American Psychological Association

Journal of Educational Psychology

0022-0663/19/$12.00

2019, Vol. 111, No. 4, 717–735
http://dx.doi.org/10.1037/edu0000307

The Function of Selection of Assessment Leads Evaluators to Artificially Create the Social Class Achievement Gap

Frédérique Autin, Anatolia Batruch, and Fabrizio Butera
Université de Lausanne

To understand the persistent social class achievement gap, researchers have investigated how educational settings affect lower versus higher socioeconomic status (SES) students' performance. We move beyond the question of actual performance to study its assessment by evaluators. We hypothesized that even in the absence of performance differences, assessment's function of selection (i.e., compare, rank, and track students) leads evaluators to create a SES achievement gap. In 2 experiments ($N = 196$; $N = 259$), participants had to assess a test supposedly produced by a high- or a low-SES student, and used assessment for selection (i.e., normative grading) or learning (i.e., formative comments). Results showed that evaluators using assessment for selection found more mistakes if the test was attributed to a low-rather than a high-SES student, a difference reduced in the assessment for learning condition. The third and fourth experiments ($N = 374$; $N = 306$) directly manipulated the function of assessment to investigate whether the production of the social class achievement gap was facilitated by the function of selection to a greater extent than the educational function. Results of Experiment 3 supported this hypothesis. The effect did not reach significance for Experiment 4, but an internal meta-analysis confirmed that assessment used for selection led evaluators to create a SES achievement gap more than assessment used for learning, thereby contributing to the reproduction of social inequalities.

Educational Impact and Implications Statement

Evaluators' knowledge about students' social class can bias their assessment, in favor of privileged students. The present research suggests that assessment in itself does not trigger such a bias, nor are teachers biased in themselves; rather it is the function given to assessment that can trigger or prevent discriminatory assessment. This research found that a social class gap in evaluation appears when assessment is used for selective purposes (i.e., gauging merit and sorting students) to a greater extent that when it is used for educational purposes (i.e., fostering learning). The findings indicate that to ensure equality in educational institutions, closer attention should be paid to the role and meaning of assessment.

Keywords: social class achievement gap, educational institutions, function of selection, evaluator, assessment practices

Supplemental materials: http://dx.doi.org/10.1037/edu0000307.supp

In most industrialized countries, educational institutions have developed with the goal of establishing a fair society in which social positions are ascribed based on individual merit, irrespective of individuals' social class (Bell, 1973; Duru-Bellat, 2006; Turner, 1961). And yet, a wealth of empirical evidence questions the fact that the educational system truly provides equal opportunities and fosters social mobility. For example, international testing such as the Program for International Student Assessment (PISA) consistently shows that, across many countries (65 involved in 2012), low socioeconomic status (SES) students are more likely to underperform compared with high-SES students (OECD, 2006, 2013a). To explain the persistent social class achievement gap,

This article was published Online First September 13, 2018.

Frédérique Autin, Anatolia Batruch, and Fabrizio Butera, Laboratoire de Psychologie Sociale de l'Université de Lausanne (UNILaPS), Institut de Psychologie, Faculté des Sciences Sociales et Politiques, Université de Lausanne.

Frédérique Autin is now at CeRCA, CNRS - Université de Poitiers, France. Anatolia Batruch is now at University of Amsterdam, Amsterdam, the Netherlands.

This work was supported by the Swiss National Science Foundation (Grant CRSII1_141872), and was conducted during FA's postdoctorate at University of Lausanne under the supervision of FB. We thank the members of the Sinergia project "The struggle for competence in academic selection: Social psychological influences on Competence Threat" and Benoit Dompnier for their help with the development of materials and their comments on the research plan and results, and Nicolas Sommet and Susanne Faber for their help during data collection.

Correspondence concerning this article should be addressed to Fabrizio Butera, Laboratoire de Psychologie Sociale (UNILaPS), Institut de Psychologie, Faculté des Sciences Sociales et Politiques, Université de Lausanne - IP-SSP, Géopolis. CH 1015 - Lausanne, Switzerland. E-mail: fabrizio.butera@unil.ch

some scholars pointed to the way educational institutions function (Bourdieu & Passeron, 1977; Croizet, Goudeau, Marot, & Millet, 2017; Stephens, Markus, & Phillips, 2014). A steadily growing research stream in social and educational psychology has investigated how educational settings create a set of barriers that hinder the success of low-SES students while supporting the performance of high-SES students (e.g., stereotype threat, cultural mismatch; Croizet & Claire, 1998; Stephens, Fryberg, Markus, Johnson, & Covarrubias, 2012). In the present article, we propose to move beyond the question of the processes affecting students' performance, and address the processes that contribute to the social class achievement gap via evaluators. We argue that the use of assessment practices with a focus on selection (i.e., compare and rank students to guide them toward different social positions) can lead evaluators to create a social class achievement gap, even in the absence of objective performance differences.

Educational Institutions and Students' Performance

Many analyses of educational institutions suggest that they play a role in perpetuating social inequalities (e.g., Bourdieu & Passeron, 1977; Croizet & Millet, 2012; Fine & Burns, 2003; Stephens, Hamedani, & Destin, 2014). The way institutions operate can be understood as a social product that not only conveys cultural ideas, values, and beliefs (Fiske, Kitayama, Markus, & Nisbett, 1998; Markus & Hamedani, 2007), but also carries traces of power relations between social groups that participate in the creation, maintenance and justification of inequalities (Adams, Biernat, Branscombe, Crandall, & Wrightsman, 2008; Jackman, 1994). Educational institutions have been created around values, norms regarding language use, bodily posture, self models, and forms of knowledge that are close to those of the middle and upper classes (Bourdieu & Passeron, 1977; Croizet et al., 2017; Stephens et al., 2014). One consequence is that students from low status groups suffer harmful effects in these institutions while the experience of individuals from dominant groups is improved (Goudeau & Croizet, 2017; Jury, Smeding, et al., 2017).

In line with these ideas, research has identified a set of characteristics of educational settings that leads low-SES students to underperform and foster the performance of high-SES students. The evaluative dimension of educational settings, by making lower social class students' stereotype of incompetence salient (Cozzarelli, Wilkinson, & Tagler, 2001; Durante & Fiske, 2017), contributes to the SES performance gap (Croizet & Claire, 1998; Croizet & Dutrévis, 2004; Désert, Préaux, & Jund, 2009; Harrison, Stevens, Monty, & Coakley, 2006; Spencer & Castano, 2007). Another line of research argues that the performance gap is fueled by the norms of independence (i.e., express yourself, follow your own path) institutionalized in American universities, that match the middle or upper class students' upbringing, but mismatch the more interdependent socialization of lower class students[1] (i.e., be responsive to others, work with them, and contribute to a community; Stephens et al., 2012, 2014).

These lines of research are important because they document how educational settings are often organized in a way that leads lower SES students to be outperformed by higher SES students (via stereotype threat, cultural mismatch). As a consequence, they mitigate the interpretation of the social class achievement gap in terms of essentialized differences between students of different social class, and pave the way to interventions aimed at reducing the effects of those barriers (see Dittmann & Stephens, 2017; Jury, Darnon, Dompnier, & Butera, 2017). For instance, Harackiewicz et al. (2014) have shown that an intervention asking students to write about their most important values serves as a buffer against social identity threat and reduces the social class achievement gap (see also Tibbetts et al., 2016). These results may lead one to think that if the barriers affecting lower SES students' performance were removed then educational institutions would offer real equality of opportunity. Yet, we propose that even in the absence of actual performance differences, other processes are at work to maintain the social class achievement gap. Thus, we now turn to a set of studies that point to sources of inequalities that go beyond students' performance.

Educational Institutions and Evaluators' Behavior

A parallel line of research has pointed out that teachers' evaluation of performance can be biased by their knowledge of a student's social background (e.g., see Malouff & Thorsteinsson, 2016). For example, Sprietsma (2013) asked German teachers to grade essays of unknown fourth-graders. Typical German or Turkish names were randomly assigned to the same essays. The essays received lower grades when the teachers thought that students with a migrant background, compared with native students, had produced them. It should be noted that students with a migrant background tend to be socioeconomically disadvantaged compared with native students (OECD, 2013a). Rangvid (2015) used large-scale data registers to compare teacher scores and external exam scores. Disparities between these scores indicate bias in teachers grading. The study showed notably that pupils with low-educated parents (an aspect of lower social class backgrounds) receive lower teacher scores than pupils with high-educated parents with similar external scores.

The above results strikingly reveal that, even if actual performance is identical, the outcome of assessment is influenced by the students' social background. However, in these studies, discrimination in grading is usually interpreted as an effect of teachers' bias: They hold prejudiced expectations based on the students' social backgrounds, which affect their behavior, even if this discrimination is not intentional. Without underplaying the impact of expectations, we propose that biased assessment cannot be isolated from the sociocultural context in which this behavior is produced. When assessing, teachers act as agents of an institution that conveys specific values and norms and promote specific practices; thus, we contend that biased assessment can be interpreted as the product of the way educational institutions are structured and operate.

[1] In the various lines of research presented in this section, social class is operationalized in different ways. In some studies, it refers to socioeconomic status (e.g., Croizet & Claire, 1998), whereas in other to sociocultural status (e.g., first/continuing generation to attend University; Stephens et al., 2012). A comparison of the various aspects of social class is beyond the scope of the present work, and has been reviewed by others (Goudeau, Autin, & Croizet, 2017; Kraus, Callaghan, & Ondish, in press). We have referred to these various lines of research without highlighting the differences in operationalization to the extent that in all of these studies social class refers to a social background that relates to more or less chances of success in education.

Two Functions of Educational Institutions

Modernized industrial societies are faced with a paradox: How to reconcile an endorsement of equality of all humans as a fundamental value, with being stratified (i.e., different occupations give unequal access to symbolic and material resources). To find justifiable ways to rank individuals, most Western societies opted for an ascription of social positions based on a characteristic seemingly naturally distributed across individuals: individual merit (Bisseret, 1974; Carson, 2007). Educational institutions became the place where individual differences can be detected, gauged, and certified, to give access to the corresponding social positions. The paradox between equality and stratification is embodied in educational institutions, which are expected to serve two different functions in society: an educational function, ensuring equality of opportunity, and a function of selection, sorting individuals by merit (Darnon, Dompnier, Delmas, Pulfrey, & Butera, 2009; Dornbusch, Glasgow, & Lin, 1996).

Educational Function

Following the Universal Declaration of Human Rights stating that "everyone has the right to education," most Western societies implemented compulsory elementary education and free access to public schools. The unrestricted access to education serves as a safeguard for equality of opportunity. The *educational function* of educational institutions refers to their role in equipping all students with knowledge, skills, and capacities for learning and helping them develop their potential. Having all individuals master basic knowledge and competence ensures that they can all take part in society (Dubet, 2004; Forquin, 1992; Parsons, 1959). Moreover, the democratization of knowledge is expected to expand opportunities and ensure that no talent is wasted; accordingly, the educational function is perceived as promoting social mobility (Bowen, Kurzweil, Tobin, & Pichler, 2005; Duru-Bellat, 2008).

Function of Selection

If mass education offers to all the opportunity to show their potential, then education institutions fulfill a *function of selection* by sorting individuals into different educational paths and ultimately different occupations. The 2012 PISA survey established that in all 65 countries, educational institutions implement some form of selection practices (OECD, 2013b), which include school admission, transfer, grade repetition, tracking into academic or vocational programs, grouping across and within classes, and combinations of these. To illustrate, tracking into different programs occurs in 75% of the countries. Across countries, 43% of the 15-year-old students are in academically selective schools, 75% attend schools that use between-classes ability grouping, and 49% within-classes grouping. The time and rigidity of this stratification varies between countries; nevertheless, all selection practices have consequences for the students' educational trajectory and, at each step of selection, a reduced proportion of the population moves to the most valued tracks.

The different educational lanes are conceptualized as a way to develop the students' potential, meet their needs and assure that they are in the right place (Chmielewski, 2014; LeTendre, Hofer, & Shimizu, 2003). Indeed, the function of selection is intertwined with the meritocratic ideal. Educational institutions are viewed as a neutral context to detect and measure the qualities of students, and select the most deserving (Carson, 2007; Lemann, 1999). Accordingly, educational institutions have been described as a social filter (Arrow, 1973), or sorting machines (Domina, Penner, & Penner, 2017), because academic credentials have become the supposedly fair basis for ascribing positions in the occupational hierarchy.

To fulfill both their functions—education and selection—educational institutions rely notably on assessment. As a consequence, the distinction between different functions exists in the theorization of assessment. Assessment can serve an educational role of promotion of learning, and a social role of estimation of merit, ranking and certification, namely the function of selection (Filer, 2002; Taras, 2005, 2009; Torrance & Pryor, 1998). We contend that beyond the consequences for students' learning, the two functions underlying assessment can have consequences in terms of social inequalities. This contention builds on research focusing on the function of assessment and inequalities in students' performance. Giving a formative framing to evaluation, by stating that critical feedback reveals the teacher's belief in the students' potential, improved low-status students' performance (Yeager et al., 2014). On the contrary, reminding students of the function of selection increased their belief in the utility of outperforming others to succeed at college and consequently increased their endorsement of such a performance-approach goal (Jury, Darnon, et. al., 2017). Endorsing this goal predicted better grades but only for higher social class students (Darnon, Jury, & Aelenei, 2017). More directly related to the present research, assessment presented as a tool for selecting the best students elicited a SES achievement gap on an exam, a gap that was closed when assessment was presented as a tool for learning (Smeding, Darnon, Souchal, Toczek-Capelle, & Butera, 2013). Similarly, simply reminding students of the function of selection of university led lower SES students to underperform compared with higher SES students (Jury, Smeding, & Darnon, 2015). We argue that, as the institutional function of assessment impacts the student's construal of the performance setting, it can likewise influence the evaluator's construal of the assessment setting. We propose that assessment for selection reflects a meritocratic ethos while assessment for learning reflects an egalitarian ethos, which beget different consequences for the reproduction of inequalities. Specifically, we propose that assessment for selection might induce more reproduction of inequalities than assessment for learning.

