

Three Experimental Accounting Studies

D O C T O R A L T H E S I S

to acquire the academic degree of
doctor rerum politicarum
(Doctor of Economics and Management Science)

submitted to the

School of Business and Economics of
Humboldt-Universität zu Berlin

by

M.Sc. Rico Chaskel

President of Humboldt-Universität zu Berlin:

Prof. Dr. Peter Frensch (komm.)

Dean of the School of Business and Economics:

Prof. Dr. Daniel Klapper

Reviewers:

1. Prof. Dr. Joachim Gassen
2. Prof. Dr. Markus Arnold

Date of Colloquium: September 16, 2022

To the reader

Providing causal evidence in business and economics is challenging and the field of accounting clearly makes no exception. A large body of empirical accounting work uses observational field data, often collected from administrative sources, to provide associative or ‘quasi-experimental’ evidence. While certainly informative and descriptive, often the findings of these studies leave causal mechanisms unexplored and thus are open to alternative interpretation of their findings.

Rico Chaskel’s work takes a different route by using online and field experiments to study the judgment and decision making of individuals. Doing so, it combines economic and psychological theories and explores settings in accounting education and practice. I applaud the author for conducting carefully designed experiments that help us to understand better how overprecision affects range estimates, how translation affects the perceived textual properties of financial disclosures and whether formative online assessments affect the learning outcomes of management accounting students.

Besides their relative diverse topics, all studies can be classified as somewhat unorthodox experimental designs. This signals not only the impressive creativity of the author but also his willingness to take certain risks. For example, the second study tries to balance the advantage of experimental studies (control over the experimental setting) with the advantage of observational studies (real-life settings and decision problems) by using real-life financial disclosures of German public firms as experimental materials. While advocates of tightly controlled laboratory studies might have their quibbles with this design choice, I appreciate the idea to assess whether translations actually affect retail investors’ perceptions in the field. It is rare to see experimental studies cater to external validity in this way.

Besides answering their respective research questions, the findings of the three studies also open up for new questions as well: What determines the participant-inherent preference for certain precision levels that Rico observes? Why do people not use all available information to increase the accuracy of their estimates? Why do translations seem to affect the readability of financial disclosures but not the perceived desirability of investments? What should we conclude from the overall low correlations between perceived and assessed linguistic measures? Why do students that self-select into online assessments do not benefit from their continuous provision while students that did not self-select into treatment seem to do?

Inspiring all these questions to me is a fine outcome of carefully conducted academic work. I hope that the work of Rico Chaskel will be widely read and used.

Joachim Gassen
Berlin, December 2022

Acknowledgements

This dissertation is the work of several years as a doctoral student of accounting at Humboldt University of Berlin. Studying accounting means to think a lot about how to convey information – yet I struggle to find the words that express my gratitude towards all those who supported me throughout the years. In that sense, dear reader, keep in mind that I am thankful beyond the words I have written here and that I treasure the memories made in the past years.

Before reflecting upon my time at the Institute of Accounting and Auditing, I would like to thank the members of my dissertation committee. In particular, I am grateful to Markus Arnold for being the second examiner and for providing helpful feedback. Of course, I also want to thank Joachim Gassen, my first examiner and supervisor. Quite literally, my research would not have been possible without him. He provided me with guidance, knowledge and also financial support for my (sometimes unconventional) research projects. The trust he placed in me to follow my ideas as well as his inspiring and possibly unlimited enthusiasm for research and teaching truly made him a great mentor.

I also want to thank Ulf Brüggemann. In my view, Ulf is a perfect example of how brilliance shines even brighter in modesty. I would like to point out that without Ulf's forgiving nature, I may have never even started my PhD: He allowed me to hand-in the application for the Master's Thesis Seminar Accounting after the official deadline (I mixed up the dates). Well, this led to me becoming part of the seminar, ultimately writing my thesis with the institute, and then becoming a PhD student there. Long story short: Ulf, thank you for being a role model not just in research, but also in character.

Then, of course, I want to thank Andrea Häußler. Andrea motivated me to give my best, in any and every task. And she leads by example: her diligence and attention to detail are inspiring. Andrea was always supportive, always honest and taught me to speak my mind. I will miss all the big and small moments we had together.

Now, to my fellow PhD students (and postdocs). I have experienced quite a few “generations”: In 2016, when I started working at the institute, Maxi, Nader, and Sarah were ever welcoming. They did not just teach me about research and teaching, but about all the little intricacies one needs to know to thrive as a PhD student. Over time, Janja, Martin, Tobi, and Tom joined the team. Their diversity in terms of research, interests, and to a sense also personality, was fantastic and I learned a great deal from each of them (and I am still learning from Tobi!). Tom is also

the co-author of my second paper. I am happy and proud that it found its way into my dissertation. Tom, thank you for being part of my journey.

Then Jule, and a little later, Bianca joined the team. With Jule I shared an office and with Bianca I shared a “digital office”, as we spent much time together in Zoom sessions during Covid. I will miss (in an Jule’s case already do) the laughter, inspiration, and not least the mutual support in every way. The moments we had together is what I will probably think about in a few years when I talk about the “good old times”.

In the last couple of years, Fikir, Jonas, Sebastian, and Simone joined the team. It has been great working with you and I am sad that I won’t be around much longer. Over all the years, one thing was always true: our team is great and I will miss that. Obviously not just professors, postdocs and PhD students, but also student assistants. Now though, when you have seen so many people come and go, it probably means that the inevitable cannot be postponed forever and thus, it is time to say good-bye. Again, thank-you all – I hope to see you around some time.

Finally, I would like to express my gratitude towards friends and family. I am well aware that it was my decision to begin that PhD, yet you also had to live with the consequences. Thank you for always being there for me, even when I was stressed or had little time. Thank you for always being understanding, supportive, and for the love you gave me. Thank you to my parents, grandparents and sister. Thank you to Stuart: Seriously, what are the odds? A hundred to one? Sometimes I still pause in amazement. Thank you to Deanie, my role model in so many ways before I even realized I had role models. And the best friend anyone could wish for. Finally, thank you to old friends and new, for joining me on my way and always being so supportive.

Abstract

This dissertation comprises three papers covering different aspects of judgment and decision making in accounting research. All papers use experimental methods. The first study contributes to the literature on overprecision, or the phenomenon that people are commonly too sure about the accuracy of their beliefs. I examine how people provide range estimates, a challenging task that requires people to balance the width of the range (i.e., its precision) with the probability of the range covering the true value (i.e., accuracy). I find that people appear to have inherent individual preferences for a certain level of precision. At the same time, they appear to predictably incorporate additional information in order to increase accuracy by either sacrificing precision or shifting their ranges altogether. Still, they do not seem to maximise accuracy, but are willing to expend some of it to provide more precise estimates. Finally, I find that even mild changes to experimental instructions can lead participants to become more precise.

The second study examines how the translation of financial disclosures changes investors' perceptions of firms as an attractive investment. It examines three possible channels through which translation could influence investment attractiveness: perceived readability, perceived tone, and perceived precision of the underlying disclosure. In a survey experiment, retail investors read real forecast reports of German firms, provided in German and English. The findings indicate that both English and German native speakers find the German version easier to read. There are no measurable differences in perceived tone or precision. Contrary to prior literature, the easier readability does not translate into higher investment attractiveness. Solely tone appears to be correlated with investment attractiveness. The results indicate that this may be driven by the fact that participants find it comparably difficult to judge readability and precision in the company disclosures, while finding it relatively easy to judge tone. Finally, correlations between the participants' perceptions and commonly applied textual measures used in archival research are low to moderate at best.

The third study analyses how offering formative online assessments throughout the semester influences student performance in the final exam. It further examines whether students perform differently depending on whether they have time-restricted access to the assessments (two weeks; *continuous learning*), or whether they can access the assessments at any time (*flexible learning*). Students had the option to voluntarily enroll for the assessments at the beginning of the semester. Two thirds of the students who did not join voluntarily were administered into treatment anyway, enabling the identification of causal intention-to-treat effects of formative online assessments relative to a control group without access to the online tests. Results indicate that offering formative online assessments can enhance student performance, but only for students who do not opt for taking the test voluntarily and who are in a continuous learning environment. The naïve treatment effect of test-taking on student performance however is significantly positive, which highlights the concern of self-selection into treatment voiced in prior studies.

Zusammenfassung

Diese Dissertation umfasst drei Studien, welche unterschiedliche Aspekte im Rahmen von Urteilsbildung und Entscheidungsfindung in der Accounting-Forschung beleuchten. Alle drei Studien nutzen experimentelle Methoden. Die erste Studie trägt zur Forschung zu Overprecision („Überpräzision“) bei. Overprecision beschreibt das Phänomen, dass sich Menschen üblicherweise zu sicher hinsichtlich der Richtigkeit ihrer Überzeugungen sind. Ich untersuche, wie Menschen Schätzungen von Spannweiten vornehmen: eine herausfordernde Aufgabe, die es erfordert, eine Balance zwischen der Größe der Spannweite (Präzision) mit der Wahrscheinlichkeit, dass sie den wahren Wert einschließt (Richtigkeit), zu finden. Die Ergebnisse zeigen, dass Menschen inhärente individuelle Vorlieben für Präzision zu haben scheinen. Gleichzeitig nutzen sie vorhersagbar zusätzliche Informationen, um die Richtigkeit ihrer Schätzungen zu erhöhen. Dafür opfern sie entweder Präzision, oder verschieben ihre Spannweitenschätzungen insgesamt. Sie maximieren dabei die Richtigkeit ihrer Schätzungen jedoch nicht, sondern geben einen Teil dessen für höhere Präzision auf. Schlussendlich zeigen die Ergebnisse, dass sogar kleine Änderungen in den experimentellen Anweisungen die Teilnehmenden dazu verleiten können, präzisere Schätzungen abzugeben.

