le cnam
Cnesco
Centre national d'étude des systèmes scolaires

UCA Université Clermont Auvergne

INSPÉ Institut national supérieur du professorat et de l'éducation Clermont Auvergne

CONFÉRENCE DE **CONSENSUS**

L'ÉVALUATION EN CLASSE, AU SERVICE DE L'APPRENTISSAGE DES ÉLÈVES

LIMITES ET BIAIS DE L'ÉVALUATION DANS LE CADRE SCOLAIRE : SYNTHÈSE DES RÉSULTATS DE RECHERCHE

#CC_EVALUATION

LES 23 ET 24 NOVEMBRE 2022

En partenariat avec :

ifé Institut français de l'éducation

Université de Lyon

ENS de Lyon

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

RESEAU-CANOPE.FR
CANOPÉ
LE RÉSEAU DE CRÉATION ET D'ACCOMPAGNEMENT PÉDAGOGIQUES

INSPÉ RÉSEAU DES INSPÉ

LIMITES ET BIAIS DE L'ÉVALUATION DANS LE CADRE SCOLAIRE : SYNTHÈSE DES RÉSULTATS DE RECHERCHE

Jean-François CHESNÉ

Lucile PIEDFER-QUENEY

Fiona JEANNEAU

Cnesco

Mars 2023

le cnam
Cnesco

Centre national d'étude des systèmes scolaires

Pour citer ce document, merci d'utiliser la référence suivante :
Chesné, J.-F., Piedfer-Quêney, L. & Jeanneau, F. (2023). *Limites et biais de l'évaluation : synthèse des travaux de recherche.* Cnesco-Cnam.

Ce texte s'inscrit dans une série de ressources publiées par le Centre national d'étude des systèmes scolaires (Cnesco) sur la thématique : **L'évaluation en classe.**

Disponible sur le site du Cnesco : www.cnesco.fr
Publié en mars 2023.
Centre national d'étude des systèmes scolaires
41 rue Gay-Lussac 75 005 Paris

Contact : cnesco@lecnam.net - 06 98 51 82 75

Sommaire

Liste des encadrés

Introduction

Les résultats de recherche sur la fiabilité (voir Encadré 1) du jugement évaluatif sont des résultats scientifiques bien établis. La plupart des études ont été répliquées et leurs conclusions font consensus parmi les chercheurs. Il semble donc important de les porter à la connaissance de tous dans le cadre d'une réflexion sur l'évaluation (en classe), et ce d'autant plus que les recherches qui se sont développées à la suite des premiers travaux de docimologie portent sur des situations d'évaluation proches de ce qui se fait dans le quotidien ordinaire des classes. Si l'on veut mettre l'évaluation au service de l'apprentissage des élèves, il paraît donc intéressant de connaître les facteurs qui peuvent influencer le jugement évaluatif.

Encadré 1 : Pertinence, validité, fiabilité : définitions

Pour définir les termes de pertinence, validité et fiabilité, nous nous appuyons sur un article de De Ketele & Gérard (2005) :

« • La **pertinence** est le caractère plus ou moins approprié de l'épreuve, selon qu'elle s'inscrit dans la ligne des objectifs visés (De Ketele *et alii*, 1989). C'est son degré de « compatibilité » avec les autres éléments du système auquel elle appartient (Raynal & Rienier, 1997, 2003).

• La **validité** est le degré d'adéquation entre ce que l'on déclare faire (évaluer telle ou telle dimension) et ce que l'on fait réellement, entre ce que l'outil mesure et ce qu'il prétend mesurer (Laveault & Grégoire, 1997, 2002).

• La **fiabilité** est le degré de confiance que l'on peut accorder aux résultats observés : seront-ils les mêmes si on recueille l'information à un autre moment, avec un autre outil, par une autre personne, etc. ? Elle nous renseigne sur le degré de relation qui existe entre la note obtenue et la « note vraie » (Cardinet & Tourner, 1985 ; Laveault & Grégoire, 1997, 2002). Il ne faut cependant pas perdre de vue que la note vraie est une abstraction, un point de convergence souhaité indépendant des évaluateurs et des circonstances. »

Pour plus de détails et des exemples, nous renvoyons le lecteur à l'article dont sont issues ces définitions (De Ketele & Gérard, 2005).

Pourquoi la question de la fiabilité de l'évaluation suscite-t-elle tant d'attention ?

Dans la recherche en éducation, **l'évaluation** scolaire est définie comme un **processus qui consiste à définir un objet d'évaluation** (ce que l'on souhaite évaluer), **à collecter de l'information** (en proposant un exercice ou une mise en situation à un élève et en observant ce qu'il fait ou produit), **à interpréter ces informations** (en portant un jugement sur ce que l'élève a fait ou produit) et **à prendre une décision**, à agir en conséquence (Allal, 2008). Cette dernière étape (prise de décision/action) recouvre aussi bien l'adaptation du cours par un enseignant en fonction du niveau qu'il a perçu chez ses élèves, que l'affectation d'un élève dans tel ou tel établissement, ou encore son orientation dans telle ou telle filière d'enseignement. Les conséquences de l'évaluation peuvent ainsi constituer des enjeux de nature et d'intensité variables.

Cependant, ce que l'on cherche à évaluer n'est pas évident : en effet, si l'on cherche à mesurer la taille d'un individu, il existe une échelle de mesure absolue, et le risque d'erreur dans la mesure est en principe très faible. Mais si l'on cherche à apprécier le niveau d'un élève en mathématiques par exemple, peut-on en dire autant ? Et d'ailleurs, qu'est-ce qui définit le niveau en mathématiques d'un élève ? Une autre série de questions vient de l'aspect ponctuel d'une évaluation : un travail réalisé à un moment donné est-il réellement révélateur des acquis d'un élève ? La fatigue, le stress, la mauvaise interprétation d'une consigne, etc. sont autant d'aléas qui peuvent impacter la performance d'un élève à un devoir. Comment les prendre en compte dans l'appréciation de cette performance ? Sont-ils d'ailleurs des éléments à prendre en compte ? Ces limites doivent d'emblée amener à relativiser l'importance accordée à l'acte d'évaluer et à son résultat quand il est pris isolément.

Par ailleurs, même si l'on considère que par la répétition des évaluations (formelles et informelles), les enseignants peuvent se faire une idée assez précise de l'avancement des élèves dans leurs apprentissages en dépit de ces biais, les résultats de recherche sur la fiabilité de l'évaluation méritent d'être connus. En effet, dans le système éducatif français, plusieurs fonctions de l'évaluation coexistent, voire se superposent, de sorte que l'évaluation en classe peut servir à soutenir l'apprentissage, en même temps qu'elle peut être prise en compte pour certifier (dans le cadre du contrôle continu) et/ou pour orienter les élèves (à travers les livrets scolaires notamment). La coexistence de ces enjeux engendre une exigence de fiabilité pour l'ensemble des pratiques évaluatives en milieu scolaire. Il semble donc utile de revenir sur les biais et les limites qui sont susceptibles d'affecter cette fiabilité.

Comment les recherches se sont-elles emparées de la question de la fiabilité de l'évaluation scolaire ?

Même si l'histoire compte quelques tentatives isolées pour déconnecter évaluation scolaire et note (comme en 1969, voir Heurdier & Prost, 2021), l'évaluation scolaire se matérialise, en France et depuis de nombreuses décennies, par une notation chiffrée (voir Encadré 2).

Encadré 2 : Comment en est-on arrivé à la note sur 20 en France ?

L'instauration de la notation est relativement récente en France. De fait, les pratiques évaluatives présentes au sein des collèges jésuites et des écoles chrétiennes au XVI^{e} et XVII^{e} siècles avaient très peu recours à cet outil. Les premiers, fondés sur la compétition et la sélection, procédaient au classement de leurs élèves, ce qui permettait de définir « le rang de chaque élève […] et [de déterminer ainsi] sa valeur scolaire » (Merle, 2015). À l'inverse, la seconde forme d'institution scolaire évaluait ses élèves par le biais d'une « appréciation globale des "compétences acquises" » (Merle, 2015).

C'est dans le cadre de concours que la note a émergé. Dans les années 1780, les prétendants à l'École de la Marine étaient plus nombreux que le nombre de places défini par le ministère de la Marine : il fallait donc réussir un concours pour intégrer l'école. Un examinateur unique parcourait la France et établissait un classement, qu'il transmettait au ministère. Si l'examinateur utilisait des lettres (de « a » à « g ») pour garder des traces de ses évaluations et réaliser le classement final, aucune lettre ni aucune note n'était attribuée de façon visible aux candidats. Pour intégrer l'École polytechnique, à partir de 1794, il fallait également réussir un concours. Les épreuves étaient organisées simultanément

dans vingt-deux villes de France. En raison du grand nombre de candidats, le recours à un examinateur unique était impossible. Les différents examinateurs établissaient donc des classements locaux, mais la mise en commun de ces classements au niveau national pour établir une liste unique soulevait d'importantes difficultés puisque les candidats étaient ordonnés (localement) mais que leur prestation n'avait pas de valeur absolue. Des mouvements de contestation de l'évaluation du concours ont accéléré l'évolution des pratiques pour conduire finalement, à partir de 1852, à l'utilisation d'une note sur 20 et d'un système de pondération des différentes épreuves du concours. Cela marque le passage d'un régime de sentence (admis/non admis) à un régime de mesure.

Progressivement, les professeurs des classes préparatoires ont adopté la notation sur 20 dans l'objectif de mieux préparer leurs élèves au concours. Les modalités de la diffusion de cette note dans l'enseignement secondaire sont moins bien connues. C'est en tout cas en 1890 qu'un arrêté a fixé l'échelle sur 20 points pour les compositions réalisées au collège et au lycée et qu'un autre arrêté l'a instaurée pour les compositions écrites du baccalauréat. Ces deux arrêtés montrent bien comment les exigences d'un examen se répercutent sur les pratiques d'évaluation en classe.

S'agissant de l'enseignement primaire, il fonctionnait au XIX^{e} siècle plutôt sur le modèle des écoles chrétiennes. Les élèves n'y étaient ni notés, ni classés. En 1880 et 1890, deux réglementations ont précisé les modalités d'évaluation du certificat d'études primaire (qui ne concernait que les meilleurs élèves), dans le but d'harmoniser les pratiques au niveau national. Ces textes ont instauré l'utilisation d'échelles sur 10 points pour les différents exercices et pour l'oral. Il n'y était pas question de « note » mais de « chiffre maximum d'appréciation ». La diffusion de ces pratiques aux évaluations en classe est mal connue, mais l'hypothèse avancée par les chercheurs est la même que pour l'enseignement secondaire, à savoir la répercussion dans les classes qui aboutissent à un examen des modalités d'évaluation de cet examen.

Si la note est massivement utilisée pour évaluer les élèves en classe aujourd'hui, il est intéressant d'avoir à l'esprit qu'elle a ses origines dans une logique de concours, c'est-à-dire de classement et de sélection, et que la valeur d'une note a d'abord été relative, correspondant alors à un rang plutôt qu'à une valeur absolue (pour plus de détails sur l'apparition et l'instauration de la notation chiffrée, voir Merle, 2015).

S'attachant à la problématique de la **subjectivité dans l'évaluation scolaire** (matérialisée par la note chiffrée) en tant que source d'inexactitudes aux conséquences potentiellement importantes, le psychologue français Henri Piéron a consacré une grande partie de ses recherches à l'analyse scientifique des examens (Merle, 2018 ; Murat, 1998). À l'aide du physiologiste Henri Laugier, Piéron fonda dans les années 1920 « **la science des examens** », appelée « **docimologie** » (« *dokime* » signifiant « épreuve » en grec et « *logos* » correspondant au discours rationnel).

Cette discipline, dont l'objet d'étude est « *l'organisation des examens, leurs contenus et leurs objectifs pédagogiques* [mais également, l'analyse des] *méthodes de correction des épreuves ainsi que le comportement des examinateurs et des examinés* » (Martin, 2002), avait l'ambition de s'attaquer à la **subjectivité de l'enseignant** dans le cadre de cette activité évaluative.

