Go to old article view

British Journal of Educational Psychology

Volume 84, Issue 1
March 2014
Pages 125–136

Original Article

Assessing does not mean threatening: The purpose of assessment as a key determinant of girls' and boys' performance in a science class

Carine Souchal, Marie-Christine Toczek, Céline Darnon , Annique Smeding, Fabrizio Butera, Delphine Martinot

First published:
1 March 2013 Full publication history

DOI:
10.1111/bjep.12012 View/save citation

Cited by:
1 article Refresh Citing literature

Am score 3

Funding Information

Abstract

Background

Is it possible to reach performance equality between boys and girls in a science class? Given the stereotypes targeting their groups in scientific domains, diagnostic contexts generally lower girls' performance and non-diagnostic contexts may harm boys' performance.

Aim

The present study tested the effectiveness of a mastery-oriented assessment, allowing both boys and girls to perform at an optimal level in a science class.

Sample

Participants were 120 boys and 72 girls (all high-school students).

Methods

Participants attended a science lesson while expecting a performance-oriented assessment (i.e., an assessment designed to compare and select students), a mastery-oriented assessment (i.e., an assessment designed to help students in their learning), or no assessment of this lesson.

Results

In the mastery-oriented assessment condition, both boys and girls performed at a similarly high level, whereas the performance-oriented assessment condition reduced girls' performance and the no-assessment condition reduced boys' performance.

Conclusions

One way to increase girls' performance on a science test without harming boys' performance is to present assessment as a tool for improving mastery rather than as a tool for comparing performances.

Enhanced Article Feedback

Background

Is gender equality attainable in scientific classes? According to several surveys, although teenage girls are on the way to fill the gender gap historically observed in mathematics (Else-Quest, Hyde, & Linn, **2010** ; for a review), their scores in science and math still average 12 points lower than boys' scores in industrialized countries (OECD, **2011**). Recent research indicates that girls underperform boys especially when they are placed in situations that activate the negative stereotype about their supposed poor scientific abilities (Appel, Kronberger, & Aronson, **2011** ; Huguet & Régner, **2007**). Activation of this negative stereotype is particularly likely when the test is presented as diagnostic of abilities, yet boys – unlike girls – benefit more in terms of performance from diagnostic tests than from non-diagnostic ones (Walton & Cohen, **2003**). The reversed effect of test diagnosticity for girls and boys poses a real dilemma in educational contexts, given that assessment – that is, situations involving tests that are used precisely because they are diagnostic of students' abilities – is a pervasive, structurally embedded educational practice and a necessary step for learning.

Is there a solution to this dilemma? Is it possible to use assessment at school in a way that harms neither of the gender groups? In the present paper, we argue that this possibility exists and, more specifically, that the threatening component of assessment for girls resides in the fact that assessment is mainly used as a tool for selection that emphasizes performance goals (i.e., desire to outperform others). If assessment is instead presented as a tool for education that focuses on mastery, learning-oriented goals (i.e., desire to improve one's own level of mastery of the task), the gender gap in scientific disciplines should be lowered.

Diagnosticity and boys' and girls' performance in science

The stereotype of girls as poor scientific performers has been the object of a great deal of research in psychology (Eccles *et al*., 1983 ; Hyde, Fennema, Ryan, Frost, & Hopp, 1990). Research on stereotype threat (Spencer, Steele, & Quinn, 1999 ; see Schmader, Johns, & Forbes, 2008 ; Schmader & Croft, 2011 ; for reviews) has shown that, when placed in a situation where they may confirm the negative stereotype about their gender group, female students may experience a psychological discomfort that results in a performance decrement. The stereotype threat effect has been examined in a wide range of science-related domains such as math performance (Ambady, Shih, Kim, & Pittinsky, 2001 ; Brown & Josephs, 1999 ; Quinn & Spencer, 2001), computer sciences, and engineering (Appel *et al*., 2011 ; Bell, Spencer, Iserman, & Logel, 2003 ; Smith, Morgan, & White, 2005 ; Smith, Sansone, & White, 2007). Huguet and Régner (2007) as well as Keller and Dauenheimer (2003) have shown that the stereotype threat effect could also appear in schoolchildren. Because abilities in science are precisely those believed to be lower for girls, stereotype threat can appear when a test is merely presented as diagnostic of these abilities (e.g., Bell *et al*., 2003 ; Gonzales, Blanton, & Williams, 2002 ; Huguet & Régner, 2007 ; Spencer *et al*., 1999 , Study 3). In other words, stereotype threat effects can be observed even in the absence of explicit stereotype-activating cues; test diagnosticity alone is sufficient to elicit stereotype threat among female participants, and accordingly, the difficulty is to remove threat from testing situations, not to create it (e.g., Inzlicht & Kang, 2010).

