CECP
CONSEIL DE L'ENSEIGNEMENT
DES COMMUNES ET DES PROVINCES

FORMATION DES ENSEIGNANTS

« EN QUOI L'ÉVALUATION PARTICIPE-T-ELLE À LA CONSTRUCTION DE L'ÉCHEC ? »

DACHET DYLAN & BAYE ARIANE

Vendredi 28 janvier 2022

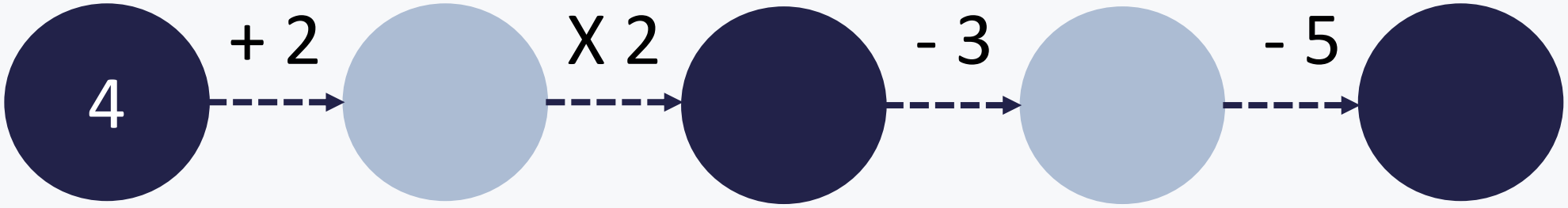
Au programme

1. Introduction : *pensez-vous qu'il soit possible d'être objectif en matière d'évaluation ?*
2. L'expérience de Gjorgjevski … et l'effet Posthumus
3. La recherche liégeoise APER
4. Des expériences plus récentes
   - Biais relatif à l'origine ethnique des élèves
   - Biais relatif au statut socio-économique de la famille
   - Biais relatif au genre
   - Biais relatif au contexte actuel
5. L'impact de la culture/fonction de l'évaluation sur l'évaluation elle-même
6. Changer l'évaluation pour changer l'école
7. Conclusion

Introduction

« *Évaluer consiste à définir des critères et des indicateurs dans le but de prendre des décisions, donc à choisir des éléments considérés comme pertinents dans le référent (ce qui devrait être, ce que l'on projette) et à déterminer durant l'action et à l'issue de celle-ci, si ces éléments sont bien présents dans le référé (ce qui est)* » (Raynal & Rieunier, 2014, p. 215)

Pensez-vous qu'il soit possible d'être objectif en matière d'évaluation ?



Réponse de Thomas

Quelle note ?

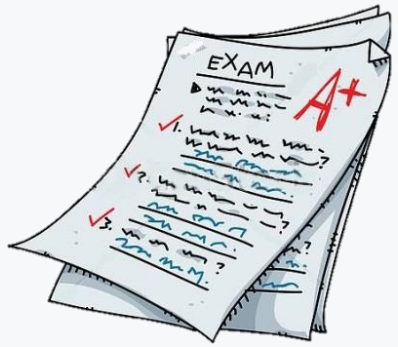
www.wooclap.com/LJRFDJ

Pensez-vous qu'il soit possible d'être objectif en matière d'évaluation ?

L'expérience de Gjorgjevski (cité par Rot et Butas, 1959)