Die zweite Studie untersucht, wie sich die Übersetzung von Finanzberichterstattung auf die Wahrnehmung einer Firma als attraktives Investment auswirkt. Sie beleuchtet drei verschiedene Kanäle, durch welche die Übersetzung sich auf die Attraktivität auswirken kann: wahrgenommene Lesbarkeit, wahrgenommene Stimmung, und wahrgenommene Präzision der zugrundeliegenden Veröffentlichung. Die Studie ist als Umfrageexperiment aufgebaut, in welchem Kleinanleger echte Prognoseberichte deutscher Firmen lesen. Die Berichte werden auf Deutsch und Englisch zur Verfügung gestellt. Die Ergebnisse zeigen, dass die deutschen Berichte sowohl von Teilnehmenden mit englischer als auch von Teilnehmenden mit deutscher Muttersprache als besser lesbar wahrgenommen werden. Es können keine messbaren Unterschiede hinsichtlich Stimmung und Präzision festgestellt werden. Im Gegensatz zu vorheriger Literatur ist die Lesbarkeit nicht mit einer höheren Investmentattraktivität korreliert. Allein die Stimmung des Textes zeigt eine Korrelation mit höherer Attraktivität. Diese Ergebnisse können dadurch getrieben sein, dass die Teilnehmenden es als vergleichbar schwierig empfinden, Lesbarkeit und Präzision der Veröffentlichungen einzuschätzen. Letztlich zeigen die Ergebnisse, dass die Korrelation zwischen der Wahrnehmung der Teilnehmenden und Textmerkmalen, die typischerweise in Archivstudien erhoben werden, nur gering bis moderat ausgeprägt ist.

Die dritte Studie untersucht, wie das Angebot von formativen Onlinetests im Laufe des Semesters die Leistungen von Studierenden in der Abschlussprüfung beeinflusst. Sie untersucht außerdem, ob die Leistung sich unterscheidet, je nachdem ob die Studierenden zeitlich begrenzten Zugang zu den Tests haben (zwei Wochen; *kontinuierliches Lernen*) oder ob sie jederzeit auf die Tests zugreifen können (*flexibles Lernen*). Die Studierenden konnten sich zu Beginn des Semesters freiwillig für die Onlinetests eintragen. Zwei Drittel der Studierenden, welche dieses Angebot nicht wahrgenommen hatten, wurde trotzdem Zugang zu den Onlinetests gegeben. Dies ermöglicht die Bestimmung des kausalen Intention-to-Treat-Effekts der formativen Onlinetests, relativ zu einer Kontrollgruppe ohne Testzugang. Die Ergebnisse zeigen, dass die formativen Onlinetests die Studienleistung erhöhen können, allerdings nur für Studierende, welche sich nicht freiwillig für die Tests gemeldet hatten und in der kontinuierlichen Lerngruppe waren. Der naive Effekt der Testteilnahme auf die Studienleistung ist hingegen signifikant positiv, was die Bedenken hinsichtlich Selbstselektion aus früheren Studien unterstreicht.

Table of contents

Introductory Summary	1
References	7
 Precision and Accuracy	 11
1 Introduction	12
2 Literature and Hypotheses Development	15
2.1 Range Estimates: a Trade-off between Accuracy and Precision	17
2.2 Hypotheses	19
3 Research Design	22
3.1 Is it Overprecision?	23
3.2 Hypotheses Tests	25
3.3 Study Conditions and Participant Selection	29
4 Results	31
4.1 Descriptive Statistics	31
4.2 Hypotheses Tests	32
4.3 Additional Analyses	39
5 Conclusion	42
References	44
Appendix	50
A1 Treatment Conditions and Variable Definitions	50
A2 Range Locations	52
A3 Recalibration Decisions	53
A4 Accuracy by Round	54
A5 Last Digits	55
A6 Coarsened Exact Matching	56
A7 Experimental Material	58
Tables	61
Figures	71

Translation and Retail Investor Perception.....	74
1 Introduction	75
2 Research Questions and Hypotheses	78
2.1 Translation and Perception of Textual Characteristics	78
2.2 Perception of Textual Characteristics and Investment Decisions.....	81
2.3 Textual Characteristics: Perception and Standard Measures	86
3 Research Approach.....	87
3.1 Study Design.....	87
3.2 Hypotheses Testing.....	91
3.2.1 Translation.....	91
3.2.2 Perception of Textual Characteristics and Investment Decisions	93
3.2.3 Textual Measures.....	97
4 Results	97
4.1 Descriptive Statistics.....	97
4.2 The Effect of Translation on the Perception of Textual Characteristics....	100
4.3 Perception of Textual Characteristics and Investment Decisions.....	100
4.4 Textual Characteristics: Perception and Standard Measures	103
5 Conclusion.....	105
References.....	107
Appendix.....	115
A1 Forecast Report - Deutsche Konsum Reit-AG (ISIN: DE000A14KRD3)	115
A2 Forecast Report - Delivery Hero (ISIN: DE000A2E4K43).....	117
A3 Survey Instructions	119
A4 Variable Definitions for Survey Questionnaire	120
A5 Variable Definitions for Demographic Questions	121
A6 Correlation Matrix	122
Tables.....	124
Figures	134

Formative Online Assessments and Student Performance.....	136
1 Introduction	137
2 Literature and Hypotheses Development	138
2.1 Participation in Formative Online Assessments and Student Performance.....	138
2.2 Continuous Learning and Student Performance	141
3 Research Design.....	143
3.1 Experimental Design.....	143
3.2 Empirical Strategy	145
4 Results	148
4.1 Descriptive Statistics.....	148
4.2 Main Tests.....	149
4.3 Discussion	151
5 Student Evaluation.....	152
6 Conclusion.....	152
References.....	154
Appendix.....	161
A1 Variable Definitions.....	161
A2 Sample Question	162
A3 E-mail Invitation	163
A4 Student Evaluation	164
A5 ANOVA (Intention-to-treat Effect)	165
Tables.....	166
Figures	171

Introductory Summary

This dissertation contributes to the larger stream of research on judgment and decision making in accounting research (Bonner [1999]) by analyzing three settings in which individuals form opinions and estimates, and even alter their course of action. Study 1 helps to understand overprecision – the phenomenon that people are too sure about the accuracy of their estimates (Moore and Healy [2008]), a bias in judgment associated with significant economic costs (Barrero [2022]). Study 2 examines how the translation of financial disclosures alone (i.e., independent of their content) contributes to a different judgment of the disclosures’ textual characteristics and ultimately the attractiveness of the issuing firm as an investment opportunity. Finally, study 3 presents an intervention that may help accounting students to improve their performance by altering their learning behavior.

Each study in this dissertation aims to make a methodological contribution to the respective question at hand. Their common denominator in this regard is that they rely on experimental manipulation. Generally, experiments are one of the methodological cornerstones in both managerial and financial accounting research (Herschung et al. [2018]; Lachmann et al. [2017]; Libby et al. [2002]). Their particular appeal lies in allowing the researcher to identify cause-and-effect relationships through the manipulation of the main independent variables (Salterio and Gondowijoyo [2017]). As such, they enable the researcher to disentangle the numerous cause-and-effect relationships we suspect and observe in practice and are thus well-suited to examine where, when, and why individuals do not behave according to standard economic theory (Libby et al. [2002]). If done well, they do not just inform through high levels of internal validity, but also allow the reader to generalize the findings to settings beyond the lab (Lachmann et al. [2017]; Libby et al. [2002]; Luft and Shields [2014]).

My first study on overprecision builds upon the phenomenon that people appear to be too certain about the accuracy of their beliefs: range estimates are typically too narrow to cover the

true value with a pre-specified probability (Moore and Healy [2008]). The consequences for business are costly and extend to various aspects of the economy: people for example appear to underestimate the volatility of demand (Ren and Croson [2013]), balance sheet and income statement items (Bar-Yosef and Venezia [2014]) and stock market returns (Ben-David et al. [2013]), to name but a few.

Yaniv and Foster [1995] argue that peoples' range estimates may be too narrow because social norms in communication lead them to balance the width of the range (*precision*) with the probability that the range covers the true value (*accuracy*). Building upon their theory, I examine the following three questions:

1. Do people have an inherent preference for a certain level of precision?
2. How do people incorporate additional information to balance accuracy and precision?
3. Is it possible to induce people to become more precise?

Answering these questions is important because it helps to learn more about whether and how it is possible to counter overprecision in judgment and decision making. The contribution through content is supported by a new experimental design that overcomes the often voiced criticism that overprecision may be an artifact of study design, rather than a behavioral bias (e.g., Gigerenzer et al. [1991]): My experiment also allows me to draw conclusions relative to an objective benchmark of how precise participants ought to be (Moore Carter et al. [2015]). In my setting, participants are supposed to estimate the number of dots on ten different pictures that they see for ten seconds each. They provide me with a lower and an upper end of a range estimate that is supposed to cover the correct number of dots for at least 9 out of the 10 pictures. The participants' compensation is not linked to their success rate in order to avoid participants gaming the task (Moore, Tenney et al. [2015]). Later, participants judge 25 range estimates to the same task with 80 % of those being computer-generated and the remaining 20 % being ranges that are calibrated as the participants' own range estimates from the first half of the

experiment. The design allows me to match participants' judgments to ranges that are calibrated as their own with judgments of other participants to *the same*, yet computer-generated ranges. I find evidence that participants judge their own range estimates significantly higher than others' estimates. This suggests that participants provide range estimates in line with their own preferred level of precision.

In order to examine whether and how people incorporate additional information to balance accuracy and precision, I provide two thirds of the participants with a "little helper": a tool that provides them with range estimates to each of their pictures that are correct in 90 % of the cases. This design allows me to know how precise participants ought to be: if they provide range estimates smaller than what the "little helper" suggests and they are correct in less than 90 % of the cases, I can infer overprecision. First, I find that participants are overprecise even with the "little helper". However, accuracy improves from near 20 % without, to about 60 % with the "little helper", indicating that participants use the additional information. Interestingly, they only use the information to widen their range estimates for relatively low numbers of dots on the picture. For larger true values, they appear to use the additional information for a decrease in range width. They still achieve significantly better accuracy rates than the participants without a helper, because the "little helper" enables them to move their estimate closer to the true value. This suggests that overprecision may not just be a problem of range estimates that are too narrow, but that ranges are often severely wrongly positioned (Moore, Carter et al. [2015]). Overall, my results indicate that participants use additional information to improve the balance between accuracy and precision.