The issue of test diagnosticity and its negative consequences for girls in scientific domains represents the first side of the assessment dilemma coin. Indeed, at school, assessment is used precisely because it is diagnostic of what students are able to do at a given moment. Given the negative consequences of test diagnosticity for girls, it may be tempting to question the very use of assessment in school and recommend eradicating assessment practices from educational structures to reduce gender inequalities in the classroom. However, would this solution benefit all students?

Several lines of research indicate that assessment is not only an institutional practice aimed at carrying out orientation choices and selection, but also a tool that takes part in and supports the process of learning (Black & Wiliam, 1998 ; Gibbs & Simpson, 2004 ; Hattie & Timperley, 2007). Research also suggests that, if girls suffer from diagnostic evaluative contexts in scientific domains, boys benefit from such contexts (e.g., Bell *et al*., 2003 ; Huguet & Régner, 2007 ; Seibt & Förster, 2004). Indeed, the ‘stereotype lift’ effect (Walton & Cohen, 2003) indicates that boys perform better on a test when it is presented as diagnostic of their ability compared to when it is not (i.e., control or low-threat group; see also Walton & Spencer, 2009). This stereotype lift effect occurs because men benefit, in the diagnostic condition, from positive stereotypic expectations and therefore from downward social comparison with the devalued group of women. Consequently, it appears that – although removing the diagnostic presentation of a test may enhance girls' performance – it may also represent a suboptimal context for boys' performance. This is the second element of the test diagnosticity dilemma.

Considering the two elements of the dilemma discussed thus far, teachers and policymakers may continue to wonder whether they should or should not use diagnostic assessment in classes. We suggest that this dilemma can be solved by reframing the purpose of assessment.

The purpose of assessment

We argue that what is threatening for low-status groups (e.g., women) is not assessment *per se*, but rather the purpose of assessment. Recent research indicates that educational systems have two main functions: to educate pupils and students and to select people – namely, to assign, or not, grades and degrees in order to orient people to various positions in the social hierarchy (Darnon, Dompnier, Delmas, Pulfrey, & Butera, **2009**; Darnon, Dompnier, & Poortvliet, **2012**; Dornbusch, Glasgow, & Lin, **1996**; Duru-Bellat, **2009**). Interestingly, assessment is a tool that can serve both an educational and a selection function (Bloom, Hastings, & Madaus, **1971**). On the one hand, assessment helps the learner achieve mastery of the task by providing formative and corrective feedback (Black & Wiliam, **1998**). On the other hand, assessment serves as a summative and certificative function in that it helps teachers decide who, among students, deserves a degree and who does not (Brookhart, **2001**, **2004**; Dornbusch *et al*., **1996**).

At the structural level, the distinction between formative and summative functions of assessment echoes a distinction that has received great attention at the individual level: the distinction between ‘mastery’ or ‘learning’ goals (desire to increase one's learning) and ‘performance’ goals (desire to perform well or not to perform poorly as compared to others; Dweck, **1986**).[1] Research in this area has documented that goals affect the way one reacts to an academic task (for reviews, see Elliot, **2005**; Hulleman, Schrager, Bodmann, & Harackiewicz, **2010**), including the reaction to failure (Diener & Dweck, **1978**; Dweck & Leggett, **1988**), intrinsic interest (Rawsthorne & Elliot, **1999**), and conflict regulation (Darnon, Butera, & Harackiewicz, **2007**; Darnon, Muller, Schrager, Pannuzzo, & Butera, **2006**).