Étape 1

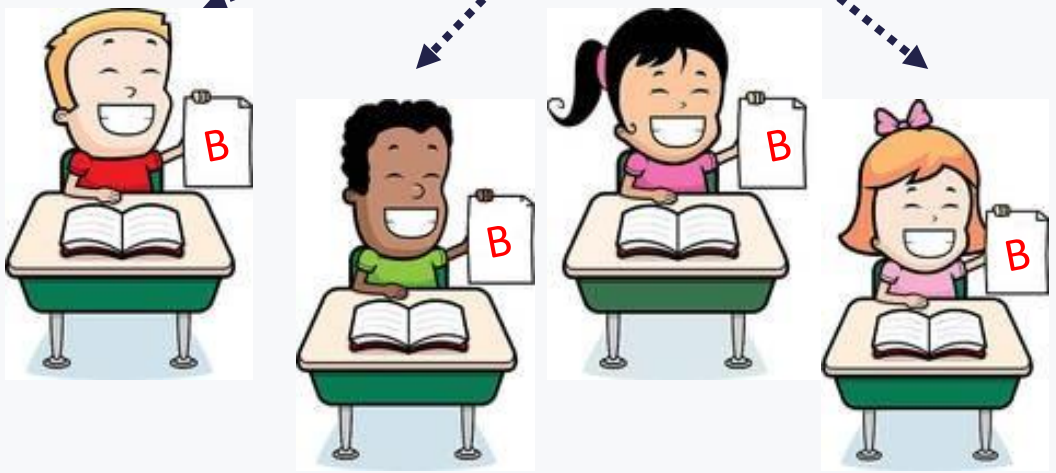
5 correcteurs



100 épreuves

Étape 2

4 autres correcteurs

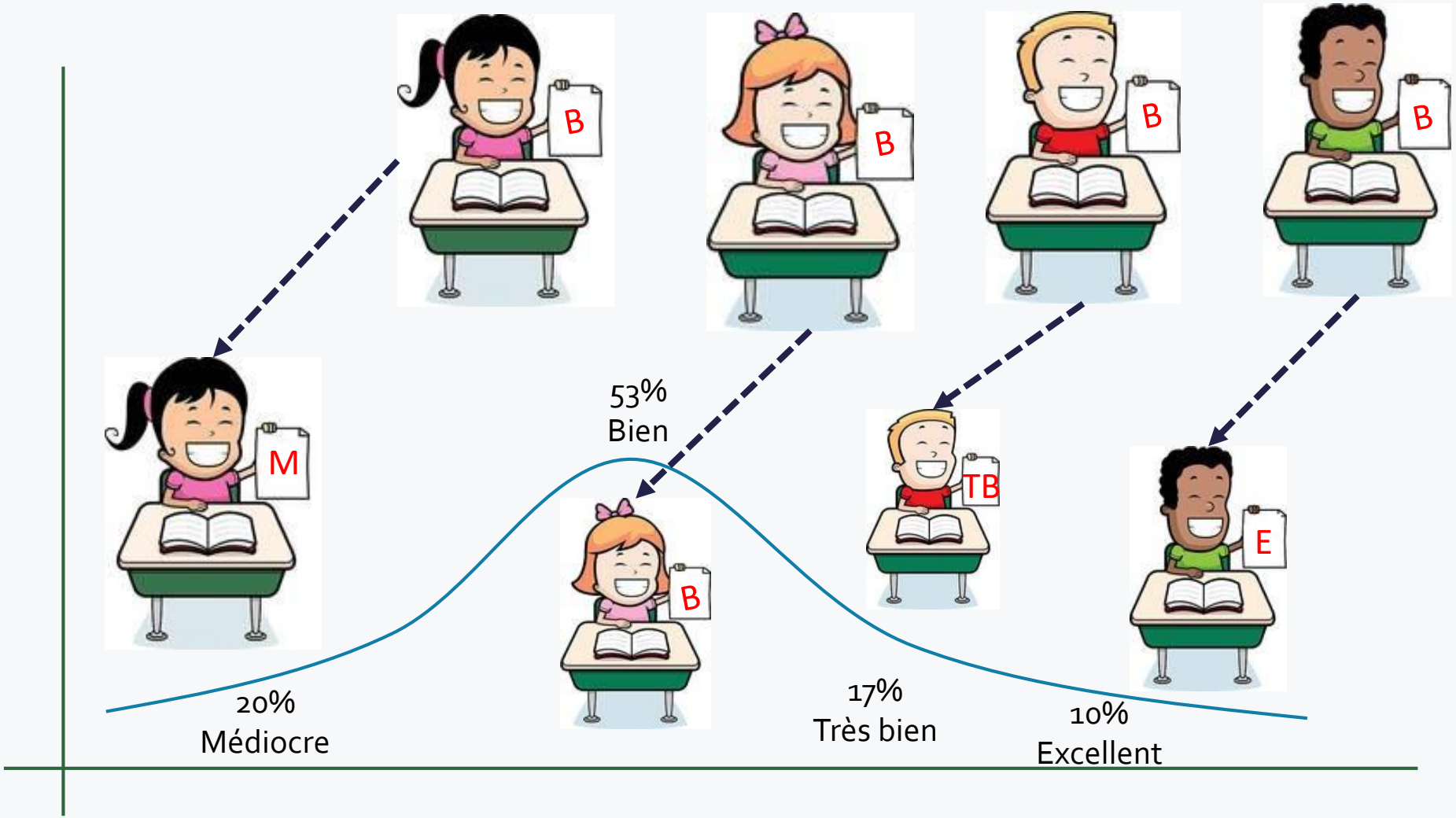


15 copies notées « Bien »

L'expérience de Gjorgjevski (cité par Rot et Butas, 1959)

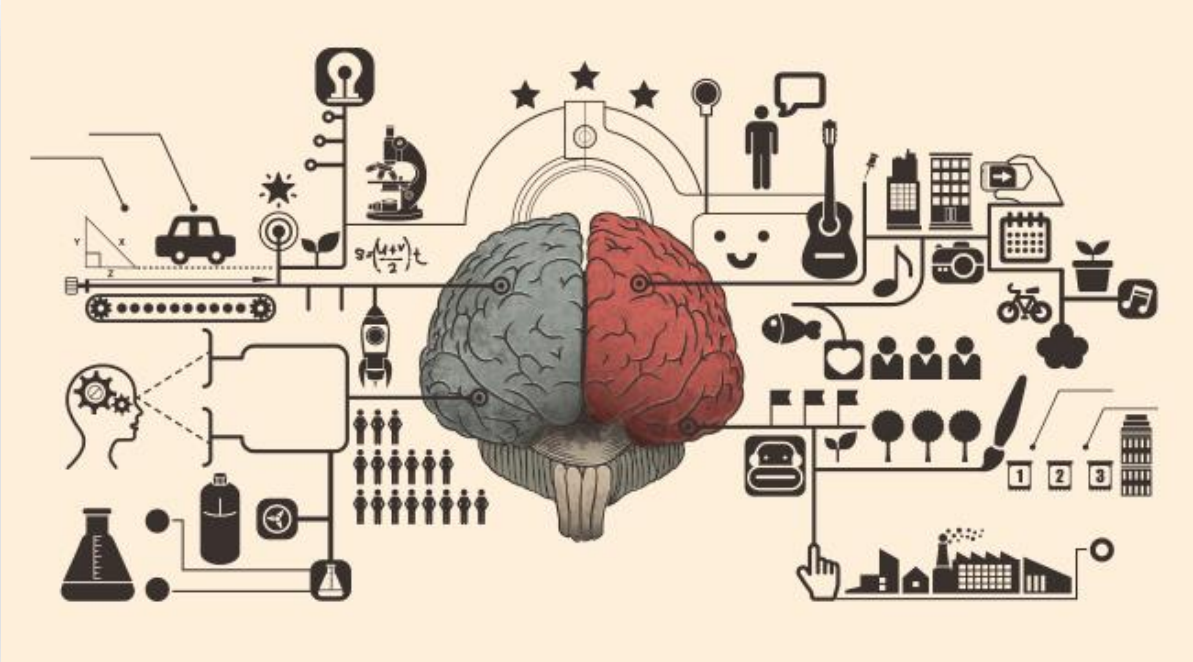
Résultats

Les nouveaux correcteurs ont spontanément adoptés des exigences nouvelles



L'effet Posthumus

« L'enseignant ***adapte 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, 1980)



La recherche liégeoise APER

A. Grisay (1984) puis Grisay, De Bal, de Landsheere (1984)

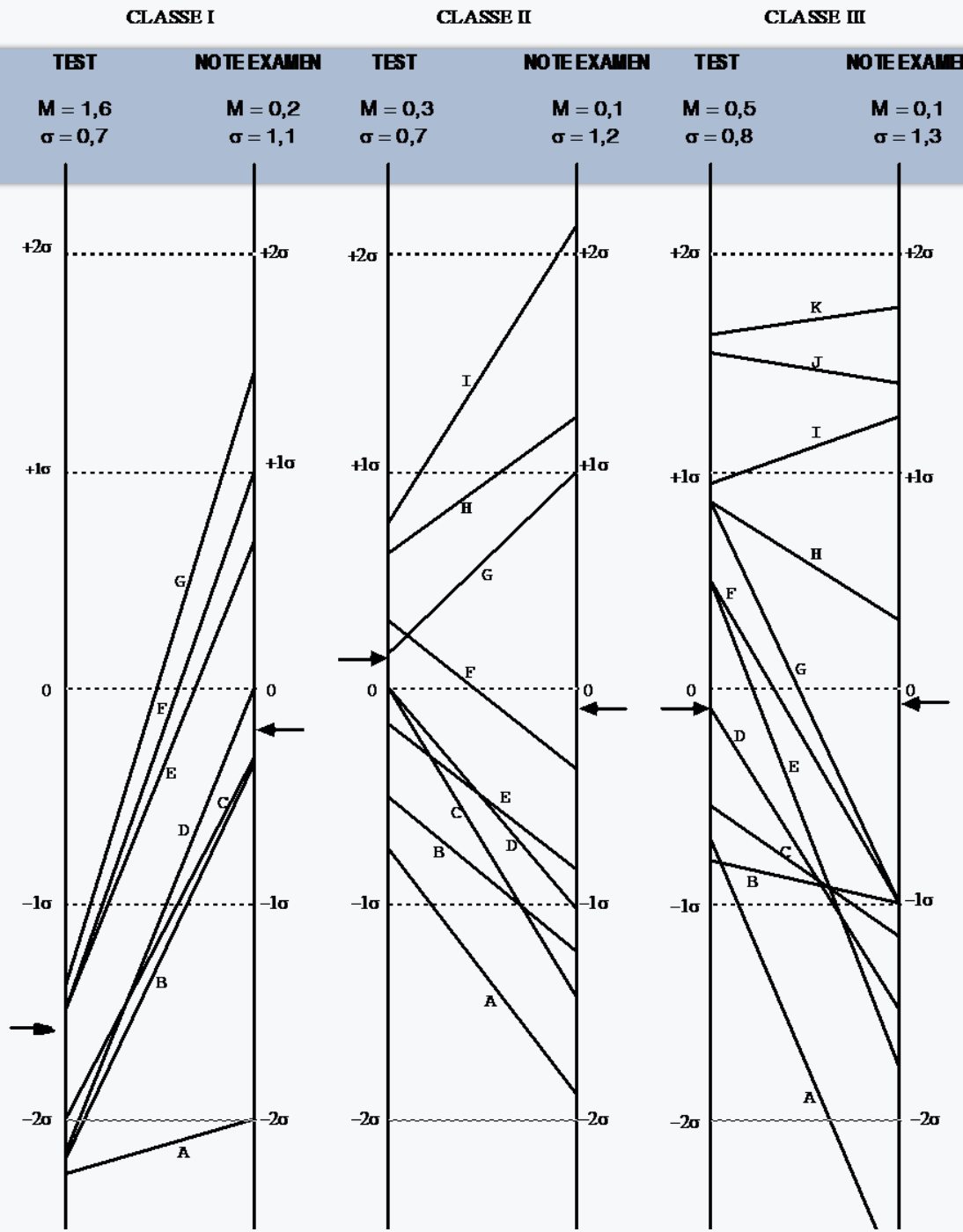
- Menée dans 52 écoles de la Communauté française de Belgique, elle porte sur les 6 années de l'enseignement primaire.
- Les chercheurs recueillent dans chaque classe les notes attribuées par l'enseignant à chacun des élèves.
- Les chercheurs soumettent les élèves à une épreuve commune ou épreuve externe.

