantillon de productions d'élèves jugées représentatives de certains niveaux de standards (ici, de A à E). Les enseignants évaluent ces travaux, puis comparent leur évaluation. Les descripteurs des tâches sont utilisés comme base pour la discussion et pour trouver un consensus au sujet de la qualité des travaux. Le processus est ensuite répété pour les travaux des autres élèves.

2) Modèle de consensus (*Conferencing model*)

Les enseignants évaluent les travaux des élèves de manière individuelle, puis font une sélection de travaux représentant les différents niveaux de standards. Les enseignants discutent de ces évaluations en groupe et tentent de trouver un consensus.

3) Modèle de validation par un expert (*Expert model*)

Les enseignants évaluent les travaux des élèves de manière individuelle, puis font une sélection de travaux représentant les différents niveaux de standards. Ils les transmettent à un expert. Ce dernier donne de la rétroaction aux enseignants confirmant leur évaluation ou proposant des ajustements justifiés dans un document.

Appréciation de la modération sociale

Les trois procédures de modération sociale présentées et utilisées dans le Queensland comportent des avantages et des désavantages (Adie et al., 2012).

Avantages

Les procédures permettent aux enseignants de développer des compétences au sujet des standards sur la base de matériel authentique. De plus, à l'exception du modèle de validation par un expert, elles permettent aux enseignants de construire ces compétences avec leurs collègues, donc de parler le même langage et de créer une culture d'évaluation commune dans un dialogue professionnel négocié. La modération sociale est donc non seulement un moyen pour les administrations de s'assurer de la qualité et de l'équité de l'évaluation des élèves (permettant une compréhension des standards et une uniformisation du jugement d'évaluation), mais elle est aussi un outil pertinent de développement professionnel continu (Harlen, 2007 ; Linn, 1993 ; Mottier Lopez et al., 2012). La modération sociale s'avère particulièrement pertinente lors de changements importants de curriculum et pour des enseignants novices ou récemment arrivés dans un nouveau contexte (Adie, 2014). En ce sens, la modération permet de développer les compétences évaluatives en classe et l'assurance des enseignants dans leurs pratiques (Klenowski & Wyatt-Smith, 2010).

Selon nous, ce développement professionnel et continu peut assurer une modération locale. Il peut aussi redonner confiance dans l'évaluation réalisée par les enseignants et les écoles, et permettre de limiter l'usage et la fréquence d'évaluations externes uniformes, dont les effets indésirables sont bien connus (Yerly, 2017). La modération sociale redonne aux enseignants et aux écoles la compétence de l'évaluation des élèves, tout en conservant un aspect de pilotage et de reddition de comptes (Colbert et al., 2012).

Limites

La plupart des recherches empiriques récentes évaluant l'impact de la modération sociale s'accordent sur les limites de cette approche (Adie, 2014 ; Adie et al., 2012 ; Mottiez Lopez & Pasquini, 2017). La modération sociale est un processus long et fastidieux dont l'effet est parfois limité sur les pratiques enseignantes. Modifier le jugement professionnel des enseignants est en effet une entreprise complexe qui mobilise d'autres processus qu'une simple exposition à des standards et à des documents officiels. Le jugement d'évaluation est le fruit de facteurs divers : historiques, culturels, épistémologiques, contextuels, politiques, etc. La modération sociale nécessite une forte collaboration entre enseignants (Adie et al., 2012). Le consensus et le compromis peuvent être parfois difficiles. De plus, les

échanges entre enseignants ne sont pas toujours équilibrés ou mutuellement bénéfiques. Le rôle de l'expert peut devenir primordial, tout comme les conditions encadrant les rencontres et les relations professionnelles entre enseignants (Linn, 1993 ; Wyatt-Smith & Colbert, 2014). Cette procédure prend également du temps et crée une surcharge dans les situations où les enseignants ne sont pas libérés sur leur temps de travail. Le matériel utilisé doit aussi être sélectionné avec soin et permettre une discussion éclairée au moyen d'indicateurs externes utiles pour atteindre le consensus. Les enseignants doivent pouvoir compter sur des standards précis de performance et sur des critères d'évaluation décrivant clairement la progression attendue des apprentissages. Sans de tels standards, la modération sociale est limitée et la marge d'interprétation devient trop grande. Enfin, à terme, la modération sociale n'est utile que dans le cas où les enseignants intègrent le processus collectif dans leurs écoles et/ou dans le cas où ils ont stabilisé leurs connaissances des standards. Finalement, la modération sociale ne génère pas des résultats de manière directe et automatique ; elle prend du temps et nécessite des moyens (humains, matériels) importants. Or, comme l'écrivent Adie et ses collaboratrices (2012), la modération sociale, même si elle ne parvient pas toujours directement à ses fins, engage les enseignants dans une réflexion à long terme et en continu sur leurs pratiques et sur leur jugement des apprentissages des élèves.

