



Negative findings of the handwriting legibility effect: the explanatory role of spontaneous task-specific debiasing

Timo Ruusuvirta¹  · Helena Sievänen¹ · Marjo Vehkamäki¹

Received: 12 September 2020 / Accepted: 8 June 2021 / Published online: 6 July 2021
© The Author(s) 2021

Abstract

Teachers' assessments of students' academic performance are susceptible to various cognitive biases. The evidence for higher grades given for legibly than illegibly handwritten texts—the handwriting legibility effect—is partially negative in the literature, however. Three explanations for the negative evidence could be offered. First, the variation in handwriting legibility could simply fail to implicitly affect the graders' behaviour. Secondly, the graders could intuitively associate handwriting illegibility with males and legibility with females. The presumed males' texts could thereby receive higher grades than females' texts, hiding the handwriting legibility effect proper. Third, the graders could spontaneously and selectively inhibit the handwriting legibility effect. In the present experiment, forty second-year teacher students graded fifth-grade students' handwritten test answers. The answers varied independently in content quality and handwriting legibility. Handwriting legibility did not measurably vary the grades. Nevertheless, handwriting legibility became gender-stereotyped and, at a trend level, varied grading confidence. However, the gender stereotypes were not measurably reflected in the grades. This set of findings is best explained by the participants' spontaneous and selective inhibition of the handwriting legibility effect. Spontaneous task-specific debiasing may, therefore, explain at least some of the previous negative findings of this effect in the literature.

Keywords Handwriting readability · Grading · Education · Gender stereotype · Cognitive bias

Teacher's assessments of academic works students produce by writing are susceptible to various cognitive biases. One major cognitive bias, a form of the halo effect (Cooper 1981; Thorndike 1920), is the handwriting legibility effect (for the

✉ Timo Ruusuvirta
timo.ruusuvirta@utu.fi

¹ Department of Teacher Education, University of Turku, 26100 Rauma, Finland

presentation effect, see, also Graham et al. 2011). This effect refers to higher grades given to legibly than illegibly handwritten academic works (for a meta-analysis, see, Graham et al. 2011). Legibility is defined by the ease of grasping the message the writing conveys, processing fluency (Szymczak 2016, for “clear display”, see, Kahneman 2011).

There are findings of the robust handwriting legibility effect with notable score average differences or even effect sizes (for reviews, see, Graham et al. 2011; Morris 2013; Meadows and Billington 2005). For example, James (1927) found in 43 high-school teachers that the handwriting quality of an essay biased their grading of the content of the essay by 8.7 points (the average of 59.8 for poor and 68.5 for good quality). Shepherd (1929) found that participant teachers gave illegibly handwritten texts lower scores than legible writings. The mean estimated effect size calculated by Graham et al. (2011) across the two experiments in Shepherd’s (1929) study was a notable 1.2. Chase’s (1968) and Soloff’s (1973) studies in teachers demonstrated a similar finding with notable effect sizes (0.7 and 0.94, respectively, as estimated by Graham et al. 2011). Markham’s (1976) or Briggs’ (1970) findings in teachers and teacher students statistically supported this evidence, although no standard deviations were available to allow the calculation of the effect sizes.

More recently, Klein and Taub (2005) found that handwriting legibility affected sixth-grade teachers’ judgments of the content of essays, the score differences corresponding to large effect sizes as indexed by Cohen’s *d*s of 1.412 for pen and 0.81 for pencil.¹ Similarly, Greifeneder et al. (2012) observed in university students higher grades for essays handwritten with high than low legibility with a high effect size ($\eta_p^2=0.70$). Greifeneder et al. (2010) observed similar findings in university students with from medium to large effect sizes (Cohen’s *d*’s ranging from 0.77 to 0.90 across different levels of content quality). However, they also observed that if the students were explicitly informed about the threat of the handwriting legibility effect to content quality assessments, this effect disappeared.