La recherche liégeoise APER

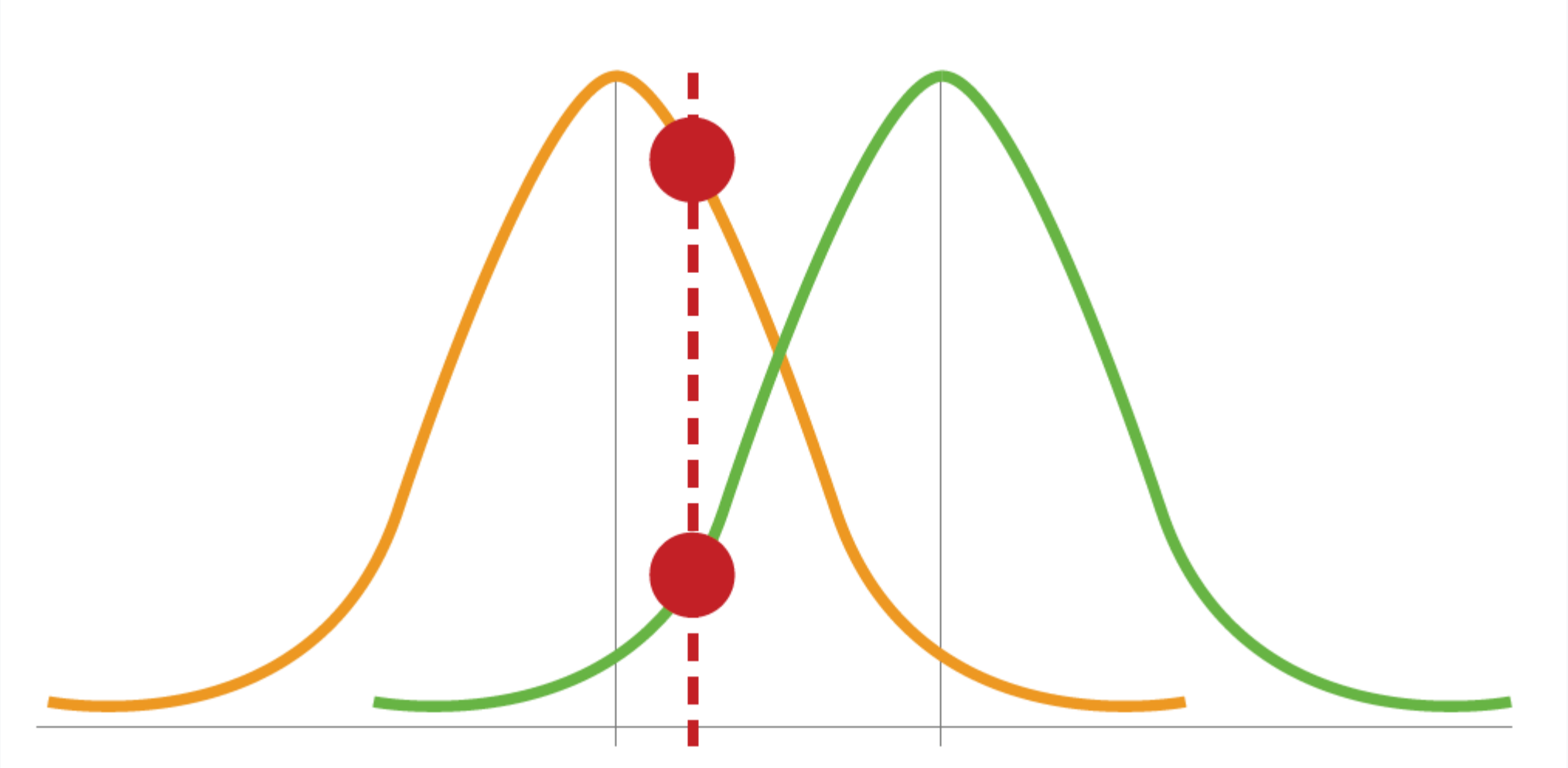
Notes standardisées des élèves de 3 classes
5e primaire

CONCLUSIONS

- À résultats équivalents au test, un élève d'une classe va redoubler et un autre appartenant à une autre classe va réussir.
- À compétences égales, un élève sera jugé fort ou faible selon la classe qu'il fréquente et donc selon la compétence des autres élèves.



La recherche liégeoise APER



Des expériences plus récentes

(Sprietsma, 2013)

L'évaluation des performances des élèves par des enseignants peut-elle être biaisée par leur connaissance du milieu social d'un élève ?

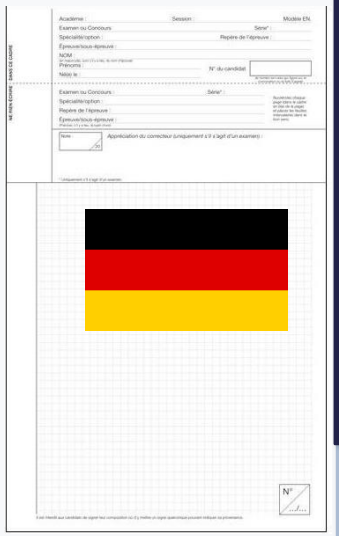
Etude expérimentale :

(1) Recueil de dissertations rédigées par deux classes anonymes

(2) Sélection aléatoire de 10 dissertations

(3) Envoie de la série chaque enseignant avec assignation aléatoire d'un prénom à chaque copie

- Des prénoms à consonance germanophone (Max, Stefan, …)
- Des prénoms à consonance turque (Hakan, Gönül, …)
  (une des communautés ethniques les plus représentées)



88 enseignants du primaire (58 écoles différentes)

Des expériences plus récentes

(Sprietsma, 2013, p. 530)

L'évaluation des performances des élèves par des enseignants peut-elle être biaisée par leur connaissance du milieu social d'un élève ?