According to Ames (**1992**), the assessment process is one of the most powerful factors for eliciting mastery versus performance goals (see also Brookhart, **1997**; Pulfrey, Buchs, & Butera, **2011**). Assessment practices that focus on normative standards increase performance goals, whereas assessment practices that emphasize the importance of progress are likely to enhance mastery goals (Butler, **2006**). In the remainder of this article, we will refer to the latter type of assessment as ‘mastery-oriented assessment’ and contrast it to performance-oriented assessment (i.e., based on normative social comparison between students).

Mastery-oriented assessment and girls' and boys' performance

As previously mentioned, diagnostic assessments may threaten girls in science disciplines. In the present research, we argue that the reason why assessment impairs girls' performance in science is not diagnosticity *per se*, but the fact that diagnosticity of assessment is most often used for selection purposes. Negatively stereotyped groups feel vulnerable in a comparative, competitive, selective environment. Interestingly, regarding this contention, research on achievement goals indicates that social comparison concerns depend on the salience of goals (Bounoua *et al*., in press; Darnon, Dompnier, Gilliéron, & Butera, 2010). Whereas social comparison threatens self-competence in a performance-goal situation (Jagacinski & Nicholls, 1987; Ryan & Pintrich, 1997), it is less the case in a mastery goal context, where others are not perceived as threats but as peers with whom collaboration is likely to occur (Poortvliet & Darnon, 2010). Moreover, mastery goals favour a focus on information relevant to task solving; meanwhile, performance goals lead individuals to focus on self-thoughts related to one's own competence (Butler, 1992). Interestingly, increased salience of (negative) self-related thoughts (Cadinu, Maass, Rosabianca, & Kiesner, 2005) and the fear of performing poorly compared to others (Brodish & Devine, 2009) are some of the mechanisms responsible for stereotype threat effects. This is probably why some authors consider that, by default, stereotype threat situations are similar to performance goals – but not mastery goals – situations (see Kaplan & Maehr, 2007; Ryan & Ryan, 2005). In line with this idea, some research suggests that women suffer most from the negative effects of performance goals (e.g., Jagacinski, Kumar, & Kokkinou, 2008). For boys, performance goals can even have a positive effect on the use of efficient learning strategies (Bouffard, Boisvert, Vezeau, & Larouche, 1995).

In the present study, a performance-oriented and a mastery-oriented assessment of a science class will be compared to a no-assessment situation. We hypothesize that girls will perform better on a science test in mastery-oriented and no-assessment situations compared to a performance-oriented assessment condition. Meanwhile, as previously discussed, boys will not suffer from a performance-goal situation, although the no-assessment situation should reduce their performance. Moreover, unlike the no-assessment situation, the mastery-oriented assessment implies test diagnosticity, which should allow for a sufficient level of visibility and motivational incentive for boys to perform well on the task. Therefore, we expect boys to perform better in the performance-oriented and the mastery-oriented assessment conditions relative to the no-assessment condition.

Method

Participants

One hundred and ninety-three high-school students from nine classes participated in this study during one of their obligatory science class. One participant was removed from the analyses because of an uncommon studentized deleted residual. The remaining participants were 120 boys and 72 girls (mean age = 15.6, *SD* = 0.74). Between 23 and 25 girls and between 39 and 41 boys were randomly assigned to each of the three experimental conditions.

Procedure

Classes were divided into two groups of students that corresponded to two of the three conditions (mastery–performance, mastery–no assessment, or performance–no assessment) and taken to two different rooms. Each group was taught by one of the two female experimenters, who introduced themselves as future teachers. They first explained that students would be taught a class on aspirin and then, depending on the condition, that they would take, or not, a test on this class.

In the performance-oriented assessment condition, the instructions were as follows:

> At the end of today's lesson, you will take a test. On the basis of this test, you will receive a grade. This test will help us compare your abilities to that of other students in the class. You have to know that this grade will count in your final semester grade.