Assessment for Selection and Inequality

Although all forms of assessment can serve both educational and selection functions, grading may especially be relevant for selective purposes, in that it allows normative assessment. Traditionally, normative assessment, or norm-referenced assessment, is conceived as allowing one to compare the performance of the person being assessed to that of other persons (Glaser, 1963). Normative assessment uses indicators such as numerical grades, letters, percentages, or value judgments (e.g., good, excellent), that can also be used in other assessment methods, but perfectly serve the purpose of normative assessment, namely comparison with a standard and across individuals (Rosenholtz & Simpson, 1984; Thorndike, 1913). These indicators summarize performance in a

number—or a letter, or a judgment—and thereby constitute an easily interpretable criterion of relative success or failure (Butler, 1987; Butler & Nisan, 1986). In industrialized countries, normative grading constitutes the most widely used method of assessment in educational and professional settings (Knight & Yorke, 2003), and is the main basis for admission to schools and programs (OECD, 2013b).

Beyond the institutional role of grading, this form of evaluation is seen as well suited to select students who are most deserving. In fact, research has shown that the more individuals believed that the function of educational institutions is to select the best students, the more they supported the implementation of normative grading; this relationship was mediated by the belief that grading fulfills equity justice principles (Autin, Batruch, & Butera, 2015). Other research showed that both teachers and students believe that grade distribution is fair as long as it follows an equity principle (Jasso & Resh, 2002; Resh, 2009; Sabbagh, Faher-Aladeen, & Resh, 2004). These elements highlight the intertwinement of normative grading, the function of selection and the meritocratic ethos, which assumes that rewards should be allocated equitably, on the basis of individual ability and hard work (Son Hing et al., 2011; Wiederkehr, Bonnot, Krauth-Gruber, & Darnon, 2015). This feature of educational institutions, however, is not without consequences in terms of inequalities.

Contexts emphasizing meritocratic selection elicit psychological and behavioral tendencies to justify and maintain social inequalities. For example, believing in meritocracy decreases perceptions of discrimination in low status groups (McCoy & Major, 2007) and perceptions of privilege in dominant groups (Knowles & Lowery, 2012). Moreover, the perceived violation of meritocratic selection is central in the opposition to social policies that challenge the status quo (Bobocel, Son Hing, Davey, Stanley, & Zanna, 1998; Faniko, Lorenzi-Cioldi, Buschini, & Chatard, 2012; Zdaniuk & Bobocel, 2011). In education, the more students and parents believe in school meritocracy the less they are willing to implement a pedagogical intervention that reduces the SES achievement gap (Darnon, Smeding, & Redersdorff, 2017).

More directly related to the effect of meritocratic assessment on bias in evaluators, Castilla and Benard (2010) found that inducing an organizational culture that emphasizes meritocracy led individuals in a managerial position to favor a male employee over a female employee who achieved similar performance. Closely related to the matter of academic assessment, a recent study had preservice teachers grade a test that was attributed either to a low- or a high-SES student. When the student was presented as being enrolled in a selective program, preservice teachers gave a lower grade to the test attributed to a low-SES student comparatively to a high-SES student. The gap in evaluation was reduced if students were supposedly enrolled in a less selective program (Batruch, Autin, & Butera, 2017). It is important to note that we do not suggest that assessment in itself necessarily produces biased evaluations; instead, we propose that assessment practices that focus on meritocratic selection, such as normative grading, may lead evaluators to reproduce inequalities in their evaluations.

Assessment for Learning and Equality

Alternative forms of assessment have long been developed, including formative assessment (Black & Wiliam, 1998), which can be defined as assessment providing specific and detailed feedback with the goal of adjusting the teaching and learning activities to the students' needs and providing relevant comments on how to overcome difficulties and make progress (Frey & Schmitt, 2007; Sadler, 1989). Formative assessment is often opposed to summative assessment, to the extent that the former intervenes during the learning process and the latter at the end of it (Bloom, Hastings, & Madaus, 1971). However, in the present research we do not focus on the temporal aspects of formative assessment, but on its function, that of providing feedback for learning. In particular, among the various existing kinds of formative assessment, we refer to qualitative feedback that points to specific learning objectives and suggests ways to improve (Bennett, 2011; Shute, 2008). Formative feedback provides useful information to students: what the expected outcome is and guidance on how to attain it (Sadler, 1989); and most importantly, feedback related to the task reduces the focus on social comparison and enhances the focus on the mastery of the task (Bloom, 1968; Butler, 1987). Because of its focus on the development of competence and knowledge, formative assessment is in line with educational institution's educational function.

Equality is central to the rationale for implementing formative assessment. It is presented as a tool to implement a corrective process and reduce the gap between individuals who are unequal before entering school (Crahay, 2012; Perrenoud, 1995). By enabling adjustment to meet the students' needs, formative feedback aims at helping all students to attain a high level of competence, irrespective of their initial abilities. Formative assessment seems to convey the institutional purpose of education centered on equality. Beyond this institutional role, research shows that people's support for formative feedback is related to a principle of corrective justice that ensures equality of outcomes through adjustment to the students' needs (Autin et al., 2015).

More important, promoting equality has positive effects on the treatment of groups. People's endorsement of egalitarian principles relates to lower levels of stereotyping (Moskowitz, Gollwitzer, Wasel, & Schaal, 1999; Moskowitz, Salomon, & Taylor, 2000) and the support for social policies aimed at reducing social inequalities (Zdaniuk & Bobocel, 2011). Invoking the concept of equality also induces a more favorable implicit evaluation of an out-group (Zogmaister, Arcuri, Castelli, & Smith, 2008). Compared with activating meritocratic values, activating egalitarian values reduces the accessibility of negative stereotypes (Wyer, 2003) and elicits more positive attitudes toward a low status group (Katz & Hass, 1988). It also diminishes the extent to which prejudice relates to avoidance activation in response to low status groups (Wyer, 2010). Thus, we propose that assessment practices oriented toward educational purposes, such as formative feedback, may prevent evaluators from reproducing inequalities in their evaluation.

Hypotheses and Overview

In the present research, we argue that institutional practices of assessment constrain the way individuals in a position of evaluator behave toward students from lower or higher social class. In two experiments, we test the hypothesis that, compared with assessment for learning, assessment for selection leads evaluators to create a larger achievement gap that reproduces existing social

inequalities (i.e., low-SES students have a lower performance than high-SES students), even though the actual performance is identical. In a third and fourth experiment, we test that it is indeed the function of assessment, rather than the form of assessment (grade vs. comments), that leads evaluators to differentially evaluate students as a function of their SES.

It should be noted that considering the difficulty to recruit practicing teachers, the studies were conducted with college students playing the role of teachers. Although this is a limitation, we decided to first test our hypotheses and paradigm with an accessible population to be able to conduct well-powered studies. We believe that the long-lasting socialization of the students in the educational institution implies that they are well aware of its functions and practices (see Darnon et al., 2009) and, therefore, able to enact them. After all, we hypothesize that it is the educational system's selective function that should drive the effects. Moreover previous studies has shown similar results with students acting as teachers and actual teachers (Batruch, Autin, Bataillard, & Butera, 2018; Rattan, Good, & Dweck, 2012; Simon, Ditrichs, & Grier, 1995).

Experiment 1

Method

Participants. A total of 220 students from a medium-size French-speaking Swiss university participated in return for a 10 CHF (10.30 USD) gift card. At the time of the study, we used a rule of at least 50 participants per cell to determine sample size (Simmons, Nelson, & Simonsohn, 2011). Data collection stopped at the end of the semester considering that the sample size requirement had been reached. Data from 17 non-French native speakers and 7 participants who failed the manipulation check were excluded. The analyses including participants who failed the check are reported in the online supplemental material. The final sample comprised 196 students (117 women, 79 men, M_{age} = 22.29, SD_{age} = 1.87). Participants were randomly assigned to one of the experimental conditions in the Assessment method (grading vs. formative comments) × Target's SES (low vs. high) between-participants design. The target's sex was also manipulated as a control and was not part of our hypotheses. It should be noted that in Switzerland experiments that do not include physical measures or vulnerable participants do not need permission from an ethical committee and the experiment was conducted in compliance with the American Psychological Association ethical guidelines. Participants were informed about confidentiality and anonymity of data, right to decline and withdraw without consequences, and whom to contact in case of questions.

Material and procedure. Students were approached in university cafeterias by one of two experimenters and asked whether they would take part in a study about assessment tools used by teachers. Participants received a booklet containing instructions about the assessment method, a description of the target (i.e., the student who produced the test) followed by a dictation to be assessed. Participants read a cover story asking them to imagine that they were a French-language teacher in a secondary school, and to assess a dictation test using a specific method.

Manipulation of the assessment method. Instructions were based on the specific properties of the assessment method reviewed above (i.e., grading vs. formative comments). Participants in the *assessment for selection* condition read that, as a teacher, they were to use a method based only on grades. They were to give students grades depending on the number of mistakes they made. The instructions also referred to the normative aspect of this assessment, that relates to the social function of certification and ranking (cf. Taras, 2009, 2005). Participants read that this method allows checking the student's level and whether he or she met the requirements. They also read that this method allows assessment of the students' learning, their standing compared with a norm that defines success and compared with the other students. This explanation was illustrated with an example of a math test graded with this method.

Participants in the *assessment for learning* condition read that they were to use a method based on formative comments only. They were to make comments to help the students learn from their mistakes. Instruction referred to the educational role of assessment. Participants read that this method explains to students how to improve and to adapt to learning situations. They also read that such method allows assessment of the students' learning and their distance from the learning goals, and propose them strategies to meet these goals. This description was illustrated with an example of a math test corrected with this method.

Manipulation of the target's SES. After reading about the assessment method, participants were presented with information about a student allegedly belonging to their class. Participants saw the student file (similar to the official student file in use) and a brief description of his or her extracurricular activities. Relevant information about the target's SES were presented among neutral information (e.g., date of birth, address, nationality—all targets were presented as Swiss). SES was manipulated via a series of indicators. The student's first name was manipulated using stereotypical names of higher-versus lower-SES girls and boys (e.g., "Louis" for a high-SES boy, "Brian" for a low-SES boy, "Charlotte" for a high-SES girl, and "Cindy" for a low-SES girl), based on Coulmont's (2011) work on the sociology of first names. Moreover, parental occupation (mother: director of marketing vs. waitress; father: architect vs. construction workman), number of siblings (1 vs. 4) and extracurricular activities (e.g., local amusement park vs. traveling to London) were also manipulated. Sex was manipulated through the student's first name and reported sex.

Dictation test. After reading the relevant information about the target, participants had to correct a dictation test. They were asked to first underline all the mistakes. Then, in the *assessment for selection* condition, participants had to give a grade in line with common practice in Swiss schools, that is, from 1 to 6, with higher numbers indicating better performance. In the *assessment for learning* condition, participants had to write a comment next to each mistake to explain the student what mistake he or she did and how to improve. The test contained 11 obvious mistakes (wrong spelling, wrong verb conjugation, and wrong name-adjective agreement) and 6 ambiguous mistakes (two possible conjugations or spellings).[2]

[2] The booklet also included a questionnaire measuring the predicted future success of the target and the perception of the assessment method and of academic performance. These measures are not relevant for the hypothesis presented here, and we did not report the results, but they are available upon request from the authors.

The booklet ended with some manipulation check items—two asking to report information presented in the description of the target and one asking to rate the socioeconomic background of the target (from 1 = *highly disadvantaged* to 7 = *highly advantaged*)—as well as sociodemographic questions, including self-reported GPA. Finally, participants were thanked and debriefed. As we anticipated that the use of formative comments would take more time than the use of grading, we recorded the time that participants took to complete the study.

Results

Perceived SES. To determine whether the description of the target affected participants' perception of his or her SES, we analyzed participants' rating of the target's socioeconomic background with the full sample except for nonnative speakers. The regression included the Assessment Method (assessment for selection coded −0.5, assessment for learning coded 0.5), Target's SES (low-SES coded −0.5, high-SES coded 0.5) and the interaction as predictors.[3] As expected, a main effect of Target's SES was obtained indicating that low-SES targets were perceived as coming from a more disadvantaged background ($M = 3.84$, $SD = .87$, 95% confidence interval, CI [3.67, 4.01]) than high-SES targets ($M = 5.92$, $SD = .75$, 95% CI [5.77, 6.07]), $b = 2.07$, 95% CI [1.85, 2.31], $t(197) = 18.14$, $p < .001$, $\eta_p^2 = .63$, 95% CI [.55, .68]). The main effect of Assessment Method and the interaction did not reach significance, respectively, $b = 0.02$, $t(197) = 0.20$, $p = .84$ and $b = 0.27$, $t(197) = 1.19$, $p = .23$. As the vast majority of participants perceived the targets as belonging to the expected social class, we decided to exclude the seven participants mentioned in the Participants section because they either perceived a low-SES target's socioeconomic background as "advantaged," or a high-SES target's socioeconomic background as "average" or "disadvantaged."