Table 5 The effect of assigned pupil origin on grades (OLS estimates)

	(1)	(2)	(3)
Name of Turkish origin	−0.13	−0.09**	−0.11**
Std error	(0.09)	(0.04)	(0.04)
Essay fixed effects	No	Yes	Yes
Teacher fixed effects	No	No	Yes
Observations	880	880	880
R squared	0.004	0.56	0.65

Note: Grades range from 1 (very insufficient) to 6 (very good)
*, **, *** indicate statistical significance at the 10, 5 and 1 % level of confidence, dependent variable: grades. Standard errors in parentheses and clustered by teacher

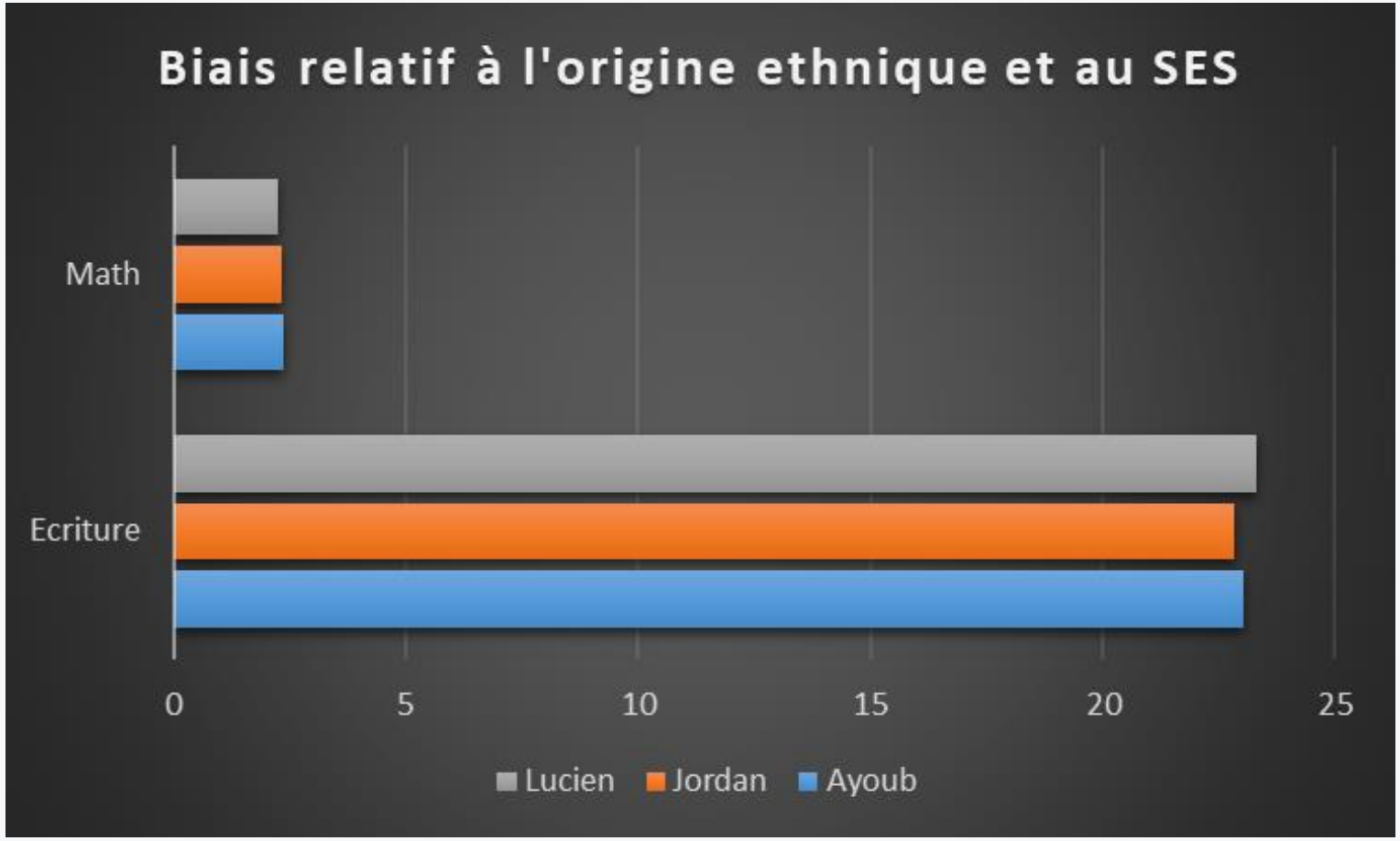
En moyenne, les élèves « d'origine » turque ont été évalués plus négativement que les élèves « d'origine » allemande, malgré que leurs copies étaient identiques.

Biais relatif à l'origine ethnique des élèves

Des expériences plus récentes

(Dachet & Baye, 2021)

L'évaluation des performances des élèves par des enseignants peut-elle être biaisée par leur connaissance du milieu social d'un élève ?



Des expériences plus récentes

(Rangvid, 2015, p. 47)

L'évaluation des performances des élèves par des enseignants peut-elle être biaisée par leur connaissance du milieu social d'un élève ?

Comparaison entre : (1) les données issues des examens étatiques externes et (2) les notes des enseignants

Table 1
Descriptive statistics by subsamples.

Exam score=7		Boys	Girls	Low parental education	High parental education	Migrants	Natives	All students
Teacher scores	*Mean*	6.61	7.09	6.40	7.37	6.51	6.89	6.86
	SD	2.13	2.02	2.14	1.99	2.21	2.08	2.09
Sample size	No. scores	762,558	838,901	387,618	335,080	103,127	1,498,332	1,601,459

Pour les élèves qui ont obtenu un score identique de 7 au test étatique, les notes des enseignants étaient plus hétérogènes : elles variaient de 6,40 pour les élèves dont les parents ont un faible niveau d'éducation à 7,37 pour les élèves dont les parents ont un haut niveau d'éducation.

Des expériences plus récentes

(Rangvid, 2015, p. 48)

L'évaluation des performances des élèves par des enseignants peut-elle être biaisée par leur connaissance du milieu social d'un élève ?

L'évaluation des élèves par les enseignants est plus hétérogène que l'évaluation des mêmes élèves par des tests étatiques. Cette hétérogénéité s'avère être au détriment des garçons, des élèves dont les parents ont un faible niveau d'éducation et des élèves issus de l'immigration.

Biais relatif au statut socio-économique des élèves

Table 2
Estimated grading gaps by gender, parental education and migrant background.

	Ability quintile				
Gender grading gap (boys)	1	2	3	4	5
(1) No controls included	−0.150***	−0.177***	−0.195***	−0.173***	−0.103***
	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
	0.021	0.038	0.040	0.031	0.019
(2) Controls included	−0.160***	−0.187***	−0.199***	−0.181***	−0.106***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)
	0.058	0.094	0.094	0.067	0.044
Parental education grading gap (low par. educ.)					
(1) No controls included	−0.277***	−0.371***	−0.345***	−0.295***	−0.240***
	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)
	0.021	0.038	0.040	0.031	0.019
(2) Controls included	−0.140***	−0.197***	−0.199***	−0.174***	−0.145***
	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
	0.058	0.094	0.094	0.067	0.044
Ethnic grading gap (migrants)					
(1) No controls included	−0.062***	−0.080***	−0.068***	−0.070***	−0.074***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
	0.021	0.038	0.040	0.031	0.019
(2) Controls included	−0.014**	−0.042***	−0.032***	−0.044***	−0.060***
	(0.005)	(0.006)	(0.006)	(0.006)	(0.006)
	0.058	0.094	0.094	0.067	0.044
# Observations	848,065	846,174	849,290	844,075	846,220

Des expériences plus récentes

(Lafontaine & Monseur, 2009)

Etude expérimentale :

(1) Copies fictives construites par un professeur de mathématiques : copie très faible, une copie assez faible, une copie moyenne et une bonne copie

(2) Seule information disponible → prénom clairement masculin (Bernard et Nicolas) ou féminin (Chloé et Emilie)