Discussion

La comparaison des avantages et des limites de chaque approche de la modération

Les procédures de modération ont toutes leurs particularités et leurs propres intérêts. Nous avons déjà relevé les limites et les avantages des deux approches de modération des résultats d'évaluation : les modérations statistique et sociale. Nous avons mis en lien les différents arguments présentés dans la littérature, surtout anglophone. Le tableau 2, traduit de Linn (1993, p. 80), permet de les synthétiser et de les comparer sur différents points.

Tableau 2
Comparaison des avantages et des limites de la modération statistique et de la modération sociale (Linn, 1993, p. 80, trad. libre)

	Temps	**Coût**	**Contrôle bureau-cratique**	**Impact sur le processus**	**Impact sur le produit**	**Développement professionnel**
Modération statistique	bas	bas	élevé	aucun	élevé	aucun
Modération sociale	élevé	élevé	bas	élevé	moyen	élevé

Pour Linn (1993), il est évident que les procédures de modération impliquant davantage le contrôle de la qualité (*quality control*), telles que la modération statistique, ont un impact principalement sur le produit d'évaluation. Néanmoins, celles impliquant davantage une assurance de la qualité (*quality assurance*), par exemple la modération sociale, exercent plus un impact indirect sur le produit, à la suite de l'amélioration du processus. Pour Linn (1993), la modération sociale a donc non seulement l'avantage de toucher le produit et le processus, mais également de contribuer au développement professionnel des enseignants. Pourtant, pour cet auteur, elle devrait toujours s'accompagner d'un contrôle statistique.

Pour Maxwell (2007), les procédures de contrôle sont surtout essentielles pour l'évaluation à enjeux élevés (*high stakes assessment*), c'est-à-dire lorsque celle-ci est déterminante dans le cursus de l'élève ou dans la réputation, voire le classement des écoles et des enseignants. Pour les autres situations d'évaluation, les procédures de qualité (*quality assurance*) sont suffisantes. Toutefois, pour Maxwell (2007), les procédures de modération ne devraient pas être prises comme seuls instruments de contrôle, mais devraient faire partie d'un système plus global.

Les recherches sur la fidélité des notations réalisées par les enseignants démontrent que les notations peuvent être améliorées non pas en diminuant la responsabilité des enseignants envers l'évaluation, mais en leur fournissant davantage de renseignements, de formation et d'information à ce sujet. Alors qu'en modération statistique, les ajustements se font sur les résultats, en modération sociale, les ajustements se font sur le

développement professionnel des enseignants. Cette différence a conduit Harlen (2007) à conclure de la façon suivante : « Les procédures qui accordent une plus grande responsabilité aux enseignants dans le processus d'évaluation et qui leur fournissent des occasions pour la modération en groupe du processus et des résultats sont pédagogiquement plus solides et donnent lieu à une évaluation fiable » (Harlen, 2007, p. 67, traduction libre).

À qui profite la modération ?

L'analyse comparative de Linn (1993) (voir Tableau 2) donne déjà un aperçu synthétique des limites et avantages de ces deux démarches. Toutefois, il nous paraît important de compléter ces données en contrastant ces éléments selon le regard des principaux acteurs de la scolarité. En effet, les enjeux ne sont pas les mêmes pour les élèves (les principaux intéressés auxquels nous associons leurs parents), les gestionnaires (directeurs d'école, administrateurs du système scolaire) ou les enseignants. Nous tentons, dans cette première partie conclusive, de répondre à la question suivante : À qui profitent les différentes démarches de modération des résultats d'évaluation ?