Number of mistakes. A preliminary analysis revealed that participants took more time to complete the study when they had to use formative comments ($M = 22.00$, $SD = 6.04$, 95% CI [20.81, 23.19]), compared with grading ($M = 14.94$, $SD = 4.30$, 95% CI [14.07, 15.82]), $b = 7.09$, 95% CI [5.59, 8.58], $t(192) = 9.37$, $p < .001$, $\eta_p^2 = .31$, 95% CI [.21, .41]. Time was not affected by the target's SES or interactions between SES and Assessment Method, respectively, $b = 0.59$, $t(192) = 0.78$, $p = .43$ and $b = -0.25$, $t(192) = -0.16$, $p = .87$. We decided to control for time in the analysis of the number of mistakes detected in the dictation test, because the time needed to assess a test could affect the number of mistakes found, but is not a variable of interest here. Following the recommendations for the inclusion of covariates, we tested for possible interactions between the covariate Time and the predictors (Judd, McClelland, & Ryan, 2011). We performed a regression analysis on the total number of mistakes with Assessment Method (assessment for selection coded −0.5, assessment for learning coded 0.5), Target's SES (low-SES coded −0.5, high-SES coded 0.5), Time (centered) and all interaction terms as predictors.[4]

Results showed a main effect of Time, with participants detecting more mistakes as they took more time to complete the study, $b = 0.17$, 95% CI [0.11, 0.24], $t(185) = 5.25$, $p < .001$, $\eta_p^2 = .13$, 95% CI [.05, .22]. The main effect of Assessment Method also reached significance indicating that participants detected more mistakes when using assessment for selection ($M = 10.16$, $SD = 2.53$, 95% CI [9.57, 10.75]) than assessment for learning ($M = 8.61$, $SD = 2.18$, 95% CI [8.11, 9.10]), $b = -1.55$, 95% CI [−2.32, −0.78], $t(185) = -3.97$, $p < .001$, $\eta_p^2 = .07$, 95% CI [.02, .16]. The target's SES also affected the number of mistakes such that participants found more mistakes in the dictation of low-SES students ($M = 9.90$, $SD = 2.54$, 95% CI [9.33, 10.47]) than in that of high-SES students ($M = 8.87$, $SD = 2.16$, 95% CI [8.34, 9.39]), $b = -1.03$, 95% CI [−1.81, −0.26], $t(185) = -2.64$, $p = .009$, $\eta_p^2 = .03$, 95% CI [.00, .10]. Time interacted with Assessment Method, $b = -0.17$, 95% CI [−0.29, −0.04], $t(185) = -2.55$, $p = .01$, $\eta_p^2 = .03$, 95% CI [.00, .10] and with the Target's SES, $b = -0.15$, 95% CI [−0.28, −0.02], $t(185) = -2.33$, $p = .02$, $\eta_p^2 = .03$, 95% CI [.00, .09]. The positive relationship between time and the number of mistakes was stronger in the grading compared with the formative comments condition and for low-SES students compared with high-SES students. The expected interaction between SES and method was not significant, $b = 0.92$, 95% CI [−0.62, 2.47], $t(185) = 1.18$, $p = .24$, $\eta_p^2 = .01$, 95% CI [.00, .05], Cohen's $d = .19$, but in the expected direction, suggesting a greater gap in the number of mistakes between low and high SES students in the assessment for selection condition than in the assessment for learning condition.

However, these effects were qualified by a three-way interaction between Time, Assessment Method and Target's SES, $b = 0.46$, 95% CI [0.20, 0.72], $t(185) = 3.52$, $p < .001$, $\eta_p^2 = .06$, 95% CI [.01, .14]. This interaction, depicted in Figure 1, was unexpected but made sense given the effect of time and was, therefore, decomposed by assessment method. In the assessment for selection condition, the positive relationship between time and the number of mistakes was significantly stronger for low-SES targets than for high-SES targets, $b = -0.38$, 95% CI [−0.60, −0.17], $t(185) = -3.50$, $p < .001$, $\eta_p^2 = .06$, 95% CI [.01, .14]. In other words, the more participants spent time assessing a dictation test with grading, the more they found mistakes, especially if the target was from a low socioeconomic background. As a result, participants who took a moderate and long time to complete the study in the assessment for selection condition found on average, respectively, 1.50 (95% CI for b [−2.68, −0.31]), and 3.90 (95% CI for b [−6.22, −1.59]) more mistakes in the dictation of a low-SES student than in the dictation of a high-SES student, respectively, $t(185) = -2.50$, $p = .01$, $\eta_p^2 = .03$, 95% CI [.00, .10], and $t(185) = -3.33$, $p = .001$, $\eta_p^2 = .06$, 95% CI [.01, .13]. In the assessment for learning condition, the positive relationship between time and the number of mistakes did not significantly differ as a function of Target's SES, $b = 0.08$, 95% CI [−0.07, 0.22], $t(185) = 1.07$, $p = .28$, $\eta_p^2 = .01$, 95% CI [.00, .05].

Supplementary analyses.

Grades. To better understand the mechanisms at work in the assessment for selection condition, we analyzed how mistakes affected the participants' grading of the test (given the design, grades were only available for the assessment for selection condition). Grades were analyzed in a regression with the Target's SES

[3] Two values are missing because participants did not fill the item.

[4] Three outliers were excluded from the analysis because of uncommon deleted studentized residual, centered leverage values, and abnormal residuals.

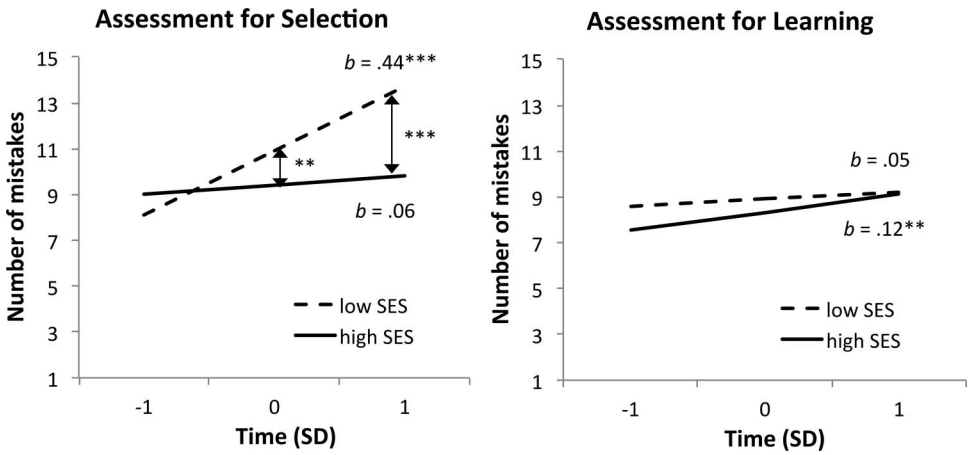

Figure 1. Experiment 1. Relationship between the time taken to complete the study and the number of mistakes found in the dictation as a function of target's socioeconomic status (SES) and assessment method. ** $p \leq .01$. *** $p < .001$.

(low-SES coded −0.5, high-SES coded 0.5), the number of mistakes (centered) and the interaction term as predictors.[5] The analysis revealed a main effect of the number of mistakes, such that the more mistakes were detected the lower the grade, $b = -0.14$, 95% CI [−0.18, −0.09], $t(87) = -6.38$, $p < .001$, $\eta_p^2 = .32$, 95% CI [.17, .45]. The main effect of SES was not significant, $b = 0.22$, 95% CI [−0.02, 0.45], $t(87) = 1.85$, $p = .07$, $\eta_p^2 = .04$, 95% CI [.00, .14]. The interaction between the number of mistakes and the target's SES was significant, $b = 0.09$, 95% CI [0.009, 0.19], $t(87) = 2.19$, $p = .03$, $\eta_p^2 = .05$, 95% CI [.00, .16]. As can be seen in Figure 2, the negative relationship between the number of mistakes and the grade was stronger for low-SES targets, $b = -0.19$, 95% CI [−0.25, −0.13], $t(87) = -6.07$, $p < .001$, $\eta_p^2 = .29$, 95% CI [.15, .43], compared with high-SES targets, $b = -0.09$, 95% CI [−0.15, −0.03], $t(87) = -2.96$, $p = .004$, $\eta_p^2 = .09$, 95% CI [.01, .22]. This suggests that mistakes led to a more negative evaluation when they were produced by low-SES students. More important, this negative evaluation resulted in more low-SES students performing below the passing grade (4, in Swiss schools). In the sample, we observed that 32.5% of the low-SES targets received a grade lower than 4 but this proportion dropped to 16.6% for the high-SES targets.

Impact of participants' characteristics. We tested whether participants' own characteristics could account for or moderate the results observed on the number of mistakes. We looked at the effect of participants' level of competence, gauged by self-reported GPA, and their own social class, indicated by whether at least one of their parents has a college degree ("continuing generation") or not ("first generation"). The analyses revealed no evidence that including the participants' level of competence or their social class moderated the observed results (see online supplemental material).

Discussion

This first experiment was designed to test the hypothesis that assessment for selection, more than assessment for learning, would lead evaluators to reproduce existing social inequalities and find lower performance for low-SES students than for high-SES students, even though the actual performance was identical. The target's SES × Assessment Method interaction that tested this hypothesis was not significant, although in the expected direction. Participants found on average 1.49 more mistakes in the low-SES test than in the high-SES test when using assessment for selection, a SES performance gap reduced to .57 in the assessment for learning condition. The size of this effect is small ($\eta_p^2 = .01$, Cohen's $d = .19$) but we believe it should not be disregarded. Indeed, students in the member countries of the Organisation for Economic Cooperation and Development (OECD, e.g., United States, Australia, Latvia, Korea, Germany, and Mexico) receive on average 9 years of compulsory education and can expect to receive 17 years of study over their lifetime (OECD, 2006). During this time, assessment is a frequent and important part of the students' experience so small biases could have a large impact in the long run.

The significant Time × SES × Assessment Method interaction, although unexpected, indicates that the hypothesized creation of an SES performance gap by participants using assessment for selection is stronger as participants spend more time on the study. We interpret this effect as a consequence of the participants' engagement in the study. It is possible that those who quickly completed the study paid less attention to the instructions and the test and were then less affected by the manipulations. The findings observed among those who spend more time on the study are in line with the idea that evaluators asked to use a traditional, normative form of assessment, artificially produce a performance gap that corresponds to the existing status asymmetry more than evaluators who use a form of assessment more oriented toward learning.

It is interesting that, when using assessment suited for selection, the mistakes produced by low-SES students were judged in a more punitive way, as indicated by an average decrease in grades of .19 points for every mistake made whereas making a mistake resulted in a loss of .09 points for high-SES students. Ultimately, we

[5] One outlier was excluded from the analysis because of uncommon Cook's distance and deleted studentized residuals.

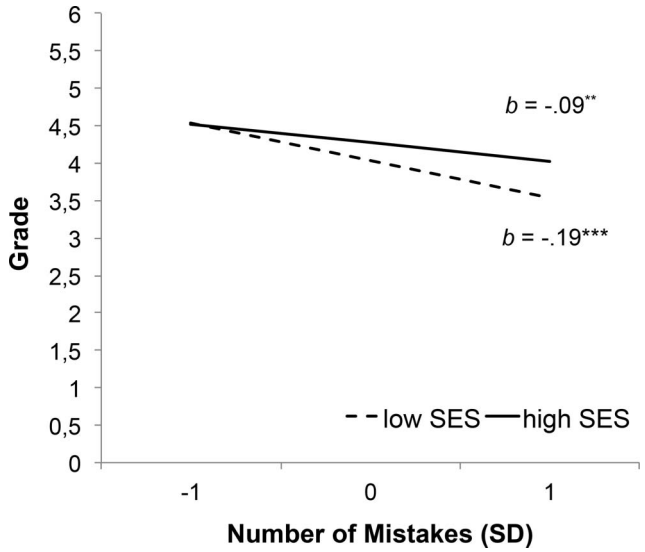

Figure 2. Experiment 1. Relationship between the number of mistakes and the grade as a function of target's socioeconomic status (SES). ** $p \leq .01$. *** $p < .001$.

observed a rate of low-SES students below the pass threshold two times higher than the rate of high-SES students. This effect is consistent with previous research showing that evaluators can redefine their assessment criteria (i.e., what is a weakness or a strength) in a way that justifies discriminatory decisions (Norton, Vandello, & Darley, 2004; Uhlmann & Cohen, 2005). For our participants who used normative grading, mistakes became less of a weakness when produced by a high-than a low-SES student. This finding further supports the idea that the practice of grading may lead evaluators to restrain the success of low status students (see also Batruch et al., 2017).

Supplementary analyses considered the impact of the participants' competence (i.e., self-reported GPA) and social class. In all cases, the interaction between target's SES, assessment method and time remained significant and was not further moderated. This rules out the idea that variations in competence could explain the results. Moreover, participants' own social class did not affect their behavior toward the target, which suggests that the bias against the lower SES students does not reflect an intergroup bias (Hewstone, Rubin, & Willis, 2002). Thus, it seems that our work does not fall in the scope of intergroup feedback. This literature showed that evaluators from a majority group do not communicate the same praise and criticism to majority and minority students (Croft & Schmader, 2012; Crosby & Monin, 2007). In our research, we hypothesized that participants endorsed their role of agents of the educational institution and acted as such, beyond their own social identity. The results supported this contention.

The unexpected interaction with the time spent on the task raises questions. The longer time needed to use formative assessment might indicate the greater cognitive and motivational costs of such a method, which requires one to identify and explain in simple words the rules underlying each mistake and to think of ways to improve. It remains possible that, even after accounting for time, the cost of formative assessment contributed to the lower number of mistakes found by participants using this method. To rule out this interpretation we conducted a second study in which we equalized the motivational and cognitive costs of the two assessment methods.

Experiment 2

Method

Participants. A total of 269 students from a medium-size French-speaking Swiss university voluntarily took part in the study. Data collection stopped at the end of the semester considering that we achieved the minimum of 50 participants per cell of the 2 (SES) × 2 (Assessment Method) design. Data from 10 participants were excluded because they were suspicious ($N = 6$) or failed the manipulation checks ($N = 4$; see online supplemental material for analysis with the full sample). The final sample consisted of 163 women, 93 men, 3 unspecified ($M_{age} = 21.55$, $SD_{age} = 2.35$). Each participant was randomly assigned to one of the experimental conditions in the Assessment method (for selection vs. for learning) × Target's SES (low vs. high) between-participants design. The target's sex was also manipulated as a control and was not part of our hypotheses.