48 professeurs de mathématiques

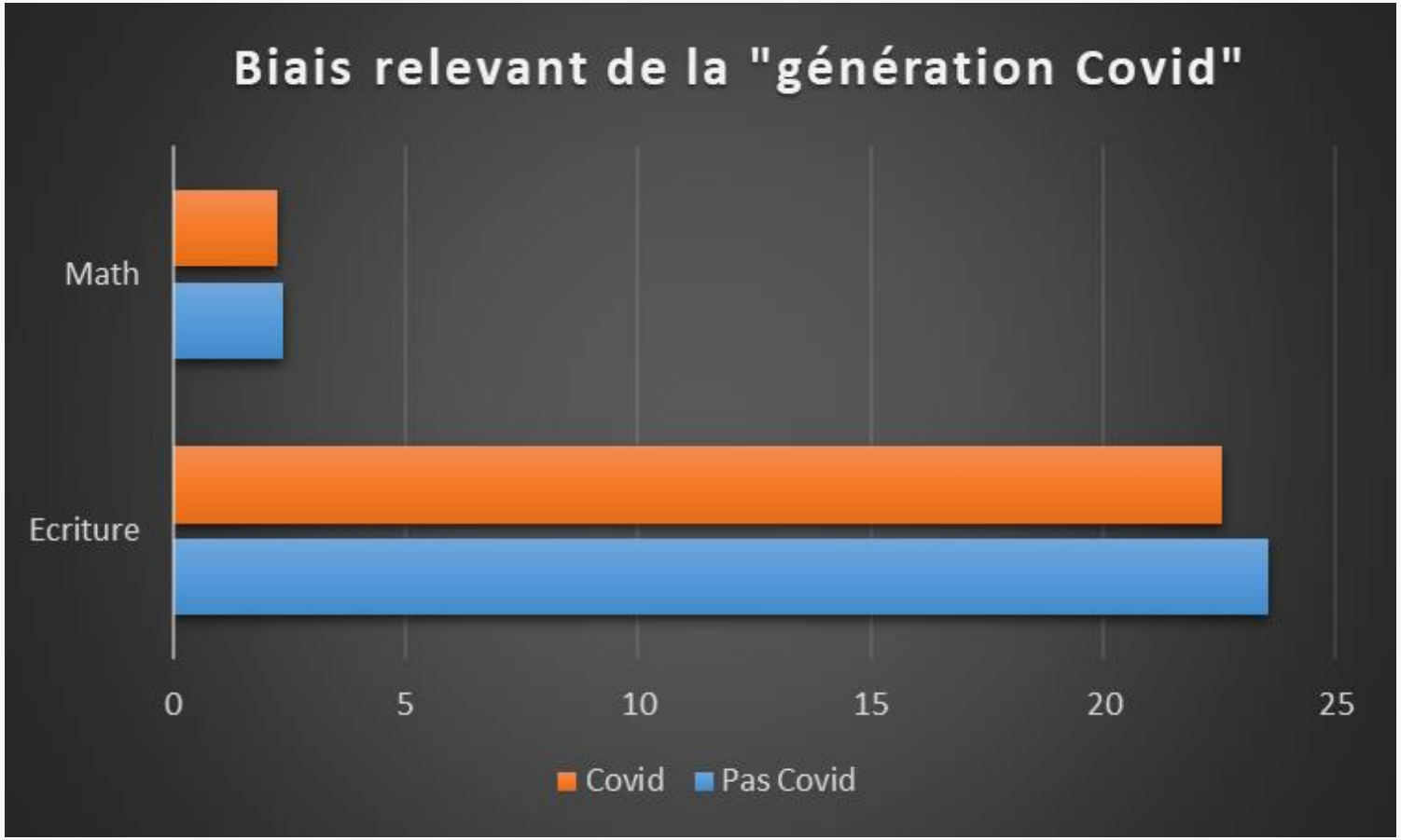
	Filles	Garçons
Moyenne	5,250	5,375

	C. très faible	C. faible	C. moyenne	Bonne C.
Moyenne filles	2,042	5,250	6,354	7,354
Moyenne garçons	1,750	4,917	6,937	7,895

Des expériences plus récentes

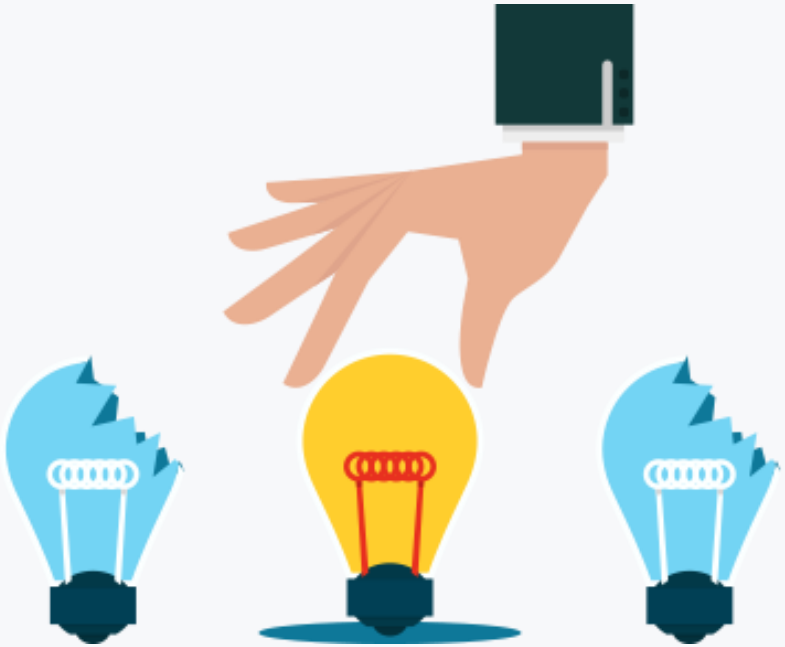
(Dachet & Baye, 2021)

L'évaluation des performances des élèves par des enseignants peut-elle être biaisée par leur connaissance du contexte de l'évaluation ?



L’impact de la culture/fonction de l’évaluation sur l’évaluation elle-même

Deux fonctions des établissements d'enseignement



L’évaluation pour la sélection ou au service de l’apprentissage

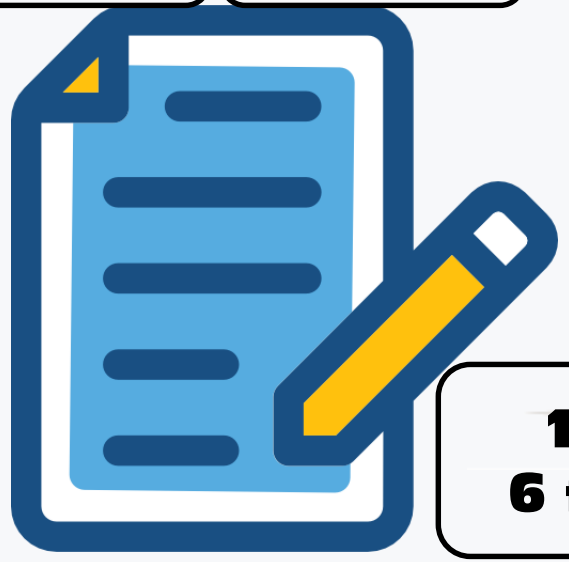
	Évaluation normative	**Évaluation formative**
Définition	utilise des indicateurs tels que des côtes, lettres ou jugements de valeur qui servent parfaitement l'objectif de comparaison avec une norme et entre individus.	fournit un feedback (ici qualitatif) spécifique et détaillé
Visée	comparer les performances de la personne évaluée à celles d'autres personnes	adapter les activités d'enseignement et d'apprentissage et fournir des commentaires pertinents
Contexte	Méritocratique	Correctif/de Maîtrise
Effet	Notation d’un test attribué à un SES - ou SES + Étudiants inscrits dans un programme sélectif SSE + > SSE – (Batruch, Autin & Butera, 2017)	Notation d’un test attribué à un SES - ou SES + Étudiants inscrits dans un programme non-sélectif SSE + = SSE – (Batruch, Autin & Butera, 2017)
Schéma		

