Please cite as: Greifeneder, R., Zelt, S., Seele, T., Bottenberg, K., & Alt, A. (2012). Towards a better understanding of the legibility bias in performance assessments: The case of gender-based inferences. *British Journal of Educational Psychology, 82*, 361-374. doi: 10.1111/j.2044-8279.2011.02029.x

Address correspondence to: rainer.greifeneder@unibas.ch

Towards a better understanding of the legibility bias in performance assessments: The case of gender-based inferences

Rainer Greifeneder, Sarah Zelt, Tim Seele,
Konstantin Bottenberg, and Alexander Alt

University of Mannheim, Germany

Structured Abstract

Background. Handwriting legibility systematically biases evaluations in that highly legible handwriting results in more positive evaluations than less legible handwriting. Because performance assessments in educational contexts are not only based on computerized or multiple choice tests but often include the evaluation of handwritten work samples, understanding the causes of this bias is critical.

Aims. This research was designed to replicate and extend the legibility bias in two tightly controlled experiments and to explore whether gender-based inferences contribute to its occurrence.

Sample(s). A total of 132 students from a German university participated in one pre-test and two independent experiments.

Method. Participants were asked to read and evaluate several handwritten essays varying in content quality. Each essay was presented to some participants in highly legible handwriting and to other participants in less legible handwriting. In addition, the assignment of legibility to participant group was reversed from essay to essay, resulting in a mixed-factor design.

Results. The legibility bias was replicated in both experiments. Results suggest that gender-based inferences do not account for its occurrence. Rather it appears that fluency from legibility exerts a biasing impact on evaluations of content and author abilities.

Conclusions. The legibility bias was shown to be genuine and strong. By refuting a series of alternative explanations, this research contributes to a better understanding of what underlies the legibility bias. The present research may inform those who grade on what to focus and thus help to better allocate cognitive resources when trying to reduce this important source of error.

Keywords: education; grading; handwriting; legibility; evaluation; fluency; bias

Towards a better understanding of the legibility bias in performance assessments: The case of gender-based inferences

Evaluations of handwritten material are subject to a series of biases, including, for instance, teacher expectations (e.g., Jussim & Eccles, 1992), writer attractiveness (Landy & Sigall, 1974), and handwriting legibility (e.g., James, 1929; Markham, 1976). Many of these biases occur because the underlying evaluative processes hinge on inference rules that are frugal but not perfectly accurate. For instance, evaluations of content are biased by composition errors (Marshall, 1967), presumably because evaluators infer content quality from writing quality—after all, when care was applied to composition, care was probably applied to content, too. This inference rule is parsimonious but not necessarily valid, thus exemplifying the way in which performance assessments may become unintentionally biased because of the nature of the underlying evaluative processes.

Conceptually, such inference rules may be described as associatively represented heuristics (Smith & DeCoster, 2000). These inference rules are learned over time, and may change by new learning. This suggests that at least some biases in performance assessment may be reduced when those who grade are aware of possible biases in evaluation. Towards this goal, our research re-examines the *legibility bias*. The legibility bias holds that legible handwritten materials are evaluated more positively than those less legible (e.g., James, 1929).

The present experiments test whether gender-based inferences contribute to the legibility bias. In particular, it may be that legible essays are evaluated more positively than less legible essays because graders form inferences from legibility to author gender, and from author gender to academic performance. Knowing whether such inferences contribute to the legibility bias (or not) is critical because such knowledge allows for a more focused “battle” against this highly consequential source of error. This appears important because evaluations of handwritten materials may have serious consequences in all stages of educational life, including regular performance assessments as well as school entrance exams, final tests, and procedures determining the eligibility for scholarships. A better understanding of what underlies the legibility bias may therefore have a potentially large impact on educational practice. In what follows, we first review earlier findings on the legibility bias, then focus on the processes presumably underlying its occurrence.

Legibility defined

We define legibility as the degree to which handwritten material is perceived as readable. Theoretically, legibility may range from illegible to highly legible. Illegible material, however, is rarely considered in empirical studies because legibility would be the sole possible information source in evaluations (e.g., Briggs, 1970; Hughes, Keeling, & Tuck, 1983). More interesting is whether legibility biases judgments when other sources of information—such as arguments, story thread, logic—are assessable. To address this question while following the lead of earlier contributions, our experiments rely only on readable materials.

The legibility bias

More than eighty years ago, James (1929) reported that British senior high school teachers evaluated legible student essays more positively than those less legible. Pointing to the bias' potential for harm, James noted that these differences were comparable to one letter grade. Unfortunately, this evidence was open to explanations other than biased performance assessment. Most prominently, evaluators may have awarded a premium for legible handwriting, or penalized less legible handwriting, because good penmanship was a virtue at that time.

Later findings support James' claim, but again were open to alternative explanations. For instance, Briggs (1970; Markham, 1976) reported a legibility bias in a sample of elementary school teachers. However, elementary school teachers teach penmanship and may therefore be expected to spontaneously take its mastery into account. Other research reported legibility biases for essay topics that lack objective content criteria, such as “Hopes and aspirations for the next decade” (e.g., Hughes et al., 1983; James, 1929). Potentially, this lack of diagnostic information invited the consideration of other information sources, such as penmanship.