Finally, I create another treatment arm orthogonal to the aforementioned that suggests to half the participants that they are free to provide smaller ranges (e.g. in comparison to what the "little helper" suggests) if they deem them fitting. This is only a mild intervention, since it merely suggests that higher precision is acceptable. Still, I find that it leads to significantly smaller range sizes for low true values.

Taken together, my results suggest that while people provide range estimates in line with their own preferred level of precision, they use additional information to actively re-balance accuracy and precision, as well as re-think the position of their range estimate. They consciously *do not* use additional information to *entirely* mitigate overprecision. Finally, my results suggest that any communication with respect to range estimates should be carefully worded as even mild interventions may lead to significant changes in estimates, potentially contributing to overprecision.

My second study is co-authored with Tom Fischer. We examine how the translation of financial disclosures changes how retail investors assess the attractiveness of the issuing firm as a (share) investment. Especially larger firms in non-English-speaking countries translate their disclosures to attract foreign investment and lower information asymmetry for foreign investors (Jeanjean et al. [2010], [2015]). Even if the translation is the same in content as the original, the perception of the translated document is not necessarily equal (Doupnik and Richter [2003]; Pan et al. [2015]).

We analyze three channels through which translation may influence a reader's perception: (1) readability, (2) tone, and (3) precision of the underlying document. We use these since they are amongst the most commonly examined textual characteristics in the accounting literature that have been shown to be correlated with investment behavior (e.g., readability: Lawrence [2013]; Miller [2010]; Rennekamp [2012]; tone: Davis et al. [2012]; Huang et al. [2014]; Levin et al. [1998]; precision: Elliott et al. [2015]; Pan et al. [2018]).

We develop a survey experiment that allows us to use real firm disclosures while still being able to randomly allocate the independent variable. The firm disclosures comprise real forecast reports from listed German Prime Standard firms. The firms issue their reports in two languages: German and English. We manually compare the reports to ensure they have the same informational content. Participants in our study have investment experience and are fluent in both German and English, with one of those being their native language. Each participant reads

six different, randomly allocated forecast reports and assesses their readability, tone, precision, and investment attractiveness through a set of survey questions. Having two reports per firm and multiple reports per participant allows us to use a strict fixed-effects structure to ensure that our results are not driven by confounders, such as firm profitability or firm size. The design allows us to causally identify the effect of translation on the textual characteristics and investment attractiveness, while ensuring comparably high external validity through the use of real company reports.

We find that translation in our setting matters for readability: German reports appear to be easier to read than their English counterparts. This finding holds for German and English native speakers alike. We do not find significant effects for tone and precision. Interestingly, and unlike prior literature, the significant effect on readability does not translate into an overall effect on investment attractiveness. Our results suggest that this may be driven by participants experiencing it comparably difficult to assess readability for the real company disclosures.

Finally, we observe that our participants' perceptions are only weakly correlated with standard measures for textual characteristics in the literature (we use the Flesch Reading Ease index for readability (Flesch [1948]; Kincaid et al. [1975]) and the Loughran and McDonald [2011] word lists to capture tone and precision). We thus contribute to the discussion on the use of these standard measures and the necessity for alternatives (e.g., Siano and Wysocki [2021]).

The third study in this dissertation is co-authored with Joachim Gassen. We examine how formative online assessments offered throughout the semester influence student performance in an entry-level cost accounting class. We furthermore analyze whether continuously offered assessments (i.e., students solve one assessment every two weeks) have a different impact than flexibly offered assessments (i.e., students are free to solve the assessments whenever they wish).

Even though prior studies have generally identified a positive relationship between formative online assessments and student performance (Sotola and Crede [2021]), the true

causal effects are unclear since ethical and practical considerations have deterred researchers from running randomized trials (Einig [2013]; Marriott and Lau [2008]). At the same time, the setting appears to be prone to unobserved confounders, many of which are difficult to control for (e.g., student motivation; Chak and Fung [2015]). We contribute to the literature by designing a randomized experiment to identify the intention-to-treat effect of offering (continuous and flexible) formative online assessments on student performance, measured as the performance in the final summative exam. We offer students to join the formative online assessments at the beginning of the semester. As expected, not all students join. We allocate two-thirds of these students to a treatment condition anyway (note that participation is still optional). The remaining third does not get access to the online tests and acts as control. Students in the treatment condition are split into a continuous learning environment and a flexible learning environment (defined as above).

Similar to prior literature, the naïve treatment effect of participation in the online assessments (i.e. actual participation) on student performance is significantly positive. The causal intention-to-treat effect (i.e. being randomly assigned to the online tests), however, indicates that solely continuously offered online assessments have a significantly positive effect on student performance, and only for students who did *not* enroll voluntarily. For this group, offering the continuous online assessments on average increases exam performance by 3.4 points (out of 60) relative to the control group.

References

- BARRERO, J. M. "The Micro and Macro of Managerial Beliefs." *Journal of Financial Economics* 143.2 (2022): 640–667.
- BAR-YOSEF, S., AND I. VENEZIA "An Experimental Study of Overconfidence in Accounting Numbers Predictions." *International Journal of Economics Sciences* 3.1 (2014): 78–89.
- BEN-DAVID, I., J. R. GRAHAM, AND C. R. HARVEY "Managerial Miscalibration." *The Quarterly Journal of Economics* 128 (2013): 1547–1584.
- BONNER, S. E. "Judgment and Decision-Making Research in Accounting." *Accounting Horizons* 13 (1999): 385–398.
- CHAK, S. C., AND H. FUNG "Exploring the Effectiveness of Blended Learning in Cost and Management Accounting: An Empirical Study." *New Media, Knowledge Practices and Multiliteracies* Springer, Singapore, 2015. 189–203.
- DAVIS, A. K., J. M. PIGER, AND L. M. SEDOR "Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language." *Contemporary Accounting Research* 29 (2012): 845–868.
- DOUPNIK, T. S., AND M. RICHTER "Interpretation of Uncertainty Expressions: A Cross-National Study." *Accounting, Organizations and Society* 28 (2003): 15–35.
- EINIG, S. "Supporting Students' Learning: The Use of Formative Online Assessments." *Accounting Education* 22 (2013): 425–444.
- ELLIOTT, W. B., K. M. RENNEKAMP, AND B. J. WHITE "Does Concrete Language in Disclosures Increase Willingness to Invest?" *Review of Accounting Studies* 20 (2015): 839–865.
- FLESCH, R. "A New Readability Yardstick." *Journal of Applied Psychology* 32 (1948): 221.

- GIGERENZER, G., U. HOFFRAGE, AND H. KLEINBILTING "Probabilistic Mental Models: A Brunswikian Theory of Confidence." In *Psychological Review* 98.4 (1991): 506–528.
- HERSCHUNG, F., M. D. MAHLENDORF, AND J. WEBER "Mapping Quantitative Management Accounting Research 2002–2012." *Journal of Management Accounting Research* 30 (2018): 73–141.
- HUANG, X., S. H. TEOH, AND Y. ZHANG "Tone Management." *The Accounting Review* 89 (2014): 1083–1113.
- JEANJEAN, T., C. LESAGE, AND H. STOLOWY "Why Do You Speak English (in Your Annual Report)?" *The International Journal of Accounting* 45 (2010): 200–223.
- JEANJEAN, T., H. STOLOWY, M. ERKENS, AND T. L. YOHAN "International Evidence on the Impact of Adopting English as an External Reporting Language." *Journal of International Business Studies* 46 (2015): 180–205.
- KINCAID, J. P., R. P. JR. FISHBURNE, R. L. ROGERS, AND B. S. CHISSOM "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. " Naval Technical Training Command Millington TN Research Branch (1975).
- LACHMANN, M., I. TRAPP, AND R. TRAPP "Diversity and Validity in Positivist Management Accounting Research—A Longitudinal Perspective over Four Decades." *Management Accounting Research* 34 (2017): 42–58.
- LAWRENCE, A. "Individual Investors and Financial Disclosure." *Journal of Accounting and Economics* 56 (2013): 130–147.
- LEVIN, I. P., S. L. SCHNEIDER, AND G. J. GAETH "All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects." *Organizational Behavior and Human Decision Processes* 76 (1998): 149–188.

- LIBBY, R., R. BLOOMFIELD, AND M. W. NELSON "Experimental Research in Financial Accounting." *Accounting, Organizations and Society* 27 (2002): 775–810.
- LOUGHRAN, T., AND B. McDONALD "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *The Journal of Finance* 66 (2011): 35–65.
- LUFT, J., AND M. D. SHIELDS "Subjectivity in Developing and Validating Causal Explanations in Positivist Accounting Research." *Accounting, Organizations and Society* 39 (2014): 550–558.
- MARRIOTT, P., AND A. LAU "The Use of On-Line Summative Assessment in an Undergraduate Financial Accounting Course." *Journal of Accounting Education* 26 (2008): 73–90.
- MILLER, B. P. "The Effects of Reporting Complexity on Small and Large Investor Trading." *The Accounting Review* 85 (2010): 2107–2143.
- MOORE, D. A., A. B. CARTER, AND H. H. J. YANG "Wide of the Mark: Evidence on the Underlying Causes of Overprecision in Judgment." *Organizational Behavior and Human Decision Processes* 131 (2015): 110–120.
- MOORE, D. A., AND P. J. HEALY "The Trouble With Overconfidence." *Psychological Review* 115 (2008): 502–517.
- MOORE, D. A., E. R. TENNEY, AND U. HARAN "Overprecision in Judgment." In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (1st ed.) John Wiley & Sons, 2015.
- PAN, L., G. MCNAMARA, J. J. LEE, J. (JOHN) HALEBLIAN, AND C. E. DEVERS "Give It to Us Straight (Most of the Time): Top Managers' Use of Concrete Language and Its Effect on Investor Reactions." *Strategic Management Journal* 39 (2018): 2204–2225.