La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Les expériences

VS

Dylan **Charles**



11 fautes claires
6 fautes ambiguës

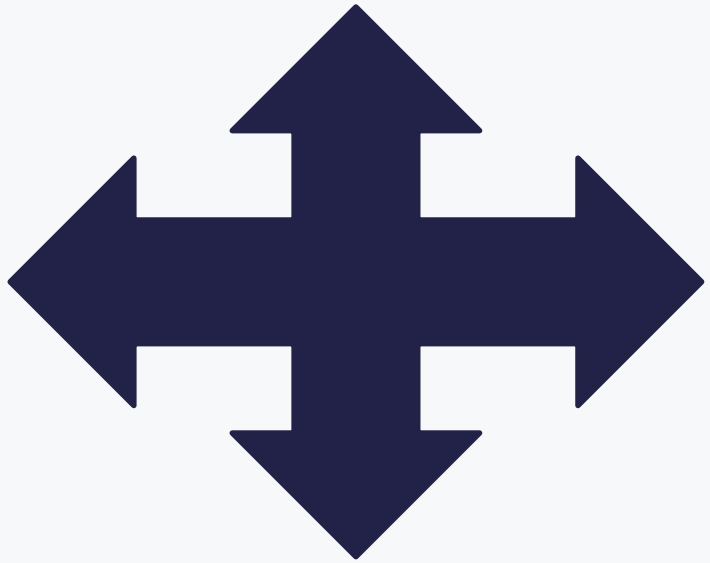
N = 196 à 374

La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Les expériences

Dylan

Dylan



Charles

Charles

La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Expérience 1

Etude expérimentale:

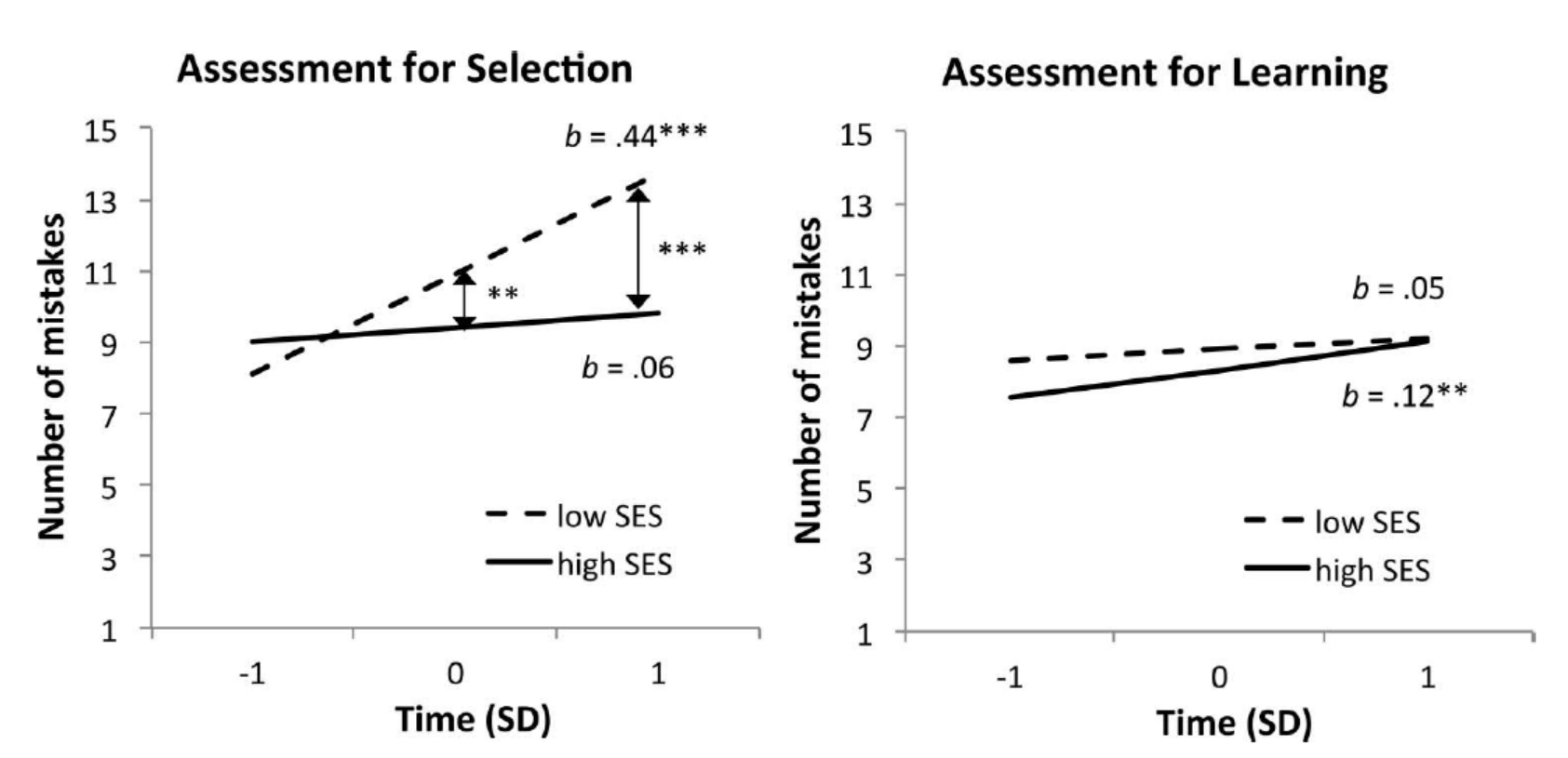
- Double manipulation :
  - **Note (normatif) vs. commentaire(s) (formatif)**
  - SES faible vs. fort (via des indicateurs tels que le prénom, le statut professionnel des parents, activités extracurriculaires, …)
- N = 196 étudiants (imaginant qu'ils sont des enseignants)
- **Correction d'une dictée (1-6) contenant 11 fautes claires et 6 fautes ambiguës**

Variables étudiées :

- Temps de correction → plus grand dans le groupe « commentaire » (M = 22, SD = 6.04) que dans le groupe « Note » (M = 15, SD = 4.30).
- Nombre de fautes repérées :
  - Note (M = 10.16, SD = 2.53) > Commentaire(s) (M = 8.61, SD = 2.18).
  - SES faible (M = 9.90, SD = 2.54) > SES fort (M = 8.87, SD = 2.16)
- Effet positif du temps sur le nombre de fautes repérées (b = 0.17).