To forestall such alternative explanations, the influence of legibility when explicit content criteria for essay evaluation are provided has been investigated more recently (Greifeneder et al., 2010). Nevertheless, legibility influenced evaluations of presumed authors and the grades assigned to their essays. Alarmingly, the legibility bias averaged half a grade point in the six point German grade system. This research further demonstrated that the legibility bias is not due to differences in handwriting beauty, because when handwriting legibility and handwriting beauty were orthogonalized, the legibility bias prevailed. It was also shown that the legibility bias is not contingent on specific samples, but is powerful in both student and teacher populations.

Greifeneder and colleagues (2010) explain the legibility bias in terms of a basic underlying cognitive process. The authors argue (a) that legible versus less legible material can be processed more fluently, (b) that individuals misperceive the fluency associated with legible material as a signal of positivity and the disfluency associated with less legible material as a signal of negativity, and (c) that this signal guides evaluations of handwritten materials and their authors. Although this explanation may not appear intuitive on first glance, its tenets have received strong support in the literature (e.g., Alter & Oppenheimer, 2009; Greifeneder, Bless, & Pham, in press). Perhaps most importantly, fluency has been shown to influence a wide variety of judgments, such as liking (e.g., Greifeneder & Bless, 2007; Reber, Winkielman, & Schwarz, 1998) or intelligence (e.g., Oppenheimer, 2006).

From this perspective, the legibility bias occurs because of an incorrect inference that fluency from handwriting legibility is diagnostic of the quality of the handwritten material. In support of this perspective, the legibility bias can be eliminated by informing participants about the biasing impact of fluency (Greifeneder et al., 2010). On a more general level, this suggests that understanding the mechanisms underlying the legibility bias may help to reduce it.

Inferences about author gender

Like many social phenomena, the legibility bias probably has multiple causes, so that inference rules other than fluency may contribute to its occurrence. One likely candidate is ascribed author gender, in that individuals may form inferences from legibility to author gender, and from author gender to academic performance. Gender inferences are a particularly plausible candidate because gender is (a) a highly salient characteristic and (b) associated with many readily available stereotypes, thus allowing for frugal inferences (e.g., Eagly, Beall, & Sternberg, 2004; Eagly & Mladinic, 1989).

A simple gender-based inference rule may build on spontaneous attributions of legible handwriting to females, and less legible handwriting to males, because females are generally believed to have more legible handwriting (e.g., Burr, 2002). Assuming that females generally perform better at school (e.g., Dwyer & Johnson, 1997; Pomerantz, Altermatt, & Saxon, 2002), this gender ascription would allow for quick performance inferences. These inferences may result in a pattern of findings comparable to the legibility bias observed in earlier research. In particular, legible handwriting would be associated with females and therefore also with more positive evaluations, and less legible handwriting with males and therefore with more negative evaluations. Gender-based inferences thus constitute plausible alternative explanations to earlier evidence.

In the interest of a more complete understanding of what causes and cures the legibility bias, it is critical to explore such alternative inferences. The present contribution attains this goal by testing—in an exploratory way—whether simple or more complex gender-based inferences may account for the legibility bias. To this end, we built on the paradigm introduced by Greifeneder and colleagues (2010) but extended it by the critical control of (Experiments 1 and 2) and influence on (Experiment 2) gender-related inferences.

Experiment 1

Experiment 1 aimed at two goals. First, to attest to the prevalence of the legibility bias, we wanted to replicate the finding that legible essays are evaluated more positively than those less legible in conditions where penmanship is unlikely to be of concern. We assumed this would be the case for university students, because the time when penmanship may be expected to affect grades—i.e., elementary school—has long passed for them.

Second, and more importantly, we wanted to explore gender-based inference rules as a contributing factor. To this end, we selected essay materials from the domain of physics, that is, a domain that is *stereotypically* associated with better performance by male students (e.g., Ehindro, 1986; Stewart, 1998). In this domain, a legible essay should be evaluated *less* positively if the legibility bias is driven by simple gender-based inferences, because more legible handwriting would suggest a female author, who would be stereotypically associated with a less competent performance in physics (vice versa for less legible handwriting). Furthermore, in order to explore gender-based inferences via mediation analyses, participants were asked to guess the presumed authors' gender. Should gender-based inferences contribute to the legibility bias, ascriptions about gender should mediate the effect of legibility on evaluations.

Method

Participants and Design. Fifty-seven University of Mannheim students (52 female; mean age 21.4 years, *SD* = 4.44) received course credit to participate in a study on "interrater reliability." To match the situation graders usually encounter, we varied both the materials' content quality and legibility. Participants were tested in groups of varying size and were asked to evaluate the authors of a good, a medium, and a poor essay. Because we wanted all participants to evaluate essays of both high and low legibility, participants were divided into two experimental groups. The first group was presented with the good, medium, and poor essay in, respectively, low, high, and low legibility (low-high-low). The second group was presented with the same content material in the identical pre-determined order, but in high, low, and high legibility (high-low-high). Together, these manipulations resulted in a 3 (content quality: good, medium, poor) x 2 (group: low-high-low vs. high-low-high legibility) mixed-factorial experimental design, with content quality as the within factor (for a concise overview, see Table 1).