Material and procedure. Students were approached in university cafeterias by the experimenter and asked to take part in a study about assessment tools used by teachers. The procedure was similar to the one followed in Experiment 1. The main difference was in the instructions about assessment. Participants in the *assessment for selection* condition read that they would have to give a grade, while participants in the *assessment for learning* condition would have to write formative comments. However, to equalize the motivational and cognitive costs of the two assessment methods, and hopefully time spent on the task, participants were asked to only underline the mistakes "for the time being," and told that they will be asked to give the grade or write the comments at a later stage (but were actually never asked to do so). The dictation test was a slightly modified version of the one used in Experiment 1 and contained 14 obvious mistakes (thus, three additional mistakes as compared with Experiment 1) and 6 ambiguous mistakes. At the end of this task, participants moved on to the following pages of the booklet. In the last section, participants answered the manipulation checks and the sociodemographic questions.[6] Finally, participants were thanked and debriefed.

Results

Perceived SES. Participants' perception of the target's SES was analyzed in a regression with Assessment Method (assessment for selection coded −0.5, assessment for learning coded 0.5), Target's SES (low-SES coded −0.5, high-SES coded 0.5), and the interaction term as predictors. The analysis was run on the sample that excluded suspicious participants ($N = 263$, but 1 missing value). As expected, Target's SES had a main effect on ratings, $b = 2.16$, 95% CI [1.97, 2.34], $t(258) = 22.94$, $p < .001$, $\eta_p^2 = .67$,

[6] The booklet also included a questionnaire measuring the predicted future success of the target, the rating of the test and tracking recommendations. These measures are not relevant for the hypothesis presented here, and we did not report the results, but they are available upon request from the authors.

95% CI [.61, .72]. The low-SES targets were perceived as coming from a more disadvantaged background (M = 3.87, SD = .78, 95% CI [3.74, 4.00]) than the high-SES targets (M = 6.03, SD = .75, 95% CI [5.90, 6.16]). The main effect of Assessment Method and the interaction between Assessment Method and Target's SES did not reach significance (b = −0.13, $t(258)$ = −1.35, p = .18 and b = 0.07, $t(258)$ = 0.38, p = .71). We excluded four participants, mentioned in the Participant section, who did not properly perceive the target's SES.

Time to complete the study. An important goal of this study was to test the hypothesis without the methodological problem related to the difference in time needed to perform the two types of assessments that we observed in Experiment 1. We analyzed the time taken by participants to complete the study in a regression including Assessment Method, Target's SES and the interaction as predictors.[7] The analysis indicated that participants took a similar amount of time when they used grades (M = 13.51, SD = 3.29, 95% CI [12.93, 14.09]) and formative comments (M = 13.15, SD = 3.92, 95% CI [12.47, 13.83]), b = −0.36, $t(254)$ = −0.78, p = .43. No SES main effect or interaction reached significance, b = −0.24, $t(254)$ = −0.54, p = .59 and b = 0.44, $t(254)$ = 0.48, p = .63.

Number of mistakes. We analyzed the number of mistakes detected by the participants in the dictation test in a regression with Assessment Method, Target's SES and the interaction as predictors. Results showed no main effect of Assessment Method, b = −0.01, 95% CI [−0.67, 0.65], $t(255)$ = −0.03, p = .98, η_p^2 = .00, 95% CI [.00, .00] and a main effect of the Target's SES, indicating that again participants detected more mistakes in the dictation of low-SES targets (M = 12.09, SD = 2.66, 95% CI [11.62, 12.55]) than in that of high-SES targets (M = 11.29, SD = 2.674, 95% CI [10.83, 11.75]), b = −.80, 95% CI [−1.46, −0.14], $t(255)$ = −2.39, p = .02, η_p^2 = .02, 95% CI [.00, .07]. The predicted Target's SES × Assessment Method interaction was significant, b = 1.37, 95% CI [0.06, 2.68], $t(255)$ = 2.05, p = .04, η_p^2 = .02, 95% CI [.00, .06], Cohen's d = .26. As shown in Figure 3, when participants used assessment for selection, they found a greater number of mistakes in the dictation attributed to low-SES students (M = 12.43, SD = 2.86, 95% CI [11.77, 13.09]) compared with high-SES students (M = 10.95, SD = 2.69, 95% CI [10.29, 11.61]), b = −1.48, 95% CI [−2.42, −0.55], $t(255)$ = −3.12, p = .002, η_p^2 = .04, 95% CI [.01, .09]. This difference was not significant when participants used assessment for learning ($M_{\text{low-SES}}$ = 11.74, $SD_{\text{low-SES}}$ = 2.41, 95% CI [11.09, 12.39]; $M_{\text{high-SES}}$ = 11.63, $SD_{\text{high-SES}}$ = 2.77, 95% CI [10.97, 12.28]), b = −0.11, 95% CI [−1.04, 0.81], $t(255)$ = −0.23, p = .81, η_p^2 = .00, 95% CI [.00, .02]. The participants' self-reported GPA and social class did not account for, or moderate these results (see online supplemental material for detailed analyses).

Discussion

This study intended to test our hypothesis without the interference of the effect of time. To do so, we asked all participants to underline the mistakes in the dictation test, and only mentioned that they would write the formative comments or provide a grade (depending on the condition) at a later stage. Results showed that after equalizing the motivational and cognitive costs of assessment, participants took approximately the same amount of time to complete the task, regardless of the condition. As in Experiment 1, participants found a greater number of mistakes in the dictation tests of low- as compared with high-SES students. More important, the predicted Assessment Method × Target's SES interaction was significant. As expected, a significant social class achievement gap was artificially produced by participants who used assessment for selection, but not by participants who used assessment for learning. Participants who used assessment for selection reported on average 1.48 more mistakes in the test of low-SES students than in the test of high-SES students. Again, the participants' characteristics did not moderate this effect.

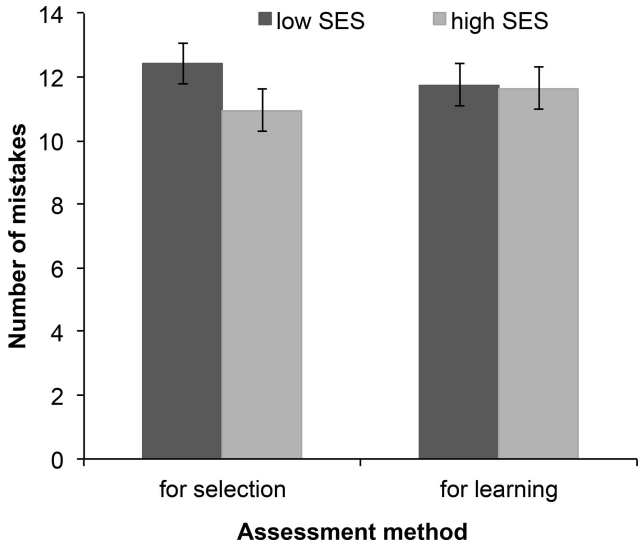

Figure 3. Experiment 2. Number of mistakes found in the dictation as a function of the target's socioeconomic status (SES) and the assessment method. Error bars represent the 95% confidence intervals.

In the theoretical development of our hypothesis, we argued that the impact of normative grading on the creation of social class inequalities is because of the fact that this method epitomizes the function of selection of educational institutions. This assumption was grounded in research showing that, from the perspective of students, assessment oriented toward selection triggered a greater SES performance gap than assessment for learning (Smeding et al., 2013) and that, from the perspective of evaluators, adherence to the function of selection related to more support for grading (Autin et al., 2015). We conducted a third experiment to directly test the hypothesis that the selective purposes of assessment, usually conveyed by normative grading, is indeed what underlies evaluators' tendency to artificially produce performance differences, whatever the form of assessment.

Experiment 3

To test this hypothesis, we manipulated the function of assessment, to induce either selective or educational purposes. We expected that evaluators would create a social class achievement gap when assessment is framed as a way to select the best students more than when it is presented as a tool to improve learning. This hypothesis could lead

[7] One missing value.

to two possible effects: The function of selection may potentiate the effect of normative grading in the creation of the social class gap and the educational function may potentiate the egalitarian effect of formative comments. In this case, a Function of Assessment × Assessment Method × Target's SES interaction should emerge. However, because we conceptualize assessment methods as tools to fulfill a specific institutional purpose it is also possible that the function of assessment overrides the effect of assessment tools (i.e., grading vs. formative comments). In this case, a Function of Assessment × Target's SES interaction should emerge.

Method

Participants. A total of 501 students from a medium-size French-speaking Swiss university voluntarily took part in the study in the cafeterias on campus or in class (data were collected in several classes, resulting in a field-related diversity similar to the data collected in cafeterias). We decided to double the sample size, and data collection was contingent on the classes we had access to and the end of the semester. Each participant was randomly assigned to one of the experimental conditions in the Assessment Method (assessment for selection vs. assessment for learning) × Target's SES (low vs. high) × Function of Assessment (selection vs. education) between-participants design. To avoid increasing the complexity of the experimental design, and as target's sex was not a variable of interest, this factor was not included in the present design; we only used boys as targets. Data from 10 participants were excluded because they expressed suspicion, were unable to assess the test or were not students. Data from 117 participants were excluded because they failed the manipulation checks regarding the Target's SES ($N = 37$), the Function of Assessment ($N = 74$) or both ($N = 6$; see online supplemental material for analysis with the full sample). The number of participants per condition ranged from 40 to 52. We believe this high number of failures can be explained by the lack of involvement from students in the collective sessions in class, a recruitment method that we did not use in the two previous experiments. Even though they were explicitly asked to carefully read the instructions, failure on the manipulations checks indicate that they did not read the main instructions or that they refused to comply with them. More important, including participants who were unable to accurately report the function of assessment might prevent us from properly testing the main hypothesis of this study: the underlying role of this structural factor in the creation of a SES gap. We considered that the validity of these data was questionable and that including them would increase noise (Oppenheimer, Meyvis, & Davidenko, 2009). The final sample of 374 students consisted of 212 female, 153 male, 9 unspecified ($M_{age} = 22.39$, $SD_{age} = 2.65$).

Material and procedure. The procedure was similar to the one followed in Experiment 2. After reading about the assessment method they have to use, and the profile of the target, participants were required to assess the dictation test. At the top of the page containing the dictation test, we presented a reminder of the assessment method and the specific instructions about the function of the assessment. Participants in the *selection* condition read that the mistakes they would find in the test would eventually help them to decide whether the student should move to next grade or not. Participants in the *education* condition read that the mistakes they would find in the test would help them to propose learning strategies that allow the student to improve. After they underlined the mistakes in the test, participants answered the manipulation checks and the sociodemographic questions.[8] Finally, participants were thanked and debriefed.

Results

Perceived SES. Perception of the students' socioeconomic background was analyzed in a regression with Assessment Method (assessment for selection coded −0.5, assessment for learning coded 0.5), Target's SES (low-SES coded −0.5, high-SES coded 0.5), Function of Assessment (selection coded −0.5, education coded 0.5), and all interactions as predictors. The analysis was conducted on the sample of nonsuspicious participants ($N = 491$, but three missing values). The Target's SES influenced the perception of socioeconomic background in the expected direction ($M_{low\ SES} = 3.88$, $SD_{low\ SES} = .92$, 95% CI [3.77, 3.99]; $M_{high\ SES} = 5.97$, $SD_{high\ SES} = .79$, 95% CI [5.86, 6.08]), $b = 2.09$, 95% CI [1.93, 2.24], $t(480) = 26.49$, $p < .001$, $\eta_p^2 = .59$, 95% CI [.54, .64]. No other effects reached significance ($ts < 1.60$, $ps > .11$). Among the participants excluded, 37 were taken out from the final sample because they did not correctly report the target's socioeconomic background.

Number of mistakes. The number of mistakes found in the test was analyzed in a regression with Assessment Method, Target's SES, Function of Assessment and all interactions as predictors.[9] The analysis revealed a significant interaction between the Function of Assessment and the Target's SES, $b = 1.34$, 95% CI [0.12, 2.56], $t(365) = 2.16$, $p = .03$, $\eta_p^2 = .01$, 95% CI [.00, .04], Cohen's $d = .20$. As shown in Figure 4, when participants thought the assessment was aimed at selecting the students, they found more mistakes in the test of a low-SES student ($M = 12.71$, $SD = 2.75$, 95% CI [12.07, 13.35]) than in the test of a high-SES student ($M = 11.59$, $SD = 2.94$, 95% CI [10.98, 12.20]), $b = -1.12$, 95% CI [−2.00, −0.24], $t(365) = -2.50$, $p = .01$, $\eta_p^2 = .02$, 95% CI [.00, .05]. This social class gap was not significant when the assessment was presented with an educational purpose ($M_{low\ SES} = 11.44$, $SD_{low\ SES} = 3.35$, 95% CI [10.84, 12.05]; $M_{high\ SES} = 11.66$, $SD_{high\ SES} = 2.84$, 95% CI [11.08, 12.14]), $b = 0.22$, 95% CI [−0.62, 1.06], $t(365) = 0.51$, $p = .61$, $\eta_p^2 = .00$, 95% CI [.00, .02]. The three-way interaction between Assessment Method, Function of Assessment and Target's SES did not reach significance, $b = 1.72$, 95% CI [−0.72, 4.15], $t(365) = 1.38$, $p = .17$, $\eta_p^2 = .01$, 95% CI [.00, .03]. These results were not impacted or further moderated by the participant's level of competence or social class (see online supplemental material).