La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Expérience 1



La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Expérience 2

Etude expérimentale:

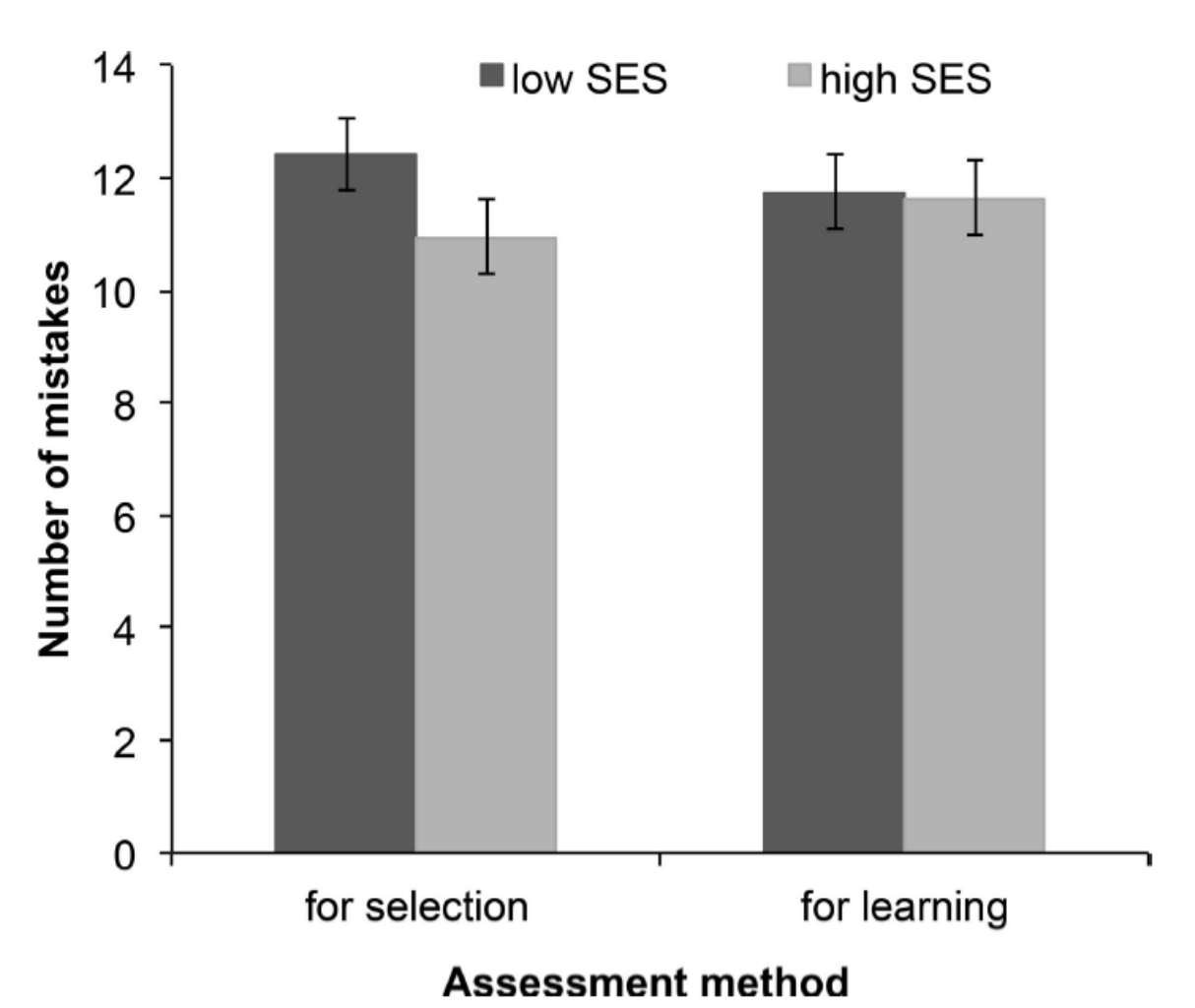
- Double manipulation :
  - **Note vs. Commentaire(s)** (sans autre contextualisation relative à l'aspect normatif et/ou formatif)
  - SES faible vs. fort (via des indicateurs tels que le prénom, le statut professionnel des parents, activités extracurriculaires, …)
- N = 269 étudiants (imaginant qu'ils sont des enseignants)
- **Souligner les erreurs dans une dictée** (et devront faire la côté et/ou le commentaire une autre fois) **contenant 14 fautes claires et 6 ambiguës.**

Variables étudiées :

- Temps de correction → aucune différence constatée.
- Nombre de fautes repérées :
  - Note = Commentaire(s)
  - SES faible (M = 12.09, SD = 2.66) > SES fort (M = 11.29, SD = 2.67)
  - Mais la situation « Note » a tendance à mener à différencier les élèves en fonction de leur SES et non la situation « commentaire(s) ».

La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Expérience 2



La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Expérience 3

Etude expérimentale:

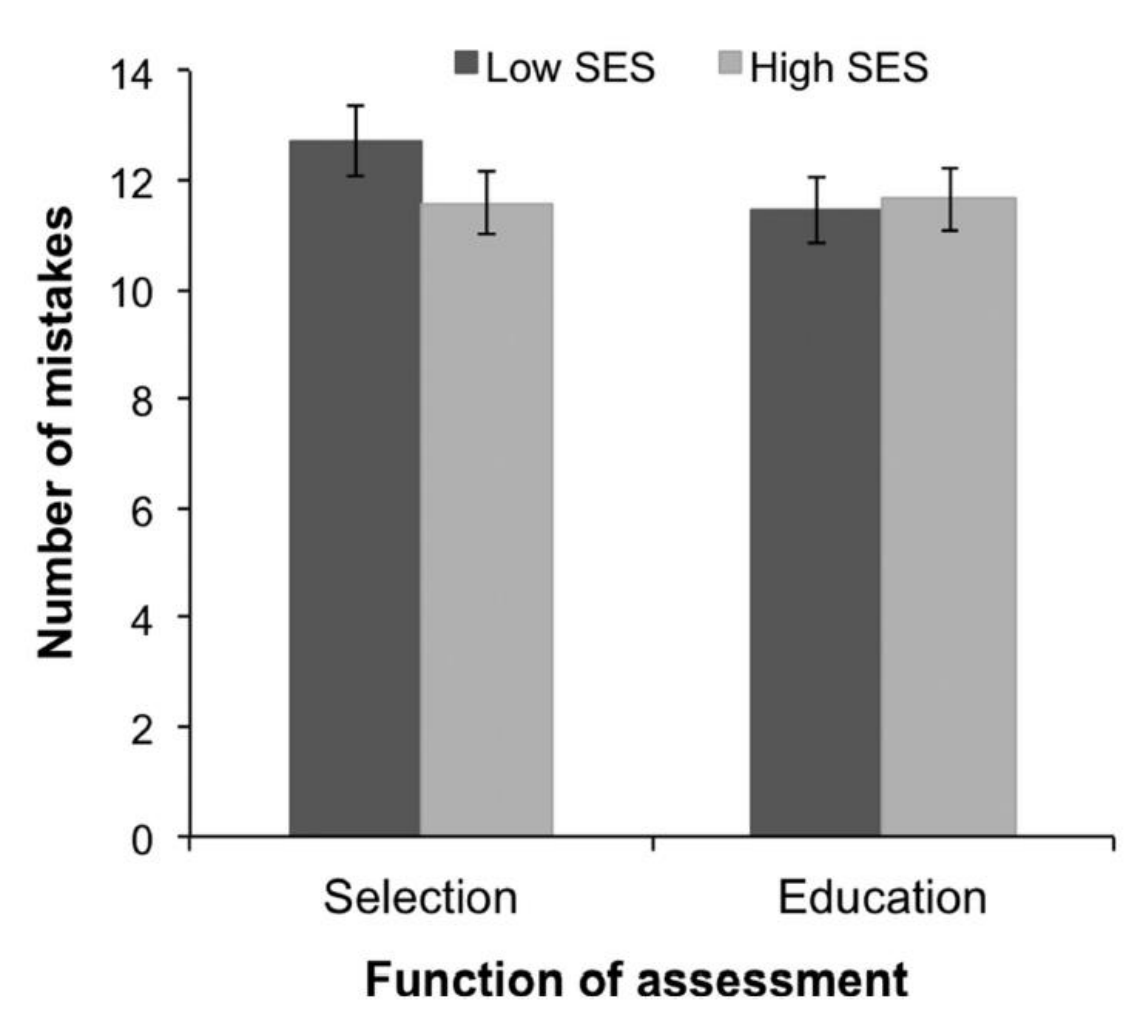
- Double manipulation :
  - Sélection (peut-il ou non être promu ?) vs. Apprentissage (quelles stratégies d'apprentissage ?)
  - SES faible vs. fort (via des indicateurs tels que le prénom, le statut professionnel des parents, activités extracurriculaires, …)
- N = 374 étudiants (imaginant qu'ils sont des enseignants)
- Souligner les erreurs dans une dictée (et devront faire la côté et/ou le commentaire une autre fois) contenant 14 fautes claires et 6 ambiguës.