This design offers several advantages. *First*, by manipulating content quality across essays, the extent to which participants are responsive to variability in content quality can be monitored, thus attesting to the design's ecological validity. In addition, the content quality manipulation allows for testing whether the legibility bias is general or restricted to certain levels of content quality. For instance, one might assume that legibility matters only in poor essays, because these lack diagnostic information. *Second*, by manipulating legibility both within and between groups, the hypothesized legibility bias can be tested more rigorously. Specifically, because every essay is presented in both high and low quality, the hypothesized effect of legibility should be manifest when comparing the evaluations for each of the three essays *between* groups (for every essay, a main effect of legibility is expected). In addition, because the assignment of legibility to the good, medium, and poor essay was systematically reversed between experimental groups (group 1: low-high-low vs. group 2: high-low-high), the direction of the hypothesized legibility effect should alternate from essay to essay. This alternation should result in an interaction effect of content quality (good, medium, poor) and group assignment across the three essays. *Third*, although essay materials were carefully pretested (see below), there is a possibility that other factors unintentionally covary with legibility. By using different sets of handwriting across the three essays, this third-variable problem can be greatly reduced.

Essay construction. Several typed essays of similar length that varied in the amount of correct content information were constructed. One good, one medium, and one poor essay were selected based on the number of correct statements (6, 4, or 2, respectively). A mixed sample of 21 students was then asked to copy the three typed essays in their usual, cursive handwriting, each on a separate, blank sheet of paper. Based on informal evaluation, for every essay a highly legible and a less legible version were selected, while making sure that all words in all essays were readable. These final six essays were subjected to formal pre-testing, which followed the above described 3 x 2 mixed-factorial design. Because the pre-test was run on desktop computers, the handwritten essays were scanned, preserving original dimensions. Pre-test participants (*N* = 48, 24 female) were seated in front of computer screens and were asked to read each of the handwritten essays at their normal reading speed. For each essay, processing latencies were recorded as a proxy for reading fluency. Subsequently, pre-test participants were asked to revisit each essay and to subjectively evaluate handwriting legibility (1, *easy to read*, to 6, *difficult to read*).

First, natural log-transformed latencies (Fazio, 1990) were subjected to planned contrast analyses. For every essay, the highly legible version was read faster than the less legible version, $|t(46)| = 1.12$, $p < .27$, $|t(46)| = 2.53$, $p < .02$, $|t(46)| = 4.52$, $p < .01$, for the good, medium, and poor essay, respectively. Note that the first comparison failed to reach conventional levels of significance; potentially, this is because participants first needed to become acquainted with the reading task. Second, participants' *subjective* ratings of legibility were submitted to the same set of planned contrast analyses. For all three essays, the highly legible version was evaluated as more legible, all $|ts(46)| > 6.70$, $ps < .01$. Together, these results attest to the suitability of the selected material for systematically producing different degrees of legibility.[1]

Procedure. In the main experiment, instructions and materials were presented in paper-pencil format. Participants were asked to carefully read a short paragraph (74 words) about a physics topic, “The emergence of lightning,” on which they would subsequently base their evaluations of essays. These essays were supposedly written by students as part of a teaching assignment. The standard paragraph and the student essays were printed on separate pages; participants read and evaluated the essays one at a time, starting with an example before working on the three target essays.

Evaluation of author abilities. Following each essay, participants were asked to evaluate the presumed author with regard to general academic competence, knowledge of other school subjects, diligence, time spent studying, verbal expressiveness, and abilities in other domains. Evaluations were assessed on six-point Likert scales (1, *high*, to 6, *low*).

Author gender. After evaluating author abilities, participants were asked to guess the presumed author’s gender on a dichotomous scale (1, *male*, 2, *female*).

Results

Author abilities. The six items assessing author abilities were individually rescaled so that higher scores indicated more positive evaluations and were averaged into one single index per essay (all Cronbach’s α > .87). These indices were subjected to a 3 (content quality: good, medium, poor) x 2 (group: low-high-low vs. high-low-high legibility) mixed-effects analysis of variance (ANOVA), with content quality as within-factor. Results revealed a main effect of content quality, such that presumed authors were evaluated more positively the higher the quality of the essays, $F(2, 110) = 83.02$, $p < .01$, $\eta_p^2 = .60$. This main effect suggests that participants took content quality into account when evaluating author abilities and attests to the experiment's ecological validity.

Recall that the experiment was designed so that the legibility bias becomes apparent in (a) an interaction effect across the three essays and (b) between-group differences for every essay. In support of these predictions, results revealed that highly legible handwriting led to more positive evaluations of the presumed author than less legible handwriting, as reflected in the hypothesized interaction effect, $F(2, 110) = 16.80$, $p < .01$, $\eta_p^2 = .23$ (main effect group, $F < 1.4$). Further planned contrasts indicated that more, as opposed to less, legible handwriting resulted in more positive evaluations for each of the three essays, all $|t\text{s}(55)| > 2.42$, $p\text{s} < .02$, $d\text{s} > 0.64$ (see Fig. 1).