- PAN, P., C. PATEL, AND R. MALA "Questioning the Uncritical Application of Translation and Back-Translation Methodology in Accounting: Evidence from China." *Corporate Ownership and Control* 12 (2015): 479–491.
- REN, Y., AND R. CROSON "Overconfidence in Newsvendor Orders: An Experimental Study." *Management Science* 59 (2013): 2502–2517.
- RENNEKAMP, K. "Processing Fluency and Investors' Reactions to Disclosure Readability." *Journal of Accounting Research* 50 (2012): 1319–1354.
- SALTERIO, S. E., AND P. M. GONDOWIJOYO "Moving beyond the Lab: Building on Experimental Accounting Researchers' Core Competencies to Expand Methodological Diversity in Accounting Research." *The Routledge Companion to Behavioural Accounting Research* Routledge (2017): 149–174.
- SIANO, F., AND P. WYSOCKI "Transfer Learning and Textual Analysis of Accounting Disclosures: Applying Big Data Methods to Small(Er) Datasets." *Accounting Horizons* 35 (2021): 217–244.
- SOTOLA, L. K., AND M. CREDE "Regarding Class Quizzes: A Meta-Analytic Synthesis of Studies on the Relationship Between Frequent Low-Stakes Testing and Class Performance." *Educational Psychology Review* 33 (2021): 407–426.
- YANIV, I., AND D. P. FOSTER "Graininess of Judgment under Uncertainty: An Accuracy-Informativeness Trade-Off." *Journal of Experimental Psychology: General* 124 (1995): 424–432.

Precision and Accuracy

Rico Chaskel

Humboldt University of Berlin

Abstract

Overprecision is a persistent bias in judgment and decision making: peoples' range estimates are typically too narrow and thus do not contain the true value. Providing range estimates requires people to balance the width of the range (*precision*) with the probability of the range covering the true value (*accuracy*). Building upon Yaniv and Foster [1995], I examine whether and how people trade-off precision and accuracy. In an experimental setting, I find that people seem to have an inherent preferred level of precision. They use additional information in order to adjust precision and achieve higher accuracy. However, they do not use the additional information to maximise accuracy, but sacrifice some of it for an increase in precision. Furthermore, I find evidence that additional information can help people to re-think the position of their range estimates and thereby become more accurate without providing wider ranges. Finally, I find that even mild changes in the wording of instructions can induce people to become more precise, highlighting the importance of careful word choices in estimation tasks.

1 Introduction

“[M]anagers are overprecise” (Barrero [2022, p. 648]) – for instance through underestimating the volatility of the firm’s return on investment (Ben-David et al. [2013]), underestimating the volatility of sales growth (Barrero [2022]), and underestimating the volatility of demand (Ren and Croson [2013]). Barrero [2022] estimates that overprecision in managerial decision making leads to losses in firm value between 0.9 and 1.4 %. Investors and analysts also appear to be overprecise, for instance when forecasting net income and earnings per share (Bar-Yosef and Venezia [2014]) and stock market movements (Deaves et al. [2010]).

Overprecision – defined as the “excessive certainty regarding the accuracy of one’s beliefs” – is one of the most persistent biases in human judgment and decision making (Moore and Healy [2008, p. 502]). It manifests in surprisingly low success rates for range estimates: in experimental settings, participants are often asked to provide estimates that cover the true value in 90 % of the cases, however, participants typically only reach about 50 % (McKenzie et al. [2008]). I contribute to the literature on overprecision by analysing how people balance the trade-off that makes the estimation difficult: the trade-off between *precision* (i.e., the width of the estimate), and *accuracy* (i.e. whether the estimate covers the true value). The trade-off between accuracy and precision (Yaniv and Foster [1997]) is difficult because it is unclear ex-ante whether a decrease in precision (e.g., choosing a range that is twice as large) leads to an increase in accuracy (i.e., a correct guess).

Yaniv and Foster [1995] argue that social norms employed in conversations lead people to sacrifice accuracy for an increase in precision. They suggest that rather than providing well-calibrated range estimates, people may offer estimates they “feel comfortable with” (Yaniv and Foster [1997, p. 25]).¹ These estimates would then typically be too precise to reach the pre-

¹ In experimental overprecision research, participants are often asked to be 90% sure (or confident) about an estimate, or to provide a 90% confidence interval around their best estimate. Note that the term “confidence interval” may be misleading since a fixed (yet unknown) value does not have a confidence interval in a statistical

defined level of accuracy. My research builds on Yaniv and Foster's [1995] theory to examine three main questions:

1. *Do people have an inherent preference for a certain level of precision?*

This question directly relates to Yaniv and Foster's [1997, p. 25] to the best of my knowledge so-far untested presumption that people provide estimates they "feel comfortable with". It helps to understand why it may be so difficult to get people to become less (over)precise (Moore, Tenney et al. [2015]).

2. *How do people incorporate additional information to balance accuracy and precision?*

Remus et al. [1995] show that people use additional information to improve their estimates, but apparently not all informational advantages lead to a reduction in overprecision (e.g., Glaser et al. [2013]). Learning more about this question helps to understand how and when additional information can help to reduce overprecision.

3. *Is it possible to induce people to become more precise?*

Moore, Tenney et al. [2015] observe that participants generally do not respond much when asked to provide less precise estimates. To the best of my knowledge, inducing people to become *more* precise has not been tested yet. However, given that people already struggle with providing accurate estimates, it is important to understand whether (over)precision can be exacerbated through communication.

I use an experimental research design employing participants acquired via Prolific.co. In the first half of the experiment, participants view ten different pictures with dots on them and have to estimate the number of dots after seeing each image for ten seconds (see appendix 7 for the experimental material). For each picture, they provide an estimate for the lower bound

sense. In overprecision research, the term typically implies that in a repeated setting, 90% of range estimates should cover the true value.

and the upper bound of their range estimate. The aim is that at least 9 out of the 10 estimates cover the true number of dots. In the second half of the experiment, participants judge the balance between accuracy and precision of range estimates to the same task. 80 % of these ranges are computer-generated, but the remaining 20 % are randomly selected, transformed² ranges that participants provided in the first half of the experiment. I match the judgment of computer-generated ranges with the judgment to the same ranges that a participant provided to their own transformed range.³ I find evidence that participants judge their own (transformed) range estimates significantly higher than another person would judge the same, computer-generated ranges. This indicates that people have an inherent preferred level of precision and could explain why range estimates tend to be rather stable (Yaniv and Foster [1997]).

In order to examine how people use additional information to balance accuracy and precision, I provide two thirds of the participants with information about perfectly calibrated range estimates. These two thirds are again split in half with one half having the option to increase precision of the additional information at low cost, and the other half at high cost. In this setting, I have the opportunity to formulate exact expectations about how precise participants should be and assess overprecision (see sect. 3.1). It furthermore allows me to examine how participants incorporate additional information. I find that participants are on average overprecise. Without additional information, they are correct in about 20 % of the cases and with additional information, they are correct in about 60 % of the cases. Furthermore, I find that participants use the additional information to predictably decrease precision for low true values. For high

² I transform the ranges as otherwise, participants may realize that they view their own estimates. Transformed means that the true value of dots has the same relative distance to the lower and upper bound of the range estimate as the range that participants provided (e.g., a range of 50 to 100 for a true value of 75 would be a transformed range of 500 to 1000 for a true value of 750). The implicit assumption is that the inherent preference for precision is stable for the range width relative to the true value instead of absolute range widths (e.g., 300 regardless of the true value).

³ As an example: Two participants judge how well the following range balances accuracy and precision: the lower bound is at 100 and the upper bound is at 200. The true value is at 190. While both participants judge the same range, for one of them it is a computer-generated range and for the other, it is the transformed range (see the previous footnote) that they provided in the first half of the experiment. Neither knows that one range was provided by the participant himself.

true values, they increase precision, but less than what the additional information would allow them. This indicates that they indeed use additional information to re-calibrate their intervals and use it to balance accuracy and precision. I also find that additional information can significantly help to re-position the estimates. Practically speaking, my data suggest that additional information can be a helpful tool to improve range estimates. Still, even perfect outside information will not lead to the desired level of accuracy since participants appear to balance it with precision. My data furthermore suggests that allowing people to re-think the position of their range estimates may lead to a bigger improvement in accuracy than asking for ever wider ranges.

Finally, I assess whether it is possible to induce people to become more precise. I create another treatment arm orthogonal to the three aforementioned arms. I suggest to half the participants that they can use more precise (i.e. smaller) ranges if they feel like this is appropriate. This is only a mild intervention, but my data suggests that this already leads to an increase in precision, particularly for low values. In my setting, it does not lead to a direct decrease in accuracy. Taken together, I show that people react if allowed to shift the relative weight between accuracy and precision. This is important, especially when high accuracy is the goal. Then, even mild suggestions to increase precision should be avoided.

Section 2 discusses the relevant literature to my paper and formulates hypotheses. Section 3 describes the research design and section 4 presents the results. Section 5 concludes.

2 Literature and Hypotheses Development

The trade-off between accuracy and precision is important to study because people quite regularly seem to provide estimates that are too precise and thus inaccurate: they are overprecise. Overprecision, just as its sibling biases overestimation and overplacement, belongs to a

larger psychological bias called overconfidence (Bazerman and Moore [2012]).⁴ Within the overconfidence spectrum, overprecision appears to be the most persistent, yet least studied type of overconfidence (Moore Tenney et al. [2015]; Moore and Healy [2008]). Its relevance in many business settings is without doubt. Barrero [2022] finds that firm managers do not seem to be overoptimistic, but overprecise and overextrapolating when estimating sales growth, leading to significant losses in firm values. Ren and Croson [2013] show that overprecision may be related to over- and understocking (i.e. correctly estimating demand). Financial analysts and firm executives alike seem to be overprecise when predicting future market returns, with those that are more overprecise typically also having poorer performances (Ben-David et al. [2013]; Deaves et al. [2010], [2019]). Fedyk et al. [2020] argue that overprecision may even be one cause of the accrual anomaly. Overall, overprecision is prevalent inside and outside the firm with numerous decisions relying on estimated numbers provided through external and internal reporting (e.g., discount factors, cash flows, and demand; see Ben-David et al. [2013]).