Les premiers travaux établissent **l'incertitude de la notation individuelle des copies d'examens**. L'expérience réalisée en 1931 par la Commission française pour l'enquête Carnegie a notamment marqué les esprits. En soumettant à la multi-correction (six correcteurs différents) cent copies anonymes ayant fait l'objet d'un écrit à l'examen final du baccalauréat, et provenant de disciplines

différentes, les chercheurs ont constaté **d'importants écarts de notation entre les correcteurs** (détails de l'expérimentation : Merle, 2018).

Après s'être saisis des résultats de l'expérience, Laugier et Weinberg les ont interprétés et analysés. Ils ont relevé des **écarts considérables** dans la globalité des disciplines étudiées, allant de 8 points d'écart en physique jusqu'à 13 points d'écart en composition française (Capelle, 2010). Face à cette observation, les deux chercheurs ont poursuivi leur analyse statistique afin d'estimer quel aurait été le nombre de correcteurs nécessaire pour obtenir une « note vraie » (Merle, 2018). À titre indicatif, ils ont estimé qu'il aurait fallu que 13 évaluateurs en mathématiques et 127 correcteurs en philosophie participent à la multi-correction pour que « *la moyenne des notes soit la plus représentative possible de la valeur du travail du candidat* » (Capelle, 2010 ; Merle, 2018)[1]. Pour Merle, ce calcul ne serait d'ailleurs en réalité que le reflet de l'« **utopie scolaire** » de la recherche d'une « note vraie » (Merle, 2018).

Cette utopie ne vaut pas seulement pour la « **fidélité inter-correcteurs** » (c'est-à-dire entre plusieurs correcteurs différents, comme dans la recherche mentionnée précédemment), mais aussi pour la **fidélité intra-correcteur**, c'est-à-dire pour un seul et même correcteur (Merle, 2018). En effet, en soumettant plusieurs fois les mêmes enseignants à la correction de 37 copies de physiologie d'un certificat d'études supérieures de sciences à des périodes différentes (intervalles de dix mois et trois ans et demi), il apparaît que les correcteurs ne notent pas de la même façon en fonction de la diversité des situations d'évaluation. En effet, d'un contexte évaluatif à l'autre, on observe que les enseignants ne sont pas toujours fidèles à leur précédente évaluation (Merle, 2018). À titre illustratif, un même correcteur sollicité dix mois plus tard pour procéder de nouveau à la correction des copies qui lui avait alors été soumises a été « **assez moyennement fidèle à lui-même,** avec un coefficient de corrélation de 0,81 entre ses deux notations » (Merle, 2018).

[1] Pour parvenir à ces résultats, Laugier et Weinberg proposent une méthode d'optimisation de la corrélation entre les notes proposées par chaque correcteur. Pour plus de détails, voir Laugier (1963), *in* Leclercq, Nicaise & Demeuse (2004).

Encadré 3 : La docimologie à ses débuts

Initialement, les **premières recherches menées sur la fiabilité de l'évaluation sont assez critiques**.

À ses débuts, la docimologie « met en évidence les problèmes sans les résoudre, du moins de manière pratique, au niveau où le problème se pose, c'est-à-dire au niveau des enseignants chargés de procéder à l'évaluation » (Leclercq, Nicaise & Demeuse, 2004).

Avant même de proposer « des pistes d'amélioration des examens et des concours », les grands précurseurs de la docimologie (Piéron, Laugier et Weinberg) vont se concentrer sur **l'analyse scientifique des examens** (Merle, 2018). Cette analyse va leur permettre de formuler plusieurs critiques des examens.

En premier lieu, ils **dénoncent le « caractère arbitraire » des examens traditionnels** : « L'aléa de l'examen tient notamment à la subjectivité de l'examinateur, par exemple au recours à des échelles de notes variables selon le correcteur » (Merle, 2018). Autrement dit, dans le cadre des examens, certains correcteurs ne vont pas utiliser toute l'échelle des notes (en effet certains d'entre eux n'ont jamais recours aux « notes extrêmes ») alors que d'autres correcteurs vont quant à eux se servir d'une échelle de notes bien plus large.

Cet « usage différencié de l'échelle des notes constitue ce que les docimologues [appellent] un **coefficient de subjectivité** » (Merle, 2018). Sont ainsi mises en évidence **les conséquences de la subjectivité du correcteur sur le résultat final**, à savoir la note finale attribuée à l'élève (Martin, 2002).

La deuxième critique formulée par les docimologues concerne la **finalité des examens et des concours** (Merle, 2018). Piéron, Laugier et Weinberg s'accordent sur le fait que les examens et les concours n'ont pas toujours recours aux mêmes objectifs en termes d'évaluation. À ce propos, Martin (2002) souligne qu'« **examens et concours traditionnels mélangent systématiquement deux types d'évaluation** :

- D'une part, il peut s'agir de "contrôler les résultats d'une formation éducative, de vérifier des acquisitions, d'évaluer le bagage de connaissances assimilées et de déterminer si, pour un écolier donné, la tâche d'un enseignement peut être considéré comme achevée" ;
- D'autre part, il peut s'agir de "déterminer les aptitudes propres d'enfants ou de jeunes gens, qui devront bénéficier d'une formation éducative particulière" » (Martin, 2002).

Finalement, cette confusion des deux objectifs poursuivis par l'évaluation peut avoir pour conséquence « **l'affectation des individus à des métiers pour lesquels ils n'ont pas forcément les aptitudes requises** » (Merle, 2018). Ainsi, le problème posé par la docimologie concerne « **la légitimité de la sélection**, c'est-à-dire **des critères sur lesquels s'opère l'affectation sociale des individus** au travers du tamis du système scolaire » (Martin, 2002).

La docimologie nous apprend-elle quelque chose sur les notes ou sur l'évaluation scolaire en général ?

La docimologie a mis en évidence les limites d'**une évaluation utilisant les échelles chiffrées**, dans la mesure où elle a analysé les différences entre les notes attribuées par différents correcteurs à des copies d'examens. Cependant, les enjeux qu'elle soulève tendent à **dépasser « l'outil-note »**. En effet, les travaux docimologiques étudient la variabilité des notes car celles-ci sont l'indicateur final qu'il est possible d'observer et de comparer. Toutefois, la question qui guide ces travaux est bien celle de la fiabilité du jugement évaluatif de l'activité d'un élève (qu'on ne peut pas observer directement) effectuant une tâche donnée.

Par ailleurs, à la suite des travaux de docimologie, d'autres disciplines, notamment la psychologie sociale, se sont à leur tour penchées sur la question de la fiabilité de l'évaluation scolaire. Ces études ont élargi le champ des examens à d'autres situations d'évaluation, plus proches des situations « ordinaires » d'évaluation en classe. Alors qu'au XX^{e} siècle, la docimologie a révélé des variations entre les corrections de plusieurs examinateurs et a suggéré de multiplier les correcteurs pour atténuer ces écarts, les études plus récentes, issues d'autres champs disciplinaires, permettent d'identifier plusieurs variables qui peuvent influencer le jugement évaluatif, et ouvrent ainsi de nouvelles perspectives pour gérer ces biais.

L'ensemble de ces travaux de recherche permettent-ils de trancher en faveur ou en défaveur de la note ?

Sans exclure totalement l'idée que l'outil utilisé pour synthétiser un jugement évaluatif (la « notation chiffrée ») puisse avoir un effet sur ce jugement, on peut se demander si des phénomènes similaires à ceux qui ont été constatés avec l'évaluation numérique/chiffrée seraient observés avec un outil différent (par exemple avec des lettres ou encore des niveaux d'acquisition). Mais il n'existe pas, à notre connaissance, d'études ayant répliqué les mêmes dispositifs avec d'autres outils. **En toute rigueur, les conclusions des travaux de recherche présentés dans ce document ne portent que sur des situations dans lesquelles les notes chiffrées sont utilisées comme outil de synthèse, mais il est impossible d'affirmer que la fiabilité de l'évaluation serait meilleure avec un autre outil de synthèse du jugement évaluatif.**

Ne prétendant pas être exhaustive, cette synthèse se propose de mettre en exergue les **principales sources de variabilité du jugement évaluatif** identifiées par les chercheurs dans le cadre d'évaluations recourant à la notation chiffrée[2]. Ajoutons que nous ne nous intéresserons pas aux évaluations standardisées, qui ont elles-mêmes leurs propres biais et limites (voir par exemple Rocher, 2015). **Elle ne s'inscrit ni dans une démarche de dénonciation de la note, ni dans une démarche de promotion de celle-ci.** Nous verrons d'ailleurs qu'il n'est pas seulement question de la subjectivité du correcteur. En réalité, les recherches ont identifié des facteurs susceptibles d'influencer la performance de l'élève, d'une part, et des facteurs susceptibles d'influencer le jugement évaluatif, d'autre part.

[2] Dans la suite de ce document, les termes « évaluation » et « notation » seront donc utilisés indifféremment.

Pour les présenter, ce document reprendra une **structure tripartite** utilisée par différents auteurs (Dieudonné, Nicaise & Demeuse, 2004 ; Merle, 2018). Celle-ci repose sur trois catégories de variables qui impactent l'évaluation de productions d'élèves :

- les variables qui relèvent du contexte scolaire ;
- les variables qui relèvent de l'élève ;
- les variables qui relèvent de l'enseignant lui-même.

Dans un premier temps, il s'agira de montrer que le contexte scolaire global et les « normes évaluatives » institutionnalisées ont une influence dans le processus d'évaluation des élèves. Dans un deuxième temps, sera présentée une autre source d'erreur, qui réside dans les caractéristiques propres aux élèves. Enfin, dans un troisième et dernier temps, il s'agira de rassembler les travaux qui montrent que les attentes et les représentations personnelles de l'enseignant influencent également la fiabilité de son jugement évaluatif.

Encadré 4 : Connaît-on les effets des classes sans notes ?

À la suite de la mise en place du premier socle commun en 2005, des expérimentations portant sur l'évaluation, et plus particulièrement sur « l'évaluation par compétences » et/ou « les classes sans notes », se sont développées au collège[3] : en 2013, le département de la recherche et du développement, de l'innovation et de l'expérimentation (DRDIE) de la Dgesco observait que ces dispositifs représentaient près du quart des 2 700 actions renseignées dans la base Expérithèque (Dgesco-DRDIE, 2013). Ces expérimentations ont souvent concerné une seule classe (dite « expérimentale »), ou parfois un niveau (en général la 6e), et se sont développées de façon hétérogène en fonction des dynamiques locales (IGEN, 2013 ; Cardie de Nantes, s. d.). Dans un certain nombre de cas, elles ont été encouragées par les académies (Saillot, 2019).