Some studies have failed to observe the handwriting legibility effect. For example, Chase (1979) found no such effect in graduate student scorers. Less legible handwriting was only found to make the scorers more susceptible to bias by the achievement expectations of the presumed writer, expectations that could be drawn from the information in a cover sheet about the writer’s previous academic achievements in other topics. Massey (1983) found that experienced examiners (the University of Oxford Delegacy of Local Examinations) were not observably biased by the quality of handwriting when assessing its content quality. Marshall (1972) also failed to find such an effect in their reasonably large sample of 480 classroom teachers. In an earlier study, Marshall and Powers (1969) had also observed quite small and confusing effects of handwriting quality in their sample of 420 prospective

¹ Cohen’s *d*’s were calculated in the present study based on the descriptive statistics provided by Klein and Taub (2005) and Marshall and Powers (1969). The calculations were based on pooled standard deviations. The design of these studies was a within-subject one. Yet, Cohen’s *d*’s could only be calculated for independent groups due to the unavailability of the correlations between the handwriting legibility levels (Morris & DeShon, 2008) in these studies. Therefore, the Cohen’s *d*’s are only approximations.

teachers. They reported that essay scores differed, confusingly, only between neatly written (5.66) and fairly neatly handwritten essays (5.02), but not between either of them and poorly written essays (5.25), the only difference corresponding in size to Cohen's d of 0.4¹ (Note also the estimated effect size of 0.38 approximated by Graham et al. 2011). Also, Eames and Loewenthal (1990) failed to observe the handwriting legibility effect in their reasonably small sample of 16 experienced psychology examiners.

The evidence of the handwriting legibility effect is thus contradictory. The positive evidence is difficult to be simply crossed off as a set of false alarms, and yet the amount of negative evidence is considerable. For the negative evidence, three possible explanations can be proposed. First, the variation in handwriting legibility is incapable of implicitly altering the participants' behaviour. Second, the handwriting legibility effect could be compensated away by another, directionally opposite gender-mediated effect of handwriting legibility. Namely, illegible handwriting could be attributed to male and legible handwriting to female writers (Burr 2002; Hartley 1991). Males could, in turn, be given higher grades than females (King 1998; Martin 1972; Spear 1984; for negative findings, however, see, Birch et al. 2016). Because this gender-mediated handwriting legibility effect and the handwriting legibility effect proper are directionally opposite, they could cancel each other out. Third, the participants could be aware of biases threatening grading and, consequently, spontaneously and selectively inhibit the behavioural manifestations of the handwriting legibility effect.

These explanations were addressed in the present experiment with the negative evidence of the handwriting legibility effect.

Materials and methods

Participants and design

Forty² first-year teacher students (30 female; age $M=22.5$ years, $SD=2.76$) of the University of Turku, Finland, formed a convenience sample in the experiment. The sample was a convenience sample. Informed consent was obtained from the participants after the nature of the experiment was explained to them. The participants were informed at the beginning of the experiment that a detailed description of the study was to be provided after the experiment (to avoid experimental bias due to knowledge of the independent variables). The participants were debriefed on the purpose of the research immediately following the

² The sample size was determined on the basis of previous studies on the handwriting legibility effect. Previous studies that have observed the handwriting legibility effect and reported the size of this effect (Greifeneder et al. 2010, 2012) suggest the expectation of relatively large (0.7–0.9) effect sizes (Cohen, 1992). A power analysis using the G*Power 3 statistical power analysis program (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that a very small sample ($N=9$) would have sufficed for detecting the effect size of 0.7 with 95% power using an F test between means with alpha at .05.

completion of their task. The research was undertaken at the University of Turku that does not require an internal research permit.

True experimental research was pursued. Each participant graded three test answers about the human circulatory system that, independently and pseudorandomly, differed in content quality (high, medium, low) and handwriting legibility (high, medium, low) across the participants.

Constructing test answers

A set of test answers was collected from test answers of a group of fifth-grade Finnish school pupils. Answers of different content qualities were selected from this set, typewritten and edited by the researchers for experimental control purposes. The answers were then copied in handwriting by a group of sixth-grade Finnish school pupils. The final set of 9 test answers were of three levels of content quality and handwriting legibility. The construction of this set is described in more detail in the following. The school principals and the teachers of the pupils had granted permission to use the pupils' test answers without personal data for the present research.

Twenty four fifth-grade Finnish students of a teacher training school wrote by hand answers to a test about human blood circulation as a part of their everyday activities in an environmental studies class. They were instructed to write everything they knew about human blood circulation. The topic had been addressed previously, so the students were familiar with the topic.

Then two of the authors (HS & MV) graded the answers in the range from 4 to 10 points. The assessment followed the criteria derived from the objectives for the teaching of the environmental studies in grades 3–6 as laid out in the Core Curriculum for Basic Education 2014 (Finnish National Board of Education 2016) and from the contents of the study book on biology and geography, Pisara (Cantell et al. 2016).