In experimental tasks, overprecision is commonly elicited by asking participants to provide estimates that are correct in 90 % of the cases. Ever since the early landmark studies by Lichtenstein and Fischhoff [1977], Alpert and Raiffa [1982], and Russo and Schoemaker [1992], studies show that participants do not meet the 90 % requirement.⁵ Instead, they appear to be correct in about 50 % of cases (McKenzie et al. [2008]). The mechanisms behind the bias are not entirely understood (Moore Tenney et al. [2015]) and potential remedies to

⁴ Moore and Healy [2008, p. 502] define overprecision as the “excessive certainty regarding the accuracy of one’s beliefs”, overestimation as the “overestimation of one’s actual ability, performance, level of control, or chance of success” and overplacement as the (false) belief to be better than others. The distinction between the three biases is meaningful. The correlations between them are low, and they indeed appear to capture distinct phenomena (Deaves et al. [2009]; Glaser and Weber [2007]; Moore and Healy [2008]). Note that the literature usually uses the term “miscalibration” synonymously with overprecision (e.g., Deaves et al. [2010]).

⁵ Implementations differ slightly across studies. In Lichtenstein and Fischhoff [1977, p. 164], participants themselves indicate a probability that their answer to a question is true. One example question is: “Absinthe is (a) a precious stone, (b) a liqueur.” While this measures overprecision in binary choice tasks, overprecision is also commonly observed in range estimates. One early example estimation task is: “The total egg production in millions in the U.S. in 1965” (Alpert and Raiffa [1982, p. 298]).

overprecision, such as feedback or monetary incentives, often show only modest improvements.⁶ In this study, I use a task that is commonly employed and consistently leads to overprecision: providing range estimates under a given level of accuracy (Klayman et al. [1999]).⁷

2.1 Range Estimates: a Trade-off between Accuracy and Precision

In principal, providing range estimates that are correct in 90 % of all cases presents an easy task: one could simply provide extremely large ranges in 9 out of 10 cases and one interval that is virtually impossible to be correct. However, this is not how people commonly behave. Instead, they appear to increase precision at the expense of accuracy. Yaniv and Foster [1995] argue that this is due to people viewing such tasks essentially as communication between a sender and a receiver. Social norms that typically apply in communication (“conversational norms”, see also Grice [1975]) imply that the receiver likely expects an informative answer. Now, providing extremely wide ranges would not provide much information and thus would not be in line with social norms. Hence, the sender trades-off precision with accuracy, ultimately resulting in overprecision in range estimates.⁸ In line with their theory, Yaniv and

⁶ Langnickel and Zeisberger [2016] suggest that people with better numeracy skills may exhibit less overprecision. Hilary and Hsu [2011] report that past success may lead people to become overprecise. Juslin et al. [2007] argue that people might be naive intuitive statisticians in that they use sampling properties as estimates for population properties. The sampling variance would thus be a biased and an overly precise estimate of the true variance. Overall however, no single theory emerged that could by itself explain the overprecision phenomenon (Moore Tenney et al. [2015]). With regard to mitigation strategies, Moore, Tenney et al. [2015] summarize the most promising approaches: feedback (e.g. Bolger and Önkal-Atay [2004]; Lichtenstein and Fischhoff [1977]; Russo and Schoemaker [1992]; Teigen and Jørgensen [2005]), having people consider more information (e.g. Koriat et al. [1980]; Remus et al. [1995]), and having people concentrate on different aspects of their response separately (e.g. asking for the 5th and 95th percentile separately; Haran et al. [2010]; Soll and Klayman [2004]; Speirs-Bridge et al. [2010]). Intriguing approaches, yet with mostly limited success, include monetary incentives (e.g., Cesarini et al. [2006]), expertise (e.g. Glaser et al. [2013]; McBride et al. [2012]; McKenzie et al. [2008]), and teamwork (e.g. Moore et al. [2017]; Plous [1995]).

⁷ Some studies point out that asking for range estimates may not be the best proxy for overprecision in decision making (see, e.g., Fellner and Krügel [2012]; Fellner-Röhling and Krügel [2014]; Glaser et al. [2013]). It appears that asking for frequency judgments in contrast to confidence intervals or probability estimates reduces measured overprecision (Gigerenzer et al. [1991]; Langnickel and Zeisberger [2016]). Note, though, that frequentist statements may not be a practical solution since they require estimators to provide multiple estimates and then declare how many of them they believe to be correct.

⁸ As noted above, other explanations have been proposed. Not all of those are reconcilable with the conversational norms hypothesis (e.g. Juslin et al. [2000]). However, no theory so far has emerged as a standard explanation.

Foster [1995] find evidence that information receivers judge the usefulness of an estimate based on a function of both its accuracy (in their setting defined as the interval's distance from the true value) and precision (the width of the interval). Yaniv and Foster [1997, p. 25f.] complement these findings by showing that individuals who provide estimates achieve similar (low) accuracy when asked for ranges that cover the true value in 95 % of all cases and when asked for ranges they “felt comfortable with.” The authors show that calibrating the estimates to meet their target of being correct in 95 % of all cases would substantially increase the intervals, likely resulting in uninformative estimates. Yaniv and Foster [1997] furthermore show students the calibrated (accurate) confidence intervals and receive feedback that such coarse intervals would make the responses worthless.

The evidence to date is not conclusive as to whether the desire to provide informative answers in the sense of Yaniv and Foster [1995] is the underlying reason for overprecision. Moore, Tenney et al. [2015] point out that adhering to social norms would imply that individuals respond to their counterparts. But even when researchers explicitly ask for accuracy, participants show overprecision. On the other hand, people may be influenced by prior experience: Cesarini et al. [2006] find that neither the people providing range estimates, nor the people receiving the estimates believe they would be accurate. Kaesler et al. [2016] argue that some people may have a predisposition for precise answers that is caused by an inherent desire to resolve uncertainty. They may avoid large intervals as this would sustain the uncertainty that large intervals imply. Finally, O'Connor et al. [2001] find that peoples' confidence intervals are not symmetrically distributed around the midpoint, which the authors deem possible to be in line with respondents trying to convey additional information. Arguably however, this does not explain why the intervals are too narrow.

Overall, it appears that people do not choose the “easy” way of providing some very wide and few implausibly small intervals just to fulfil the task instructions. It rather seems as though

they view the estimation task as a joint optimization problem: they intend to be accurate, but they also want to be precise. In my paper, I explore how people balance accuracy and precision.

2.2 Hypotheses

Providing range estimates is a difficult task as long as people aim to balance accuracy and precision, starting with understanding what the experimenter expects. Løhre and Teigen [2017] point out that people may confound high confidence with high precision. Furthermore, they may struggle to translate a correctly calibrated response into confidence levels (Gigerenzer et al. [1991]). In addition to that, the task itself is difficult because people can only ever influence precision without knowing the exact impact on accuracy. Precision is a continuous variable that participants can directly control. Accuracy – i.e., whether the range covers the true value or not – for a single question on the other hand is binary: the guess is either correct or not. Hence, from a participants' perspective, any change to precision only influences the *probability* for an accurate answer. In essence, even large decreases in precision may not lead to an improvement in accuracy, which makes the trade-off difficult. Taking into account that the best estimate of a person may be far off the true value (Moore, Carter et al. [2015]), even a person who aspires to balance accuracy and precision, may end up only increasing their range width without increasing accuracy.

Taken together, a challenging task with unclear benefits of being less precise may lead people to resort to Yaniv and Foster's [1997, p. 25f.] observation and provide ranges they feel "comfortable with". In fact, studies suggest that range estimates appear to be rather stable (Langnickel and Zeisberger [2016]; Teigen and Jørgensen [2005]). My first hypothesis thus directly tests whether people provide ranges they "feel comfortable" with:

H1: People provide estimates with their preferred level of precision.

Resorting to one's preferred level of precision does not imply that people are unwilling or incapable of changing their estimates. For instance, Remus et al. [1995] show that people, albeit imperfectly, are able to incorporate outside information in order to improve the accuracy of their estimates.⁹ Improving accuracy can work through two channels. First, participants may improve accuracy at the expense of precision: they provide wider ranges. The second channel is less obvious but nonetheless important if people strive for a balance between accuracy and precision (Yaniv and Foster [1995]): people may re-position their range estimate. In other words, they do not provide wider intervals, but place them differently. This channel relates to findings by Moore, Carter et al. [2015]. They provide evidence that people may not be overly precise, but their estimates are so far off the true value, that even range estimates that are theoretically *too wide* would not be able to cover it.¹⁰ Attempts that solely aim to decrease peoples' precision may thus not be sufficient since the position of the range estimate is so far off the true value.

⁹ Remus et al. [1995] ask participants to forecast time series. They provide participants either with perfect information about the future development of the time series, with imperfect information that contains probabilities about the future development, or with no information.

¹⁰ Moore, Carter et al. [2015] explicitly discuss the trade-off between accuracy and precision by designing an experiment for which an optimal confidence interval is estimable. They find that participants' probability distributions are wider than what theory would imply. At the same time, confidence intervals appear to be centered around the wrong value. While participants' probability distributions were wider than theoretically justified (i.e., indicating underprecision), they were still not wide enough to cover the actual value. Hence, participants still fail to account for their own error distributions (Moore and Schatz [2017]). Moore, Carter et al. [2015] note that conversational norms are not a likely explanation for this observation because the theoretical ranges are too wide and thus would not indicate a desire for precision. While this is compelling, the correct theoretical range width is based on the position of the true value and thus is not static across participants. In other words: If a participant's estimate for the correct value is 800 % too large, then it would only be reasonable to also expect their confidence interval to be larger than for a participant whose best guess is closer to the smaller correct value. For example, Moore, Carter et al. [2015], p. 112, ask: "Suppose you are planning to participate in a lottery game. Each day there is a 60% chance you will win \$1 and a 40% chance that you will lose \$1. How much money will you end up with after 500 days?". The most likely outcome would be $(0.6 - 0.4) * 500 = 100$ with a 90% confidence interval of about 35. However, these 35 are dependent on the distribution. If, for instance, a participant's best estimate for the most likely outcome was 200, then the parameters of the equation would necessarily need to change (e.g. by implicitly increasing the runs to 1,000), thereby changing the theoretical confidence interval as well (in this case to about 50). Arguably, this is still below the values observed in Moore, Carter et al. [2015], but it illustrates that one cannot fully infer correct confidence intervals from theoretical distributions if the best estimate itself does not even remotely reflect the theoretically correct value.