À notre avis, les principaux bénéficiaires de la modération, dans toutes ses formes, devraient être principalement les élèves. Toute procédure de modération devrait être, pour eux, garante de plus d'équité et d'un meilleur alignement. Toutefois, cette première catégorie d'acteurs n'a aucun contrôle sur ces procédures. Ce sont les gestionnaires qui ont ce contrôle. Pour les gestionnaires, il importe de pouvoir contrôler et améliorer « à distance » les résultats d'évaluation scolaire. Aussi, pour eux, d'un point de vue pragmatique, l'efficacité, voire l'efficience de la démarche (rationalisation des coûts par rapport aux résultats) sont primordiales. Les enseignants, quant à eux, sont touchés directement par ces procédures. Elles ont un impact sur leurs pratiques, mais aussi sur la représentation qu'ils se font de leur profession. En effet, la modération sous ses différentes formes peut leur offrir des occasions de développement professionnel, mais peut aussi être la cause de désagréments (p. ex., perte d'autonomie et de confiance).

Dans le tableau 3, nous synthétisons les avantages et les limites de la modération statistique et de la modération sociale pour chacune des trois catégories d'acteurs. Ces données sont ensuite mises en comparaison afin de répondre à notre questionnement.

Tableau 3

Avantages et limites de la modération pour les acteurs

	Gestionnaires	**Enseignants**	**Élèves et parents**
Modération statistique	+ impact automatique et total sur le produit + démarche peu coûteuse + démarche courte + contrôle important sur la démarche	+ aucun effort à fournir	+ effet direct + équité totale
	- aucun impact sur le processus évaluatif - pas de développement professionnel du personnel enseignant - impact non durable sur le produit (artificiel)	- ajustement non durable - pas de développement professionnel - démarche statistique peu connue des enseignants - peu de confiance dans les compétences des enseignants	- effet provisoire (artificiel) - procédure difficile à comprendre - baisse de la confiance dans l'école/les enseignants
Modération sociale	+ impact durable sur le produit + impact sur le processus évaluatif + développement professionnel + confiance dans les compétences du personnel/meilleure image de l'école + adéquation de la démarche avec la formation et les programmes	+ ajustement durable + développement professionnel des enseignants + confiance dans les compétences des enseignants + processus participatif	+ procédure invisible + effet durable + confiance dans les enseignants + effet indirect sur l'élève
	- impact partiel sur le produit - démarche coûteuse - démarche longue - peu de contrôle sur la démarche et ses résultats	- demande du temps/de l'énergie	- équité partielle

Note. + = avantages; - = limites.

Le cas des gestionnaires

La modération statistique possède des caractéristiques particulières qui peuvent intéresser les gestionnaires. Elle permet d'avoir un impact automatique et total sur le produit, car les résultats de toutes les classes sont traités. Cette démarche est aussi plus rapide et moins coûteuse avec des examens à choix multiples ou des tests qui ne requièrent pas l'embauche de correcteurs professionnels. En outre, les gestionnaires gardent un contrôle total sur la démarche. Ce qui constitue un avantage en modération statistique peut devenir un inconvénient. Par exemple, la modération sociale peut demander davantage de temps et de financement, et est difficilement contrôlable en totalité. La formation des enseignants à la modération sociale comporte également un coût. Or, comme l'a mis en avant Harlen (2007), elle permettrait des changements plus durables, moins artificiels. Elle serait aussi plus pertinente sur le plan pédagogique pour engager de véritables ajustements de la part des enseignants. Toutefois, il s'agit dans tous ces cas d'hypothèses qu'il importe de mieux documenter.

Le cas des enseignants

Contrairement à la modération statistique, qui n'apporte que peu d'avantages aux enseignants (voire aucun s'ils ne reçoivent aucune forme de rétroaction sur les résultats), la modération sociale peut leur être profitable à long terme. La modération statistique exclut les enseignants de la démarche et leur transmet le sentiment qu'on ne fait pas confiance à leurs compétences ou qu'on réduit leur rôle à celui d'un technicien exécutant. La modération sociale leur donne plutôt la possibilité de développer leurs compétences en participant à la démarche et en créant de nouvelles connaissances et une culture commune avec leurs pairs (et pourquoi pas avec les gestionnaires qui participent à la démarche). D'ailleurs, la modération statistique est assez lointaine des pratiques évaluatives, voire des programmes défendus par les facultés d'éducation. En effet, cette pratique s'oppose par exemple aux pratiques pédagogiques contemporaines d'évaluation formative et d'évaluation des compétences. Toutefois, si la modération sociale accorde un rôle important aux enseignants en reconnaissant leurs compétences et leur professionnalisme, elle leur demande un investissement important. Bien sûr, son impact sur le produit n'est pas total.