Discussion

This study sought to test the hypothesis that the selective (rather that educational) role of assessment is the mechanism that leads evaluators to create a performance gap that corresponds to existing

[8] The booklet also included a questionnaire measuring the predicted future success of the target and the rating of the test. These measures are not relevant for the hypothesis presented here, and we did not report the results, but they are available upon request from the authors.

[9] One outlier was excluded because of uncommon deleted studentized residual.

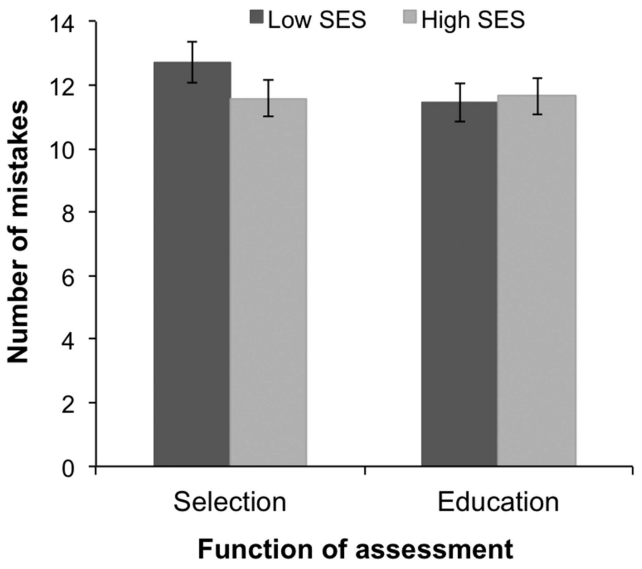

Figure 4. Experiment 3. Number of mistakes found in the dictation as a function of the target's socioeconomic status (SES) and the function of assessment. Error bars represent the 95% confidence intervals.

status asymmetries in the absence of actual differences in performance. Results supported our hypothesis: The Function of Assessment × Target's SES interaction revealed that when assessment was presented as a way to select the best students, evaluators found on average 1.12 more mistakes in a dictation supposedly produced by a low-SES student than in that attributed to a high-SES student. This artificial performance gap was reduced when assessment was presented as a way to help students improve. The results did not show a significant moderating effect of the assessment method (i.e., assessment for selection vs. learning). These findings suggest that it is not so much normative grading and formative comments per se that lead evaluators to, respectively, create or not a social class performance gap, but rather the function attributed to the assessment tools.

However, a high number of participants had to be excluded in particular for not properly reporting the function of assessment. This raises concerns about the design and suggests a possible conflict between instructions about the assessment method and the function of assessment. For example, it might have been difficult for participants to understand or comply with the instruction of both looking for mistakes to help improve learning and decide whether the student should move to the next grade. This is actually not surprising if we consider that the function of selection is positively associated with support for grading whereas negatively associated with support for formative comments (Autin et al., 2015). Contradicting the usual associations might have led to unpredictable consequences on the participants' behavior.

Therefore, we designed a fourth study to test the hypothesis that it is the function of assessment that triggers or not the creation of a SES performance gap. To avoid confusion between the assessment tools and their functions, we kept the assessment tool constant. Because of the link between normative grading and a selective function and between formative comments and educational functions, we decided not to use these assessment methods. Rather, we relied on a less common procedure based on the highlighting of sections of the student's work. Participants had to highlight in two different colors the positive and negative aspects of an essay (see Croft & Schmader, 2012, for a similar design). The fourth study aims to replicate the previously observed findings on a different measure of evaluation.

Moreover, this assessment tool provides information about both positive and negative feedback, which could shed light on how the SES performance gap is created. Indeed, in the previous studies, we observed the creation of a difference between the low- and high-SES student but could not definitely determine whether this difference resulted from negative behavior against the low-SES student or advantage given to the high-SES student. Indeed, inequalities were traditionally framed as the product of discrimination, bias against low status groups but they actually also result from favoritism, bias for high status groups (e.g., Adams et al., 2008). Some even argue that in societies where hostility toward low status groups and intergroup conflicts are not acceptable, favoritism is more prevalent (Brewer, 1999; DiTomaso, 2015; Greenwald & Pettigrew, 2014). Disentangling the processes at play behind the creation of the social class achievement gap is not the central question addressed in the present paper, yet we could expect that when evaluating an essay with a selective purpose, if participants discriminate against low-SES students, they will provide more negative feedback to this student compared with a high-SES student. If participants favor high-SES students, they might provide more positive feedback to this student compared with a low-SES student. When assessment is used with an educational purpose, these differences should be attenuated.

Finally, the fourth study investigated a potential conflation between the educational versus selection function of assessment and a growth versus fixed mindset. A sizable literature has shown that individuals can adopt a growth mindset that refers to the belief that one's qualities are malleable and expandable through learning, or a fixed mindset that corresponds to the belief that qualities are unchangeable (Dweck, 2012). Rattan et al. (2012) showed that instructors with a fixed theory of intelligence, compared with a malleable theory, attribute low ability to low-performing students and give them less engaging feedback. Because the induction of the educational function focuses evaluators on improvement and learning, it might be associated with a growth mindset. Conversely, the function of selection focuses evaluators on the student's stance relative to the requirement and might relate to a fixed mindset. We included a measure of the evaluators' perception of the malleability of students' intelligence to test whether the function of assessment affects their mindset.

Experiment 4

Method

Participants. A total of 335 students in a French university participated in the study, in exchange for course credit ($N = 227$) or were recruited in a university library ($N = 108$). We aimed for at least 50 participants per cell; as we anticipated attrition, we oversampled. Twenty-eight participants were excluded for not being able to report the function of assessment ($N = 7$), the SES of the target ($N = 19$) or both ($N = 3$; see online supplemental material for analyses on the full sample). The final 306 participants

(246 women, 57 men, 3 unspecified, M_{age} = 19.69, SD = 3.80) were randomly assigned to one of the experimental conditions in the Function of Assessment (selection vs. education) × Target's SES (low vs. high) between-participants design.

Material and procedure. Participants had to imagine that they were a history teacher who has to assess an essay produced by an eighth grade male student using a new assessment tool. They had to highlight in one color (yellow) the parts of the essay that were well written (i.e., clear, logical, and important for the structure of the text) and in another color (orange) the parts that needed to be revised (i.e., unclear, misplaced regarding the logical organization of the text, spelling, syntax, or grammar errors). An example was provided.

Manipulation of the function of assessment. To emphasize the *function of selection*, half of the participants read that their evaluation of the student's skills counted toward his semester GPA. Evaluating the essay would give them information to decide whether the student should move to the next grade or not by identifying his strengths and weaknesses in this kind of exercise. To focus the other half of the participants on the *educational function*, they read that their evaluation of the student's skills was part of a learning program. Evaluating the essay would give them information to help the student improve his learning by identifying strategies to make progress in this kind of exercise.

Manipulation of the student's SES. Participants were then asked to read the file of the student who supposedly produced the essay. The files were similar to the ones used in Experiment 3 but adapted to the French context. The target's SES was manipulated by changing the student's name, parental occupation, and extracurricular activities.

Implicit theories of intelligence. Eight items were adapted from Souchal and Toczek (2010) to measure participants' conception of students' intelligence. Four items referred to an entity theory (e.g., "Students have a certain level of intelligence and no matter what they do, it cannot change") and four to an incremental theory (e.g., "Students' intelligence grows with every new experience they live"). A factor analysis revealed one factor (value = 2.69, 33.6% of explained variance) including the four entity items and two incremental items (reversed). We then computed a score of entity by averaging the scores on these items (Cronbach's α = .78).

After completing the questionnaire, participants assessed the essay. They were briefly reminded of the Function of Assessment and of the instructions regarding the use of highlighters to provide positive and negative feedback. The essay was a picture of a 20 line-long handwritten text inspired from actual essays. The number of characters highlighted in each color was computed as indicators of the quantity of positive (yellow) and negative (orange) feedback.

After the assessment of the essay,[10] participants reported the function of the assessment, two types of information presented in the student file and estimated his background on a 7-point scale (1 = *highly disadvantaged* to 7 = *highly advantaged*). They provided sociodemographic information (age, sex, parental level of education, and occupations), were thanked and debriefed.

Results

Perceived SES. Participants' perception of the target's SES was analyzed on the full sample in a regression with the Function of Assessment (selection coded −0.5, education coded 0.5), Target's SES (low-SES coded −0.5, high-SES coded 0.5), and the interaction term as predictors.[11] Results showed a main effect of Target's SES, b = 1.75, 95% CI [1.56, 1.94], $t(328)$ = 18.34, $p < .001$, η_p^2 = .51, 95% CI [.43, .57]. The low-SES target was perceived as coming from a more disadvantaged background (M = 3.98, SD = 0.86, 95% CI [3.84, 4.11]) than the high-SES target (M = 5.72, SD = .88, 95% CI [5.59, 5.86]). The Function main effect and interaction did not reach significance, b = −0.09, $t(328)$ = −0.91, p = .36 and b = −0.09, $t(328)$ = 1.13, p = .26.

Ratio of negative feedback. We computed the number of characters highlighted in each color as indicators of negative and positive feedback. We calculated a ratio of negative feedback relative to the total amount of positive and negative feedback such that higher scores indicate more negativity in the evaluation. This ratio was created to have a negative evaluation indicator that is comparable to evaluation in our previous studies (i.e., finding mistakes in a test). This ratio was analyzed in a regression with the Function of Assessment (selection coded −0.5, education coded 0.5), Target's SES (low-SES coded −0.5, high-SES coded 0.5), and the interaction as predictors.[12] The results showed no main effect of the Target's SES, b = −0.03, 95% CI [−0.05, 0.00], $t(300)$ = −1.82, p = .07, η_p^2 = .01, 95% CI [.00, .05] or the Function of Assessment, b = −0.004, 95% CI [−0.03, 0.02], $t(300)$ = −0.29, p = .77, η_p^2 = .00, 95% CI [.00, .02]. The expected interaction between SES and Function did not reach significance, b = 0.03, 95% CI [−0.03, 0.08], $t(300)$ = 0.91, p = .36, η_p^2 = .00, 95% CI [.00, .03], Cohen's d = .10. However, this interaction was in the expected direction with a larger difference in ratio between low and high SES students when the assessment was meant to select ($M_{low\ SES}$ = .42, $SD_{low\ SES}$ = .12, 95% CI [.39, .44] vs. $M_{high\ SES}$ = .38, $SD_{high\ SES}$ = .13, 95% CI [.35, .41]) rather than to improve learning ($M_{low\ SES}$ = .40, $SD_{low\ SES}$ = .12, 95% CI [.38, .43] vs. $M_{high\ SES}$ = .39, $SD_{high\ SES}$ = .13, 95% CI [.36, .42]).

Positive and negative feedback. The number of characters highlighted was analyzed in a 2 (Function of Assessment: selection vs. education) × 2 (Target's SES: high vs. low) × 2 (Type of Feedback: negative vs. positive) mixed analysis of variance (ANOVA) with the last factor as a within-participant factor.[13] The analysis revealed a main effect of the Type of feedback, $F(1, 299)$ = 189.34, $p < .001$, η_p^2 = .39, indicating that participants gave more positive feedback (M = 418, SD = 189) than negative feedback (M = 272, SD = 133). This effect was qualified by an interaction with the Target's SES, $F(1, 299)$ = 7.05, p = .008, η_p^2 = .023, 90% CI [.00, .06]. Participants gave more positive feedback to a high SES student (M = 446, SD = 204, 95% CI [413, 479]) compared with a low SES student (M = 392, SD = 171, 95% CI [365, 419]), $F(1, 299)$ = 5.98, p = .015, η_p^2 = .02, 90% CI [.00, .05]. This difference between SES was not significant for negative feedback, $F(299)$ = .38, p = .54, η_p^2 = .00, 90% CI [.00, .01]. No other effect reached significance Fs < 2.53, ps >

[10] The booklet also included an evaluation of the essay on a 10-point scale. The results are available upon request from the authors.

[11] Three participants did not fill the item.

[12] Two outliers were removed from the analysis because of elevated cooks' distances and studentized deleted residuals.

[13] Three participants were excluded because of abnormal residuals, elevated cooks' distances, and studentized deleted residuals.

.11, $\eta s_p^2 < .01$, including the expected interaction between the Type of feedback, SES and the Function of Assessment, $F(299) = 0.03$, $p = .86$, $\eta_p^2 = .00$, 90% CI [.00, .01].

Entity theory of intelligence. The score of belief in an entity theory of intelligence was analyzed in a regression with the Function of Assessment (selection coded −0.5, education coded 0.5), Target's SES (low-SES coded −0.5, high-SES coded 0.5), and the interaction as predictors.[14] The main effect of Function that would indicate an impact of that induction on mindset did not reach significance $b = 0.02$, 95% CI [−0.14, 0.18], $t(298) = 0.25$, $p = .80$, $\eta_p^2 = .00$, 95% CI [.00, .01]. The main effect of SES and the interaction were also nonsignificant, respectively, $b = 0.05$, 95% CI [−0.12, 0.21], $t(298) = 0.55$, $p = .59$, $\eta_p^2 = .00$, 95% CI [.00, .02], and $b = -0.23$, 95% CI [−0.55, 0.10], $t(298) = -1.36$, $p = .17$, $\eta_p^2 = .01$, 95% CI [.00, .04].

Supplementary analyses investigating the impact of the participants' own social class were conducted (see online supplemental material for the details) and showed no change or moderation of the described results.