The answers were then divided into three categories by content quality, excellent, good and poor. The excellent answer was chosen from among answers with a score of 10. It was typed with a computer and further edited in length down to 126 words. The good test answer was similarly made up in 121 words by combining elements from several test answers of similar quality so that it achieved a score 7.5. The poor test answer was similarly made up in 121 words so that it achieved a score 5.

Then, sixth-grade school pupils ($N=16$) wrote the typed answers by hand, so that each student wrote all three answers. These handwritten answers were assessed for legibility by 9 adult females who were kindly willing to help with the assessment stage of material production. A score was given to describe how easy each answer was to read (range from “difficult to read” 1 via “mediocre to read” 3 to “easy to read” 5). Based on these scores, three handwritings, one high, one medium and one low, were selected for being used in conjunction with answers of excellent, good and poor content quality.

Materials

The materials for the participants included the general information of the experiment, detailed instructions, a summary of the human circulatory system and three test answers, each followed by an assessment form. The order of the three test answers in the set was pseudorandom across the participants. In the end, the materials included a questionnaire about the participant's personal attributes (age, gender) and about the gender of the presumed writer of each test answer (this was not asked about earlier to prevent handwriting legibility to engage the participants' attention during grading).

In the assessment forms, three elements of an answer were graded with a score that could range from "unacceptable" 0 to "excellent" 6 with steps of 1. These three elements were (a) the basic structure and functions of the human circulatory system, (b) the composition and functions of blood and (c) the role of physical exercise in health. Then, an overall score was the answer as a whole. This score could range from "unacceptable" 4 to "excellent" 10 with steps of 1. The confidence of grading also received a score that could range from "unconfident" 0 to "highly confident" 10 with steps of 1. Finally, it was to be assessed whether a female or male student had written the test answer. The writer's presumed grade in mathematics was asked last of but not analysed here.

Finally, the legibility of each answer was assessed with a score that could range from "very difficult to read" 1 to "very easy to read" 5 with steps of 1. Also, the participants were asked about their gender, age and study year. Finally, a sample of the participant's own writing was obtained. This sample was not analysed here.

Procedure

The participants were given 20 min to complete the task at their own pace. They were instructed to strictly follow the ordering of the sub-tasks in their individual sets of materials.

Analysis

The effects of content quality and handwriting legibility were tested as within-subject factors in repeated-measures analyses of variance (ANOVA) performed using version 24 of SPSS software (SPSS Inc., Chicago, IL, USA). Greenhouse–Geisser-adjusted degrees of freedom were used whenever the sphericity assumption (Mauchly's test) was violated. Only p -value was then reported as corrected. Subsequent pairwise comparisons were uncorrected. Partial eta square (η^2p) was used as a measure of effect size in ANOVAs and Pearson-correlation corrected Cohen's d (to allow its use in within-subject designs) in subsequent pairwise comparisons (Morris and DeShon 2002). All statistical tests were two-tailed with an alpha level of 0.05. There were two missing values in the dataset (one in the basic structure and functions of the human circulatory system with excellent quality and the other in the basic