Le cas des élèves et de leurs parents

Les élèves et leurs parents tiennent un rôle passif aussi bien dans la modération statistique que sociale. Toutefois, c'est pour eux que l'enjeu est le plus important. Réalisée correctement, la modération statistique assure instantanément une certaine équité entre les élèves. Néanmoins, cette transformation soudaine d'un résultat scolaire peut aussi provoquer chez eux de la surprise ou du mécontentement et surtout les amener, par manque de connaissance et/ou d'information, à s'interroger sur le fonctionnement du système scolaire et sa qualité. En effet, peuvent-ils faire confiance à un système qui modifie ses résultats a posteriori ? En outre, l'impact de la modération statistique et son avantage d'équité ne sont que de très courte durée. Une fois passée la modération statistique de son évaluation, l'élève retrouvera bien vite les pratiques plus « subjectives » de ses enseignants.

La modération statistique est séduisante et peut paraître une démarche efficiente surtout aux yeux des gestionnaires. Toutefois, elle n'a un effet qu'à courte durée et est très peu profitable (voire pas du tout) pour les enseignants si elle n'est pas accompagnée d'autres démarches de formation. À terme, elle risque même de causer certains dégâts. En effet, elle peut entraîner la démotivation et la déresponsabilisation des enseignants, mais aussi étendre encore davantage le manque de confiance de la population envers l'école et ses enseignants.

La modération statistique peut être bénéfique pour les élèves, surtout pour ceux qui ont été « sous-évalués ». Toutefois, elle ne leur permet pas d'approfondir les critères d'évaluation, de mieux connaître ce qui est attendu de leur part et, en bout de piste, de mieux situer leurs résultats en cours d'année. Bref, la modération statistique ne permettrait pas plus aux élèves qu'aux enseignants de développer leur jugement d'évaluation et leurs compétences à s'autoévaluer.

Finalement, ce sont donc deux visions qui s'opposent quant à la modération des résultats d'évaluation. La modération statistique propose une vision managériale (descendante ou *top-down*) axée sur l'efficacité et la production immédiate de résultats et sur le contrôle des enseignants. La modération sociale offre quant à elle une vision pédagogique axée sur la participation et l'amélioration du processus et sur la confiance envers le professionnalisme et les compétences des enseignants.

Conclusion

Quel avenir pour la modération ?

Modération statistique : que promettent les avancées en psychométrie ?

Les nombreux changements se produisant dans le monde de l'éducation font en sorte que le portrait actuel de la modération statistique est appelé à changer et à évoluer. Les méthodes statistiques utilisées pour réaliser la modération jusqu'à présent reposent sur les modèles de la théorie classique des scores (TCS) et datent déjà de plusieurs années. Les puissants moyens de mise en équivalence offerts par la théorie de réponse aux items (TRI) font en sorte que les méthodes « classiques » de modération statistique sont à reconsidérer à l'éclairage des possibilités offertes par ces nouveaux modèles et par les nouvelles technologies. Wilmut et Tuson (2005) reconnaissaient le besoin de mettre à jour les connaissances dans ce domaine.

Qu'il s'agisse des modèles de la TRI ou de la TCS, aucun modèle statistique ne pourra faire abstraction des conditions d'utilisation de la modération statistique que nous avons décrites précédemment. Les postulats de la TRI (unidimensionnalité et indépendance locale) sont encore plus forts et plus exigeants que ceux de la TCS. Ils nécessitent par ailleurs des échantillons plus grands que les méthodes classiques, car la TRI repose sur l'estimation des paramètres d'items.

L'intérêt grandissant pour la notation de performances complexes dans le cadre d'une approche par compétences fait en sorte que la modération statistique risque de devenir de plus en plus complexe et difficile d'application. Comment justifier l'utilisation de modèles unidimensionnels de la TRI dans le cas de performances fort probablement multidimensionnelles ? Enfin, l'utilisation d'échelles de notes qualitatives ou d'échelles qui comportent un nombre limité d'échelons, par exemple la notation de A à E, même si elle demeure possible (Kim & Cohen, 2002), complique davantage l'utilisation de tels modèles.