Discussion

This fourth study first aimed at testing the hypothesis that the function of selection, while keeping the assessment tool constant, triggers the creation of a SES performance gap, compared with the educational function. The analysis on the ratio of negative feedback showed a pattern that was congruent with this hypothesis, but the effect was not significant. The results on the number of characters highlighted showed an overall favoritism of the higher social class student. Irrespective of the function of assessment, participants provided more positive feedback to the high-SES student than the low-SES student. This result is in line with the idea that nowadays the creation of inequalities relies on a favorable bias for high status group members who are offered more positive experience (e.g., DiTomaso, 2015). However, further replication of this effect is needed as we initially predicted that it would appear only when the context emphasizes on selection. Moreover, a possible way to disentangle favoritism toward the high-SES student from negative treatment toward the low-SES student could be to include a control condition with no information about the target's SES (i.e., anonymous). This would provide information about whether it is the low-SES condition that triggers more negative assessment than the control or the high-SES condition that triggers more positive assessment, or whether both discrimination and favoritism are at play.

A secondary goal of this study was to examine whether the function of assessment could impact the evaluators' mindset, with selective purposes fostering a more entity theory of the student's abilities than educational purposes. The results are not in line with this proposition. Previous research showing changes in mindset used direct intervention by telling participants that intelligence is fixed or can grow (Rattan, Savani, Chugh, & Dweck, 2015). It could be that information about the function of assessment is not sufficiently powerful to affect the mindset. Yet, some studies suggest that mindsets are also sensitive to subtle information such as praise or generic statements about categories (i.e., talking about boys in general instead of a boy in particular; Cimpian & Markman, 2011; Mueller & Dweck, 1998). At this stage, the impact of the function of assessment on the belief in an entity theory of intelligence remains an open question, although the effect size ($\eta_p^2 = .00$, 95% CI [.00, .01]) could suggest a possibly negligible effect.

Finally, it should be recognized that the sample of this experiment presents an imbalance in terms of gender and recruitment location. Such an imbalance makes it difficult to test the effects of these variables, but as the personal characteristics of the participants do not seem to alter the observed effects (see online supplemental material), we believe that this asymmetry should not be a source of concern.

Meta-Analysis

To understand more precisely the size of the effect of interest, we ran an internal meta-analysis on the four experiments (Cumming, 2013) and estimated the effect size of the moderation of the SES performance gap by the orientation of assessment toward selection (i.e., grading or selection function) or education (i.e., formative comments or educational function). We computed the standardized mean difference corresponding to the difference of simple effects of SES between the selection and education assessment practices $[(\bar{X}lowSES\_selection - \bar{X}highSES\_selection) - (\bar{X}lowSES\_education - \bar{X}highSES\_education)]/2^{*}sp$ (where *sp* is the pooled *SD*; Westfall, 2015). In Experiment 1, we used the effect size of the SES × Assessment Method interaction at a moderate time spent on the study. In Experiment 4, we used the effect size of the SES × Function of Assessment interaction on the ratio of negative feedback, as this measure is the functional equivalent to the number of mistakes measured in the previous three studies. We used a weighted random-effects model (Cumming, 2013). A weighted model lowers the contribution of studies with higher variance around the effect size. Random effects models take into account the heterogeneity between studies and postulates that different studies can estimate different effect sizes.

The analysis revealed a small and significant effect size, $d = 0.19$, $p = .002$, 95% CI [0.07; 0.30]. The variance index between the four studies was not significant, suggesting low heterogeneity between studies $Q(df = 3) = 0.84$, $p = .84$. This internal meta-analysis provides evidence that evaluators artificially create a greater SES performance gap when assessment is used to select rather than foster learning. The effect size is small but we nonetheless believe it should be interpreted in light of the length of education and the frequency of assessment. Very small differences in repeated evaluations can have important consequences on the overall experiences and educational outcomes of students when they accumulate over time.

General Discussion

A growing line of research has addressed the question of the cultural and structural determinants underlying the social class achievement gap (e.g., Croizet & Claire, 1998; Stephens et al., 2012). This endeavor has been particularly valuable in revealing the sociocultural influences that contribute to the social class inequalities. However, the majority of these studies have focused on the psychological processes (e.g., stereotype threat, cultural

[14] Four outliers were excluded because of abnormal residuals, elevated cooks' distances, and studentized deleted residuals.

mismatch) that impact the academic performance of students. In the present research, we argue that, in addition, a new stream of research should emerge that addresses how evaluators' behavior contributes to the social class achievement gap, independent of the students' actual performance. We proposed that the endemic use of normative grading in education, given its strong association with the meritocratic ideal and the function of selection of educational institutions, leads evaluators to reproduce existing social class asymmetries. In contrast, assessment with an educational function and an egalitarian ethos should reduce the impact of the student's social class on evaluation. More specifically, we hypothesized that evaluators would differentially assess the work produced by low- and high-SES students when using assessment for selection, even in the absence of any objective differences. This tendency should be reduced when using assessment for learning.

Experiments 1 and 2 showed consistent evidence that when evaluators used an assessment method oriented toward selection (i.e., normative grading; cf. Autin et al., 2015), they actively detected more mistakes for low-SES students than for high-SES students. This effect emerged in a dictation test that objectively contained the same number of mistakes in all conditions (with a moderation by the time spent on the task in Experiment 1). The creation of such an artificial social class achievement gap was not observed when evaluators used an assessment method oriented toward education (i.e., formative comments). We believe that a strong asset of these results is the use of a behavioral measure—the number of mistakes that participants actually found in the test—that did not allow participants to control the social desirability of their responses (Dompnier, Darnon, & Butera, 2009).

Experiments 3 and 4 manipulated the mechanism that we assumed to explain the results observed in the first two experiments: the function of assessment. Indeed, we expected that evaluators reproduce in their assessment existing social class asymmetries when using normative grading because this form of assessment epitomizes the function of selection of educational institutions. The results support this hypothesis. In Experiment 3, making the selection function of assessment salient led evaluators to find significantly more mistakes for low-SES students than for high-SES students, regardless of the assessment method they used. Such a differential treatment was no longer significant when the educational function of assessment was made salient. In Experiment 4, the replication of this finding with a different assessment tool (i.e., highlighting sections in the student's essay) showed a consistent but not significant pattern in the ratio of negative feedback. And accordingly, we conducted a small-scale meta-analysis to test the overall support to our main hypothesis received from the four studies. Overall, the results of the meta-analysis support the hypothesis that even in the absence of objective differences in performance, social class inequalities can be perpetuated by evaluators who re-create an achievement gap, especially when the selective purpose of assessment is put to the fore.

Experiment 4 secondarily aimed at specifying how the social class achievement gap was created in the selection context, by using both negative and positive feedback. This question goes beyond the scope of the present article, as the primary goal here was to document the creation of a SES performance gap. Yet, we postulated that inequalities are the byproduct not only of negative treatment against low status individuals but also of the privilege of high status individuals (e.g., Adams et al., 2008). The results showed an overall effect of favoritism toward high-SES students. This unexpected effect calls for further investigation but is consistent with an understanding of the mechanisms of inequalities that emphasizes the implication of favoritism. For example, it has been observed that high-status individuals receive advantages, and especially better evaluation from both high- and low-status actors (DiTomaso, Post, Smith, Farris, & Cordero, 2007). Previous research documented how higher social class students benefit from many aspects of the educational institutions such as valued forms of knowledge, language, or posture, compatible self-models and boosting evaluative settings (Bourdieu & Passeron, 1977; Croizet et al., 2017; Lareau, 2011; Stephens, Markus, et al., 2014). The present results suggest that evaluator's behavior during assessment might also be one of the privileges that enhance the academic experience of higher social class students.

Contribution to Ongoing Debates

The first contribution of the present research is to participate in the growing effort to bring the study of social class to the core of social and educational psychological investigations (S. T. Fiske & Markus, 2012). Alongside previous research directly studying student performance (e.g., Goudeau & Croizet, 2017; Jury et al., 2015), the present article unveils a new path through which social class inequalities are reproduced in schools via the assessment of performance. Our research suggests that, even if the educational system could offer a matching and nonthreatening environment to all students, evaluators could still artificially create a social class achievement gap when they assess with selective purposes.

The second contribution pertains to research in sociology of education. Observing that evaluators create a social class achievement gap is in line with the classic sociological theory of social reproduction, and in particular with the proposition that agents of institutions play an important role in the reproduction of social inequalities (Bourdieu & Passeron, 1977). Our research provides experimental evidence of the evaluators' role in social reproduction and, more importantly, identified an institutional factor that leads evaluators to create social inequalities: the selection versus educational function of assessment.

Finally, our findings are consistent with previous research in educational sciences showing discrimination by evaluators in assessing the same product attributed to students of different backgrounds (e.g., Sprietsma, 2013). However, the existing literature had only investigated this phenomenon in settings using grading, and our comparison with alternative forms of assessment offers new insights. Not only because this comparison shows that discriminatory behavior is not inherent to assessment, but also—and especially—because it shows that evaluators do not always act in a biased manner. Regarding the discriminatory behavior previously observed in grading, our results suggest that it would be better interpreted as the product of the selective purposes conveyed by such assessment practices, rather than biased individual evaluators. This result is consistent with qualitative work showing that changing the tools (e.g., replacing grades with formative comments) is not enough to change the vision of assessment, and that ultimately teachers use all forms of assessment primarily with quantifying and ranking purposes (McNair, Bhargava, Adams, Edgerton, & Kypros, 2003). Moreover, teacher's use of assessment has been related to the requirement of educational institutions

stemming from their societal role of selection (Gewirtz, 2000; Hall, Collins, Benjamin, Nind, & Sheehy, 2004; Popham, 2001). Our research concurs with an analysis of assessment practices as contingent on the institutional function they serve. Through the prism of selection, all forms of assessment might produce inequalities in evaluation.

Overall, the present research underscores the need for a sociocultural approach to understanding inequalities (Adams et al., 2008; Markus & Stephens, 2017) that highlights the intertwining of individuals with institutions in their production (see also Kraus & Park, 2017). Social class inequalities cannot be reduced to consequences of direct and intentional actions of biased individuals (i.e., biased evaluators), or to byproducts of agentless institutions that mechanically exclude lower social class students and favor higher social class students (i.e., biased schools). We rather propose that the educational institutions' logic shapes evaluators' behavior, and in turn low- and high-SES students' experience. More specifically, the functions of educational institutions seem to affect the way evaluators make sense of the assessment situation and cause them to differentially evaluate students' performance based on their social class.

Limitations and Conclusion

Several limitations of the present research should be acknowledged. First, the studies were conducted with students who were put in the position of a teacher, and not real teachers. Replicating these findings with teachers would certainly increase their ecological validity, but we do not expect any remarkable difference. Indeed, our theoretical approach is precisely that institutional norms and functions shape their agents' behaviors; thus, the observed effects should be reproducible with actual teachers, as they have been socialized in the very context that we have experimentally induced in this research. Furthermore, previous research using role-playing paradigms showed that participants adjust their attitudes to the role (Covington & Omelich, 1979; Harari & Covington, 1981; Houston & Holmes, 1975), and our results are consistent with those obtained in research conducted with teachers (e.g., Hinnerich, Höglin, & Johannesson, 2015; Rangvid, 2015; Sprietsma, 2013).

Second, more research is needed to understand the psychological mechanisms at play in the discriminatory behavior of evaluators. The present research proposed a sociocultural approach and therefore, in the last two experiments, we manipulated a structural-level mechanism (i.e., the function of selection vs. education) believed to underlie the creation of the social class achievement gap. However, future research may also be interested in individual-level variables induced by both functions of assessment. We argued that assessment for selection relates to a meritocratic ethos whereas the assessment for learning relates to an egalitarian ethos (Autin et al., 2015). Egalitarianism and meritocratic values have contrasted consequences in terms of stereotyping and attitudes toward groups (e.g., Wyer, 2003) and are involved in the reduction/maintenance of inequalities (e.g., Costa-Lopes, Dovidio, Pereira, & Jost, 2013). The perception of egalitarian versus meritocratic values is, therefore, a possible mechanism underlying the effect of assessment for learning versus selection on the creation of a social class performance gap. Downstream mechanisms could also be explored. For example, assessment as an apparatus of meritocratic selection might give participants a greater sense of objectivity than assessment to foster learning. Feelings of objectivity are known to increase bias in decisions (Uhlmann & Cohen, 2007). Assessment for selection also requires a more firm, nonambiguous response (to give a grade, to decide about grade repetition) than assessment for learning. The desire for clear-cut answers—known as *need for closure*—relates to biases in thinking (Webster & Kruglanski, 1997), and high need for closure can be involved in discrimination in grading (Kruglanski & Freund, 1983). It is, therefore, possible that the discriminatory behavior observed in the present studies is partly because of the higher need for closure triggered by the selection context compared with the educational context. It is also possible that different processes simultaneously occur when evaluators focus on educational purposes. For example, participants might feel more accountable for their evaluation when told that they have to write formative comments or suggest learning strategies to improve, and accountability reduces bias in decisions (Lerner & Tetlock, 1999; Uhlmann & Cohen, 2007).

Finally, in the absence of a control condition for the function of assessment, it is impossible to conclude from the present experiments whether the effects are driven by assessment for selection or assessment for learning. However, in the current sociocultural and educational context it seems unlikely to have a control assessment condition that does not conjure an institutional function. Given that normative grading is by far the most widespread assessment method in OECD countries (Knight & Yorke, 2003), and that the social class achievement gap is also omnipresent in these countries (OECD, 2013a), we believe that an "ecological" control condition (e.g., "please, assess this dictation test") would probably be interpreted as the assessment for selection conditions. Moreover, creating a "neutral" control condition (without any form of assessment; e.g., "please, find all the mistakes in this dictation to prove your skills") would be pointless, as in such a condition the participant would no longer be an evaluator. The difficulty to have a control condition explains why, like the majority of research investigating the factors contributing to inequalities in education (see Croizet et al., 2017; Jury, Darnon, et. al., 2017), we compared a situation where these factors are at play to a situation where they are actively countered.