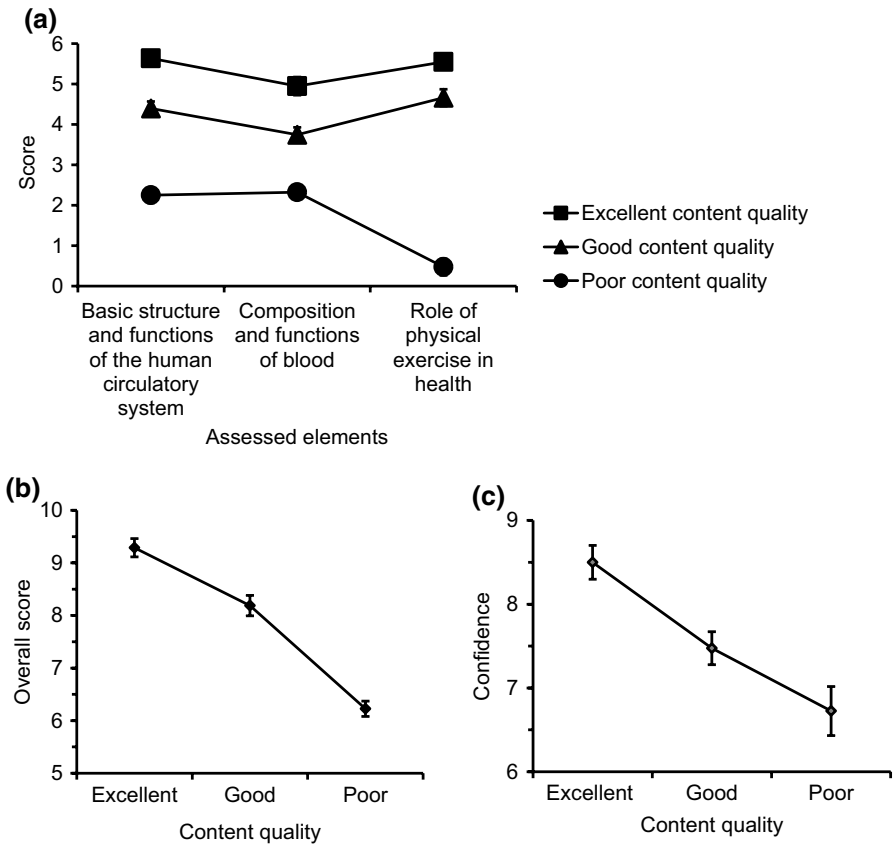


Fig. 1 Results for content quality. (a) Mean scores of content quality across content quality levels and test answer elements. (b) Mean scores of overall content quality across content quality levels. (c) Mean scores of grading confidence across content quality levels. Error bars (that in Panel (a) are overlapped by markers) refer to the standard error of the mean

structure and functions of the human circulatory system with poor content quality). These values were replaced with the average of the existing values of the variable.

Results

Content quality

Scores for elements

The participants graded each of the three elements of the test answers with a score. A repeated-measures ANOVA performed on this score with content quality (excellent, good, poor) and assessed elements (the basic structure and functions of the

human circulatory system, the composition and functions of blood, the role of physical exercise in health) as within-subject factors revealed a statistically significant main effect of content quality, $F(2, 78) = 158.4$, $p < 0.001$, $\eta_p^2 = 0.80$ (Fig. 1a). Subsequent comparisons indicated a significant difference between any pair of the three factor levels (Paired t -tests, $p < 0.001$). A statistically significant interaction between content quality and assessed elements, $F(4, 156) = 32.2$, $p < 0.001$, $\eta_p^2 = 0.45$, indicated a distinctly low score for “role of physical exercise in health” with poor content quality. Statistically significant (Paired t -tests, $p \leq 0.002$) differences were found between excellent and good and between good and poor content quality for each of the three assessed elements.

Overall scores

The participants also graded each test answer as a whole with an overall score. A repeated-measures ANOVA performed on the overall score revealed a statistically significant main effect of content quality (excellent, good, poor), $F(2, 78) = 89.8$, $p < 0.001$, $\eta_p^2 = 0.70$ (Fig. 1b). Subsequent pairwise comparisons indicated a statistically significant difference between any pair of content qualities (Paired t -tests, $p < 0.001$), Cohen’s d values being 0.82 (corrected with correlation of 0.33) between excellent and good and 1.37 (corrected with correlation of 0.11) between good and poor content qualities.

Grading confidence

A repeated-measures ANOVA performed on the participants’ grading confidence revealed a statistically significant main effect of content quality (excellent, good, poor), $F(2, 78) = 31.8$, $p < 0.001$, $\eta_p^2 = 0.30$ (Fig. 1c). Subsequent pairwise comparisons indicated statistically significant (Paired t -tests, $p \leq 0.032$) differences between excellent and good and between good and poor content quality, Cohen’s d values being 0.61 (corrected with correlation of 0.10) between excellent and good and 0.36 (corrected with correlation of 0.09) between good and poor content quality.

Presumed gender of the writer

The observed probabilities of the presumed genders of the writers did not differ from the expected 0.5 with excellent, good or poor content quality (Binomial exact test, $p \geq 0.268$).

To assess whether the presumed gender of the writer was reflected in grading, females and males were grouped (15 males and 25 females) from a subsample of cases with medium handwriting legibility (to keep handwriting legibility constant). A mixed-model ANOVA on scores for content quality with assessed elements (the basic structure and functions of the human circulatory system, the composition and functions of blood, the role of physical exercise in health) as a within-subject factor and gender (female, male) as a between-subject factor revealed no significant main effect of gender, $F(1, 38) = 0.10$, $p = 0.753$, or its interaction with assessed elements, $F(1, 76) = 0.83$, $p = 0.406$ (Fig. 2a).