À notre connaissance, il n'existe pas de synthèse des expérimentations de classes sans notes. Plusieurs académies proposent des « bilans » (datant généralement de la première moitié des années 2010) et l'on trouve quelques articles portant sur des recherches-actions menées localement, mais il n'existe pas de données agrégées. Cela s'explique notamment par le fait que les initiatives n'ont pas été coordonnées, si bien que ni l'organisation concrète des classes sans notes ni les dispositifs de suivi ou d'évaluation de ces classes ne sont comparables. Toutefois, à la lecture de plusieurs « bilans », nous proposons de dégager quelques points qui semblent converger :

- **Les préoccupations à l'origine des expérimentations de classes sans notes sont souvent similaires.** La faible motivation des élèves, un climat scolaire dégradé, un stress important lié à l'évaluation et à la peur de l'échec, des difficultés à faire état de progrès dans les acquis des élèves avec les notes… sont autant de constats cités par différents documents (articles de recherches et

[3] Dans ce contexte, « évaluation des compétences » ou « par compétences » et « évaluation sans notes » sont souvent considérées comme synonymes. Cela peut en partie s'expliquer par le fait que le socle commun de connaissances et de *compétences* s'est accompagné d'un outil de restitution des résultats (le « livret personnel de *compétences* ») qui ne fonctionnait pas avec des notes chiffrées mais avec un code binaire (acquis/non acquis) : compétences et absence de note ont donc pu être associées aux yeux de différents acteurs de la communauté éducative (enseignants, familles, etc.). Pourtant, il semble important de souligner que les compétences renvoient à un *objet* d'évaluation, tandis que les notes renvoient à un *outil* d'évaluation (le niveau d'acquisition des premières pourrait tout à fait être exprimé à travers les secondes) (Genelot, 2017).

documents issus d'académies) comme ayant motivé la mise en place de ces classes. Les établissements accueillant un public scolaire en difficulté (notamment en éducation prioritaire) semblent avoir été parmi les premiers à s'essayer aux classes sans notes, notamment car le maintien de l'engagement des élèves dans les apprentissages (en termes de motivation mais aussi de présence en cours) y est un enjeu particulièrement fort (IGEN, 2013 ; Cardie de Créteil, s. d. ; Service Santé et Cardie du Rectorat de Poitiers, s. d. ; Cardie de Nantes, s. d.).

- **La mise en œuvre de classes sans notes demande du temps et des efforts.** Dans les documents consultés, lorsque la mise en œuvre des classes sans notes est évoquée, elle apparaît comme un processus assez long, qui exige un temps important d'explicitation pour y faire adhérer les enseignants, les élèves et les parents d'élèves, ainsi que du temps laissé aux enseignants pour se concerter, travailler ensemble et/ou se former au nouveau système d'évaluation. Fonctionner sans notes oblige à se détacher d'habitudes et de représentations largement ancrées et ne se fait pas sans résistances (Service Santé et Cardie du Rectorat de Poitiers, 2018 ; Cardie de Nantes, s. d.).

- **Les quelques bilans dressés en académie semblent globalement positifs, même s'ils ne concluent pas sur les effets de la suppression des notes sur l'apprentissage des élèves.** En termes de climat scolaire, de développement de l'autonomie et de la responsabilisation des élèves, d'entraide et de recul de l'absentéisme, ces bilans sont positifs. Ces constats convergent en grande partie avec les éléments rassemblés par l'Inspection générale (IGEN, 2013). En revanche, ils reconnaissent ne pas disposer de suffisamment de recul et de données pour observer des effets sur les résultats des élèves. Lorsque des enseignants ont été interrogés (par questionnaire), ils ne semblent pas tous convaincus que la suppression des notes a permis de faire progresser les élèves (les réponses ne convergent pas d'une enquête à l'autre) ; en revanche, l'idée que la suppression des notes améliore le climat scolaire, les comportements, l'entraide, la compréhension des éléments à retravailler, et que cela diminue le stress, la peur de l'échec et la comparaison entre élèves, semble partagée. Il est intéressant de noter que même si les élèves, dans ces enquêtes, relèvent des effets positifs (sur la peur de l'échec par exemple), ils souhaitent majoritairement revenir à un système de notation chiffrée (Cardie de Nantes, s. d. ; Cardie de Créteil, s. d. ; Collège des Bauges (académie de Grenoble), s. d. ; Service Santé et Cardie du Rectorat de Poitiers, s. d.).

- **D'après les études de chercheurs consultées, la suppression des notes aurait peu (voire pas) d'effets, et quand ils existent, ces derniers seraient différenciés selon les élèves.** Une première étude menée auprès d'élèves de CM2 (sans notes), de 6e (avec et sans notes) et de 5e (avec des notes) permet d'observer que l'évaluation est source de stress pour les élèves dès le CM2, alors même que les notes chiffrées ne sont pas utilisées. En revanche, les auteurs constatent que chez les élèves de 6e en difficulté, la peur de l'échec est nettement moins présente dans la classe sans notes que dans celle qui en utilise. Pour ce qui est de la motivation, les chercheurs observent que dans la classe qui utilise des notes, les évaluations motivent essentiellement les élèves en réussite scolaire, tandis qu'elles motivent indistinctement tous les élèves de la 6e sans notes (Bénit & Sarremejane, 2019). D'autres chercheurs se sont intéressés spécifiquement à l'effet d'appartenir à une classe sans notes sur le sentiment d'efficacité personnelle. Les résultats d'une étude menée auprès de 579 élèves indiquent que l'utilisation d'un système de notation alternatif à la note chiffrée n'a pas d'effet sur le sentiment d'efficacité personnelle des filles, mais que les garçons des

classes sans notes ont un meilleur sentiment d'efficacité personnelle que leurs homologues des classes utilisant des notes chiffrées (Cartierre *et al.*, 2016). Enfin, deux chercheurs ont cherché à voir si des différences existaient entre des élèves de 3e dont le collège n'utilise pas de notes chiffrées et des élèves de 3e dont le collège utilise des notes chiffrées en matière de motivation, de sentiment d'efficacité personnelle, de tendance à se comparer aux autres élèves et en termes de résultats scolaires. L'analyse des réponses au questionnaire qu'ils ont fait passer auprès de 800 élèves (issus de 11 collèges dont 6 sans notes) croisée avec l'analyse des résultats de 15 000 élèves au brevet des collèges les amène à conclure qu'il n'y a aucune différence entre les deux cohortes d'élèves. Ce résultat leur permet d'avancer que la suppression des notes ne nuit pas aux résultats scolaires (Goudeau & Autin, s. d.).

Pourquoi est-il difficile de conclure à un effet (positif ou négatif) des classes sans notes sur l'apprentissage des élèves ? On l'a dit, si les expérimentations ont été nombreuses, elles n'ont pas été recensées de manière exhaustive (Lapostolle & Genelot, 2015) et elles ont pu prendre des formes variées qu'il serait délicat de comparer. Notons aussi que la plupart du temps, ces expérimentations ont vu le jour sur la base du volontariat des personnels, ce qui introduit un biais difficile à prendre en compte. Par ailleurs, si l'on constate un intérêt marqué pour le suivi de ces expérimentations au début des années 2010 (notamment au sein des académies), peu de bilans récents et effectués sur un temps long sont disponibles (nous n'en avons pas trouvé). En outre, une hypothèse est soulevée par certains chercheurs : la suppression des notes à un moment ponctuel de la scolarité (uniquement en 6e par exemple), lorsque que la notation chiffrée continue par ailleurs d'être utilisée en dehors des classes expérimentales (dans les autres classes de l'établissement, dans les autres niveaux) et continue d'être une référence pour les élèves, n'est peut-être pas suffisante pour produire les effets positifs attendus (Goudeau & Autin, s.d.). D'ailleurs, même en l'absence de note, il est courant que les élèves (et les enseignants) recourent à des systèmes de conversion pour revenir à des repères plus familiers (IGEN, 2013), ce qui suggère que, même si l'outil est modifié, la « logique » d'évaluation peut rester la même.

Plusieurs textes consultés font état, **à défaut d'effets nets visibles sur les résultats des élèves, d'évolutions observées dans les équipes impliquées dans de tels projets** : ces expérimentations seraient l'occasion pour les équipes pédagogiques d'interroger l'évaluation et notamment ses fonctions, ce à quoi elle doit servir, la façon dont elle est construite par chaque professeur (avec quels critères, et quelle transparence de ces critères pour les élèves), la façon dont ses résultats sont communiqués aux élèves et aux parents et dans quels buts ils le sont, etc. (Saillot, 2019 ; IGEN, 2013 ; Collège des Bauges (académie de Grenoble), s. d.).

I. L'influence du contexte scolaire dans l'exercice de notation des élèves

A. La réputation de l'établissement

Les établissements dans lesquels sont scolarisés les élèves influencent l'estimation du niveau de leurs copies. À ce titre, plusieurs études ont montré qu'un « **effet-établissement** » est attaché au processus d'évaluation des élèves.

- La recherche menée par Noizet et Caverni (1978) illustre cet aspect : un effet d'assimilation établissement-élève est à l'œuvre lorsque l'évaluateur a connaissance de la provenance des copies qu'il corrige (Merle, 2018). En soumettant à 16 enseignants différents 12 copies de sciences naturelles qui ont été fictivement associées à un établissement dont la réputation est variable, les chercheurs ont constaté que « **la provenance des copies, par construction fictive, exerce un effet sur l'évaluation du niveau des copies** », en faveur des copies associées à un établissement favorisé (la différence était faible, de l'ordre de moins d'un point sur 20, mais elle était statistiquement significative) (Merle, 2018).

- L'étude de Duru-Bellat et Mingat (1988) confirme que **le poids de l'établissement entre en jeu lors de la notation des élèves**. En comparant les notes obtenues par un certain nombre d'élèves de collège à des évaluations en classe et à des épreuves standardisées, des écarts entre les deux types d'évaluation apparaissent pour les élèves de certains établissements. En effet, on constate que ceux appartenant à un établissement de plus faible niveau obtiennent des notes plus élevées quand ils sont évalués en classe que lors d'épreuves standardisées, et inversement.

Si elles indiquent toutes deux une influence de la réputation de l'établissement d'un élève sur le jugement évaluatif porté sur sa copie, ces deux études semblent mener à des constats contradictoires : dans un cas, la bonne réputation favorise la copie, dans l'autre, elle la défavorise. Le contexte a toute son importance pour comprendre cela. Dans l'étude de Noizet et Carverni (1978), les correcteurs recherchent une certaine objectivité à travers la notation ponctuelle d'une copie d'élève dans un contexte d'examen (le baccalauréat). Dans l'étude menée par Duru-Bellat et Mingat (1998), le jugement évaluatif a lieu dans le cadre de l'évaluation qui se déroule dans le quotidien de la classe. On peut alors imaginer que l'enjeu pour les correcteurs n'est pas (seulement) la justesse du jugement qu'ils portent sur chaque copie à un moment donné, mais bien l'appréciation globale qu'ils vont se faire, au fil des productions évaluées, et les potentiels usages pédagogiques qu'ils peuvent faire de ce jugement évaluatif pris isolément.

Ces études étant relativement anciennes, il serait intéressant de disposer de résultats plus récents pour savoir si la connaissance de l'établissement d'appartenance d'un élève a toujours un effet, et si oui dans quel sens (les correcteurs sont-ils plus sévères quand ils savent qu'un élève vient d'un « bon » établissement, ou bien est-ce que leurs attentes les amènent à rechercher des traces de réussite dans sa copie ? À l'inverse, sont-ils plus indulgents avec un élève scolarisé dans un établissement dont le niveau est réputé faible, ou sont-ils plus enclins à lui attribuer de mauvaises notes ?). Il serait également intéressant de comparer ces effets selon que le contexte est celui d'une évaluation à fort enjeu (un examen par exemple) ou à faible enjeu (évaluation en classe).

B. Le cadre de la classe

Les recherches scientifiques en docimologie se sont également attachées à démontrer que la classe dans laquelle se trouve l'élève influence l'appréciation de son travail. En effet, les productions d'un élève ne seront pas évaluées de la même manière en fonction du niveau global de sa classe.

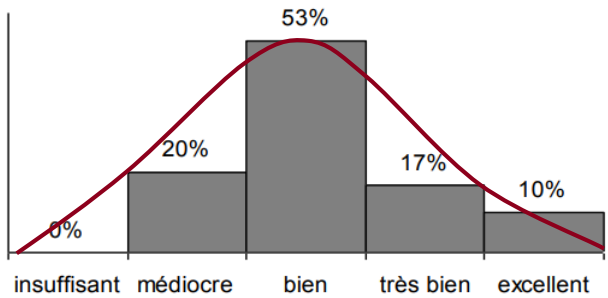
Si le processus de notation varie en fonction du niveau des autres élèves, il est également conditionné par ce que Merle appelle des « **normes d'évaluation** » (Merle, 2018). On peut donc distinguer plusieurs phénomènes :

- **La « loi de Posthumus »** : « un enseignant tend à ajuster le niveau de son enseignement et ses appréciations des performances des élèves de façon à conserver, d'année en année, approximativement la même distribution (gaussienne) de notes » (De Landsheere, 1992 *in* Merle, 2018).