To conclude, this line of research challenges the idea that educational institutions perform a meritocratic selection based solely on an objective assessment of individuals' qualities. To the contrary, this selection function might actually contribute to the reproduction of social inequalities by leading evaluators to create a social class achievement gap. There is a growing literature demonstrating how the way educational institutions operate is involved in the reproduction of social inequalities (e.g., Croizet et al., 2017; Jury, Darnon, et. al., 2017). Our research adds to this literature by identifying a structural factor—the function of selection of assessment—that shapes the role of evaluators in the creation of a social class achievement gap, even when there are no actual differences in performance.

References

Adams, G., Biernat, M., Branscombe, N. R., Crandall, C. S., & Wrightsman, L. S. (2008). Beyond prejudice: Toward a sociocultural psychology of racism and oppression. In G. Adams, M. Biernat, N. R. Branscombe, C. S. Crandall, & L. S. Wrightsman (Eds.), *Commemorating Brown: The social psychology of racism and discrimination* (pp. 215–246). Wash-

ington, DC: American Psychological Association. http://dx.doi.org/10.1037/11681-012

Arrow, K. J. (1973). Higher education as a filter. *Journal of Public Economics, 2,* 193–216. http://dx.doi.org/10.1016/0047-2727(73)90013-3

Autin, F., Batruch, A., & Butera, F. (2015). Social justice in education: How the function of selection in educational institutions predicts support for (non)egalitarian assessment practices. *Frontiers in Psychology, 6,* 707. http://dx.doi.org/10.3389/fpsyg.2015.00707

Batruch, A., Autin, F., Bataillard, F., & Butera, F. (2018). School selection and the social class divide: How tracking contributes to the reproduction of inequalities. *Personality and Social Psychology Bulletin*. Advance online publication. http://dx.doi.org/10.1177/0146167218791804

Batruch, A., Autin, F., & Butera, F. (2017). Re-establishing the social-class order: Restorative reactions against high-achieving, low-SES pupils. *Journal of Social Issues, 73,* 42–60. http://dx.doi.org/10.1111/josi.12203

Bell, D. (1973). *The coming of post-industrial society*. London, UK: Heinemann.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18,* 5–25. http://dx.doi.org/10.1080/0969594X.2010.513678

Bisseret, N. (1974). L'idéologie des aptitudes naturelles [The ideology of natural aptitudes]. In N. Bisseret (Ed.), *Les inégaux ou la sélection universitaire* [Being unequal or of academic selection] (pp. 13–52). Paris, France: Presses Universitaires de France.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7–74. http://dx.doi.org/10.1080/0969595980050102

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1,* 1–5.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill.

Bobocel, D. R., Son Hing, L. S., Davey, L. M., Stanley, D. J., & Zanna, M. P. (1998). Justice-based opposition to social policies: Is it genuine? *Journal of Personality and Social Psychology, 75,* 653–669. http://dx.doi.org/10.1037/0022-3514.75.3.653

Bourdieu, P., & Passeron, J. (1977). *Reproduction in education, culture and society*. London, UK: Sage.

Bowen, W. G., Kurzweil, M. A., Tobin, E. M., & Pichler, S. C. (2005). *Equity and excellence in American higher education*. Charlottesville, VA: University of Virginia Press.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues, 55,* 429–444. http://dx.doi.org/10.1111/0022-4537.00126

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79,* 474–482. http://dx.doi.org/10.1037/0022-0663.79.4.474

Butler, R., & Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology, 78,* 210–216. http://dx.doi.org/10.1037/0022-0663.78.3.210

Carson, J. (2007). *The measure of merit: Talents, intelligence, and inequality in the French and American republics, 1750–1940*. Princeton, NJ: Princeton University Press.

Castilla, E. J., & Benard, S. (2010). The paradox of meritocracy in organizations. *Administrative Science Quarterly, 55,* 543–676. http://dx.doi.org/10.2189/asqu.2010.55.4.543

Chmielewski, A. K. (2014). An International comparison of achievement inequality in within- and between-school tracking systems. *American Journal of Education, 120,* 293–324. http://dx.doi.org/10.1086/675529

Cimpian, A., & Markman, E. M. (2011). The generic/nongeneric distinction influences how children interpret new information about social others. *Child Development, 82,* 471–492. http://dx.doi.org/10.1111/j.1467-8624.2010.01525.x

Costa-Lopes, R., Dovidio, J. F., Pereira, C. R., & Jost, J. T. (2013). Social psychological perspectives on the legitimation of social inequality: Past, present and future. *European Journal of Social Psychology, 43,* 229–237. http://dx.doi.org/10.1002/ejsp.1966

Coulmont, B. (2011). *Sociologie des prénoms* [Sociology of first names]. Paris, France: La Découverte.

Covington, M. V., & Omelich, C. L. (1979). Are causal attributions causal? A path analysis of the cognitive model of achievement motivation. *Journal of Personality and Social Psychology, 37,* 1487–1504. http://dx.doi.org/10.1037/0022-3514.37.9.1487

Cozzarelli, C., Wilkinson, A. V., & Tagler, M. J. (2001). Attitudes toward the poor and attributions for poverty. *Journal of Social Issues, 57,* 207–227. http://dx.doi.org/10.1111/0022-4537.00209

Crahay, M. (2012). *L'école peut-elle être juste et efficace?* [Can school be fair and efficient?] (2nd ed.). Bruxelle, Belgium: De Boeck Superieur.

Croft, A., & Schmader, T. (2012). The feedback withholding bias: Minority students do not receive critical feedback from evaluators concerned about appearing racist. *Journal of Experimental Social Psychology, 48,* 1139–1144. http://dx.doi.org/10.1016/j.jesp.2012.04.010

Croizet, J.-C., & Claire, T. (1998). Extending the concept of stereotype and threat to social class: The intellectual underperformance of students from low socioeconimic backgrounds. *Personality and Social Psychology Bulletin, 24,* 588–594. http://dx.doi.org/10.1177/0146167298246003

Croizet, J.-C., & Dutrévis, M. (2004). Socioeconomic status and intelligence: Why test scores do not equal merit. *Journal of Poverty, 8,* 91–107. http://dx.doi.org/10.1300/J134v08n03_05

Croizet, J.-C., Goudeau, S., Marot, M., & Millet, M. (2017). How do educational contexts contribute to the social class achievement gap: Documenting symbolic violence from a social psychological point of view. *Current Opinion in Psychology, 18,* 105–110. http://dx.doi.org/10.1016/j.copsyc.2017.08.025

Croizet, J.-C., & Millet, M. (2012). Social class and test performance: From stereotype threat to symbolic violence and vice versa. In M. Inzlicht & T. Schmader (Eds.), *Stereotype threat: Theory, process, and application* (pp. 188–201). New York, NY: Oxford University Press.

Crosby, J. R., & Monin, B. (2007). Failure to warn: How student race affects warnings of potential academic difficulty. *Journal of Experimental Social Psychology, 43,* 663–670. http://dx.doi.org/10.1016/j.jesp.2006.06.007

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge/Taylor & Francis Group.

Darnon, C., Dompnier, B., Delmas, F., Pulfrey, C., & Butera, F. (2009). Achievement goal promotion at university: Social desirability and social utility of mastery and performance goals. *Journal of Personality and Social Psychology, 96,* 119–134. http://dx.doi.org/10.1037/a0012824

Darnon, C., Jury, M., & Aelenei, C. (2017). Who benefits from mastery-approach and performance-approach goals in college? Students' social class as a moderator of the link between goals and grade. *European Journal of Psychology of Education, 2017,* 1–14. http://dx.doi.org/10.1007/s10212-017-0351-z

Darnon, C., Smeding, A., & Redersdorff, S. (2017). Belief in school meritocracy as an ideological barrier to the promotion of equality: School meritocracy and promotion of equality. *European Journal of Social Psychology, 48,* 523–534.

Désert, M., Préaux, M., & Jund, R. (2009). So young and already victims of stereotype threat: Socio-economic status and performance of 6 to 9 years old children on Raven's progressive matrices. *European Journal of Psychology of Education, 24,* 207–218. http://dx.doi.org/10.1007/BF03173012

DiTomaso, N. (2015). Racism and discrimination versus advantage and favoritism: Bias for versus bias against. *Research in Organizational Behavior, 35,* 57–77. http://dx.doi.org/10.1016/j.riob.2015.10.001

DiTomaso, N., Post, C., Smith, D. R., Farris, G. F., & Cordero, R. (2007). Effects of structural position on allocation and evaluation decisions for scientists and engineers in industrial R&D. *Administrative Science Quarterly, 52,* 175–207. http://dx.doi.org/10.2189/asqu.52.2.175

Dittmann, A. G., & Stephens, N. M. (2017). Interventions aimed at closing the social class achievement gap: Changing individuals, structures, and construals. *Current Opinion in Psychology, 18,* 111–116. http://dx.doi.org/10.1016/j.copsyc.2017.07.044

Domina, T., Penner, A., & Penner, E. (2017). Categorical inequality: Schools as sorting machines. *Annual Review of Sociology, 43,* 311–330. http://dx.doi.org/10.1146/annurev-soc-060116-053354

Dompnier, B., Darnon, C., & Butera, F. (2009). Faking the desire to learn: A clarification of the link between mastery goals and academic achievement. *Psychological Science, 20,* 939–943. http://dx.doi.org/10.1111/j.1467-9280.2009.02384.x

Dornbusch, S. M., Glasgow, K. L., & Lin, I.-C. (1996). The social structure of schooling. *Annual Review of Psychology, 47,* 401–429. http://dx.doi.org/10.1146/annurev.psych.47.1.401

Dubet, F. (2004). *L'école des chances: Qu'est-ce qu'une école juste* [School of opportunity: What is a fair school]. Paris, France: Seuil.

Durante, F., & Fiske, S. T. (2017). How social-class stereotypes maintain inequality. *Current Opinion in Psychology, 18,* 43–48. http://dx.doi.org/10.1016/j.copsyc.2017.07.033

Duru-Bellat, M. (2006). *L'inflation scolaire: les désillusions de la méritocratie* [School inflation: The disenchantment of meritocracy]. Paris, France: Le Seuil.

Duru-Bellat, M. (2008). Recent trends in social reproduction in France: Should the political promises of education be revisited? *Journal of Education Policy, 23,* 81–95. http://dx.doi.org/10.1080/02680930701754104

Dweck, C. S. (2012). Mindsets and human nature: Promoting change in the Middle East, the schoolyard, the racial divide, and willpower. *American Psychologist, 67,* 614–622. http://dx.doi.org/10.1037/a0029783

Faniko, K., Lorenzi-Cioldi, F., Buschini, F., & Chatard, A. (2012). The influence of education on attitudes toward affirmative action: The role of the policy's strength. *Journal of Applied Social Psychology, 42,* 387–413. http://dx.doi.org/10.1111/j.1559-1816.2011.00892.x

Filer, A. (2002). Socio-historical and cultural contexts of assessment policy. In A. Filer (Ed.), *Assessment: Social practice and social product* (pp. 7–10). New York, NY: Routledge.

Fine, M., & Burns, A. (2003). Class notes: Toward a critical psychology of class and schooling. *Journal of Social Issues, 59,* 841–860. http://dx.doi.org/10.1046/j.0022-4537.2003.00093.x

Fiske, A. P., Kitayama, S., Markus, H. R., & Nisbett, R. E. (1998). The cultural matrix of social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1 and 2, pp. 915–981). New York, NY: McGraw-Hill.

Fiske, S. T., & Markus, H. R. (2012). *Facing social class: How societal rank influences interaction*. New York, NY: Russell Sage Foundation.

Forquin, J.-C. (1992). *École et Culture: Le point de vue des sociologues britanniques* [School and Culture: The point of view of British sociologists]. Bruxelle, Belgium: De Boeck Université.

Frey, B. B., & Schmitt, V. L. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics, 18,* 402–423. http://dx.doi.org/10.4219/jaa-2007-495

Gewirtz, S. (2000). Bringing the politics back in: A critical analysis of quality discourses in education. *British Journal of Educational Studies, 48,* 352–370. http://dx.doi.org/10.1111/1467-8527.00152

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519–521. http://dx.doi.org/10.1037/h0049294

Goudeau, S., Autin, F., & Croizet, J. C. (2017). Etudier, mesurer et manipuler la classe sociale en psychologie sociale: Approches economiques, symboliques et culturelles [Studying, measuring and manipulating social class in social psychology: Economic, symbolic and cultural approaches]. *Revue Internationale de Psychologie Sociale, 30,* 1–19. http://dx.doi.org/10.5334/irsp.52

Goudeau, S., & Croizet, J.-C. (2017). Hidden advantages and disadvantages of social class: How classroom settings reproduce social inequality by staging unfair comparison. *Psychological Science, 28,* 162–170. http://dx.doi.org/10.1177/0956797616676600

Greenwald, A. G., & Pettigrew, T. F. (2014). With malice toward none and charity for some: Ingroup favoritism enables discrimination. *American Psychologist, 69,* 669–684. http://dx.doi.org/10.1037/a0036056

Hall, K., Collins, J., Benjamin, S., Nind, M., & Sheehy, K. (2004). SATurated models of pupildom: Assessment and inclusion/exclusion. *British Educational Research Journal, 30,* 801–817. http://dx.doi.org/10.1080/0141192042000279512

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Giffen, C. J., Blair, S. S., Rouse, D. I., & Hyde, J. S. (2014). Closing the social class achievement gap for first-generation students in undergraduate biology. *Journal of Educational Psychology, 106,* 375–389. http://dx.doi.org/10.1037/a0034679

Harari, O., & Covington, M. V. (1981). Reactions to achievement behavior from a teacher and student perspective: A developmental analysis. *American Educational Research Journal, 18,* 15–28. http://dx.doi.org/10.3102/00028312018001015

Harrison, L. A., Stevens, C. M., Monty, A. N., & Coakley, C. A. (2006). The consequences of stereotype threat on the academic performance of White and non-White lower income college students. *Social Psychology of Education, 9,* 341–357. http://dx.doi.org/10.1007/s11218-005-5456-6

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology, 53,* 575–604. http://dx.doi.org/10.1146/annurev.psych.53.100901.135109

Hinnerich, B. T., Höglin, E., & Johannesson, M. (2015). Discrimination against students with foreign backgrounds: Evidence from grading in Swedish public high schools. *Education Economics, 23,* 660–676. http://dx.doi.org/10.1080/09645292.2014.899562

Houston, B. K., & Holmes, D. S. (1975). Role playing versus deception: The ability of subjects to simulate self-report and physiological responses. *The Journal of Social Psychology, 96,* 91–98. http://dx.doi.org/10.1080/00224545.1975.9923266

Jackman, M. R. (1994). *The velvet glove: Paternalism and conflict in gender, class, and race relations*. Berkeley, CA: University of California Press.