Cela dit, il ne faut pas sous-estimer les progrès remarquables de la psychométrie depuis l'introduction de la TRI, ni les possibilités présentées par l'introduction de plus en plus répandue du testing assisté par ordinateur (TAO) pour assurer la comparabilité des résultats et une forme d'équi-

té. Or, dans le cas du TAO comme dans celui de la TRI, la mise en pratique de ces nouvelles avancées ne paraît pas devoir se faire à grande échelle dans un avenir immédiat.

Une des limites les plus importantes de la modération statistique tient à ce que son impact se limite à la comparabilité des résultats. Dans le contexte où les cadres d'évaluation des juridictions scolaires aspirent à plus d'équité, non seulement sur le plan des résultats mais aussi des chances de réussite, la modération statistique s'inscrit davantage dans un cadre d'évaluation sommative, plutôt que dans un cadre d'évaluation formative dans lequel l'évaluation jouerait un rôle de soutien à l'apprentissage des élèves et de régulation du système éducatif.

Modération sociale : encore de nombreux défis à surmonter

En modération statistique, le choix des élèves qui feront partie du groupe de modération a une influence sur l'ajustement des scores individuels (*the "company you keep" factor*; Wilmut et Tuson, 2005). Or, un phénomène similaire ne joue-t-il pas sur la modération sociale ? Dans quelle mesure la qualité de la modération sociale dépend-elle de la façon dont les équipes d'enseignants sont constituées et de la somme des compétences individuelles des enseignants qui en font partie (un autre *"company you keep" factor*) ? Il est irréaliste de tenir pour acquis que les équipes d'enseignants sont également compétentes et efficaces, et que les standards sont facilement interprétables par tous. Aussi, la qualité et la lisibilité des « standards » que les enseignants sont appelés à interpréter pour atteindre un consensus peuvent être remises en question. Il en va de même pour la qualité de l'accompagnement, qui ne leur est pas toujours fourni (p. ex., expertise et pratique des experts), et pour les conditions de réalisation (p. ex., temps). Bref, alors que les postulats de la modération statistique sont assez bien connus, ceux de la modération sociale, eu égard au fonctionnement des équipes d'enseignants et d'experts, ne sont pas toujours clairs quant à la façon dont ils affectent la comparabilité des résultats scolaires.

Devant les défis que présente la modération sociale, Wyatt-Smith et Colbert (2014) font le bilan de 40 années d'expérience de l'État australien du Queensland dans ce domaine. Elles formulent plusieurs recommandations visant à améliorer la composition et le fonctionnement des panels de modération et à assurer un haut niveau de compétence des enseignants qui en font partie. Voici quelques exemples de ces recommandations :

- Organiser des séances de perfectionnement professionnel sur une base continue, tant pour les présidents des panels de modération que pour les membres ;
- Créer un système d'accréditation qui certifie la compétence des membres des panels de modération ;
- Rendre disponible du matériel de formation, notamment des copies-types d'élèves dont les résultats sont très près du seuil de réussite (cas frontières), avec commentaires expliquant et justifiant les décisions de réussite ou d'échec lorsque des renseignements de différentes natures doivent être combinés et des compromis réalisés ;
- Établir un calendrier des années de service des enseignants en modération sociale de manière à assurer la formation de panels de modération équilibrés et d'égales compétences, et ce, d'une année à l'autre ;
- Consigner et analyser les rapports de modération de manière à mieux comprendre comment les enseignants interprètent et utilisent les standards, et comment l'évaluation s'aligne sur le curriculum et l'enseignement.

Ces recommandations attestent non seulement des défis à relever pour mettre en place un système de modération sociale de qualité, mais aussi de la difficulté à assurer que différents groupes de modération sociale sont de compétences équivalentes entre eux et d'année en année. De trop grands écarts dans la composition des différents panels peuvent réduire la comparabilité des résultats et nuire à l'atteinte des objectifs d'équité recherchés. De plus, la mise en place d'un système fonctionnel de modération sociale entraîne de nombreux coûts récurrents en dégagement de tâche et en développement professionnel des enseignants pour que ceux-ci s'acquittent adéquatement de leurs responsabilités.