Ce constat a été popularisé par Antibi sous le nom de « constante macabre »[4] pour souligner qu'il serait toujours produit un **pourcentage d'élèves en échec** au sein de chaque classe (Antibi & Luciani, 2003). Autrement dit, les enseignants ajusteraient leur notation (consciemment ou non) pour retrouver cette distribution et obtenir une moyenne située plus ou moins autour de 10/20 (moyenne qui marque le seuil entre échec et réussite). Cette adaptation du niveau de difficulté de l'évaluation et/ou du niveau de sévérité de la correction rend moins visibles les différences de niveau qui existent d'une classe à une autre. Cela signifie aussi que les résultats des évaluations ne rendent pas forcément compte du niveau réellement atteint par les élèves au regard des attendus des programmes officiels (Merle, 2018).

À titre illustratif, **l'expérience de Gjorgjevski** met également en lumière ce phénomène. En soumettant plusieurs fois un ensemble de copies à des correcteurs différents, le chercheur constate une **répartition des notes sous la forme d'une courbe dont la forme rappelle celle d'une courbe gaussienne** (Rot & Butas, 1959, *in* Leclercq, Nicaise & Demeuse, 2004).

Graphique 1 : Répartition des appréciations des performances des élèves



Source : Leclercq, Nicaise, Demeuse, 2004.
Note : Nous avons ajouté la courbe « en cloche » (rouge) sur le graphique initial afin de mettre en évidence une répartition des notes qui rappelle une distribution gaussienne.

[4] Rejoignant l'idée de la répartition des notes au sein d'une classe sous la forme d'une courbe gaussienne (une « courbe en cloche »), cet effet suppose que les élèves sont nécessairement divisés en trois groupes distincts : les « bons » élèves, les « moyens » élèves (constituant la majorité d'entre eux), et les « mauvais » élèves.

La notation des élèves est affectée par le niveau « réputé » de la classe dans laquelle se situe un élève :

- **L'expérience** menée par Bonniol, Caverni et Noizet (1972, cités par Merle, 2018) met en évidence cet effet. Après avoir fictivement inscrit sur des copies une information concernant le niveau de la classe dans laquelle se trouve l'élève évalué, les chercheurs font le constat suivant : **les classes les mieux réputées sont celles qui obtiennent les meilleures notes** (pour plus de détails sur l'étude, voir Merle, 2018). Ainsi, le jugement évaluatif est influencé par la connaissance qu'a le correcteur du niveau de la classe de l'élève.

Encadré 5 : Effet classe, effet-établissement : un phénomène d'assimilation

L'influence de la classe et de l'établissement dans le processus d'évaluation des élèves renvoie à ce que certains auteurs appellent un « **effet d'assimilation** » (Noizet et Caverni, 1978). La connaissance du **niveau de la classe** et celle de la **réputation de l'établissement** dans lesquels se trouve l'élève sont autant **d'éléments intégrés par l'enseignant lorsqu'il évalue**. D'une manière volontaire ou non, l'enseignant peut être amené à situer la production d'un élève sur une échelle en partie déterminée par le niveau général de la classe, notamment par l'ensemble des productions des autres élèves ; il peut également ajuster son appréciation du travail d'un élève en fonction de la réputation de l'établissement.

L'« **effet d'assimilation** » est également identifié dans le **cadre du traitement anonyme de copies**. Lors de la correction d'une copie, l'enseignant peut en effet faire le lien entre la production qu'il évalue et les précédentes notes attribuées à l'élève. Lors d'une expérience évaluative au cours de laquelle plusieurs copies ont été soumises à une multi-correction, il a été montré que lorsque le second correcteur a connaissance de la note attribuée par le premier correcteur à la même copie, il a tendance à intégrer la première appréciation de la copie dans son jugement professoral (Merle, 2018) (détails de cette expérience, voir II.A.).

C. Le traitement des copies des élèves

Entre le début et la fin de l'exercice de notation, les attentes du professeur peuvent être amenées à évoluer. Malgré l'utilisation d'un barème initial, l'enseignant peut tout de même être contraint de **revoir ses exigences** à la hausse ou à la baisse, en fonction de l'ensemble des copies qu'il corrige. En d'autres termes, il **s'adapte nécessairement au niveau global de ses élèves**.

Cet « **effet de contraste** » entre les différentes copies, mis en évidence par divers chercheurs (voir par exemple Bonniol, 1965 ; De Landsheere, 1992 cités par Merle, 2018) impacte fortement l'appréciation de chacune d'entre elles et **limite** ainsi **la fiabilité de la note qu'obtiendront les élèves**. Il a été démontré qu'une copie corrigée après une autre jugée « excellente », aura tendance à être sous-évaluée, et inversement.

En s'intéressant à **l'ordre de correction des copies**, les chercheurs suggèrent également que la place occupée par chacune d'elles au sein d'un paquet influence la note attribuée.

- En 1965, Bonniol fait l'expérience suivante : il soumet à 18 correcteurs un ensemble de mêmes copies. La première partie des correcteurs a corrigé les copies dans un certain ordre ; la seconde, dans l'ordre inverse. Le résultat est le suivant : « **chaque copie est notée différemment selon sa position à l'égard des autres copies** » (Merle, 1998, 2018).

Cet effet, qui influencerait l'enseignant dans sa correction, peut être lié à différents phénomènes. Si l'on exclut volontairement les causes attachées à l'enseignant lui-même (concentration, fatigue, volonté de sur-notation ou sous-notation, etc.) et celles liées à l'élève (caractéristiques scolaires et sociales), causes qui seront développées par la suite, il apparaît que l'ordre de correction des copies relève d'un **effet d'ordre**, propre au contexte évaluatif, qui peut venir s'ajouter à l'« effet d'assimilation » mentionné ci-dessus.

Encadré 6 : Les questions à choix multiples (QCM) permettent-elles une évaluation objective ?

La standardisation des évaluations, c'est-à-dire le fait d'évaluer des élèves avec le même sujet (c'est-à-dire avec les mêmes consignes) et dans les mêmes conditions de passation (comme c'est le cas dans un examen national, par exemple), a pour objectif de garantir l'objectivité de l'évaluation. Or, les travaux de docimologie ont montré que cela n'était pas suffisant, puisque des écarts apparaissent au moment de la correction. L'utilisation de questions à choix multiples (QCM) peut sembler constituer une réponse à ce constat, car elles permettent de supprimer les aléas dans la correction : dans la mesure où les réponses attendues sont définies en amont, il ne peut pas y avoir de différence entre deux correcteurs, ni entre deux corrections espacées dans le temps réalisées par un même correcteur. Pourtant, les QCM soulèvent d'autres problématiques. Citons-en quelques-unes :

- Tout d'abord, si la correction de QCM peut être standardisée, l'étape de conception de QCM fait clairement intervenir l'évaluateur et aucun de ses choix n'est neutre : formulation de la question, options de réponse, distracteurs (réponses fausses faites pour induire en erreur), etc. Il est donc illusoire de penser que l'évaluation par QCM est plus objective. Autrement dit, « si la fidélité de correction est assurée, la validité des QCM n'en est pas assurée pour autant » (Hadji, 1992). Pour concevoir des QCM satisfaisantes de ce point de vue, il faudrait vérifier leur validité d'un point de vue didactique *a priori* (voir par exemple Grapin & Sayac, 2017) et réaliser des « pré-tests » suivis d'analyses des réponses pour s'assurer de leur qualité psychométrique (voir par exemple Braibant *et al.*, 2014).

- Par ailleurs, les QCM sont plus ou moins adaptées en fonction de ce que l'on souhaite évaluer :

    - Les QCM ne peuvent concerner que des questions pour lesquelles il n'y a pas d'ambiguïté dans la réponse et ne sont donc pas pertinentes pour tous les objets d'évaluation. On n'utilisera pas de QCM, par exemple, si l'on attend que des élèves argumentent un point de vue (Leclerq, 1986).

    - Les QCM ne permettent pas d'évaluer la façon dont un élève aurait formulé une réponse, puisqu'il choisit seulement parmi des options déjà proposées (en droit, par exemple, on peut estimer qu'il est important d'évaluer la façon dont des étudiants expriment leurs réponses ; voir Huang, 2017). Cela dit, ce peut être un moyen de se concentrer sur la vérification de connaissances en évitant que certains élèves soient justement pénalisés par leur expression écrite (Leclerq, 1986).

- Les QCM peuvent aussi induire des comportements qui biaisent les résultats des élèves :
    - Un élève peut répondre correctement en sélectionnant une option de réponse au hasard (ce qui n'est pas le cas avec des questions ouvertes) (Leclerq, 1986 ; Newble *et al*., 1979 ; Heck & Stout, 1998). Une abondante littérature (notamment anglophone) existe sur la question des QCM à points négatifs, qui visent à corriger ce risque.
    - Le fait que des options de réponses soient proposées peut mettre les élèves sur la piste et amener à un taux de réussite plus important que si les questions avaient été ouvertes. Autrement dit, il peut y avoir un risque de surestimer les connaissances ou compétences des élèves (Newble *et al.*, 1979 ; Heck & Stout, 1998).
    - Le fait que des options de réponses soient proposées peut aussi transformer la nature de l'activité. En mathématiques, par exemple, on peut tenter de résoudre le problème (ou le calcul) qui est posé et sélectionner la réponse qui correspond au résultat que l'on obtient, mais on peut aussi tester les réponses proposées et sélectionner celle qui permet de valider la question (c'est le cas si la QCM porte sur un programme de calcul, une opération à trou, une équation, la conversion d'une fraction en un nombre décimal, etc.). Ainsi, une QCM peut évaluer une autre activité que celle qui était visée initialement.

Notons que toutes les études ayant comparé les résultats d'élèves (ou le plus souvent, d'étudiants) à deux tests équivalents (l'un utilisant des QCM et l'autre utilisant des questions ouvertes, mais pour évaluer les mêmes choses), n'arrivent pas à des conclusions identiques : dans certains cas, les élèves « sur-performent » avec les QCM, dans d'autres, les performances sont similaires. Toutefois, la plupart des auteurs de ces études invitent à varier les modalités d'évaluation, considérant que les QCM ne peuvent pas constituer l'unique façon d'évaluer l'acquisition de connaissances ou la maîtrise de compétences. Enfin, certaines études attirent l'attention sur le risque pour les élèves de mémoriser de fausses réponses après y avoir été exposés dans des QCM, ce qui pose question si l'on aborde l'évaluation dans la perspective de soutenir l'apprentissage des élèves (Leclerq, 1986).

D. Autres facteurs de différenciation (discipline, exercice et stéréotype)

Les travaux de recherche sont parvenus à mettre en évidence un certain nombre de facteurs complémentaires qui exercent eux aussi une influence lors du processus d'évaluation des élèves.

Des travaux ont par exemple cherché à savoir si les écarts de notation entre les élèves variaient d'une discipline à une autre. Les premières études (Laugier et Weinberg, 1936) ont relevé une variabilité du jugement évaluatif entre différents correcteurs quelle que soit la discipline (en tout cas, aucune des disciplines qu'ils ont étudiées n'échappe à ce constat), avec des ampleurs toutefois différentes d'une discipline à l'autre. D'autres recherches plus récentes ont quant à elles mis en évidence de réelles différences dans les résultats des élèves à une même tâche, selon la manière dont on la présente comme relevant d'une discipline ou d'une autre.