Jasso, G., & Resh, N. (2002). Exploring the sense of justice about grades. *European Sociological Review, 18,* 333–351. http://dx.doi.org/10.1093/esr/18.3.333

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. New York, NY: Routledge.

Jury, M., Darnon, C., Dompnier, B., & Butera, F. (2017). The social utility of performance-approach goals in a selective educational environment. *Social Psychology of Education, 20,* 215–235. http://dx.doi.org/10.1007/s11218-016-9354-x

Jury, M., Smeding, A., & Darnon, C. (2015). First-generation students' underperformance at university: The impact of the function of selection. *Frontiers in Psychology, 6,* 710. http://dx.doi.org/10.3389/fpsyg.2015.00710

Jury, M., Smeding, A., Stephens, N. M., Nelson, J. E., Aelenei, C., & Darnon, C. (2017). The experience of low-SES students in higher education: Psychological barriers to success and interventions to reduce social-class inequality. *Journal of Social Issues, 73,* 23–41. http://dx.doi.org/10.1111/josi.12202

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structures.

Journal of Personality and Social Psychology, 55, 893–905. http://dx.doi.org/10.1037/0022-3514.55.6.893

Knight, P. T., & Yorke, M. (2003). *Assessment, learning and employability*. Maidenhead, UK: Open University Press.

Knowles, E. D., & Lowery, B. S. (2012). Meritocracy, self-concerns, and Whites' denial of racial inequity. *Self and Identity, 11,* 202–222. http://dx.doi.org/10.1080/15298868.2010.542015

Kraus, M. W., Callaghan, B., & Ondish, P. (in press). Social class as culture. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology*. New York, NY: Guilford Press.

Kraus, M. W., & Park, J. W. (2017). The structural dynamics of social class. *Current Opinion in Psychology, 18,* 55–60. http://dx.doi.org/10.1016/j.copsyc.2017.07.029

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay inferences: Effects of impressional primacy, ethnic stereotyping and numerical anchoring. *Journal of Experimental Social Psychology, 19,* 448–468. http://dx.doi.org/10.1016/0022-1031(83)90022-7

Lareau, A. (2011). *Unequal childhoods: Class, race, and family life*. Oakland, CA: University of California Press.

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. New York, NY: Farrar, Straus & Giroux.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin, 125,* 255–275. http://dx.doi.org/10.1037/0033-2909.125.2.255

LeTendre, G. K., Hofer, B. K., & Shimizu, H. (2003). What is tracking? Cultural expectations in the United States, Germany, and Japan. *American Educational Research Journal, 40,* 43–89. http://dx.doi.org/10.3102/00028312040001043

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 60,* 245–256. http://dx.doi.org/10.1177/0004944116664618

Markus, H. R., & Hamedani, M. G. (2007). Sociocultural psychology: The dynamic interdependence among self systems and social systems. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 3–39). New York, NY: Guilford Press.

Markus, H. R., & Stephens, N. M. (2017). Editorial overview: The psychological and behavioral consequences of inequality and social class: A theoretical integration. *Current Opinion in Psychology, 18,* iv–xii. http://dx.doi.org/10.1016/j.copsyc.2017.11.001

McCoy, S. K., & Major, B. (2007). Priming meritocracy and the psychological justification of inequality. *Journal of Experimental Social Psychology, 43,* 341–351. http://dx.doi.org/10.1016/j.jesp.2006.04.009

McNair, S., Bhargava, A., Adams, L., Edgerton, S., & Kypros, B. (2003). Teachers speak out on assessment practices. *Early Childhood Education Journal, 31,* 23–31. http://dx.doi.org/10.1023/A:1025180617689

Moskowitz, G. B., Gollwitzer, P. M., Wasel, W., & Schaal, B. (1999). Preconscious control of stereotype activation through chronic egalitarian goals. *Journal of Personality and Social Psychology, 77,* 167–184. http://dx.doi.org/10.1037/0022-3514.77.1.167

Moskowitz, G. B., Salomon, A. R., & Taylor, C. M. (2000). Preconsciously controlling stereotyping: Implicitly activated egalitarian goals prevent the activation of stereotypes. *Social Cognition, 18,* 151–177. http://dx.doi.org/10.1521/soco.2000.18.2.151

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75,* 33–52. http://dx.doi.org/10.1037/0022-3514.75.1.33

Norton, M. I., Vandello, J. A., & Darley, J. M. (2004). Casuistry and social category bias. *Journal of Personality and Social Psychology, 87,* 817–831. http://dx.doi.org/10.1037/0022-3514.87.6.817

OECD. (2006). *Education at a glance: OECD indicators 2006*. Paris, France: OECD Publishing.

OECD. (2013a). *PISA 2012 results: Excellence through equity: Giving every student the chance to succeed* (Vol. II). Paris, France: OECD Publishing.

OECD. (2013b). *PISA 2012 results: What makes schools successful? Resources, policies and practices* (Vol. IV). Paris, France: OECD Publishing.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45,* 867–872. http://dx.doi.org/10.1016/j.jesp.2009.03.009

Parsons, T. (1959). The school class as a social system: Some of its functions in American society. *Harvard Educational Review, 29,* 297–318.

Perrenoud, P. (1995). *La pédagogie à l'école des différences* [Pedagogy at the school of differences]. Paris, France: ESF.

Popham, W. J. (2001). Teaching to the test? *Educational Leadership, 58,* 16–21.

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

Rattan, A., Good, C., & Dweck, C. S. (2012). "It's ok—Not everyone can be good at math": Instructors with an entity theory comfort (and demotivate) students. *Journal of Experimental Social Psychology, 48,* 731–737. http://dx.doi.org/10.1016/j.jesp.2011.12.012

Rattan, A., Savani, K., Chugh, D., & Dweck, C. S. (2015). Leveraging mindsets to promote academic achievement: Policy recommendations. *Perspectives on Psychological Science, 10,* 721–726. http://dx.doi.org/10.1177/1745691615599383

Resh, N. (2009). Justice in grades allocation: Teachers' perspective. *Social Psychology of Education, 12,* 315–325. http://dx.doi.org/10.1007/s11218-008-9073-z

Rosenholtz, S. J., & Simpson, C. (1984). The formation of ability conceptions: Developmental trend or social construction? *Review of Educational Research, 54,* 31–63. http://dx.doi.org/10.3102/00346543054001031

Sabbagh, C., Faher-Aladeen, R., & Resh, N. (2004). Evaluation of grade distributions: A comparison of Jewish and Druze students in Israel. *Social Psychology of Education, 7,* 313–337. http://dx.doi.org/10.1023/B:SPOE.0000037547.11163.36

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119–144. http://dx.doi.org/10.1007/BF00117714

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153–189. http://dx.doi.org/10.3102/00346543007313795

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. http://dx.doi.org/10.1177/0956797611417632

Simon, S., Ditrichs, R., & Grier, J. B. (1995). The simulated class as a method for studying teacher decision making. *Computers in Human Behavior, 11,* 163–180. http://dx.doi.org/10.1016/0747-5632(94)00020-I

Smeding, A., Darnon, C., Souchal, C., Toczek-Capelle, M.-C., & Butera, F. (2013). Reducing the socio-economic status achievement gap at University by promoting mastery-oriented assessment. *PLoS ONE, 8,* e71678. http://dx.doi.org/10.1371/journal.pone.0071678

Son Hing, L. S., Bobocel, D. R., Zanna, M. P., Garcia, D. M., Gee, S. S., & Orazietti, K. (2011). The merit of meritocracy. *Journal of Personality and Social Psychology, 101,* 433–450. http://dx.doi.org/10.1037/a0024618

Souchal, C., & Toczek, M.-C. (2010). Buts de réussite, conceptions de l'intelligence, différences de performances liées à l'appartenance socio-économique des élèves: De nouvelles hypothèses explicatives? [Achievement goals, conceptions of intelligence, differences of performance based on socio-economic status: New hypotheses]. *Les Sciences*

de l'Education pour l'Ere Nouvelle, 43, 13–35. http://dx.doi.org/10.3917/lsdle.431.0013

Spencer, B., & Castano, E. (2007). Social class is dead. Long live social class! Stereotype threat among low socioeconomic status individuals. *Social Justice Research, 20,* 418–432. http://dx.doi.org/10.1007/s11211-007-0047-7

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

Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S., & Covarrubias, R. (2012). Unseen disadvantage: How American universities' focus on independence undermines the academic performance of first-generation college students. *Journal of Personality and Social Psychology, 102,* 1178–1197. http://dx.doi.org/10.1037/a0027143

Stephens, N. M., Hamedani, M. G., & Destin, M. (2014). Closing the social-class achievement gap: A difference-education intervention improves first-generation students' academic performance and all students' college transition. *Psychological Science, 25,* 943–953. http://dx.doi.org/10.1177/0956797613518349

Stephens, N. M., Markus, H. R., & Phillips, L. T. (2014). Social class culture cycles: How three gateway contexts shape selves and fuel inequality. *Annual Review of Psychology, 65,* 611–634. http://dx.doi.org/10.1146/annurev-psych-010213-115143

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies, 53,* 466–478. http://dx.doi.org/10.1111/j.1467-8527.2005.00307.x

Taras, M. (2009). Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education, 33,* 57–69. http://dx.doi.org/10.1080/03098770802638671

Thorndike, E. L. (1913). *Educational psychology* (Vol. 1). New York, NY: Teachers College.

Tibbetts, Y., Harackiewicz, J. M., Canning, E. A., Boston, J. S., Priniski, S. J., & Hyde, J. S. (2016). Affirming independence: Exploring mechanisms underlying a values affirmation intervention for first-generation students. *Journal of Personality and Social Psychology, 110,* 635–659. http://dx.doi.org/10.1037/pspa0000049

Torrance, H., & Pryor, J. (1998). Introduction. In H. Torrance & J. Pryor (Eds.), *Investigating formative assessment: Teaching, learning and assessment in the classroom* (pp. 1–7). Berkshire, UK: Open University Press.

Turner, R. H. (1961). Modes of social ascent through education: Sponsored and contest mobility. In A. H. Halsey, J. Floud, & J. Anderson (Eds.), *Education, economy and society* (pp. 121–139). New York, NY: Free Press.

Uhlmann, E., & Cohen, G. L. (2005). Constructed criteria: Redefining merit to justify discrimination. *Psychological Science, 16,* 474–480.

Uhlmann, E. L., & Cohen, G. L. (2007). "I think it, therefore it's true": Effects of self-perceived objectivity on hiring discrimination. *Organizational Behavior and Human Decision Processes, 104,* 207–223. http://dx.doi.org/10.1016/j.obhdp.2007.07.001

Webster, D. M., & Kruglanski, A. W. (1997). Cognitive and social consequences of the need for cognitive closure. *European Review of Social Psychology, 8,* 133–173. http://dx.doi.org/10.1080/14792779643000100

Westfall, J. (2015, May 27). *Follow-up: What about Uri's 2n rule?* Retrieved from http://jakewestfall.org/blog/index.php/2015/05/27/follow-up-what-about-uris-2n-rule/

Wiederkehr, V., Bonnot, V., Krauth-Gruber, S., & Darnon, C. (2015). Belief in school meritocracy as a system-justifying tool for low status students. *Frontiers in Psychology, 6,* 1053. http://dx.doi.org/10.3389/fpsyg.2015.01053

Wyer, N. A. (2003). Value conflicts in intergroup perception: A social cognitive perspective. In G. V. Bodenhausen & A. J. Lambert (Eds.), *Foundations of social cognition* (pp. 263–289). Mahwah, NJ: Erlbaum.

Wyer, N. A. (2010). Salient egalitarian norms moderate activation of out-group approach and avoidance. *Group Processes & Intergroup Relations, 13,* 151–165. http://dx.doi.org/10.1177/1368430209347326

Yeager, D. S., Purdie-Vaughns, V., Garcia, J., Apfel, N., Brzustoski, P., Master, A., . . . Cohen, G. L. (2014). Breaking the cycle of mistrust: Wise interventions to provide critical feedback across the racial divide. *Journal of Experimental Psychology: General, 143,* 804–824. http://dx.doi.org/10.1037/a0033906

Zdaniuk, A., & Bobocel, D. R. (2011). Independent self-construal and opposition to affirmative action: The role of microjustice and macrojustice preferences. *Social Justice Research, 24,* 341–364. http://dx.doi.org/10.1007/s11211-011-0143-6

Zogmaister, C., Arcuri, L., Castelli, L., & Smith, E. R. (2008). The impact of loyalty and equality on implicit ingroup favoritism. *Group Processes & Intergroup Relations, 11,* 493–512. http://dx.doi.org/10.1177/1368430208095402

Received December 21, 2017
Revision received July 3, 2018
Accepted July 19, 2018 ■