Intérêt pour une approche hybride de la modération

Si une certaine forme de modération est nécessaire pour assurer davantage d'équité entre les élèves, il nous semble pertinent, au regard des arguments avancés et à l'image de Harlen (2007) et de Linn (1993), de viser une démarche plus pédagogique dont les résultats sont plus durables et non artificiels. Depuis longtemps, une approche « hybride » de la modération a été évoquée par Linn (1993). Dans sa perspective, le contrôle statistique agit comme garde-fou ou signal d'alarme, voire comme une occasion d'engager les enseignants dans des démarches collectives de régulation

de leurs pratiques d'évaluation. Toutefois, suffit-il d'informer les enseignants de l'écart entre la note institutionnelle et la note modérée pour que ceux-ci sachent quoi faire pour y remédier par eux-mêmes ?

La pertinence de l'utilisation conjointe des deux types de modération est à rechercher dans des cadres d'évaluation où ils sont utilisés de manière à s'articuler dans une forme d'interaction ou de corégulation donnant lieu à des ajustements entre enseignants, évaluateurs et concepteurs des programmes d'études. Selon Matters (2006), tous les États et territoires australiens (à l'exception de Victoria) ont développé à la fois des approches statistiques et sociales de la modération. Le cas de l'État de l'Australie-Méridionale est, pour Maxwell (2007), le modèle de modération le plus élaboré. En effet, il implique plusieurs types d'assurance et de contrôle de la qualité. Les écoles doivent fournir les travaux des élèves et les résultats à un panel d'experts pour validation. Les enseignants, en petits groupes, comparent leurs évaluations sur la base des travaux de leurs élèves. Des « modérateurs » de l'administration visitent les écoles afin de valider l'évaluation des travaux. Enfin, la modération statistique s'applique uniquement pour les élèves qui ont également été testés par une épreuve externe. Ces derniers résultats ne sont utilisés comme standards que pour les écoles qui ne sont pas engagées dans une des trois procédures de modération précédemment citées. Le cas de l'État de l'Australie-Méridionale démontre que les deux types de modération peuvent s'articuler et que chaque juridiction peut développer une approche différenciée de la modération selon son cadre d'évaluation et ses priorités.

Que faut-il exiger de la modération ?

Au-delà des objectifs de comparabilité et d'équité poursuivis par toute approche de modération (statistique, sociale ou même combinée), quel est, à long terme, l'impact de ces approches sur l'efficacité du système éducatif et sur la réussite scolaire des élèves ?

C'est sur le plan de l'impact, direct et indirect, de chacune des approches de la modération qu'il faut juger de leur valeur respective. Pour l'instant, tant la modération sociale que la modération statistique ne permettent d'assurer la comparabilité et l'équité des résultats que pour des élèves appartenant à la même cohorte. Qu'en est-il de la comparabilité et de l'équité des résultats d'élèves de différentes cohortes ? Sans items communs pour les épreuves de calibration d'année en année, il est impossible d'effectuer les transformations de mise en équivalence des résultats qui

permettraient de situer les résultats des élèves sur la même échelle, ni d'apprécier si la modération a un impact positif sur la régulation du travail des enseignants et, par ricochet, sur l'amélioration des résultats d'apprentissage des élèves.

C'est sur la réussite scolaire du plus grand nombre, nous semble-t-il, que doivent porter les efforts d'équité des systèmes éducatifs. S'il est difficile de remettre en question la comparabilité des résultats pour assurer l'équité des évaluations sommatives/certificatives des élèves, il importe également de se demander si la mobilisation des ressources peut aller au-delà de la comparabilité et améliorer réellement les compétences professionnelles du personnel enseignant. À cet égard, il serait important de vérifier si les bénéfices attendus de la modération sociale sur le développement du jugement professionnel d'évaluation des enseignants et des élèves se généralisent au-delà des épreuves finales et perdurent tout au long de l'année scolaire dans l'ensemble des activités d'évaluation. La modération sociale gagnerait énormément en crédibilité si elle s'avérait – données probantes à l'appui – réellement plus efficace à cet égard que la modération statistique et s'il était démontré que ses effets sur le jugement d'évaluation des enseignants se transfèrent aux capacités d'autoévaluation des élèves (*assessment as learning*; Earl, 2003). Au-delà de la comparabilité et en l'absence de résultats de recherche quant à l'impact véritable de différentes approches de modération, c'est à l'aune d'un impact durable sur l'amélioration de la réussite scolaire des élèves qu'il faudra juger de l'utilité des différentes approches de modération.