Obviously, people can also use additional information to *increase* precision. In this regard, one perhaps surprising result in prior research is that experts do not seem to be better at using their information advantage for providing accurate responses than novices (e.g. Glaser et al. [2013]; McKenzie et al. [2008]). It appears that experts use their advantage and provide estimates closer to the true value, but with such narrow intervals, that overprecision still prevails (McKenzie et al. [2008]). This, however, does not imply that people would not *try* to use their advantage to also improve accuracy as the relationship between accuracy and precision is not linear.

Under the sustained assumption that estimation tasks present a joint optimization problem to participants who want to balance accuracy and precision (Yaniv and Foster [1995]), I expect people to *predictably* respond to additional information that allows them to re-calibrate the estimates.¹¹ As long as people consider the trade-off between accuracy and precision (Yaniv and Foster [1995]), I expect the dominant strategy to be non-random and to depend on the precision and position of the outside information relative to participants' prior beliefs. Section 3 describes how I operationalize my predictions. I formulate the following hypotheses:

H2a: People *predictably* respond to additional information by decreasing precision.

H2b: People *predictably* respond to additional information by increasing precision.

H2c: People *predictably* respond to additional information by re-positioning their ranges.

The hypotheses so far argue that people provide estimates in line with their own preferred level of precision when they do not have additional information and that people use additional information to re-calibrate their responses. Results in line with the predictions would imply

¹¹ *Additional* information in my setting means information that participants did not have prior to the experiment, that gives them a noisy signal about a possible range estimate that covers true value.

that people actively balance accuracy and precision. What is still unclear to this point is whether participants do so in response to task instructions. As Moore, Tenney et al. [2015] point out, it appears very difficult to induce people to become *less* precise (and hopefully more accurate). The authors argue that this is an indication that people do not seem to respond to task instructions. However, I argue that this finding could be explained by people *believing* they are already appropriately calibrated. Hence, in order to test whether people respond to instructions, I will do the opposite and test whether it is possible to make people *more* precise (and less accurate).

H3: It is possible to increase (over)precision in participants' range estimates through task instructions.

3 Research Design

Strictly speaking, my hypotheses do not rely on people being overprecise as they are only concerned with how people make range estimates in the presence or absence of external information and instructions. Still, my findings draw much of their appeal from the fact that people are overprecise. Research on overprecision has faced criticism from scholars who argue that the bias we observe may to some part be an artifact of research design (Gigerenzer et al. [1991]; Juslin et al. [2000]). Thus, a central challenge in overprecision research is to determine whether participants are actually overprecise. In order to contribute to the overprecision literature, I need to adapt my research design accordingly. Hence, this section will first explain how I capture overprecision and then continue with the hypotheses tests.

3.1 Is it Overprecision?

Measuring overprecision comes with several experimental challenges:

1. It has to be clear how precise participants ought to be (Moore, Carter et al. [2015]). A comparison of true values with participant ranges can only be valid if the researcher knows *ex ante* how precise a participant should be.
2. The results should not be driven by participants' limited understanding of confidence intervals (Gigerenzer et al. [1991]).
3. Participants should not blindly guess. At the same time, additional information must be of similar quality and has to be interpretable in the same way for every participant. Otherwise, the incorporation of additional information would systematically vary with participants' (unobservable) cognitive abilities (Glaser et al. [2013]).
4. Confidence intervals around participants' best estimates are usually not symmetrical (O'Connor et al. [2001]). Thus, measuring overprecision has to be possible within the given question design, but without any assumptions about properties of participants' underlying estimate distributions.

My setting aims to address all these challenges. Participants in my experiment see ten different pictures with dots on them (see appendix 7). The true number of dots is any random integer between 100 and 1,100 and unknown to participants. Each picture is shown for 10 seconds. After the picture disappears, participants have to provide a range to each of the pictures so that at least 9 out of the 10 estimates cover the correct number of dots.¹² This task has two main advantages: first, it is relatively unusual, so it is unlikely that some participants have an

¹² In order to alleviate concerns of mixing probabilistic statements (for one question) with frequentist statements (for all questions), I ask for 9 correctly answered questions out of 10 (instead of 90 % confidence for a given question, Teigen and Jørgensen [2005]).

advantage over others (e.g. like for trivia knowledge). Second, I can manipulate the true value, which is important for ensuring that participants perceive a different level of difficulty across the questions (Gigerenzer et al. [1991]; Klayman et al. [1999]).

Each participant randomly receives a set of true values (i.e., dots). Key to my design is that each participant also receives a range estimate to each of the true values. I (truthfully) tell participants that “it may be difficult to quickly assess the number of dots on each picture. This is why I provide you with some help. In each round, you will have a ‘little helper.’ The little helper will give you a suggestion for a range estimate. It is programmed so that it is correct in 9 out of 10 cases.” I ensure this to be true by pre-defining the sets of true values as well as the suggested ranges. The width of the suggested ranges is uniformly distributed between 100 and 500. Their exact position is random, but with 9 out of 10 sets covering the true value. For the one set that does not cover the true value, I also assign a random position with the limitation that it cannot be farther away from the true value than 100. This ensures that participants do not receive range suggestions that are obviously wrong. Participants may not perceive the intervals to be particularly precise. However, they allow participants to calibrate their responses. Without any additional knowledge (which participants cannot have in my setting), the given intervals present a calibrated response that is correct in 90 % of the cases. If, however, participants provide smaller ranges that contain the true value in less than 90 % of the cases, I can directly infer overprecision. I therefore know how precise participants ought to be (challenge 1). By showing participants a suggested range, I also do not need to rely on the participants’ understanding of what a confidence interval is (challenge 2). Furthermore, the information signal through the range is of similar quality and in only one way interpretable (challenge 3). Finally, I do not need to know participants’ estimate distributions. Any participant with smaller ranges who is incorrect in more than one instance is identified as overprecise (challenge 4).

3.2 Hypotheses Tests

For my first hypothesis (**H1**), I expect that people provide estimates with their preferred level of precision. I test this by splitting the experiment in two parts:

1. In the first part, participants have to provide the above-mentioned ten range estimates of which at least 9 are supposed to cover the true value.
2. In the second part, participants judge 25 different range estimates with five sets of five estimates each (see appendix 7 for an example). The following statement inspired by Yaniv and Foster [1997] is used to elicit the participants' judgment: *The given range estimate is useful in that it balances accuracy and precision well.* Participants state how much they agree on a 7-point Likert scale ranging from “strongly disagree” to “strongly agree”.

All five estimates have the *same* underlying true value *within* each of the five sets in the second part of the experiment. The true value *differs across* the sets. Out of the five estimates per set, four are computer-generated with two of them having a range that covers the true value, and the other half not.¹³ One range per set, however, is not computer-generated, but a transformed range provided by the participant in the first half of the experiment.¹⁴ The selection of participant ranges from the first half of the experiment is random. As there are five sets, I only

¹³ The algorithm is similar to the one for the *helper* described above: First, I randomly select a true value between 100 and 1,100. Now, I generate four different ranges to that true value. The algorithm randomly generates a range width between 100 and 500 (uniformly distributed). The range position is also random; I sample until I have two ranges that cover the true value and two ranges that do not cover the true value.

¹⁴ Obviously, I cannot show participants the same range that they have provided in the first half of the experiment. I therefore transform the range I show to be just as calibrated as the participant's range in the first half. Being of similar calibration means that the relative distance of the upper and lower end of the range to the true value, relative to the true value, is the same. The transformation for the lower range would be: $x_{second} - \frac{x_{first} - input_{lower}}{x_{first}} * x_{second}$, with x being the true value in the first and second half of the experiment, and the input being the participant estimate for the lower end of the range. For example, assume that in the first half of the experiment, the true (yet unknown) number of dots on the picture was 200. The participant estimated a range of 100 to 300. The true value in the second half of the experiment is 400. Therefore, the lower end of the range that I show participants is $400 - \frac{200-100}{200} * 400 = 200$. Similarly, the upper end of the range is $400 + \frac{300-200}{200} * 400 = 600$.

use five out of ten ranges that participants provided in the first half of the experiment. Furthermore, the order of the ranges shown to participants in the second half of the experiment is also random. Given these precautions, it is highly unlikely that participants recognize that they evaluate a transformed range that they provided themselves. In summary, participants view five sets of five ranges. Within each set, one range is a transformed range that the participant provided themselves, and four ranges are computer-generated so that two cover the true value and two do not. Taken together, my approach allows me to let people judge their own transformed range estimates without them knowing.

The key is now to compare a participant’s judgment to their own range estimate with another participant’s judgment that was computer-generated.¹⁵ If participants have an inherent preference for a certain level of precision, I expect that participants judge their own range estimates higher than participants who judge *the same* computer-generated range.¹⁶ I use coarsened exact matching to match participants’ own ranges with other participants’ ranges. The matching procedure and results are presented in section 4.

As hypothesis **H1** covers range estimates without any additional information, I restrict the sample to participants who did not have additional information provided by the “little helper”. Arguably, this is the setting that is most likely to capture ranges that participants “feel comfortable with” (Yaniv and Foster [1997]).

¹⁵ Note that the range needs to be computer-generated, because otherwise I would compare a participant’s judgment of their own transformed range with another participant’s judgment of their respective transformed range. This would violate the notion behind H1, that the comparison is between the judgment of a range created by oneself and a range not created by oneself.