Right before the assessment, students were reminded of the fact that the purpose of the assessment was to see how they were doing compared to others. In the mastery-oriented assessment condition, the instructions were as follows:

> At the end of today's lesson, you will take a test. On the basis of this test, you will receive a grade. This test will help you memorize and understand the lesson well. You will see that, even during the test, you will continue to learn. You have to know that this grade will count in your final semester grade.

Before the assessment, they were reminded of the fact that the purpose of the assessment was to help them in the learning process. Finally, in the no-assessment condition, the instructions were as follows:

> At the end of today's lesson, you will have to answer some questions. It is important to note that you will not be evaluated on this lesson. Indeed, you will not take any test about today's session.

Before the assessment, students were told:

> Now please answer some questions about today's lesson. I will explain to you why I am interested in your answers to these questions after the whole session. Just keep in mind that, as I told you before, today's lesson will not be evaluated.

The lesson lasted approximately 30 min and presented a variety of information about aspirin (its history, chemical formula, etc.). The two experimenters were trained to provide the exact same content in the class and spend the same amount of time on each part of the lesson. After the class, the experimenter restated the experimental instructions, and then, all students took a test and answered the manipulation check measures. Students were then thanked, debriefed, and explained that this test would actually not count in their final grade.

Materials

Manipulation checks

Participants were asked to report whether they thought they would be evaluated. If they answered 'yes', they were asked the extent to which they thought this evaluation was 'designed to help you in the learning process', 'designed to help you memorize the content of the class', 'aimed at helping you understand well the class', or 'part of the learning process' ($\alpha = .82$; $M = 4.81$; $SD = 1.32$; mastery-oriented assessment) or 'designed to measure your competences compared to those of the other pupils', 'designed to identify whether you are better or worse than the other pupils', 'designed to compare you to other pupils', or 'designed to see how you do compared to others' ($\alpha = .90$; $M = 3.35$; $SD = 1.63$; performance-oriented assessment).

Performance

The test contained 10 multiple-choice questions related to the lesson. For each question, participants had to choose the correct answer among the three suggested (e.g., Which product must be mixed with acetic anhydride to obtain aspirin? Phenol, soda, or salicylic acid?). Scores could range from 0 to 10 ($M = 8.6$; $SD = 1.45$).

Results

Manipulation checks

All participants answered the question of whether they would be evaluated or not according to the experimental instructions they received. The two assessment conditions were further compared. Participants from the mastery-oriented assessment condition more often perceived the assessment to be helpful in the learning process ($M = 5.06$; $SD = 1.41$) than participants from the performance-oriented assessment condition ($M = 4.56$; $SD = 1.16$), $F(1, 121) = 4.54$, $p < .04$, $\eta^2 = .04$. Symmetrically, participants from the performance-assessment condition ($M = 3.96$; $SD = 1.59$) perceived the assessment to be aimed at comparing students to each other more than participants from the mastery-oriented assessment condition ($M = 2.73$; $SD = 1.42$), $F(1, 118) = 19.89$, $p < .001$, $\eta^2 = .14$.[2]

Performance

Regarding performance, we expected a drop for girls in the performance-oriented assessment condition and for boys in the no-assessment condition. To test the model summarizing our predictions, the variance was decomposed into two orthogonal contrasts: one one-degree-of-freedom planned comparison testing the model and the other testing the remaining variance. If the model fits the data, the first contrast should be significant, but not the second (Judd & McClelland, 1989). The performance-assessment condition for girls and the no-assessment condition for boys were each coded −2, as a drop of performance was expected for these conditions only. The four remaining conditions were each coded +1. The contrast testing our predictions is presented in Table 1. The second contrast tested the remaining effects after the model was removed. Because preliminary analyses revealed an experimenter effect, $F(1, 180) = 6.18$, $p < .02$, $\eta^2 = .04$, indicating that performance was higher with one of the experimenters ($M = 8,89$, $SD = 1.4$) than with the other ($M = 8,37$, $SD = 1.47$), this variable and its interactions with other variables were included in the analyses.