- En ce sens, les chercheurs se sont aperçus que lorsque les enseignants présentent de façon différente une même tâche à leurs élèves, les performances de ces derniers varient (Huguet, Burnot & Monteil, 2001). En proposant à des garçons âgés de 10 à 15 ans de réaliser une tâche de mémorisation et de reproduction d'une figure complexe pouvant aussi bien mobiliser leurs capacités en géométrie qu'en dessin, ils se rendent compte que le choix de la présentation de

l'exercice exerce une influence sur les performances des élèves. Il s'avère que les élèves qui ont un faible niveau en mathématiques, et plus précisément en géométrie, obtiennent de moins bons résultats lorsque le professeur présente la tâche comme visant à mesurer leurs compétences en géométrie ; de leur côté, les élèves ayant un bon niveau en géométrie réussissent mieux dans cette condition. Notons que les performances de tous les élèves sont équivalentes lorsque la tâche est présentée comme relevant du domaine du dessin. Ces résultats s'expliqueraient par le fait que la situation d'évaluation renvoie l'élève à l'image qu'il se fait de lui-même à partir de ses expériences passées : **la performance de l'élève est inhibée par le fait que la tâche lui est présentée comme relevant d'un domaine dans lequel il a déjà échoué ou rencontré des difficultés**. Plus largement, les chercheurs ont constaté **que les élèves sous-performaient lorsqu'ils étaient face à des incohérences entre leurs antécédents scolaires et une situation d'évaluation qui rend visible leur performance auprès d'autres élèves**. Dans l'étude de Monteil (Monteil, 1998, 1991 *in* Huguet *et al.*, 2001), le dispositif consistait à faire publiquement (devant les autres élèves) un feedback positif ou négatif à des élèves suite à une tâche liée à une nouvelle leçon. Les élèves s'attendaient ensuite à être interrogés pendant la leçon (contexte de visibilité sociale forte), ou non (contexte de visibilité sociale faible). Dans les deux cas, les élèves réalisaient à nouveau une tâche liée à la leçon à l'issue de celle-ci. Les résultats sont présentés dans le Tableau 1 ci-dessous. Les incohérences entre le passé scolaire de l'élève d'une part et le feedback reçu à l'approche d'une situation évaluative (en contexte de forte visibilité sociale) d'autre part augmenteraient l'attention portée par l'élève à lui-même, au détriment du traitement des stimuli pertinents pour la réalisation de la tâche et donc de la performance. Ce phénomène est appelé *idiosyncratic effects* dans la littérature anglophone.

Tableau 1 : Résultats de l'étude de Monteil (1998, 1991) d'après Huguet *et al.* (2001)

Niveau de l'élève en mathématiques	Feedback reçu suite à la tâche préliminaire	Performance à la tâche réalisée après la leçon selon le contexte de visibilité sociale
Faible	Négatif (cohérent avec les antécédents scolaires)	Pas de différence de performance, quel que soit le contexte de visibilité sociale
Faible	Positif	Meilleure si le contexte de visibilité sociale est faible
Fort	Négatif	Meilleure si le contexte de visibilité sociale est faible
Fort	Positif (cohérent avec les antécédents scolaires)	Meilleure si le contexte de visibilité sociale est fort

Un autre phénomène peut venir altérer les performances des élèves en fonction de la présentation de la situation évaluative : **la menace du stéréotype**. Amplement développé par Steele et Aronson à la fin du XX^{e} siècle (1995), ce phénomène correspond à la stigmatisation d'une personne en raison de son appartenance à un groupe ou à une catégorie spécifique. Autrement dit, la situation évaluative active un stéréotype que les élèves ont sur eux-mêmes (par exemple, « les filles sont moins bonnes en mathématiques que les garçons, or je suis une fille »), ce qui va les faire sous-performer par rapport à ce dont ils auraient été capables si ce stéréotype n'avait pas été activé.

- La **menace du stéréotype** a été pour la première fois mise en avant lors de l'étude menée en 1995 par Steele et Aronson. L'une de leurs expériences a consisté à faire varier la présentation d'une tâche à un échantillon d'étudiants composé de personnes blanches et de personnes noires. En présentant à une partie des étudiants la tâche comme relevant des capacités intellectuelles (situation évaluative) et à l'autre partie comme étant une tâche de résolution de problèmes (situation non-évaluative), les chercheurs remarquent que les étudiants noirs sont plus sensibles à la présentation de la tâche. Ayant conscience des stéréotypes sociaux relatifs à la population noire américaine, **ces étudiants réussissent moins bien lorsqu'on leur présente une tâche en contexte évaluatif**.
  Ainsi, ce n'est qu'au nom de cette menace du stéréotype et de la présentation de la situation évaluative que les performances de ces étudiants sont détériorées (Steele & Aronson, 1995).

Dans le cadre scolaire, des **stéréotypes de genre** ont également pu être identifiés. Ces derniers proviennent des représentations que l'on se fait du niveau scolaire, des caractéristiques et des performances des filles et des garçons.

- Huguet et Régner (2007) ont repris le dispositif de l'étude comparant les performances de différents élèves à une tâche selon sa présentation comme relevant du dessin ou de la géométrie (Huguet, Burnot & Monteil, 2001) pour tester l'hypothèse d'une menace du stéréotype lié au genre. En présentant à des élèves (filles et garçons) âgés de 11 à 13 ans l'exercice de mémorisation et de restitution d'une figure complexe, ils ont observé que les filles réussissent moins bien que les garçons lorsque la tâche qu'elles doivent effectuer leur est présentée comme relevant de la géométrie, tandis qu'elles réussissent mieux que les garçons dans la condition « dessin » (Huguet & Régner, 2007)[5]. La moindre réussite des filles dans un contexte menaçant s'expliquerait par la charge cognitive supplémentaire générée par cette menace (pensées parasites, impact de ces pensées sur la mémoire de travail et sur les ressources cognitives disponibles pour réaliser la tâche).
  Cette étude confirme les conclusions d'études antérieures et a le mérite d'en étendre la portée à un contexte plus proche d'une situation ordinaire de classe (dans les études précédentes, les élèves réalisaient la tâche en étant isolés, tandis qu'ils la réalisent ici au sein d'un groupe-classe). En outre, les chercheurs font également varier le niveau de pression évaluative. Ils observent que la menace du stéréotype joue même lorsque l'enjeu évaluatif n'est qu'implicitement suggéré (c'est-à-dire dès lors que les filles considèrent que l'exercice mesure leur niveau en mathématiques). Par ailleurs, cette étude teste aussi l'impact de la composition du groupe d'élèves au sein duquel la tâche est réalisée : lorsque le groupe est uniquement composé de filles, l'effet de la menace du stéréotype disparaît. Cela ne signifie pas que c'est la présence d'élèves du sexe opposé qui crée le stéréotype (puisque d'autres

[5] Les chercheurs ont également cherché à comprendre comment la menace du stéréotype pouvait interagir (voire se cumuler) avec l'effet des antécédents scolaires (Régner *et al*., 2016). D'après leurs résultats, les deux phénomènes peuvent coexister, sans nécessairement se cumuler. Toujours avec le même dispositif (mémorisation et reproduction d'une figure complexe en indiquant que la tâche mesure des capacités en géométrie ou en dessin), ils observent ainsi que la menace du stéréotype a diminué la performance des filles ayant un bon niveau en géométrie (alors que dans la condition « dessin », les filles et les garçons performent de façon équivalente), tandis que les antécédents scolaires ont affecté la performance des garçons ayant un faible niveau en géométrie (alors que dans la condition « dessin », les élèves performent de façon équivalente quel que soit leur niveau en géométrie).

études l'avaient observé dans un contexte de passation individuelle). L'explication résiderait plutôt dans le fait qu'en condition non mixte, les filles peuvent s'identifier plus facilement à des modèles de réussite féminin (en condition mixte, elles ont tendance à citer plutôt des garçons parmi les « bons élèves »)[6].

- Dans le prolongement de ces résultats, l'étude menée par Bagès, Martinot et Toczek (2008) est parvenue à mettre en évidence la **réduction de la menace du stéréotype**, lorsque les filles ont connaissance d'un modèle féminin dans la discipline dans laquelle elles sont évaluées.
  En l'espèce, juste avant la réalisation d'une tâche mathématique, **l'enseignant a parlé à la classe d'une grande mathématicienne**. Ainsi, en raison de la présentation aux élèves de la **réussite d'un modèle féminin**, il apparaît que la « différence liée au sexe […] est minimisée lorsque les filles sont en présence d'une femme expliquant sa réussite par ses efforts » (Bagès *et al.*, 2008).

Ainsi, la présentation d'une tâche évaluative peut influencer les performances des élèves.

Cependant, à partir d'une revue systématique de littérature, Flore et Wicherts (2015) appellent à la prudence concernant les résultats sur la menace du stéréotype de genre. En effet, les auteurs suggèrent l'existence d'un biais de publication à propos des recherches portant sur cet effet chez les enfants et adolescents ; autrement dit, les études concluant à l'existence de la menace du stéréotype de genre sont plus susceptibles d'être publiées dans des revues scientifiques que celles concluant à l'absence de manifestation de cet effet.

Mais on peut toutefois faire l'hypothèse que les caractéristiques propres aux élèves sont autant d'informations qui peuvent par ailleurs exercer une influence sur le jugement professoral dans le cadre du processus évaluatif.

[6] Les chercheurs expliquent dans leur article pourquoi la constitution de groupes non mixtes ne constituerait pas une solution, malgré ces résultats : la menace du stéréotype serait diminuée en condition d'évaluation mais elle aggraverait potentiellement les stéréotypes dans les autres phases ; d'autres résultats ont montré que même dans un contexte de passation individuelle, la menace du stéréotype peut produire des effets ; d'autres dispositifs peuvent diminuer ce phénomène sans passer par la constitution de groupes non mixtes (par exemple, en expliquant le principe de la menace du stéréotype aux élèves ; en favorisant une vision incrémentale (*versus* fixiste) de l'intelligence, etc.).

II. L'influence des caractéristiques propres aux élèves sur le jugement évaluatif

Alors que les premières enquêtes en docimologie ont essentiellement porté sur le contexte scolaire et l'ordre de correction des copies, les études menées dans les années 1970 ont cherché à savoir si la connaissance d'informations scolaires et/ou extra-scolaires spécifiques à chaque élève, pouvait influencer le jugement évaluatif.

- Analysant la psychologie de l'évaluation scolaire, Caverni et Noizet parviennent au constat suivant : « **l'évaluateur ne note jamais une copie en soi** » (Merle, 1998). En d'autres termes, la notation par l'enseignant est biaisée par les informations qu'il possède sur l'élève qu'il évalue.

A. La connaissance d'informations scolaires relatives à l'élève

Les informations scolaires relatives à l'élève font globalement référence au niveau qui lui est imputé en fonction de son parcours scolaire, de ses précédentes notes, de son éventuel retard ou redoublement.

- Dans le cadre de leur recherche**,** Caverni, Fabre et Noizet (1975) ont fait l'expérience suivante : après avoir inscrit aléatoirement un score sur différentes copies d'élèves, ces mêmes copies ont été soumises à de nouveaux correcteurs pour qu'à leur tour ils les évaluent. Le **second processus de notation a été influencé par le premier**. Pensant avoir eu connaissance du niveau scolaire de l'élève qu'ils évaluaient (alors même que celui-ci était fictif), les **évaluateurs** ont assimilé dans leur jugement les **notes précédemment attribuées** : ils ont tendance à attribuer de meilleures notes à des copies dont les notes aléatoires étaient élevées et de moins bonnes notes à celles dont les notes aléatoires étaient faibles.
- Les élèves seraient donc en fait « **victimes de leur niveau scolaire »** (Duru-Bellat et Mingat, 1993). Ainsi, une « **contagion des résultats** » antérieurs des élèves est observable (Crahay, Mottier Lopez & Marcoux, 2019). À ce propos, De Landsheere avait déjà constaté qu'« un premier travail médiocre incline à penser que le second le sera aussi ; si cela se vérifie, la tendance à accorder une note médiocre au troisième s'accroît encore, et ainsi de suite » (De Landsheere, 1980, *in* Crahay, Mottier Lopez, Marcoux, 2019).