¹⁶ To continue the previous example, I would expect that the participant who sees their own transformed range of 200 to 600 with a true value of 400 judges this range more favourably than another participant who sees the exact same but computer-generated range for the same true value. Note that any finding in this design cannot be explained by the position of the range relative to the true value, since both range estimates are the same. They only differ in who made the range estimate: the participant themselves, or a computer. The participants do not know that some ranges are their own transformed ranges. Any difference in judgment must therefore arise from participants’ inherent preferences. Note that the computer-generated ranges were determined prior to the experiment. Thus, I could not ensure that all participant-generated ranges would have a computer-generated equivalent. However, the number of participants and ranges is large enough to create sufficient overlap for matching.

Hypotheses **H2a-c** argue that participants predictably change the precision as well as the position of their range estimate. In order to test the hypotheses, I make use of the “little helper”. About two thirds of the participants are in the *helper* condition. They receive range estimate suggestions as described above. Following the findings of Remus et al. [1995], I expect the information received in the *helper* condition to significantly improve accuracy for all true values. However, I expect the precision of estimates to vary depending on the true value: An increase in the true value is likely to be related to an increase in participants’ range width.¹⁷ At the same time, the range width proposed by the little helper is uniformly distributed between 100 and 500. Therefore, each unit increase in the true value increases the likelihood that the range suggested by the helper is smaller than the range participants would have chosen without the helper. If participants incorporate additional information in line with **H2a**, I expect the helper to decrease precision (i.e. widen the interval) for relatively smaller true values. In line with **H2b**, I expect larger true values to be associated with an increase in precision. If I observe an increase in accuracy for *all* true values and thus importantly also for those ranges that do not increase, then I have evidence that participants shift their range estimate in response to the helper (**H2c**). The exact switching point (low versus high values) is ultimately an empirical question.

Results in line with these expectations would suggest that participants re-calibrate their responses through managing their level of precision and position of range estimates in response to additional information. However, they would not allow for a direct conclusion as to whether participants balance accuracy and precision – it is not clear whether participants aim to maximize precision while holding the level of 90 % accuracy constant, or whether participants are re-calibrating both accuracy and precision. In order to shed more light on this question, I

¹⁷ The reason is the existence of a natural lower bound for the true value (zero), since low estimated values offer less room for the lower end of the range estimate than larger values do.

introduce another treatment condition. For five out of ten images in the *helper* condition, participants have the ability to “recalibrate” their helper. Participants are (truthfully) informed that recalibration increases precision without any change in accuracy. However, recalibration takes time. I split participants in the helper condition in two equal-sized groups: for one group, recalibration takes ten seconds to complete (*low cost* condition). For the other group, recalibration takes 30 seconds (*high cost* condition). The increase in precision does not differ between the two groups. Participants have to decide for or against recalibration before they see any range suggested by the helper. Note that they have no obvious incentive to use the recalibration option as compensation is not linked to accuracy or precision.

Now, having these two conditions, I know the calibrated ranges that the *helper* shows participants as well as the ranges that participants enter. I expect that participants in the *low cost* condition use recalibration more often than participants in the *high cost* condition. If so, the average increase in *precision* of the helper is larger for participants in the *low cost* condition than for participants in the *high cost* condition. In other words: The increase in precision by the helper would thus be relatively higher for participants in the *low cost* condition. Recall that participants know that any increase in the helper’s precision from recalibration does not affect accuracy. Therefore, any increase in participant precision that is lower than the increase in the helper’s precision between the *high cost* and *low cost* condition, implies that participants are willing to sacrifice precision in the hope of increasing accuracy. This would be direct evidence that participants balance accuracy and precision in their responses. Analogously to the discussion above, I expect the increase in precision to be visible for higher true values, but not for lower true values. In summary, I compare differences in changes in precision from the *high cost* to the *low cost* condition between ranges shown by the helper and participants’ ranges. Significant differences suggest that participants actively balance accuracy and precision.

The final hypothesis **H3** argues that it is possible to increase (over)precision. In order to test this hypothesis, I create one final, simple treatment condition. In the *high precision condition*, participants are being told that “[i]n the end, selecting a range is entirely up to you: If you feel like a range is too large, then it may make sense to choose a smaller range”. I do not ask participants for smaller ranges directly, since this would send mixed signals about the ultimate aim of the task. The second group of participants does not receive this statement. If the third hypothesis holds, I expect accuracy in the *high precision* condition to be lower and precision to be higher than in the control condition.

Overall, I have a 2x3 full factorial design with one treatment arm being the *high precision / low precision condition*, and the other arm the *no helper / helper with low cost / helper with high cost* condition (see appendix 1 for a link to the experiment, as well as more information on treatment conditions and variable definitions).

3.3 Study Conditions and Participant Selection

The experiment is executed online via an interactive Shiny app. I invite participants using Prolific.co. Prolific specializes in matchmaking between researchers and participants for online studies.¹⁸ Data was collected between October 12 and October 14, 2021. Since the experiment is entirely in English, I pre-screen participants to be fluent in the language. Participants are paid 3 GBP for an estimated median time of 20 minutes to complete the experiment. They do not receive monetary incentives to produce accurate estimates. While incentives can be a good way to clarify instructions, they invite people to game the task and maximize payouts by simply providing extreme ranges (Moore, Tenney et al. [2015]). This would strongly work against participants to provide ranges they “feel comfortable with”, which is crucial, especially for

¹⁸ See Palan and Schitter [2018] and Peer et al. [2017] for a comparison of Prolific with its main competitor in the field, Amazon MTurk.

testing H1. Arguably, even without monetary incentives participants could feel invited to game the task. However, there is no evidence of such behavior in prior literature.

The task involves the first and second half of the experiment as outlined above. After completing both tasks, participants were invited, but not forced, to provide demographic information about themselves (see appendix 1 for all variables). About 95 % of participants provide demographic information. I also include two attention checks. The questions read:

1. “What was your task in the first half of the experiment?” with the correct answer: “Be accurate: Estimate ranges that cover the correct number of dots in at least 9 out of 10 cases”.
2. “What was your task in the second half of the experiment?” with the correct answer: “Judging the usefulness of different ranges by assessing the balance between accuracy and precision”.

The attention checks are strict. It was not possible for participants to go back and check the correct answer. Especially the analyses related to H2 and H3 require that participants answered the first attention check correctly. Arguably, I lose a sizeable number of participants. However, it is vital for the experiment that participants answer both questions correctly, as only then the results of my experiment are interpretable with respect to overprecision. See table 1 for the sample selection.

[Table 1]

As an additional check, I screen participants’ ranges for implausible results. I define an implausible result if the lower end of the range is above the upper end of the range.¹⁹ If that happened more than once, I excluded the participant. If it happened once, I excluded only the

¹⁹ It was not prohibited to enter such ranges, as I wanted to have an additional option to identify participants who did not take the task seriously.

observation. Furthermore, I excluded participants whose upper limit is 50 or lower more than once as these ranges are implausibly low. Finally, I correct select values for which participants indicate that they accidentally entered a wrong guess in the open text field at the end of the experiment.

4 Results

4.1 Descriptive Statistics

The final sample consists of 180 participants who provide 1,796 range estimates in the first half of the experiment and judge 4,489 ranges in the second half of the experiment. The median time of completion is about 16 minutes (see table 2). 47 % of participants identify as female. The average age is 27 years. Participants on average agree that it was expected of them to use the helper, if available, and on average slightly agree that they were expected to recalibrate when possible.

[Table 2]

In terms of precision, the median range width in the *no helper* condition is 100, which is considerably smaller than in the *helper* condition with a range width of 180. This however, is also considerably smaller than the median range suggested by the helper at 260. Overall, only 5 % of all participants provided ranges larger or as large as the range suggested by the helper in 9 or 10 cases. Accuracy on the other hand is low at an average of about 47 % for the whole sample and 61 % (20 %) for participants in the *(no) helper* condition. These results are in line with prior findings in the overprecision literature (McKenzie et al. [2008]). Participants in my experiment thus appear to be overprecise.

Table 3 shows the results of two-sided t-tests for all treatment conditions. Figure 1 shows graphical evidence. In line with my expectations, participants in the *high precision condition*

appear to be slightly less accurate with a mean *accuracy* of 45 % in the treatment and 49 % in the control condition ($p = 0.06$). At the same time, participants also appear to be slightly more precise ($p = 0.06$). Note that results for precision (i.e., the range width) are log-transformed due to the shape of the distribution of range estimates.

[Table 3]

[Figure 1]

Furthermore, participants in the *helper* condition are about 40 percentage points more accurate than participants without the helper ($p < 0.01$). Precision is also significantly lower in the *helper* condition ($p < 0.01$). Furthermore, participants in the *high cost* condition appear to be less precise ($p = 0.06$), but without any significant increase in accuracy. Finally, participants in the *low cost* condition use the option to recalibrate on average 2.9 times, while participants in the *high cost* condition use the option only 2.2 times (also see appendix 3 in this regard). Accordingly, the ranges shown to participants in the *high cost* condition are on average less precise than in the *low cost* condition ($p = 0.01$) when recalibration was available. Participants' precision in the *high cost* condition is not significantly different from precision in the *low cost* condition ($p = 0.15$).

The results up to this point are based on the range level and can only be a first insight since the t-tests do not account for clustered standard errors (each participant provides 10 ranges). The next sections will thus dig deeper into answering the hypotheses.

4.2 Hypotheses Tests

In order to shed light on hypothesis **H1**, I first analyze which factors drive participants' judgment of ranges. Table 4 and figure 2 present the results. It is evident that both accuracy and precision matter for participants' judgments. However, while precision matters regardless

of whether the range covers the true value or not, its effect on participant judgment is between three and four times higher on participants' judgment when the range covers the true value than when it does not. Figure 2 highlights how important it appears to be for participants that the range covers the true value (note that this includes the 80 % computer-generated ranges that are by construction close to the true value, even if they do not cover it). Participants in the *helper* condition view the ranges provided as more positively than participants in the *no helper* condition, which is explained by their own transformed ranges being on average far better calibrated. There is no difference in judgment for the other treatment conditions.