Author gender. Participants' beliefs about author gender were subjected to three independent χ^2-tests (one per essay). Results show that for every essay legible handwriting was largely attributed to female authors, and less legible handwriting to male authors, all $\chi^2\text{s}(1) > 37.22$, $p\text{s} < .01$.[2] In order to test whether inferences based on gender ascription contribute to the legibility bias, mediation analyses were performed following Baron and Kenny (1986). We analyzed separately for the good and the medium essays whether the effect of legibility (independent variable) on author abilities (dependent variable) is mediated by ascribed author gender. (Analyses could not be performed for the poor essay, because legibility and ascribed author gender were perfectly correlated.) For both essays, (a) legibility predicted ratings of author abilities, $|\beta\text{s}| > .30$, $|t\text{s}| > 2.40$, $ps < .02$; (b) legibility was highly correlated with the presumed mediator, ascribed author gender, all $C\text{s} > .62$, $ps < .01$ (contingency coefficients); but (c) ascribed author gender did not predict ratings of author abilities when simultaneously controlling for the direct relationship between legibility and author abilities, $|\beta\text{s}| < .13$, $|t\text{s}| < 1$.

Discussion

Results of Experiment 1 suggest that handwriting legibility systematically biases evaluations of those who are assumed to have authored the essays. The effect occurs for good, medium, and poor essays, and is significant between groups (for every essay) and across essays (interaction effect). Moreover, the legibility bias occurred despite the fact that essay materials were selected from the domain of physics. Unlike general performance at school, the domain of physics is stereotypically associated with better performance by males (e.g., Ehindro, 1986), so that a simple gender-based inference rule would yield the opposite of a legibility bias. Because we still observed a strong legibility bias, one may conclude that simple gender-based inferences do not contribute to the bias' occurrence. In support of this speculation, ascribed gender did not mediate the legibility bias.

Experiment 2

By using materials from the domain of physics, Experiment 1 ensured against simple gender-based inference rules. More complex gender-based inference rules, however, remain viable. For instance, females have been shown to be subject to stereotype threat in natural sciences—that is, they may perform poorly because they are concerned about being evaluated based on an existing negative group stereotype (e.g., Quinn & Spencer, 2001; Steele, 1997). Since knowledge about stereotype threat has been widely disseminated (e.g., Dewar, 2010), participants may have heard about stereotype threat before coming to the laboratory. In this case, participants may have evaluated essays in legible handwriting more positively because they attributed legible essays to females and then compensated for the presumed threat, such as by applying more lenient standards when evaluating the stereotyped group (for evidence on shifting standards, see Biernat, Manis, & Nelson, 1991). Critically, this would result in a pattern comparable to the legibility bias even with essay materials from the domain of physics.

This more complex explanation hinges on a stereotypical belief that is specific to the natural sciences. If this stereotypical belief played a critical role, more legible essays would be evaluated more positively than less legible essays when the stereotypical belief applies—such as in physics (e.g., Ehindro, 1986). But when the subject domain is associated with a different stereotypical belief—such as in the domain of education—the legibility bias should not occur. Building on this logic, participants in Experiment 2 were asked to evaluate not only the three physics essays, but also three education essays. If participants in Experiment 1 evaluated the legible physics essays more positively because of gender-specific stereotypical beliefs about who has legible handwriting and who performs well or poorly in natural sciences, evaluations of the education essays should not (or in the opposite direction) be biased by legibility.

Experiment 1 assessed evaluations of author abilities as the dependent variable and demonstrated that the legibility bias may affect how we perceive others, thus illustrating the bias' potential for harm. This choice was motivated by the fact that evaluations about others are often spontaneously formed (e.g., Osgood, Suci, & Tannenbaum, 1957) and may thus occur even when those who grade focus on content. In addition to evaluations of people, however, it appeared desirable to collect direct evidence of the legibility bias in evaluations of content. To this end, Experiment 2 first assessed grades assigned to the handwritten material and then evaluations of author abilities.

Method

Participants and Design. Twenty-seven University of Mannheim students (22 female; mean age 21.8 years, *SD* = 2.00) participated in return for 2 EUR and a chocolate bar. As we wanted all participants to evaluate essays from the domains of both physics and education, participants first worked on a good, a medium, and a poor physics essay (just as in Experiment 1), and then on a good, a medium, and a poor education essay. Again, participants were randomly assigned to two groups. The first group was presented with the six essays in low, high, low, high, low, high legibility. The second group was presented with the same content material in the identical pre-determined order, yet in high, low, high, low, high, low legibility. These manipulations resulted in a 6 (content quality: good, medium, poor, good, medium, poor) x 2 (group: low-high-low-high-low-high vs. high-low-high-low-high-low legibility) mixed-factorial design, with content quality as within factor, and essay topic—physics versus education—nested within.

Procedure. Procedures were similar to those of Experiment 1, except for the following two changes. First, after evaluating the three physics essays, participants were asked to read a new short standard paragraph, “The permissive parenting style” (70 words), as a prototypical education topic. Participants then read and evaluated essays 4 to 6 on this topic. Second, for every essay, participants assigned a grade to the respective essay before evaluating the abilities of the author.