Réception : 8 décembre 2016

Version finale : 16 juillet 2017

Acceptation : 19 juillet 2017

NOTES

1. Linn (1993) préfère parler de *linking* ou *scaling*. En français, les termes « liaison » et « échelonnage » décrivent une famille de procédures statistiques visant à assurer non seulement la comparabilité des scores, mais leur équivalence. Dans le cas de la modération statistique, seule la comparabilité des scores est assurée.
2. Pour plus de renseignements sur la modération statistique au Québec, voir le site web du ministère de l'Éducation et de l'Enseignement supérieur (anciennement le MELS) : www.education.gouv.qc.ca/eleves/examens-et-epreuves/traitement-des-resultats/conversion-et-moderation
3. La problématique de l'alignement a été traitée notamment par Looney (2011) et par Laveault (2009a, 2009b).
4. P^re^ Claire Wyatt-Smith (communication personnelle, septembre 2015).

RÉFÉRENCES

Adie, L. (2014). The development of shared understandings of assessment policy: Travelling between global and local contexts. *Journal of Education Policy*, *24*(4), 532-545. doi: 10.1080/02680939.2013.853101

Adie, L. E., Klenowski, V., & Wyatt-Smith, C. (2012). Towards an understanding of teacher judgement in the context of social moderation. *Educational Review*, *64*(2), 223-240. doi: 10.1080/00131911.2011.598919

Allal, L., & Mottier Lopez, L. (2014). Teachers' professional judgment in the context of collaborative assessment practice. In C. Wyatt-Smith, V. Klenowski & P. Colbert (Eds.). *Designing assessment for quality learning* (pp. 151-165). London: Springer.

Board of Studies, Teaching & Educational Standards NSW. (2011). *Explanation of aligning and moderating procedures for the Higher School Certificate.* Sydney: New South Wales Education Standards Authority. Retrieved from www.boardofstudies.nsw.edu.au/hsc-results/moderation.html

Burton, E., & Linn, R. L. (1994). *Comparability across assessments: Lessons from the use of moderation procedures in England*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York: Routledge.

Colbert, P., Wyatt-Smith, C., & Klenowski, V. (2012). A systems-level approach to building sustainable assessment cultures: Moderation, quality task design and dependability of judgement. *Policy Futures in Education*, *10*(4), 386–401. doi: 10.2304/pfie.2012.10.4.386

Earl, L. (2003). *Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oaks, CA: Corwin Press.

Elley, W. B., & Livingstone, I. D. (1972). *External examinations and internal assessments.* Wellington: NZCER.

Harlen, W. (2007). *Assessment of learning*. London: SAGE Publications.

Hong Kong Examinations and Assessment Authority (HKEAA). (2010). *Moderation of school-based assessment scores in the HKDSE*. Hong Kong: HKEAA. Retrieved from www.hkeaa.edu.hk/DocLibrary/Media/Leaflets/HKDSE-SBA-Moderation Booklet_r.pdf

Johnson, S. (2011). *A focus on teacher assessment reliability in GCSE and GCE*. Coventry, UK: Office of Qualifications and Examinations Regulation.

Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*, 25-41. doi: 10.1177/0146621602026001002

Klenowski, V., & Wyatt-Smith, C. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assessment Matters*, *2,* 107-131. Retrieved from http://researchbank.acu.edu.au/fea_pub/583

Laveault, D. (2009a). L'évaluation en classe : des politiques aux pratiques. *Mesure et évaluation en éducation*, *32*(3), 1-22. doi: 10.7202/1024929ar

Laveault, D. (2009b). L'amélioration de l'efficacité du système éducatif : sur quels indicateurs s'appuyer ? Dans X. Dumay & V. Dupriez (dir.), *L'efficacité dans l'enseignement : promesses et zones d'ombre* (pp. 177-196). Bruxelles : De Boeck.