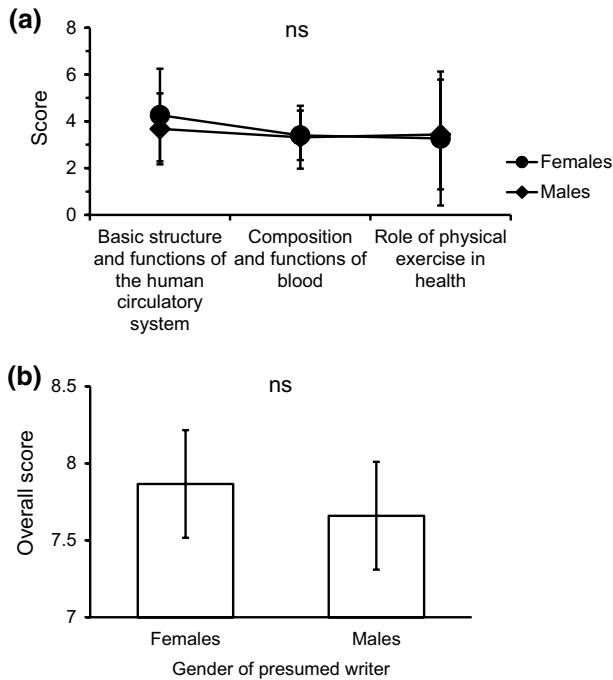


Fig. 2 Results for a writer's presumed gender. Cases with medium handwriting legibility are only included. (a) Mean scores of content quality across genders and test answer elements. *ns* refers to statistically non-significant ($p > .05$) main effect of gender or its interaction with a main effect of elements (b) Mean scores of overall content quality across genders. *ns* refers to statistically non-significant ($p > .05$) gender difference

Consistently, no significant gender differences in the overall score for content quality were found in the participants' grades (paired t -test, $t(38)=0.50$, $p=0.619$) (Fig. 2b).

Handwriting legibility

Scores for elements

The participants graded each of the three elements of the test answers with a score. A repeated-measures ANOVA on this score with handwriting legibility (high, medium, low) and assessed elements (the basic structure and functions of the human circulatory system, the composition and functions of blood, the role of physical exercise in health) as within-subject factors revealed neither main effect of handwriting legibility, $F(2, 78)=0.5$, $p=0.61$, nor an interaction between handwriting legibility and assessed elements, $F(4, 156)=0.6$, $p=0.61$ (Fig. 3a).

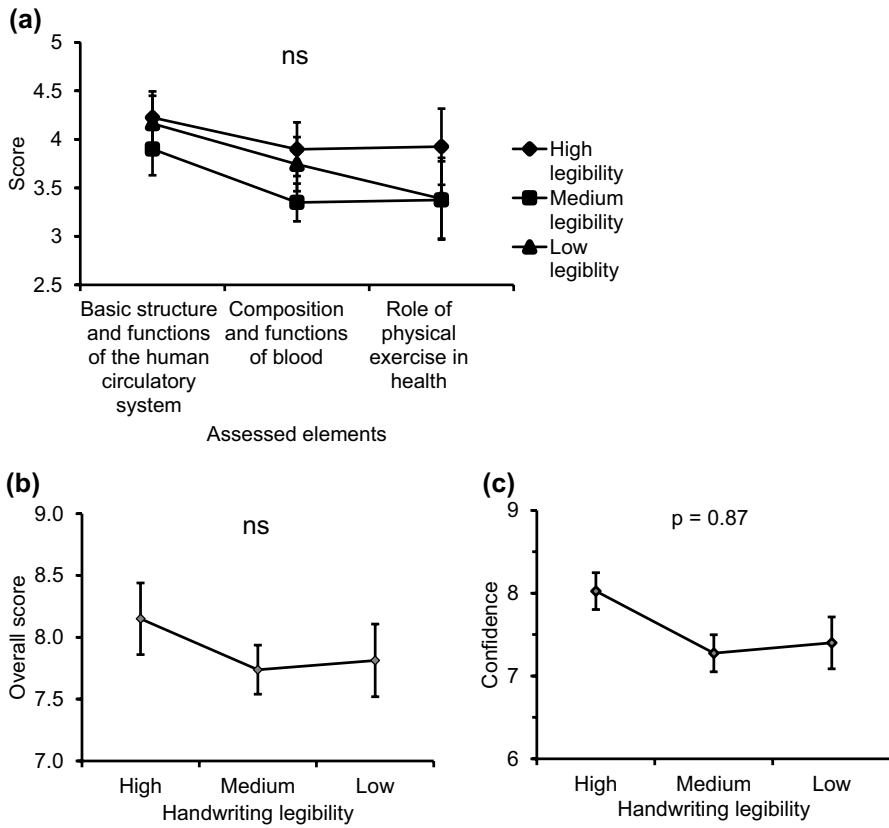
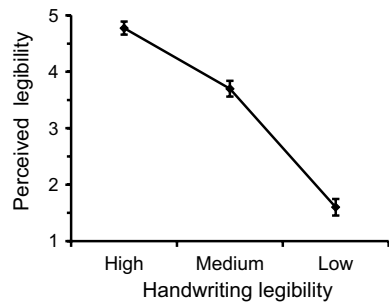