Essay construction. To construct education essays, a series of typed essays of similar length that varied in the amount of correct content information with respect to the standard paragraph was composed. One good, one medium, and one poor essay were selected based on the number of correct statements (6, 4, or 2, respectively). A mixed sample of 30 students was then asked to copy these essays in their usual cursive handwriting, each on a separate, blank sheet of paper. Based on informal evaluation, for every essay, a highly legible and a less legible version were selected, with the constraint that all words in all essays were readable.

Evaluations. For every essay, participants first assigned a grade on a scale from 1, *excellent*, to 6, *insufficient* (standard German grade scale), then evaluated the abilities of the author (six items), and finally guessed the author’s gender.

Results

Assigned grades. Grades were individually rescaled so that higher values indicate more positive evaluations and were subjected to a 6 (content quality: good, medium, poor, good, medium, poor) x 2 (group: low-high-low-high-low-high vs. high-low-high-low-high-low legibility) mixed-factorial ANOVA.[3] The better the essay's content quality, the better the grades, as reflected in a significant content quality main effect, $F(5, 120) = 52.06$, $p < .01$, $\eta_p^2 = .68$. More importantly, the more legible the handwriting, the better the essay's evaluation, as reflected in the hypothesized interaction effect, $F(5, 120) = 2.61$, $p < .03$, $\eta_p^2 = .10$ (main effect group, $F < 1$). Figure 2a reveals that this legibility effect occurred for both physics essays and education essays.[4]

Author abilities. The six items assessing author abilities were individually rescaled so that higher scores would indicate more positive evaluations and were combined to form one single index for each of the six essays (all Cronbach's $\alpha > .73$). The assumption of equal variances between groups was violated for none of the essays, all *F*s < 1.35, *p*s > .25. The six indices were subjected to the described 6 x 2 mixed-factorial ANOVA. A significant content quality main effect reflects that participants took content quality into account, $F(5, 125) = 44.29$, $p < .01$, $\eta_p^2 = .64$. More importantly, evaluations were more positive the more legible the handwriting was, as apparent in a significant interaction, $F(5, 125) = 3.83$, $p < .01$, $\eta_p^2 = .13$ (main effect group, $F < 1$). Figure 2b reflects that this legibility effect occurred for both physics and education essays.

Author gender. Participants' beliefs about author gender were subjected to six independent χ^2-tests (one per essay). Results show that for every essay, legible handwriting was more often ascribed to female authors, whereas less legible handwriting was more often ascribed to male authors, all χ^2s(1) > 13.71, *p*s < .01. Mediation analyses were performed following Baron and Kenny (1986). Separately for each essay, we analyzed whether the effect of the legibility manipulation (independent variable) on assigned grades (dependent variable) is mediated by ascribed author gender. Sobel tests (1982) were non-significant for all essays (essay 3: Sobel's Z = 1.68, p > .09; all other essays, Sobel's Zs < 0.80, ps > .45). Similarly, we analyzed separately for each essay whether the effect of the legibility manipulation (independent variable) on author abilities (dependent variable) was mediated by ascribed author gender. Again, Sobel tests were non-significant for all essays (Sobel's Zs < 1.30, ps > .20).[5] These results suggest that the legibility bias is not mediated by differences in ascribed gender.

Discussion

Experiment 2 was designed to experimentally explore whether gender-based inferences contribute to the legibility bias. Because gender-based inferences build on assumptions about who performs well and who performs poorly in a specific domain, using domains that are opposite in gender stereotypicality is a pivotal test. We selected the domain of physics, which is stereotypically male, and the domain of education, which is stereotypically female. Should the legibility bias be governed by domain specific gender-based inferences, legible essays should be evaluated more positively than less legible essays in only one of the two domains. In contrast to this gender-based prediction, the legibility bias was strong across domains.

Although this finding questions whether gender-based inferences are causing the legibility bias, one may argue that participants may draw different inferences depending on subject domain. For instance, one could assume that in the domain of physics, participants might draw inferences taking stereotype threat into account, whereas in the domain of education, participants are flexible enough to use a different inference rule. Should this be true, however, ascribed gender should mediate the effect of legibility on dependent variables for every essay. However, except for a tendency in Essay 3, this was not the case. As in Experiment 1, this failure to find evidence for mediation argues against gender-based inferences as a cause of the legibility bias.

General discussion

Evaluations of handwritten materials are subject to a series of biases, including that legible essays are systematically evaluated more positively than essays that are less legible (e.g., James, 1929). This legibility bias is problematic because performance assessments in many educational settings are not based only on computerized or multiple choice tests, but often include the evaluation of handwritten work samples. It is therefore important to further understand what causes the legibility bias. This research contributes to that goal by (a) replicating the standard finding in tightly controlled experiments and (b) testing potential underlying mechanisms.

Two experiments demonstrate that legible handwritten material may result in more positive evaluations than less legible material. This legibility bias occurred independent of performance level (good, medium, poor) and independent of subject domain (physics vs. education). These experiments tested further whether the legibility bias is due to gender-based inferences about who has legible handwriting and who performs well. However, both experimental and correlational evidence suggest that the legibility bias does not arise from simple or more complex gender-based inferences. The present experiments thus strongly extend earlier results which were open to gender-based alternative explanations (e.g., Greifeneder et al., 2010; James, 1929). Moreover, the present results extend our understanding of what causes the legibility bias. Although knowing what to focus attention on is important, we believe knowing what to neglect is also critical, given the limitations of attentional resources (Miller, 1956). These findings may therefore help to more successfully allocate cognitive resources when countering the legibility bias in performance assessment.