B. La connaissance d'informations sociales et ethniques relatives à l'élève

En plus de l'influence du niveau scolaire dans le processus de notation, différents travaux ont constaté une **corrélation entre la note et le milieu social de l'élève** (Moinet, 2018).

1. L'influence de l'origine sociale

- L'étude menée par Pourtois révèle l'influence que l'origine sociale peut avoir dans la notation des élèves. En indiquant sur la copie d'un élève une information relative à son milieu social, le chercheur constate que certains correcteurs ont tendance à attribuer de **meilleures notes aux**

élèves issus des milieux sociaux les plus favorisés (Pourtois *et al*., 1978 *in* Leclercq, Nicaise & Demeuse, 2004).

- Face à un tel constat, Merle (2007) consacre d'ailleurs une partie de ses réflexions aux **fiches de renseignement** qui sont parfois demandées aux élèves en début d'année scolaire (pour plus de détails, voir Merle, 2007). Parce qu'y figurent des éléments qui permettent à l'enseignant d'avoir connaissance du milieu social de l'élève, le jugement professoral en devient, consciemment ou non, biaisé. En d'autres termes, **ces fiches peuvent contribuer à la construction d'attentes différenciées selon les élèves**.

L'environnement familial des élèves semble également avoir un effet sur l'enseignant lors du processus évaluatif.

De fait, lors de l'appréciation d'une copie, les attentes d'un enseignant sont influencées par diverses variables liées au milieu d'origine (Duncan, 1967, *in* Moinet, 2018), relevant aussi bien du leur que de celui de l'élève. Le jugement professoral peut alors varier en fonction de **la convergence ou de la divergence des deux milieux sociaux**. Ainsi, chacune de ces situations peut être aussi bien **favorable** que **défavorable** pour l'élève :

- *a priori*, elle lui est **favorable** lorsque ce dernier appartient au même milieu social que son professeur (convergence des milieux) ;
- au contraire, elle lui est **défavorable** lorsque l'élève et le professeur appartiennent à un milieu social différent (divergence des milieux).

2. L'influence de l'origine ethnique

- En soumettant à différents correcteurs des copies d'élèves auxquelles ont été fictivement attribués des noms de famille différents (consonance française ou étrangère), les chercheurs ne constatent pas d'emblée de différence significative. Toutefois, lorsque ces derniers observent cette fois-ci le **patronyme des enseignants**, ils se rendent compte que ceux d'origine française ont tendance « à surévaluer les copies dont l'auteur fictif a un patronyme d'origine étrangère » (et *vice versa*) (Amigues, Bonniol & Caverni, 1975, *in* Merle, 2018).
  Face à un tel résultat, Merle (2018) propose la justification suivante : cet ajustement de notation résulterait d'un « processus de compensation de préjugés défavorables dont l'évaluateur peut avoir conscience ». Autrement dit, on observe que certains enseignants ont des **comportements qui visent à compenser des biais** dont ils ont conscience. Cela confirme que **l'information, détenue par un enseignant sur un élève, est susceptible d'influencer le processus d'évaluation, dans un sens ou dans un autre**.

- Une étude plus récente produit également des résultats indiquant une influence de l'information concernant l'origine ethnique des élèves sur l'évaluation de leurs copies (Sprietsma, 2013, *in* Autin, Batruch & Butera, 2019). Après avoir attribué de manière aléatoire des noms typiques allemands ou turcs à des copies de dissertation, il a été demandé à des enseignants allemands de noter ces rédactions d'élèves, sur lesquelles ces nouveaux noms étaient indiqués.

Les résultats démontrent que **les élèves considérés comme autochtones obtiennent des notes plus élevées que leurs camarades présupposés issus de l'immigration**. Autrement dit, les enseignants ont attribué des notes plus basses dès lors qu'ils pensaient que les élèves étaient issus de l'immigration.

À ce propos, Autin, Batruch et Butera (2019) précisent dans leur analyse que ces résultats ne sont autres que le **reflet de préjugés internes à notre société**. Dans la mesure où les élèves issus de l'immigration sont généralement défavorisés sur le plan socio-économique par rapport aux autres élèves, alors cela apparaît comme une discrimination dans la notation mais ce n'est en réalité qu'un « **effet des préjugés des enseignants** ».

C. L'influence des caractéristiques individuelles de l'élève

À côté des caractéristiques scolaires et sociales des élèves, d'autres recherches ont établi que **certains traits personnels des élèves influençaient également l'enseignant lors de l'évaluation**. Ainsi, le genre, le comportement et l'apparence physique des élèves ont été caractérisés comme des variables pouvant biaiser la notation de leurs travaux.

1. Le genre

- Pour confirmer les biais de genre dans la notation, Terrier (2014) rend compte d'une analyse effectuée au sein de l'académie de Créteil qui a consisté à **comparer les notes attribuées à 4 519 élèves en classe de sixième, de façon anonyme et non anonyme.** En comparant « l'écart moyen entre la note non anonyme et la note anonyme pour les filles et ce même écart pour les garçons », la chercheuse met en évidence des **écarts de notation**. Même s'il apparaît que les divergences dépendent des disciplines étudiées, un **biais de genre dit « positif » en faveur des filles est constaté.** Ainsi, « **à notes anonymes égales, les notes des enseignants sont en moyenne 6,2 % plus élevées pour les filles que pour les garçons en mathématiques** » (Terrier, 2014).
- Une étude menée par Duru-Bellat et Mingat avait déjà abouti à des résultats analogues. Après avoir comparé les notes obtenues en mathématiques et en français par différents élèves à des épreuves standardisées avec celles qui leur ont été attribuées par leur enseignant dans le cadre d'une évaluation en classe, les deux chercheurs font le constat suivant **: les filles sont globalement mieux notées que les garçons par leurs professeurs** (Duru-Bellat & Mingat, 1993).

2. Le comportement

- De la même façon que les enseignants peuvent avoir des attentes de réussites différenciées selon les informations dont ils disposent sur leurs élèves, ils peuvent aussi développer des attentes comportementales différentes selon les caractéristiques des élèves. Réputées comme étant plus calmes et plus studieuses, les **filles font ainsi l'objet d'un biais positif de notation**. En raison d'un comportement souvent jugé « meilleur », elles se rapprochent du « comportement idéalement attendu par le professeur » (Felouzis, 1993 ; Duru-Bellat, 1995, *in* Leclercq, Nicaise, Demeuse, 2004).

- Or, l'influence du comportement de l'élève dans le processus de notation avait déjà été mise en évidence à travers l'étude de Bennet *et al.* (1993) menée aux États-Unis auprès d'élèves scolarisés en fin d'école maternelle et début d'école élémentaire. Ainsi, il s'avère que « **plus le comportement en classe est apprécié par le professeur, meilleur est le jugement scolaire que celui-ci porte sur l'élève** » (Bennet *et al*, 1993, *in* Merle, 2018).

3. L'apparence physique

- Dans de nombreuses situations évaluatives, les chercheurs ont révélé que les **élèves jugés « plus beaux »** et/ou « **plus proches des idéaux médiatiques** » par leurs enseignants ont tendance à être **mieux notés** que leurs camarades (Leyens & Yzerbit, 1997 ; Merle, 1998).

- Cet **effet de l'apparence physique** a également été le résultat **de l'expérience menée par Nilson et Nias (1977)** auprès d'élèves scolarisés en France : en associant au hasard une photo d'élève plus ou moins attrayante à un livret scolaire, il en ressort que ceux associés à un « visage plus agréable » obtiennent une notation plus favorable (Nilson & Nias, 1977, *in* Merle 1998).

III. L'influence des représentations et des attentes personnelles des enseignants comme évaluateurs

Après avoir porté une attention particulière sur l'influence des caractéristiques attachées à chaque élève dans le processus évaluatif, il s'agit désormais d'observer **les effets** des différentes représentations qu'un enseignant se fait d'un élève.

Étant très libres dans le processus d'évaluation en classe, tous les enseignants ne notent pas leurs élèves selon les mêmes critères et les mêmes exigences. Dans le cadre des évaluations dans le quotidien de la classe, les enseignants peuvent alors adapter en partie leur notation au regard de leurs propres représentations et attentes personnelles.

Les recherches qui s'intéressent aux limites et aux biais du processus d'évaluation des élèves sont parvenues à identifier à la fois une **différenciation évaluative intra et inter-correcteurs** :

- La **première situation** désigne les différences de notation qui sont le fait d'un seul et même correcteur. Cela signifie qu'un même enseignant peut évaluer différemment selon les situations ;
- La **seconde situation** désigne, quant à elle, la variabilité de la notation entre plusieurs correcteurs. Du fait de la divergence des attentes et exigences d'un correcteur à l'autre, ces derniers peuvent évaluer différemment des élèves, tout en étant dans la même situation.

A. Les attentes des enseignants : quelques grands effets identifiés par la recherche

Différentes recherches sont également parvenues à identifier **plusieurs effets** qui reflètent la divergence des attentes des enseignants quant aux prestations de leurs élèves.

- **L'effet Matthieu** : initialement conceptualisé par Robert King Merton en 1968 pour désigner un phénomène de domination (entre scientifiques renommés et leurs pairs moins réputés), cet effet a par la suite été étendu à d'autres situations, notamment au contexte éducatif.
  Dans le cadre scolaire, l'effet Matthieu désigne le fait **que les représentations d'un enseignant peuvent le conduire à agir de façon à renforcer les écarts de performance entre les élèves**. En d'autres termes, les enseignants auront tendance, consciemment ou non, à stimuler un élève qu'ils perçoivent comme un bon élève, et à l'inverse, le faire moins avec un élève d'un niveau inférieur : les pratiques enseignantes viendraient ainsi amplifier cet effet Matthieu. Une même situation peut ainsi profiter à certains et pas à d'autres, notamment dans une situation évaluative.

- **L'effet Pygmalion**, ou les **« prophéties autoréalisatrices positives »** : révélé par l'étude réalisée par Rosenthal et Jacobson (1968, cités par Leclerq, Nicaise & Demeuse, 2004), ce résultat suggère que la **notation des élèves est corrélée aux attentes des enseignants**. À la suite de la création artificielle d'attentes chez les professeurs lors d'un exercice de notation des élèves, il en ressortirait un véritable changement du comportement des élèves. Ceux qui ont été associés à une **attente** dite « **positive** » obtiennent de **meilleurs résultats scolaires** que leurs camarades. Autrement dit, cela signifie que, dans une sorte de « spirale vertueuse »,

les élèves s'adaptent aux attentes de leur enseignant pour les satisfaire au mieux et ainsi obtenir de meilleures notes.

- **L'effet Golem** ou les « **prophéties autoréalisatrices négatives** » : à l'inverse de l'effet Pygmalion, l'effet Golem traduit le fait que les attentes des professeurs peuvent également avoir une **influence néfaste** sur les élèves. On se trouve alors dans des situations dans lesquelles « les élèves ont intériorisé leurs faibles capacités scolaires du fait des jugements dévalorisants de leur enseignant » (Gauthier, 2020).

Ces différents effets ne se manifestent pas systématiquement, et leur portée est sujette à discussion au sein de la communauté scientifique, y compris par Rosenthal lui-même (Rosenthal, 1987). Certains auteurs ont notamment échoué à répliquer les effets Pygmalion et Golem, comme par exemple, Trouilloud *et al*. (2002).

B. Une différenciation évaluative liée aux caractéristiques propres aux enseignants

- La recherche menée par Chatel sur la notation des épreuves de sciences économiques et sociales (SES) au baccalauréat indique que les pratiques de notation varient d'un correcteur à un autre en **fonction de leurs caractéristiques propres** (Chatel, 1998 *in* Merle, 2018). Cette étude est parvenue à identifier plusieurs variables susceptibles d'être à l'origine de ces écarts de notation :
    - l'âge du professeur : les enseignants les plus jeunes notent plus sévèrement que leurs collègues plus âgés ;
    - son grade : les enseignants agrégés accordent aux élèves de meilleures notes que leurs collègues maîtres auxiliaires[7] ;
    - son genre : les femmes notent plus largement que les hommes ;
    - son origine sociale : par rapport aux enseignants dont les parents exerçaient une profession intermédiaire, ceux dont les parents étaient employés ou ouvriers notent plus sévèrement les élèves[8] ;
    - sa formation initiale : les économistes accordent des notes plus élevées à leurs élèves que leurs collègues provenant d'autres disciplines ;
    - son académie d'exercice : les enseignants en exercice dans les académies de Lille et de Rouen accordent de meilleures notes à leurs élèves par rapport à ceux exerçant dans l'académie de Versailles[9].