Laveault, D. & Grégoire, J. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (3e éd.). Bruxelles : De Boeck.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *1*(1), 83-102. doi: 10.1207/s15324818ame0601_5

Looney, J. W. (2011). Alignment in complex education systems: Achieving balance and coherence. *OECD Education Working Papers*, *64*. doi: 10.1787/5kg3vg5lx8r8-en

Matters, G. (2006). *Statistical moderation and social moderation around Australia.* Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment, Singapore. Retrieved from https://works.bepress.com/gabrielle_matters/42

Maxwell, G. S. (2001). *Moderation of assessments in vocational education and training.* Brisbane: Queensland Government Departement of Employment and Training.

Maxwell, G. S. (2007). *Implications for moderation of proposed changes to senior secondary school syllabuses*. Brisbane: Queensland Studies Authority.

Ministère de l'Éducation de l'Ontario (MEO). (2007). *Harmonisation de l'évaluation : une collaboration pour évaluer de façon cohérente et équitable le travail des élèves*. Toronto : Gouvernement de l'Ontario, Secrétariat de la littératie et de la numératie. Repéré à www.edu.gov.on.ca/fre/literacynumeracy/inspire/research/Teacher_Moderation_fr.pdf

Ministère de l'Éducation de l'Ontario (MEO). (2010). *Faire croître le succès : évaluation et communication du rendement des élèves fréquentant les écoles de l'Ontario* (1re éd.). Toronto : Gouvernement de l'Ontario. Repéré à www.edu.gov.on.ca/fre/policyfunding/growSuccessfr.pdf

Ministère de l'Éducation, du Loisir et du Sport (MELS). (2011). *Sanction des études.* Québec : Gouvernement du Québec, Direction de la sanction des études.

Ministère de l'Éducation, du Loisir et du Sport (MELS). (2015). *Guide de gestion – Édition 2015 : Sanction des études et épreuves ministérielles – Formation générale des jeunes ; formation générale des adultes ; formation professionnelle*. Québec : Gouvernement du Québec. Repéré à www.education.gouv.qc.ca/fileadmin/site_web/documents/dpse/sanction/Guide-sanction-2015_fr.pdf

Ministère de l'Éducation, du Loisir et du Sport (MELS). (2016). *Examens et épreuves : conversion et modération.* Québec : Gouvernement du Québec. Repéré à www.education.gouv.qc.ca/eleves/examens-et-epreuves/traitement-des-resultats/conversion-et-moderation

Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods and prospects*. Princeton, NJ: Educational Testing Service.

Mottier Lopez, L., & Pasquini, R. (2017). Professional controversies between teachers about their summative assessment practices: A tool for building assessment capacity. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 228-249. doi: 10.1080/0969594X.2017.1293001

Mottier Lopez, L., Tessaro, W., Dechamboux, L. & Villabona, F. M. (2012). La modération sociale : un dispositif soutenant l'émergence de savoirs négociés sur l'évaluation certificative des apprentissages des élèves. *Questions vives : Recherches en éducation*, *6*(18), 159-175. doi: 10.4000/questionsvives.1235

Nusche, D., Laveault, D., MacBeath, J., & Santiago, P. (2012). *OECD Reviews of evaluation and assessment in education: New Zealand 2011.* Paris: OECD Publishing. doi: 10.1787/9789264116917-en

Queensland Studies Authority (QSA). (2007). *September 2007 trial common assessment task: Teacher guidelines*. Brisbane: QSA.

Wilmut, J., & Tuson, J. (2005). *Statistical moderation of teacher assessments: A report to the Qualifications and Curriculum Authority*. London: Qualifications and Curriculum Authority.

Wilson, M. (2004). Assessment, accountability and the classroom: A community of judgment. *Yearbook of the National Society for the Study of Education*, *103*(2), 1-19. doi: 10.1111/j.1744-7984.2004.tb00046.x

Wyatt-Smith, C., & Colbert, P. (2014). *An account of the inner workings of standards, judgement and moderation: A previously untold evidence-based narrative*. Brisbane: Australian Catholic University. Retrieved from http://research.acer.edu.au/cgi/viewcontent.cgi?article=1006&context=qld_review

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 59-75. doi: 10.1080/09695940903565610

Yerly, G. (2017). Évaluation des apprentissages en classe et évaluation à large échelle: quel est l'impact des épreuves externes sur les pratiques évaluatives des enseignants? *Mesure et évaluation en éducation*, *40*(1), 33-60.