Going beyond the observed evidence, at least four aspects deserve short discussion. First, it should be noted that conclusions about the non-existence of effects are logically not valid in null-hypothesis-testing. The conclusion that gender-based inferences do not contribute to the legibility bias should therefore be treated with caution. Given, however, that this conclusion was reached based on multiple experiments, each using multiple essays, and with both correlational and experimental methodology, there seems to be a strong set of convergent findings.

Second, our samples largely consisted of female participants, reflecting characteristics of the student population from which participants were drawn. As a result of this, participant gender could not be included as an explanatory variable in statistical analyses. Future research may fruitfully aim for balanced samples to investigate, for instance, whether the legibility bias in general, or the use of gender-based inferences in particular, differ between female and male participants.

Third, we used only subjective assessments of legibility and did not rely on standardized tests or coded motor schemes (e.g., Peeples & Retzlaff, 1991). This choice appears sensible because what likely biases performance assessments of non-expert graphologists are subjective perceptions of legibility. Still, future research may fruitfully employ more objective measures of legibility to discern, for instance, which handwriting features (e.g., size and slant) are most strongly related to the legibility bias.

Finally, it should be noted that the observed legibility effect was always smaller than the effect of factual content. This is noteworthy because despite its strength, legibility did not overpower content. The bias exerted by legibility was, however, still alarmingly strong, suggesting that its potential role in unfair treatment should not be underestimated.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219-235. doi:10.1177/1088868309341564

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182. doi:10.1037/0022-3514.51.6.1173

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology, 60*, 485-499. doi:10.1037/0022-3514.60.4.485

Briggs, D. (1970). The influence of handwriting on assessment. *Educational Research, 13*, 50-55. doi:10.1080/0013188700130107

Burr, V. (2002). Judging gender from samples of adult handwriting: Accuracy and use of cues. *The Journal of Social Psychology, 142*, 691-700. doi:10.1080/00224540209603929

Dewar, G. (2010). *Stereotype threat: How your child's beliefs about people can hinder his performance in school ... and life*. Retrieved 25.11.2010, from http://www.parentingscience.com/stereotype-threat.html

Dwyer, C. A., & Johnson, L. M. (1997). Grades, accomplishments, and correlates. In W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 127-156). Mahwah, NJ: Erlbaum.

Eagly, A. H., Beall, A. E., & Sternberg, R. J. (2004). *The psychology of gender* (2nd ed.). New York, NY US: Guilford Press.

Eagly, A. H., & Mladinic, A. (1989). Gender stereotypes and attitudes toward women and men. *Personality and Social Psychology Bulletin, 15*, 543-558. doi:10.1177/0146167289154008

Ehindro, O. J. (1986). Correlates of physics achievement: The role of gender and non-induced student expectations. *Journal of Experimental Education, 54*, 189-192.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 74-97). Thousand Oaks, CA, US: Sage Publications.

Greifeneder, R., Alt, A., Bottenberg, K., Seele, T., Zelt, S., & Wagener, D. (2010). On writing legibly: Processing fluency systematically biases evaluations of handwritten material. *Social Psychological and Personality Science, 1*, 230-237. doi:10.1177/1948550610368434

Greifeneder, R., & Bless, H. (2007). Relying on accessible content versus accessibility experiences: The case of processing capacity. *Social Cognition, 25*, 853-881. doi:10.1521/soco.2007.25.6.853

Greifeneder, R., Bless, H., & Pham, M. T. (in press). When do people rely on affective and cognitive feelings in judgment? A review. *Personality and Social Psychology Review*. doi:10.1177/1088868310367640

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. *Journal of Educational Measurement, 20*, 65-70. doi:10.1111/j.1745-3984.1983.tb00190.x

James, H. W. (1929). The effect of handwriting upon grading. *The English Journal, 16*, 180-185.

Jussim, L., & Eccles, J. S. (1992). Teacher expectations: II. Construction and reflection of student achievement. *Journal of Personality and Social Psychology, 63*, 947-961. doi:10.1037/0022-3514.63.6.947

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. *Journal of Personality and Social Psychology, 29*, 299-304. doi:10.1037/h0036018

Markham, L. R. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal, 13*, 277-283. doi:10.2307/1162390

Marshall, J. C. (1967). Composition errors and essay examination grades re-examined. *American Educational Research Journal, 4*, 374-385.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81-97. doi:10.1037/h0043158

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology, 20*, 139-156. doi:10.1002/acp.1178

Osgood, C., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Peeples, E. E., & Retzlaff, P. D. (1991). A component analysis of handwriting. *Journal of General Psychology, 118*, 369-374.