Fig. 3 Results for handwriting legibility. (a) Mean scores of content quality across handwriting legibility levels and test answer elements. (b) Mean scores of overall content quality across handwriting legibility levels. (c) Mean scores of grading confidence across handwriting legibility levels. Ns refers to statistically non-significant ($p > .05$) main effect of handwriting legibility or its interaction with content quality. $p = .087$ refers to the statistical significance of the main effect of handwriting legibility. Error bars refer to the standard error of the mean

Overall scores

The participants also graded each test answer as a whole with an overall score. A repeated-measures ANOVA on this overall score revealed no statistically significant main effect of handwriting legibility (high, medium, low), $F(2, 78) = 0.55$, $p = 0.577$ (Fig. 3b).

Fig. 4 Results for perceived handwriting legibility. Mean scores of perceived handwriting legibility across handwriting legibility levels. Error bars refer to the standard error of the mean



Grading confidence

A repeated-measures ANOVA performed on the participants' grading confidence revealed a marginally statistically significant main effect of handwriting legibility (high, medium, low), $F(2, 78) = 2.5$, $p = 0.087$, $\eta_p^2 = 0.06$ (Fig. 3c).

Presumed gender of the writer

The gender distribution differed significantly from expected 0.5 with low (34 males against 6 females) and with high (3 males vs. 37 females) handwriting legibility (Binomial exact test, both $p < 0.001$). This distribution (15 males vs. 25 females) did not differ significantly from the expected 0.5 (Binomial exact test, $p > 0.01$) with medium handwriting legibility.

Perceived handwriting legibility

A repeated-measures ANOVA performed on the participants' perceived level of handwriting legibility revealed a statistically significant main effect of handwriting legibility (high, medium, low), $F(2, 78) = 135.5$, $p < 0.001$, $\eta_p^2 = 0.78$, subsequent pairwise comparisons indicating statistically significant ($p < 0.001$) differences between consecutive levels of handwriting legibility, Cohen's d values being 0.99 (corrected with correlation of 0.09) between high and medium and 1.81 (corrected with correlation of 0.18) between medium and poor (Fig. 4).

Discussion

The participants, first-year teacher students, were instructed to grade fifth-year primary school student's handwritten test answers. The answers varied independently in content quality and handwriting legibility. The participants were found to succeed in grading (Fig. 1a and b), although with lower grading confidence towards lower content quality (Fig. 1c). The grades were found not to be systematically altered by handwriting legibility (Fig. 3a and b)—a negative finding

of the handwriting legibility effect. Nevertheless, the participants' grading confidence was found to decrease at a trend level towards lower handwriting legibility (Fig. 3c), and they were found to stereotype different levels of handwriting legibility by the presumed gender of the writer (low legibility associated with males and high legibility with females). However, these gender stereotypes could not be found to be observably reflected in the grades (Fig. 2a and b).

The negative evidence (Fig. 3a and b) of the handwriting legibility effect (Graham et al. 2011) called for further analysis of the data to address the plausibility of the three explanations for the negative evidence. In the following, these explanations are addressed.