  La recherche de Chatel fait aussi état de caractéristiques sociodémographiques propres qui n'affectent pas significativement les notes données à leurs élèves, comme le niveau de diplôme le plus élevé de l'enseignant.

[7] Les différences de notation observées entre les maîtres auxiliaires et les enseignants certifiés ne sont toutefois pas significatives. Autrement dit, on considère que la probabilité que les différences de notation observées entre ces groupes d'enseignants soient dues au hasard est trop élevée pour s'autoriser à conclure à l'existence d'un lien causal entre origine sociale et notation.

[8] Les différences de notation observées entre les enseignants dont les parents exerçaient une profession intermédiaire et ceux dont les parents étaient cadres ou agriculteurs/commerçants/artisans ne sont pas significatives.

[9] Les différences de notation ne sont cependant pas significatives entre les enseignants exerçant dans l'académie de Versailles et ceux exerçant dans les académies de Montpellier et de Paris.

C. Les arrangements personnels

Outre les biais évaluatifs qui viennent d'être présentés, *«* **le fait de noter un élève est également une action proprement rationnelle** qui trouve ses fondements à la fois dans **les intérêts et les valeurs propres à l'action enseignante** » (Leclercq, Nicaise & Demeuse, 2004).

Merle (1996) avait déjà mis en lumière cet effet spécifique à l'action enseignante qu'il a dénommé « **arrangement évaluatif** ». Ce dernier se traduit par le lien entre le jugement professoral et « un ensemble quotidien "d'arrangements" et de "bricolage" des notes, intentionnels ou non » (Merle, 1996, *in* Leclercq, Nicaise & Demeuse, 2004). Selon le sociologue, les arrangements évaluatifs seraient de trois ordres distincts (pour une analyse détaillée de ces trois types d'arrangements, voir Merle, 1996) :

- Les arrangements dits « **internes** » : ces derniers seraient **pris en faveur de la classe et des élèves d'une manière générale ou individuelle**. Les objectifs visés par ces ajustements sont de nature diverse :
    - conserver un bon climat de travail au sein de la classe ;
    - harmoniser la notation pour ne pas sanctionner les élèves en difficulté.

On observe notamment que les **enseignants exerçant dans des zones d'éducation prioritaire** ont recours à certains de ces arrangements évaluatifs. Afin de favoriser un climat serein dans leur classe, ces enseignants ont tendance à travailler « dans une logique de réussite à court terme, parfois même dans l'instantané » et à privilégier « un enseignement comme un traitement très individualisé des comportements au détriment des apprentissages collectifs » (Butlen *et al*., 2015).

- Les arrangements dits « **externes** » : la notation des élèves peut être ajustée en fonction de **l'établissement scolaire** dans son ensemble et plus globalement pour **satisfaire l'ensemble de la communauté éducative** (collègues enseignants, parents d'élèves, etc.). L'objectif majeur n'est autre que le **maintien de la réputation de l'école ou de l'établissement** et les taux de réussite des élèves conséquents. À titre illustratif, ces « bricolages » peuvent consister à :
    - ne pas tenir compte des notes d'un contrôle qui se révèlent être très basses ;
    - abaisser ou augmenter le coefficient de certaines évaluations.

- Les arrangements dits « **pour soi** » : cette dernière sorte d'arrangement reflètent **l'intérêt de l'évaluateur**. Autrement dit, en tant que sujet, le correcteur a des représentations personnelles qu'il peut volontairement décider de faire intervenir lors de l'évaluation des élèves. La notation sera ainsi dépendante de :
    - l'idéal pédagogique de l'enseignant ;
    - son parcours scolaire ;
    - son origine sociale ;
    - ses valeurs personnelles, etc.

  (Leclercq, Nicaise & Demeuse, 2004).

Encadré 7 : Les arrangements évaluatifs en EPS

La notion d'arrangement évaluatif est notamment connue dans le champ de l'éducation physique et sportive (EPS). Se détachant quelque peu de la définition donnée par Merle, l'arrangement évaluatif en EPS a pour objectif premier de considérer le contexte dans lequel a lieu le processus évaluatif, dans un **souci d'équité des élèves.**

En raison de la diversité des activités physiques et sportives et des aptitudes des élèves acquises en dehors des cours (vitesse, coordination, souplesse, etc.), de tels arrangements permettent de compenser les écarts de performance, ces derniers étant sources d'**inégalités au niveau des résultats** (Fayaubost, Gibon & Rossi, 2021).

En effet, en imposant des critères de notation et des barèmes en EPS, l'institution scolaire reflète des différences de niveaux sportifs, qui certes renvoient à des performances, mais ne rendent pas nécessairement compte des apprentissages des élèves en EPS. Par exemple, dans le cadre scolaire, la pratique de la natation peut, dans certains cas, refléter les performances des élèves dans cette discipline sans pour autant relever d'un processus d'apprentissage. En effet, dans la mesure où certains élèves pratiquent la natation en dehors du cadre scolaire, il ne s'agira pas forcément d'apprentissages scolaires, mais plutôt de performances, résultant de compétences acquises auparavant et désormais maîtrisées par l'élève en question.

Ainsi, dans cette hypothèse, l'arrangement évaluatif peut consister à **prendre en compte d'autres indicateurs pour évaluer les apprentissages de l'élève**, tels que ses progrès, son implication ou encore sa participation aux différentes activités proposées dans le cadre scolaire. Cet arrangement permet *in fine* d'attribuer une note la plus juste possible au regard des performances des autres élèves.

Les enseignants voient donc toute **la nécessité d'adapter leur notation**, et le font souvent de « façon "secrète voire clandestine" », pour ainsi **pallier la comparaison sociale et physique**, et les **nombreuses causes de discriminations** (liées notamment aux attendus d'excellence de la performance sportive des élèves en EPS).

Comme le montrent les **travaux de Cogérino** (1998), les arrangements évaluatifs en EPS sont alors « le résultat d'un décalage entre la représentation qu'ont les enseignants de l'évaluation équitable et les procédures évaluatives exigées dans les textes officiels » (Fayaubost *et al.*, 2021).

Par ailleurs, il existe des pratiques d'arrangements évaluatifs qui sont directement mises en oeuvre par l'institution elle-même. Intervenant à la fin de chaque année scolaire pour des évaluations qui ont une visée certificative (baccalauréat général et technologique, baccalauréat professionnel, certificat d'aptitude professionnelle), les commissions académiques de l'harmonisation des notes (CAHN) procèdent à une harmonisation des résultats obtenus par les élèves. En EPS, par exemple, cette harmonisation des notes des élèves vise à réduire les écarts entre les filles et les garçons, entre les activités proposées aux élèves et entre les établissements.

(Pour plus de détails sur l'existence et l'utilisation des arrangements évaluatifs en EPS, voir la partie 1.2.2 « Equité et arrangements évaluatifs » dans le dossier « Une démarche EPIC pour apprendre et enseigner en EPS » par le groupe EPIC, Fayaubost *et al.,* 2021).

Conclusion

Il ressort de cette synthèse des travaux de docimologie et des autres travaux de recherche qui ont poursuivi les questionnements autour des limites et biais de l'évaluation que **l'évaluation scolaire, comme toute évaluation, n'est pas une science exacte**. Comme cela vient d'être mentionné, de **nombreuses variables** peuvent influencer l'évaluation, et ce à différentes étapes :

- au moment où l'élève « performe » ;
- au moment où le jugement se forme chez l'enseignant ;
- au moment où l'enseignant doit matérialiser son jugement par un indicateur synthétique (tel qu'une note).

Identifier et connaître les variables susceptibles d'influencer le résultat d'une évaluation est important, mais n'amène pas nécessairement à un « rejet absolu de la notation subjective » (Leclercq, Nicaise & Demeuse, 2004) :

- D'une part, un jugement évaluatif objectif semble **impossible** (Leclercq, Nicaise & Demeuse, 2004). En tant qu'évaluateur, l'enseignant est un être humain, avec des caractéristiques personnelles qui lui sont propres. Il ne peut ainsi fonctionner « comme une machine à évaluer, sans préjugés ni préférences, sans erreurs ni omission, sans lassitude, sans ennui » (Perrenoud, 2004)[10].
- D'autre part, refuser tout crédit au jugement professoral reviendrait à **remettre en cause « l'éthique et l'expertise des évaluateurs »**. Cela signifierait également **nier la complexité de l'exercice évaluatif**, lequel doit être satisfait par les enseignants dans le cadre de leur profession (Perrenoud, 2004).
- En outre, dans certaines situations, il est possible de considérer comme souhaitable le fait qu'un évaluateur prenne en compte des éléments extérieurs à la prestation d'un élève (par exemple, lors d'une décision de passage ou de maintien ; ou encore lors de l'évaluation d'une épreuve de rattrapage au baccalauréat – c'est d'ailleurs dans ce but que les évaluateurs disposent du livret scolaire du candidat).

Les biais soulevés dans cette synthèse peuvent constituer des **points de vigilance** pour les enseignants selon deux dimensions :

- d'une part, dans la perspective de s'assurer que leur évaluation soutient effectivement et de manière équitable l'apprentissage de leurs élèves (on pense par exemple aux effets de la présentation de la tâche évaluée, aux effets Matthieu, Pygmalion ou Golem, etc.) ;
- d'autre part, lorsque les évaluations qu'ils réalisent dans le quotidien de leur classe sont susceptibles d'être utilisées pour des fonctions extérieures à la régulation de l'enseignement et de l'apprentissage en classe, telle que l'orientation ou la certification ; c'est-à-dire quand le résultat attribué à une évaluation peut avoir d'importantes conséquences sur la suite du parcours scolaire des élèves (on pense par exemple aux effets classe et établissement, à l'ordre de correction des copies, etc.).

[10] Les QCM et autres corrections automatiques ne font pas intervenir de facteur humain, et peuvent donc, en ce sens, constituer un recours au problème de la subjectivité. Elles présentent toutefois d'autres faiblesses ; à ce sujet, voir par exemple Leclercq (1986).

Références

Allal, L. (2008). Évaluation des apprentissages. In A. van Zanten (Éd.), *Dictionnaire de l'éducation* (pp. 311-314). Presses universitaires de France.

Antibi, A. & Luciani, S. (2003). *La constante macabre ou Comment a-t-on découragé des générations d'élèves ?* Math'Adore.

Autin, F., Batruch, A. & Butera, F. (2019). The function of selection of assessment leads evaluators to artificially create the social class achievement gap. *Journal of Educational Psychology*, *111*(4), 717-735. https://doi.org/10.1037/edu0000307

Bagès, C., Martinot, D. & Toczek, M.-C. (2008). Le rôle modérateur de l'explication donnée à la réussite d'un modèle féminin sur la performance des filles en mathématiques : Une étude exploratoire. *Les Cahiers Internationaux de Psychologie Sociale*, *80*(4), 3-11. https://doi.org/10.3917/cips.080.0003

Bénit, S. & Sarremejane, P. (2019). L'expérience de l'évaluation scolaire chez les écoliers et collégiens. Contribution à la connaissance des processus motivationnels à l'école. *McGill Journal of Education*, *54*(1), 1060862ar. https://doi.org/10.7202/1060862ar

Braibant, J.-M., Lecroart, I., & Billat, E. (2014). *Comment améliorer la qualité des examens QCM à l'université sur base d'une analyse des items ? L'exemple de l'UCL.* 26ème colloque de l'ADMEE Europe, Marrakech. https://dial.uclouvain.be/pr/boreal/object/boreal:177475

Butlen, D., Charles-Pézard, M. & Masselot, P. (2015). Apprentissage et inégalités au primaire : le cas de l'enseignement des mathématiques en éducation prioritaire. Cnesco.