Table 1. Contrast of interest

	No assessment	Performance-oriented assessment	Mastery-oriented assessment
Girls	1	−2	1
Boys	−2	1	1

The 2 (sex) × 3 (assessment condition) × 2 (experimenter) ANOVA indicated that the interaction between condition and sex was significant, $F(2, 180) = 3.21$, $p < .05$, $\eta^2 = .03$. More importantly, the contrast testing the model was significant, $F(1, 180) = 9.40$, $p < .003$, $\eta^2 = .05$, whereas the contrast testing the residual was not, $F(10, 180)= 1{,}36$, *n.s*. Thus, in line with predictions, girls' performance suffered in the performance-oriented assessment condition, whereas boys' performance suffered in the no-assessment situation. The only condition in which both boys and girls performed at their optimal level was the mastery-oriented assessment condition. Means are presented in Figure 1.

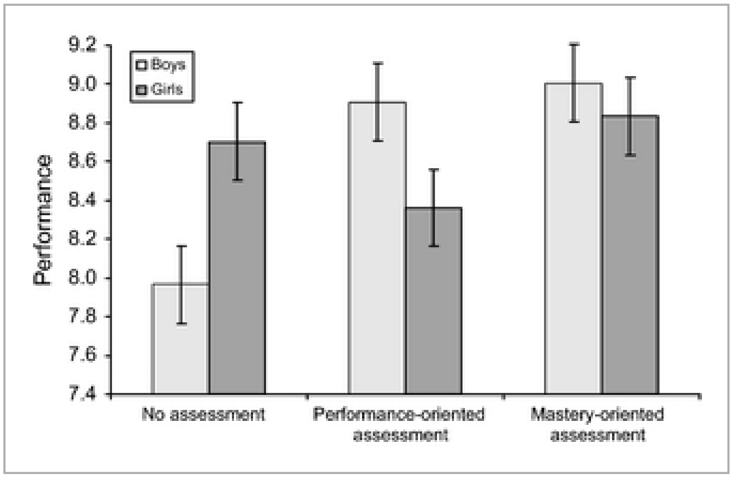

Figure 1.

Open in figure viewer

Performance at the science test as a function of assessment type and gender.

Discussion

Extensive literature has shown that merely presenting a test as diagnostic of scientific abilities is sufficient to observe a gender gap favouring boys' performance (e.g., Bell *et al*., **2003** ; Davies, Spencer, Quinn, & Gerhardstein, **2002** ; Gonzales *et al*., **2002**). This finding poses a fundamental moral question in terms of assessment practices as they are used in schools. One can indeed argue that one way to reduce inequalities (including gender inequalities) between students could be to remove the assessment process from school. However, this solution is certainly neither optimal nor realistic as assessment is an integral part of the pedagogical process and may be used to promote learning (Brookhart, **1997**; Crooks, **1988**). Moreover, the stereotype lift effect (Walton & Cohen, **2003**) suggests that no-assessment situations may harm the performance of groups associated with a positive stereotype, such as boys in science, when compared to diagnostic situations.

Is it possible, then, to solve this dilemma and assess students' scientific performance in a way that allows both girls and boys to perform at an optimal level? In the present research, we argued that one way to increase girls' performance on a science test without harming boys' performance is to present assessment as a tool for improving mastery rather than as a tool for comparing performances. Our results supported the hypothesis that in the mastery-oriented assessment condition, both boys and girls performed at a similarly high level, whereas the performance-oriented assessment condition reduced girls' performance and the no-assessment condition reduced boys' performance.

The first contribution of the present research is that it provides empirical evidence supporting the hypothesis that assessment in science is threatening for girls, not because it is diagnostic of abilities *per se*, but because diagnosticity may be used to compare abilities and *in fine* select students, which is bound to be detrimental to negatively stereotyped group members. Indeed, our results showed that, with a mastery-oriented assessment – a diagnostic assessment – girls performed better than with a performance-oriented assessment and at an equally high level as in the no-assessment condition. The second contribution is that this study demonstrates that there is no need to eliminate assessment altogether to favour girls' performance, which would impair boys' performance. Using assessment as a learning tool provides enough diagnosticity for boys to perform well and eliminates the threatening reference to comparative selection, allowing girls not to underperform.