Pomerantz, E. M., Altermatt, E. R., & Saxon, J. L. (2002). Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology, 94*, 396-404. doi:10.1037/0022-0663.94.2.396

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57*, 55-71. doi:10.1111/0022-4537.00201

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science, 9*, 45-48. doi:10.1111/1467-9280.00008

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108-131. doi:10.1207/S15327957PSPR0402_01

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equations models. In S. Leinhart (Ed.), *Sociological methodology* (pp. 290-312). San Fransisco: Jossey-Bass.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*, 613-629. doi:10.1037/0003-066X.52.6.613

Stewart, M. (1998). Gender issues in physics education. *Educational Research, 40*, 283-293.

Wilcox, R. (1993). Robustness in ANOVA. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 345-374). New York: Dekker.

Footnotes

[1] Due to space limitations, only *t*-values from planned contrast analyses are reported. Detailed descriptive statistics and full ANOVA results are available from the authors. The pre-test was computer-based, whereas Experiments 1 and 2 were assessed in paper-pencil format. Because individuals may behave differently when reading from computer screen compared to paper, *absolute* fluency levels may differ between computer-based versus paper-based testing. The observed *relative* differences in reading fluency, however, which is the critical factor in this research, should be similar across materials.

[2] χ^2-values are reported as summary statistics in Experiments 1 and 2. Detailed results are available from the authors.

[3] One precondition for ANOVA is equal variances between groups. This precondition was violated for the first, second, and fourth essay, *F*s > 4.85, *p*s < .04 (for the third, fifth, and sixth essay, *F*s < 1.00). We still elected to calculate ANOVA, because ANOVA has been shown to be very robust against violations of the equal variance assumption (e.g., Wilcox, 1993) and to allow for consistent presentation across dependent variables and experiments. Additional non-parametric tests yielded similar results.

[4] It is not logically possible to test whether the legibility bias is *equally* pronounced in the two subject domains. The conclusions reached are therefore only descriptive and should be viewed with caution. This caveat notwithstanding, inspection of Figure 2 reveals a clear pattern of results in support of a legibility bias for both subject domains.

[5] Sobel *Z* was computed based on regression coefficients and is reported as a summary statistic due to space limitations. Detailed Baron-Kenny calculations yielded similar results.

Acknowledgements

Authors 2 to 5 have contributed equally and are listed in reversed alphabetical order. This research was supported by a post-doctoral grant from the Deutsche Forschungsgemeinschaft.

Tables

Table 1

Design of Experiment 1

	Content quality		
	Essay 1: Good	Essay 2: Medium	Essay 3: Poor
Essay legibility in group 1	Low	High	Low
Essay legibility in group 2	High	Low	High

Figure Legends

Figure 1. Mean evaluations of author abilities with standard errors for every essay in Experiment 1. Higher ratings indicate more positive evaluation. Means are not arranged by experimental group, but re-arranged by handwriting legibility. Evaluations of legible essays are displayed as white bars, evaluations of less legible essays as black bars.

Figure 2. Mean assigned grades (2a) and mean evaluations of author abilities (2b) with standard errors for every essay in Experiment 2. Higher ratings indicate more positive evaluation. Means are not arranged by experimental group, but re-arranged by handwriting legibility. Evaluations of legible essays are displayed as white bars, evaluations of less legible essays as black bars.

Figure 1

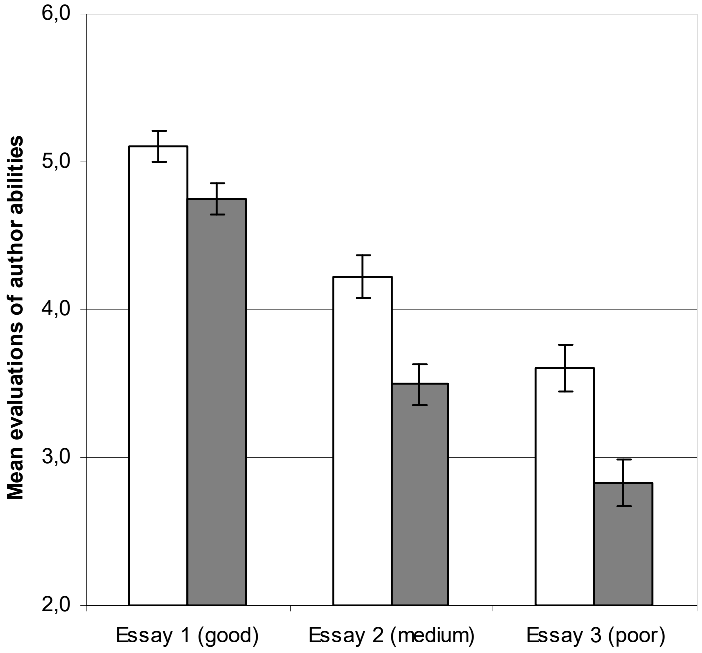

Mean evaluations of author abilities
6,0
5,0
4,0
3,0
2,0
Essay 1 (good)
Essay 2 (medium)
Essay 3 (poor)