Variables étudiées :

- Nombre de fautes repérées :
  - Sélection X SES faible : M = 12.71 ; SD = 2.75
  - Sélection X SES fort : M = 11.59 ; SD = 2.94
  - Apprentissage X SES faible : M = 11.44 ; SD = 3.35
  - Apprentissage X SES fort : M = 11.66 ; SD : 2.84

La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Expérience 3



La fonction de sélection de l'évaluation (Autin, Batruch & Butera, 2019)

Expérience 4

Etude expérimentale:

- Double manipulation :
  - Sélection (points du semestre) vs. Apprentissage (quelles stratégies d'apprentissage ?)
  - SES faible vs. fort (via des indicateurs tels que le prénom, le statut professionnel des parents, activités extracurriculaires, …)
- N = 306 étudiants (imaginant qu'ils sont des enseignants d'histoire)
- Évaluer une rédaction en utilisant un nouvel outil d'évaluation → souligner en jaune ce qui est bien écrit et en orange ce qui doit être retravaillé.

Variables étudiées :

- Ratio de feedbacks négatifs formulés :

	Sélection	Apprentissage
Faible SES	M = 0.42 SD = 0.12	M = 0.40 SD = 0.12
Haut SES	M = 0.38 SD = 0.13	M = 0.39 SD = 0.13

Méta-analyse

Estimation de l'ampleur de l'effet de la modération de l'écart de performance dû au SSE de l'élève en fonction de l'orientation de l'évaluation (sélective vs. éducative):

D = 0.19

p = 0.002

« *Cette méta-analyse interne prouve que les évaluateurs créent artificiellement un plus grand écart de performance SSE lorsque l'évaluation est utilisée pour sélectionner plutôt que pour favoriser l'apprentissage. L'ampleur de l'effet est faible, mais nous pensons néanmoins qu'elle doit être interprétée à la lumière de la durée de l'éducation et de la fréquence de l'évaluation. De très petites différences dans les évaluations répétées peuvent avoir des conséquences importantes sur les expériences globales et les résultats scolaires des étudiants lorsqu'elles s'accumulent au fil du temps* » (Autun, Batruch & Butera, 2019, p. 729).

Changer l'évaluation pour changer l'école

Conception initiale de Bloom versus conception élargie de l'évaluation formative dans la littérature scientifique francophone (Allal & Mottier Lopez, 2005)

Conception initiale de Bloom	Conception élargie
Insertion de l'EF après la phase d'enseignement	Intégration de l'EF durant tout l'apprentissage
Utilisation de tests formatifs	Utilisation de divers moyens de recueil d'informations
Feed-Back + correction ➔ remédiation	Feed-back + adaptation de l'enseignement ➔ régulation
Gestion de l'EF par l'enseignant	Participation active des élèves à l'EF
Maîtrise des objectifs par tous les élèves	Différenciation de l'enseignement et des objectifs
Remédiation bénéfique aux élèves qui ont été évalués	Régulation à 2 niveaux : pour les élèves évalués et pour les futurs élèves

Changer l'évaluation pour changer l'école

Étude expérimentale de Nunziati (1990) :

- Marseille
- 1974-1977
- Impact positif quand les élèves sont en mesure de s'approprier les critères d'évaluation des enseignants, d'autogérer leurs erreurs et de maîtriser des outils d'anticipation et de planification de l'évaluation en classe.

Évaluation formative	Évaluation formatrice
Peut être à référence critériée	
Adapter la phase d'enseignement/d'apprentissage	Impliquer l'élève dans l'ensemble du processus d'évaluation formative Apprendre à s'autoévaluer correctement
Gérée exclusivement par l'enseignant	Gérée exclusivement par l'élève
Coresponsabilité entre les enseignants et les élèves dans des modalités dynamiques et complémentaires	

Changer l'évaluation pour changer l'école (Crahay, 2019)

Évaluation formative formelle

Conçue pour produire des traces et indices sur l'apprentissage de l'élève
➔ *épreuves papier-crayon débouchant sur un feed-back critérié explicite*

Recueillir	**Interpréter** (analyser)	**Agir**
Des informations sur les élèves dans un temps planifié	Les informations recueillies (souvent hors de la présence des élèves)	en planifiant une action visant à soutenir l'élève dans l'atteinte des buts d'apprentissage

Évaluation formative informelle

Traces et indices générées pendant l'activité quotidienne de classe
➔ Pleinement intégrées aux processus d'enseignement et d'apprentissage en contexte de classe

Solliciter	**Reconnaitre**	**Utiliser**
Des réponses verbales des élèves afin de récolter de l'information	Les réponses des élèves au regard des concepts enseignés	Immédiatement l'information dans le cours des activités continues de la classe

Conclusion

- Évaluations subjectives et impactées par de nombreux biais (place de la copie, l'effet de halo, l'effet de genre, l'effet pygmalion, la stéréotypie, …)

➔Docimologie

La lutte contre l'échec scolaire appelle un double changement :

1. Abandon de l'évaluation normative ➔ articulation de l'évaluation formative et sommative à référence critériée.
2. Intégration de la notion de régulation au cœur de l'évaluation formative.
3. Utilisation de systèmes de renforcement positif.

Bibliographie

Allal, L. et Mottier Lopez, L. (2005). Formative assessment of learning : A review of publications in French. In *Formative Assessment – Improving Learning in Secondary Classrooms* (pp. 241-264). Paris, France : OECD-CERI Publication (What works in innovation in education).

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

Crahay, M. (2019). *Peut-on lutter contre l'échec scolaire ?* (4th ed.). Louvain-la-Neuve, Belgique : de Boeck supérieur.

De Landsheere, G. (1980). *Examens et évaluation continue. Précis de docimologie*. Bruxelles, Belgique : Labor.

Grisay, A. (1984). *Trébucher sur le seuil de l'école. Enquête sur le problème du redoublement et de l'échec scolaire au premier cycle de l'enseignement primaire*. Liège, Belgique : Laboratoire de Pédagogie expérimentale de l'Université.

Nunziati, G. (1990). Pour construire un dispositif d'évaluation formatrice. Cahiers pédagogiques, 280, 47-64.

Rangvid, B. S. (2015). Systematic differences across evaluation schemes and educational choice. *Economics of Education Review, 48*, 41–55. http://dx.doi.org/10.1016/j.econedurev.2015.05.003

Rot, N., Butas, Z. (1959). Les distributions des notes scolaires comparées eux distributions des résultats obtenus aux tests de connaissances. *Le travail humain, XXII*, 1-2.

Sprietsma, M. (2013). Discrimination in grading: Experimental evidence from primary school teachers. *Empirical Economics, 45*, 523–538. http://dx.doi.org/10.1007/s00181-012-0609-x