The first explanation for the negative evidence, the inability of the variation in handwriting legibility to implicitly alter the participants' behaviour, had to be rejected. The participants were found to spontaneously stereotype texts written with different levels of handwriting legibility by gender. They attributed high-legibility writing to female and low-legibility writing to male writers (see also, Burr 2002; Hartley 1991), presumably reflecting the participants' previous experiences of higher-legibility handwritings produced by females than males (Weintraub et al. 2007; Graham et al. 1998). Furthermore, a statistical trend for the participants' lower grading confidence towards lower handwriting legibility (Fig. 3c) even tentatively suggests that handwriting legibility directly affected the process of grading albeit not the grades themselves. It is possible that the participants needed higher cognitive effort to grade low- than high-legibility test answers (Kahneman 2011). Their decreased grading confidence towards lower content quality (Fig. 1c) is more difficult to explain. It is possible that the test answers with lower content quality were cognitively more demanding to grade as they were poorer matches to the predefined elements of an ideal test answer specified in the instructions.

The second explanation for the unobservability of the handwriting legibility effect—higher grades given to (illegibly-writing) males than (legibly-writing) females (King 1998; Martin 1972; Spear 1984)—had to be rejected as well. The logic of this explanation as such was straightforward. The handwriting legibility effect involves lower grades for low-legibility than high-legibility handwritings (Graham et al. 2011). However, genders, if stereotyped from these legibility levels (Burr 2002; Hartley 1991), should lead to just the opposite. That is, low-legibility writers, just because they are males, should earn higher grades than high-legibility writers, just because they are females (King 1998; Martin 1972; Spear 1984), thereby compensating away the handwriting legibility effect proper (Graham et al. 2011). Despite the present data showed that participants stereotyped handwriting legibility by gender, the gender stereotypes were not observably reflected in the grades (Fig. 2a and b).

Out of three explanations, one remains. Most probably, the participants spontaneously and selectively inhibited the effects of handwriting legibility on grading in particular. The handwriting legibility effect (Graham et al. 2011) has been found to disappear if participants become explicitly aware of the threat of this bias (Greifeneder et al. 2010). In the present study, the participants were not informed about possible bias by handwriting legibility. However, the grading task was well-structured and

informed, which as such could effectively support the participants' objectivity in their assessments.

There are also some limitations of the present findings. First, the participants explicitly expressed their gender stereotypes for handwriting legibility not until the grades had been given. Therefore, it remains unclear whether or not the participants implicitly engaged in such stereotyping during grading. Note, however, that if they did not, the alternative hypothesis (of the handwriting-induced gender effect) for the unobservability of the handwriting legibility effect must have been rejected anyway. Secondly, the finding that gender stereotypes were not measurably reflected in grades was only from a part of the dataset, namely, from cases with mediocre handwriting legibility (to control for handwriting legibility variation). This obviously reduced statistical power, which emphasizes the inconclusiveness of this negative finding. Note also that the genders were likely to be less clearly inferable from medium- than high- or low-legibility handwritings and, thereby, possibly less effective in biasing grading. Thirdly, despite the test answers had some typical elements of informational content and handwriting legibility faced by class teachers in their professions, grading here took place as a part of an experiment in a non-natural environment. Therefore, the external validity of the present findings remains unclear. There is a need for future studies on whether, and if so to what extent, teachers would be similarly inclined to resist the handwriting legibility effect in their professional settings.

Conclusions

Teacher students participants were found not to measurably grade higher for test answers written with legible than illegible handwriting (the handwriting legibility effect). Nevertheless, they spontaneously stereotyped different levels of handwriting legibility by gender and, at a trend level, were less confident in grading the content of handwriting towards the lower legibility of this writing. Therefore, the handwriting legibility was unlikely absent simply due to the insufficient variation in handwriting legibility for implicitly altering the participants' behaviour. The finding that the gender stereotypes were not measurably reflected in grades, in turn, suggests that a directionally opposite handwriting legibility-induced gender effect on the grades was unlikely to compensate away the handwriting legibility effect proper. Thus, most probably, the participants voluntarily inhibited the handwriting legibility effect. Such voluntary debiasing could also explain at least some of the previous negative findings of this effect.

Author contributions All authors contributed to the study conception and design. Material preparation and data collection were performed by HS and MV. The analysis was performed by TR, HS and MV. The first draft of the manuscript was written by TR and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by University of Turku (UTU) including Turku University Central Hospital. No funding was received to assist with the preparation of this manuscript.

Availability of data and material (data transparency) The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors have no conflicts of interest or competing interests to disclose.

Ethics approval The University of Turku does not require research permits for projects conducted internally.