Capelle, C. (2010). Pratiques de correction sur copies d'examen et nouveaux usages instrumentés, *EducPros*, 1-16.

Cardie Créteil. (s. d.). *Journée d'étude des classes sans notation chiffrée*.

Cardie de Nantes. (s. d.). *Classe sans note : Bilan à partir des fiches expérithèque (mars 2013)*. https://ww2.ac-poitiers.fr/cardie/sites/cardie/IMG/pdf/bilan_des_experimentations_par_cardie_nantes.pdf

Cartierre, N., Finez, L. & Genelot, S. (2016, juin). L'auto-handicap comportemental dans des classes sans note : Effet de médiation du sentiment d'efficacité personnelle des collégiens. *11e Congrès International de Psychologie Sociale en Langue Française - CIPSLF 2016*. https://hal-univ-bourgogne.archives-ouvertes.fr/hal-01352916

Caverni, J.-P., Fabre, J.-M. & Noizet, G. (1975). Dépendance des évaluations scolaires par rapport à des évaluations antérieures : études en situation simulée. *Travail humain (Paris)*, 38(2), 213-222.

Collège des Bauges (académie de Grenoble) (s. d.). *Expérimentation classe sans notes 2015/2016—Synthèse des 3 enquêtes : Élèves, parents, professeurs*. http://www.ac-grenoble.fr/college/bauges/sites/default/files/pdf/synth%C3%A8se%20enqu%C3%AAtes%20classe%20sans%20notes%202015_0.pdf

Crahay, M., Mottier Lopez, L. & Marcoux, G. (2019). Chapitre 6. L'évaluation des élèves : Docteur Jekyll and Mister Hyde de l'enseignement. In M. Crahay, *Peut-on lutter contre l'échec scolaire ?* (pp. 357-425). De Boeck Supérieur. https://doi.org/10.3917/dbu.craha.2019.01.0357

De Ketele, J.-M. & Gérard, F.-M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et Éducation en Évaluation*, *28*(3), 1-26.

Direction générale de l'enseignement scolaire – Département de la recherche et du développement, de l'innovation et de l'expérimentation Dgesco – DRDIE. (2013). *Bilan du réseau de l'innovation, 2012-2013*. https://fr.calameo.com/books/00008737062a12433a674

Duru-Bellat, M. & Mingat, A. (1988). Le déroulement de la scolarité au collège : Le contexte « fait des différences ». *Revue Française de Sociologie*, *29*(4), 649-666. https://doi.org/10.2307/3321516

Duru-Bellat, M. & Mingat, A. (1993). *Pour une approche analytique du fonctionnement du système éducatif* (1re éd). Presses universitaires de France.

Fayaubost, R., Gibon, J., & Rossi, D. (2021). *Une démarche ÉPIC pour apprendre et enseigner en EPS* (Groupe Évaluation par indicateurs de compétence (ÉPIC), Éd.). Aeeps.

Flore, P. C. & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology*, *53*(1), 25-44. https://doi.org/10.1016/j.jsp.2014.10.002

Gauthier, P. (2020). *Effet Rosenthal et filtre affectif*.

Genelot, S. (2017). *Evaluer « "des" » compétences ou évaluer « "par" » compétences. Le cas des classes sans notes.* https://mediaserveur.u-bourgogne.fr/permalink/v12583ea0e4c0s107c36/

Goudeau, S. & Autin, F. (s. d.). Existe-t-il des différences entre les collégiens évalués avec ou sans notes ? Synthèse de l'étude comparée des impacts des classes sans note et des classes avec notes dans l'académie de Poitiers.

Grapin, N., & Sayac, N. (2017). Évaluer la maîtrise de la numération écrite chiffrée : Choix du format QCM et validité d'items d'évaluations externes. *Éducation et didactique*, *11*(3), 55-72. https://doi.org/10.4000/educationdidactique.2836

Hadji, C. (1992). Chapitre III. De l'évaluation comme saisie objective… In C. Hadji, *L'évaluation des actions éducatives* (pp. 77-109). Presses universitaires de France. https://www.cairn.info/l-evaluation-des-actions-educatives--9782130448310-page-77.htm

Heck, J.L. & Stout, D.E. (1998). Multiple-Choice vs. Open-Ended exam problems: Evidence of their impact on student performance in introductory finance, *Financial Practice and Education*, 8(1), 83–93.

Heurdier, L. & Prost, A. (2021). *Les politiques de l'éducation en France* (3e éd). la Documentation française. https://ww2.ac-poitiers.fr/cardie/sites/cardie/IMG/pdf/synthe_se_e_tude_sur_la_suppression_des_notes.pdf

Huang, V. (2017). An Australian study comparing the use of multiple-choice questionnaires with assignments as interim, summative law school assessment, *Assessment & Evaluation in Higher Education*, 42(4), 580-595, https://doi.org/10.1080/02602938.2016.1170761

Huguet, P., Brunot, S. & Monteil, J. M. (2001). Geometry versus drawing: Changing the meaning of the task as a means to change performance. *Social Psychology of Education: An International Journal, 4*(3-4), 219–234. https://doi.org/10.1023/A:1011374700020

Huguet, P. & Régner, I. (2007). Stereotype threat among schoolgirls in quasi-ordinary classroom circumstances. *Journal of Educational Psychology*, *99*(3), 545-560. https://doi.org/10.1037/0022-0663.99.3.545

Inspection générale de l'Éducation nationale - IGEN (2013). *La notation et l'évaluation des élèves éclairées par des comparaisons internationales* (N° 013-072 ; p. 76). https://www.vie-publique.fr/sites/default/files/rapport/pdf/134000726.pdf

Lapostolle, G. & Genelot, S. (2015). Petite histoire d'une expérience innovante dans l'académie de Dijon. Du rôle du chef d'établissement dans la naissance, l'installation et la diffusion d'une pratique de «classes sans note» (2006 à 2014). *Spirale - Revue de recherches en éducation*, *55*(1), 3-16. https://doi.org/10.3406/spira.2015.1734

Laugier, H. & Weinberg, D. (1936). Commission française pour l'enquête Carnegie sur les examens et concours. La correction des épreuves écrites au baccalauréat. Maison du livre.

Leclerq, D. (1986). *La conception des Questions à Choix Multiple*. Bruxelles : Labor.

Leclercq, D., Nicaise, J. & Demeuse, M. (2004). Docimologie critique : Des difficultés de noter des copies et d'attribuer des notes aux élèves. *In* M. Demeuse (Éd.), *Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation* (pp. 273-292). Les éditions de l'Université de Liège.

Leyens, J.-Ph. & Yzerbit, V. (1997). Psychologie sociale. Mardaga.

Martin, J. (2002). Aux origines de la « science des examens » (1920-1940). *Histoire de l'éducation*, *94*, 177-199. https://doi.org/10.4000/histoire-education.817

Merle, P. (1996). Le jugement professoral au quotidien : l'arrangement évaluatif. In P. Merle, *L'évaluation des élèves. Enquête sur le jugement professoral* (pp. 74-144). Presses universitaires de France.

Merle, P. (1998). Sociologie de l'évaluation scolaire. Presses Universitaires de France.

Merle, P. (2007). *Les notes : Secrets de fabrication* (1re éd). Presses universitaires de France.

Merle, P. (2015). L'école française et l'invention de la note. Un éclairage historique sur les polémiques contemporaines, *Revue française de pédagogie*, (193), 77-88.

Merle, P. (2018). *Les pratiques d'évaluation scolaire : Historique, difficultés, perspectives*. Presses universitaires de France.

Moinet, A. (2018). *La notation scolaire : Inconvénients et alternatives*. https://hal.archives-ouvertes.fr/hal-01700229

Murat, F. (1998). Les différentes façons d'évaluer le niveau des élèves en fin de collège : Évaluation et notation des élèves. *Éducation & Formations*, *53*, 35-49.

Newble, D. I., Baxter, A., Elmslie, R. G. (1979). A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Med Educ*. 13(4), 263-8. https://doi.org/10.1111/j.1365-2923.1979.tb01511.x. PMID: 470647

Noizet, G. & Caverni, J-P. (1978). Psychologie de l'évaluation scolaire. Presses universitaires de France.

Perrenoud, P. (2004). Évaluer des compétences. *L'Éducateur*, *La note en pleine évaluation*, 8-11.

Régner, I., Selimbegović, L., Pansu, P., Monteil, J.-M. & Huguet, P. (2016). Different Sources of Threat on Math Performance for Girls and Boys: The Role of Stereotypic and Idiosyncratic Knowledge. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00637

Rocher, T. (2015). Mesure des compétences. *Éducation & formations*, *n° 86-87*(02), 37. https://doi.org/10.48464/ef-86-87-02

Rosenthal, R. (1987). Pygmalion effects: Existence, magnitude, and social importance. *Educational Researcher*, *16*(9), 37-40.

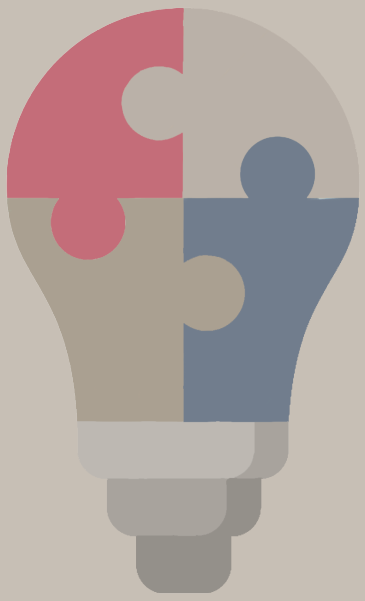
Saillot, É. (2019). Évaluation par compétences : Les préoccupations des enseignants en matière d'évaluation au cours d'une expérimentation des « classes sans notes » dans un collège français. *Mesure et évaluation en éducation*, *42*(2), 35-61. https://doi.org/10.7202/1071515ar

Service Santé et Cardie du Rectorat de Poitiers. (s. d.). Résultats du questionnaire auprès des élèves et leurs professeurs. Classes sans notes dans l'académie de Poitiers—Année 2012-2013. https://ww2.ac-poitiers.fr/cardie/sites/cardie/IMG/swf/classesansnote.swf

Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797-811. https://doi.org/10.1037/0022-3514.69.5.797

Terrier, C. (2014). *Un coup de pouce pour les filles ? Les biais de genre dans les notes des enseignants et leurs effets sur le progrès des élèves*. Notes de l'IPP.

Trouilloud, D. O., Sarrazin, P. G., Martinek, T. J. & Guillet, E. (2002). The influence of teacher expectations on student achievement in physical education classes: Pygmalion revisited. *European Journal of Social Psychology*, *32*(5), 591-607. https://doi.org/10.1002/ejsp.109



le cnam
Cnesco
Centre national d'étude des systèmes scolaires

CENTRE NATIONAL D'ÉTUDE DES SYSTÈMES SCOLAIRES
CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS
41 rue Gay-Lussac - 75005 PARIS
06 98 51 82 75 - cnesco@lecnam.net
www.cnesco.fr

UCA Université Clermont Auvergne
INSPÉ Institut national supérieur du professorat et de l'éducation Clermont Auvergne

UNIVERSITÉ CLERMONT AUVERGNE
INSPÉ DE L'ACADÉMIE DE CLERMONT-FERRAND
36 avenue Jean-Jaurès - 63407 CHAMALIÈRES
04 73 31 71 50
https://inspe.uca.fr

RETROUVEZ LES DERNIÈRES ACTUALITÉS DU CNESCO :

www.cnesco.fr Cnesco Cnesco Cnesco