Some limitations should be noted. First, although stereotype threat and lift effects may explain the results, we manipulated the processes argued to be the origin of both girls' underperformance in science (the comparative and selective aspect of assessment) and boys' (the lack of diagnosticity); but we have no direct evidence that stereotypes were involved. It should be highlighted, however, that according to previous research, test diagnosticity alone is sufficient to elicit stereotype threat (Bell *et al*., **2003**; Gonzales *et al*., **2002**; Huguet & Régner, **2007**; Spencer *et al*., **1999**, Study 3). Moreover, it should be noted that even if we did not have a domain identification measure, science and mathematics – along with French – represent the cornerstone of the academic curriculum in France. Because it has proved difficult to devalue a domain that is highly valued (Crocker & Major, **1989**; Steele, Spencer, & Aronson, **2002**), students should be domain-identified to at least some extent, resulting in threat effects among girls. However, to strengthen our explanation, future research may replicate the present results while manipulating variables directly related to gender stereotypes. A second concern is that, in the present study, students received goal manipulations before the learning phase; thus, it is hard to know whether the threat occurred during learning, during testing, or both (Appel *et al*., **2011**; Rydell, Rydell, & Boucher, **2010**). Future research should examine whether assessment manipulations presented after the learning phase produce similar results. Moreover, goal measures should be included to make sure the assessment inductions resulted not only in different perception of the assessment but also on different goal states. Finally, the present research focused on the approach forms of mastery and performance goals, but future research should also examine the effects of performance-avoidance-oriented assessment. Because performance-avoidance goals are generally associated with threat and anxiety, such a condition should impair both boys' and girls' performance.

Notwithstanding these limitations, the present results have important practical implications. Interestingly, several methods have been proposed to reduce the performance gap between boys and girls in scientific domains. For example, some authors have proposed that promoting self-affirmation (Martens, Johns, Greenberg, & Schimel, **2006** ; Miyake *et al* ., **2010**), informing participants about the stereotype threat effect (Johns, Schmader, & Martens, **2005**), presenting same gender role models (Marx & Roman, **2002**), or role models who have been successful thanks to regular efforts (Bagès & Martinot, **2011**) could lead girls to perform as well as boys in a scientific domain. These studies are encouraging. However, all these interventions consist of helping students cope with the threat; thus, they are all focused on individuals. Classroom practices that generate the threat are rarely questioned. In the current research, we do not document how to individually cope with the threat, but rather how the educational system could change the meaning – and, most importantly, the purpose – attributed to assessment so as not to threaten students. In particular, our research questions the selection function of the educational system and the practices used to exert this function. We believe that, as long as educational institutions have to select and classify people, it will be hard to convince students to focus on the learning of the lessons and not be threatened by diagnostic situations. Therefore, educational institutions should make clear that their role is to educate students and design and use assessment practices accordingly. Only in such a context will students understand that they are in school to learn and not to 'make it through the filter'.

Acknowledgement

This research was supported by the French Agence Nationale de la Recherche ('Gender Inequality Maintenance') the Conseil régional Auvergne and the Swiss National Science Foundation.

[1] Goal research also makes a distinction between approach and avoidance goals within mastery and performance goals (Elliot & McGregor, **2001**). Because they are the most relevant regarding our hypotheses, in the present research, we will focus on the approach-oriented goals, namely performance-approach and mastery-approach goals.

[2] Differences in degrees of freedom are due to missing values on this variable.

References

Citing Literature

Help **Browse by Subject** **Browse Publications** **Resources**

Agents | Advertisers | Cookies | Contact Us | About Us

Privacy | Site Map | Terms & Conditions | Media

WILEY

Wiley.com **About Wiley** **Wiley Job Network**