Consent to participate Written informed consent was obtained from the participants after the nature of the study was explained to them.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Birch P, Batten J, Batey J (2016) The influence of student gender on the assessment of undergraduate student work. *Assess Eval High Educ* 41:1065–1080
- Briggs D (1970) The influence of handwriting on assessment. *Educ Res* 13:50–55
- Burr V (2002) Judging gender from samples of adult handwriting: accuracy and use of cues. *J Soc Psychol* 142:691–700
- Cantell H, Jutila H, Laiho H, Lavonen J, Pekkala E, Saari H (2016) *Pisara 5, Biologia ja maantieto*. Sanoma Pro, Helsinki
- Chase CI (1979) The impact of achievement expectations and handwriting quality on scoring essay tests. *J Educ Meas* 16:39–42
- Chase CI (1968) The impact of some obvious variables on essay test scores. *J Educ Meas* 3:315–318
- Cohen J (1992) A power primer. *Psychol Bull* 112:155–159
- Cooper WH (1981) Ubiquitous halo. *Psychol Bull* 90:218–244
- Eames K, Loewenthal K (1990) Effects of handwriting and examiner's expertise on assessment of essays. *J Soc Psychol* 130:831–833
- Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* 39:175–191
- Finnish National Board of Education (2016) *National Core Curriculum for Basic Education 2014*. Next Print, Helsinki
- Graham S, Berninger V, Weintraub N, Schafer W (1998) Development of handwriting speed and legibility in grades 1–9. *J Educ Res* 92:42–52
- Graham S, Harris KR, Hebert M (2011) It is more than just the message: presentation effects in scoring writing. *Focus except Child* 44:1–12
- Greifeneder R, Alt A, Bottenberg K, Seele T, Zelt S, Wagener D (2010) On writing legibly: processing fluency systematically biases evaluations of handwritten material. *Soc Psychol Personal Sci* 1:230–237
- 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. *Br J Educ Psychol* 82:361–374

- Hartley J (1991) Sex differences in handwriting: a comment on spear. *Br Edu Res J* 17:141–145
- James A (1927) The effect of handwriting on grading. *Engl J* 16:180–205
- Kahneman D (2011) *Thinking, fast and slow*. Farrar, Straus and Giroux, New York
- King JL (1998) The effects of gender bias and errors in essay grading. *Educ Res Q* 22:13–25
- Klein J, Taub D (2005) The effect of variations in handwriting and print on evaluation of student essays. *Assess Writ* 10:134–148
- Markham LR (1976) Influences of handwriting quality on teacher evaluation of written work. *Am Educ Res J* 13:277–283
- Marshall JC (1972) Writing neatness, composition errors, and essay grades re-examined. *J Educ Res* 65:213–215
- Marshall JC, Powers JM (1969) Writing neatness, composition errors, and essay grades. *J Educ Meas* 6:97–101
- Martin WD (1972) The sex factor in grading composition. *Res Teach Eng* 6:36–47
- Massey A (1983) The effects of handwriting and other incidental variables on GCE “A” level marks in English Literature. *Educ Rev* 35:45–50
- Meadows M, Billington L (2005) A review of the literature on marking reliability. *Manch AQA*. Available: https://filestore.aqa.org.uk/content/research/CERP_RP_MM_01052005.pdf. Accessed 24 Jul 2021
- Morris KJ (2014) Does paper presentation affect grading: examining the possible educational repercussions of the quality of student penmanship. *Honors theses* 30. https://digitalcommons.salemstate.edu/honors_theses/30. Accessed 24 Jul 2021
- Morris SB, DeShon RP (2002) Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods* 7:105–125
- Shepherd EM (1929) The effect of the quality of penmanship on grades. *J Educ Res* 19:102–105
- Soloff S (1973) Effect of non-content factors on the grading of essays. *Grad Res Educ Relat Discip* 6:44–54
- Spear MG (1984) The biasing influence of pupil sex in a science marking exercise. *Res Sci Technol Educ* 2:55–60
- Szymczak P (2016) Grading for translation quality or legibility? A challenge to objective assessment of translation quality in handwritten samples. *J Transl Educ Transl Stud* 1:50–61
- Thorndike EL (1920) A constant error on psychological rating. *J Appl Psychol* 4:25–29
- Weintraub N, Drory-Asayag A, Dekel R, Jokobovits H, Parush S (2007) Developmental trends in handwriting performance among middle school children. *OTJR* 27:104–112