Figure 2a

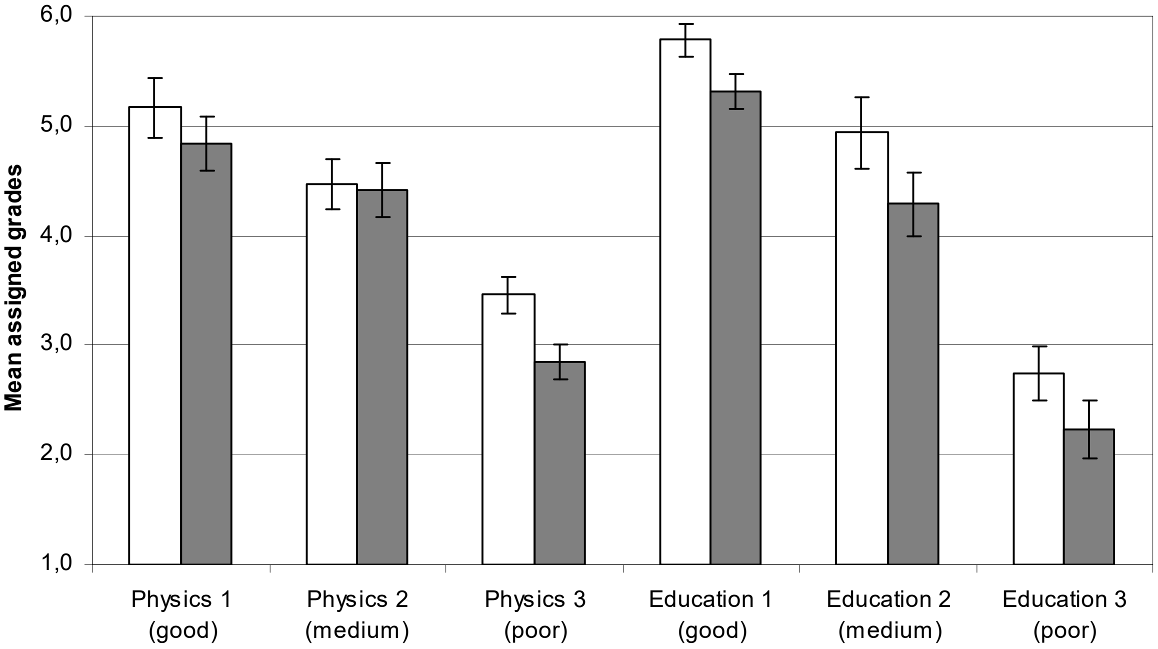
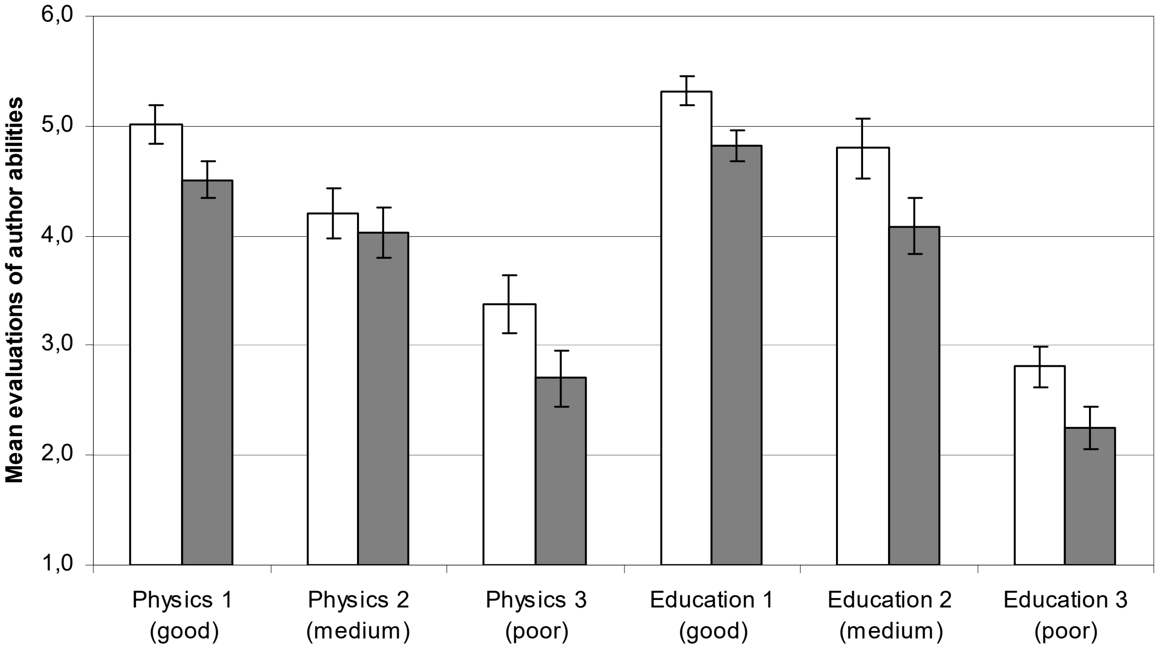

6,0
5,0
4,0
3,0
2,0
1,0
Mean assigned grades
Physics 1 (good)
Physics 2 (medium)
Physics 3 (poor)
Education 1 (good)
Education 2 (medium)
Education 3 (poor)

Figure 2b


6,0
5,0
4,0
3,0
2,0
1,0
Mean evaluations of author abilities
Physics 1 (good)
Physics 2 (medium)
Physics 3 (poor)
Education 1 (good)
Education 2 (medium)
Education 3 (poor)