[Table 4]

[Figure 2]

The data implies that participants' judgments of ranges is distinct. For instance, it matters substantially whether a range of similar width just covers the true value or it does not. Even small shifts in the range position and, to a lesser extent, the range width, can have a large influence on the judgment. Therefore, I decide against propensity score matching for matching participants' own ranges with computer-generated ranges.²⁰ Due to the peculiar nature of the data, it would be ideal to match on the exact range position and the exact position of the true value. Unfortunately, this reduces the number of matches to unusable levels. Therefore, I apply coarsened exact matching. I match on the lower end of the range, the upper end of the range, and the position of the true value. Additionally, I apply an exact match on whether the range covers the true value or not. Otherwise, it could happen that even a small unit change in the position of the true value could move it over the range bounds, influencing the participant judgment severely. With regard to the balance between good matches and power, I decide for better, but less matches in order to avoid false-positive results. Matching results are presented

²⁰ Regardless, I try propensity score matching in different specifications, but the matching statistics also suggest not to go forward with this approach.

in appendix 6. The statistics, as well as a manual inspection of matches, suggests that matching worked well.

As for the hypothesis test, technically, a t-test between participant judgments for own and computer-generated ranges would be sufficient. Still, I decide to run regression analyses that also control for accuracy and precision (first set of regressions) and for matching variables (second set of regressions) in case matching did not work perfectly. Table 5 presents the regression results.

[Table 5]

The first and fourth regression in table 5 are the strictest in that they only include participants who answered all attention checks correctly. The number of matches is rather low at 70. The relationship between seeing the own range and judgment is, albeit as expected in sign, insignificant. However, for the second half of the experiment it is not crucially important that participants answered the attention check to the first half of the experiment correctly. After all, not answering this attention check correctly may actually be more in line with participants providing a range they “feel comfortable with”. Therefore, I decide to run the analyses again on matched data that contains all participants who only answered the second attention check correctly. I also run the analyses with all participants. Both substantially increases the number of observations. Results are significant on the five percent level with a magnitude that is similar to the results in the first and fourth regression.²¹ This supports the notion that results in the first and fourth regression may be insignificant due to limited power. Overall, I view the results in table 5 to support hypothesis **H1** in that people have and apply an inherent level of precision.²²

²¹ Seeing the own range on average increases judgment of about 0.4 points on a 7 point Likert scale.

²² Note that results are naturally also depending on the matching algorithm. The results are not the same across matching techniques. However, this is to be expected. I choose the matching algorithm with the best matching statistics.

I now turn to hypotheses **H2a-c**. I will thus first discuss results for the data set containing all true values and then split the sample for low and high true values as explained in section 3.2.

[Table 6]

Table 6 shows OLS regression results covering all experimental conditions separately, as well as combined and including control variables. Since the treatments are randomized across participants, control variables are not strictly necessary. I still use participants' age, their liking of math, as well as their gender as control variables in case randomization was not entirely successful.²³ Accuracy, as the dependent variable for the first four logit regressions, is a binary variable equal to one if the range provided by the participant covers the true value and zero otherwise. Precision, as the dependent variable for the last four OLS regressions, is the log-transformed width of the range provided by participants. I use clustered standard errors on the participant level. The results indicate that there is no overall significant association between the *high precision condition* and precision or accuracy. Participants in the *helper* condition have a significant (and sizeable 40 percentage point) increase in accuracy, but do not show a significant change in precision. I do not find a significant effect of being in the *high cost* condition on accuracy or precision.

It may seem puzzling that participants in the *helper* condition are able to improve their accuracy dramatically without changes to precision. However, this is still in line with expectations as I do not expect an overall significant effect of the helper on precision. Figure 3 implies that accuracy is similarly higher across all correct values for people in the *helper* condition. However, participants' precision is not. For low true values, the *helper* appears to decrease participants' precision relative to people without a helper. For higher true values however,

²³ I choose their liking of math as research suggests that people with better numeracy skills are better at overprecision tasks (Langnickel and Zeisberger [2016]). Some studies also indicated a relationship between age (e.g. Crawford and Stankov [1996]; Kaesler et al. [2016]) and gender (e.g. Lundeberg et al. [1994]) with precision and accuracy.

people in the *helper* condition on average increase precision. They still provide far more accurate estimates than participants without a helper. This implies that they tend to shift the position of the interval for larger true values. In summary, the graphical evidence is in line with **H2a-c**.

Finally, figure 3 (third graph) shows that despite increasing range width in the *no helper* condition, it is still by far not enough to be close to the true value. On the contrary, participants in the *helper* condition, who only modestly increase their range widths for larger true values (see second graph), are very close to the true value across all estimates. As such, the graphical evidence indicates that overprecision may not just be an issue of small ranges, but of severely wrongly positioned ranges.

[Figure 3]

I define the breakpoint between low and high true values as the intersection of the slopes from a linear regression of precision on the true value, once for participants in the *no helper* condition and once for participants in the *helper* condition (similar to the graphical evidence as shown in figure 3, graph 2). Intuitively, the intersection represents the point at which participants in the *helper* condition switch from being less precise to being more precise relative to participants in the *no helper* condition. The breakpoint is at a true value of 375. Now, I run the same regressions as for table 6 separately for true values below and above 375. The results are presented in table 7.²⁴

[Table 7]

Table 7 reveals an interesting pattern. Across all true values, the *helper* condition improves accuracy significantly. However, as expected, it is only for small true values that the helper

²⁴ The breakpoint by definition splits the sample into a group where precision is on average higher for the helper condition and a group where precision is lower for the helper condition. However, this does not automatically imply that the differences between the conditions are significant (evidently, as the results show). Second, accuracy does not automatically follow precision since ranges are, as discussed above, often severely wrongly positioned. The results thus help to distinguish the effects of the *helper* on precision, accuracy, and the range position.

condition significantly decreases precision. Interestingly, the *helper* does not seem to increase precision for high true values. Still, given that accuracy is significantly improved, the *helper* must have changed participants' range position.

I now assess whether participants balance accuracy and precision. I begin by showing univariate statistics in table 3, Panel C. Across all true values, the ranges that participants see in the *helper* condition do not differ when participants have no option to recalibrate. However, ranges are significantly lower in the *low cost* condition when recalibration is available. This is in line with the expectation that participants recalibrate more often in the *low cost* condition (see table 10 in appendix 3). Results for *accuracy* are not displayed for the *helper*, as they are 0.9 across all conditions by construction. Participants seem to follow a range decrease by the helper in general, as shown by the smaller range estimates between conditions when recalibration was available versus when it was not available ($p = 0.04$ and 0.06). There is an insignificant increase in precision in the *low cost* relative to the *high cost* condition.

I present multivariate statistics in table 8. In line with my hypotheses, I split the sample into high and low true values. I then assess range estimates via the following regression:

$$\begin{aligned} \log(\text{precision}_t) = & \alpha + \beta_1 \text{helper}_t + \beta_2 \text{high effort}_t + \beta_3 \text{smaller option}_t + \\ & \beta_4 \text{helper}_t * \text{high effort}_t + \beta_5 \text{helper}_t * \text{smaller option}_t + \beta_6 \text{high effort}_t * \\ & \text{smaller option}_t + \beta_7 \text{helper}_t * \text{high effort}_t * \text{smaller option}_t + \varepsilon_t \end{aligned}$$

The dependent variable is the logarithm of the range width provided by the helper or the participant to the same underlying true value. The second and fourth regressions use accuracy as the dependent variable (defined as above). *Helper* in this regression is a binary variable equal to one if the range was provided by the helper. *High cost* is a binary variable equal to one for estimates provided in the high cost condition. *Smaller option* is a binary variable equal to one for estimates provided when recalibration was possible (not when it was used). Standard errors

are clustered on the participant level (with the helper being considered a different participant for each set of estimates that one participant makes).

[Table 8]

There are no significant differences for low true values with the exception that the helper's estimates are less precise and more accurate. However, there is an interesting pattern for high true values: Again, ranges provided by the helper are significantly larger than ranges provided by participants (β_1 , $p < 0.01$). As expected, the interaction between *helper* and *smaller option* is negative and significant (β_5 , $p < 0.01$), indicating that the helper shows significantly smaller ranges whenever recalibration is possible. The interaction between *helper*, *high cost*, and *smaller option* is positive and significant on the 10 percent level (β_7 , $p = 0.07$). This indicates that the ranges that the helper shows participants in the *low cost* condition are indeed smaller than the ranges that the helper shows participants in the *high cost* condition when recalibration is possible. Participants on the other hand also reduce their range width in the *low cost* condition when recalibration is possible (β_3 , $p < 0.01$). They, however do not show any difference in precision between the *high* and the *low cost* condition (β_6 , $p = 0.60$). This means that although people see significantly smaller ranges in the *low cost* condition compared to the *high cost* condition, they do not translate this into smaller ranges for themselves. Therefore, the findings indicate that while people use outside information to increase precision (β_3), they do not incorporate all increases in precision that the outside information offers. Notably, having the option to recalibrate is significantly associated with higher accuracy ($p = 0.09$) for high true values. Participants thus do not just become more precise, but also more accurate.

Overall, my results are in line with my hypotheses: Participants predictably adjust their range estimates and re-position their estimates when presented with additional information. The data suggest that instead of aiming for a rate of 90 % accuracy, participants balance accuracy and precision.

For my final hypotheses **H3**, I examine the impact of the *high precision* manipulation. The t-tests presented in table 3 offer univariate support for the hypothesis. The results show that not only become participants more precise ($p = 0.06$, two-sided), but they also become less accurate ($p = 0.06$, two-sided) through the experimental manipulation. Table 6 shows that the results do not continue to hold on conventional significance levels in a multiple regression with standard errors clustered on the participant level. However, table 7 presents a more nuanced picture. The regression results show a significant increase in precision for low true values in the *high precision condition*. However, this is not accompanied by a significant decrease in accuracy, even though the signs are in line with expectations.

Overall, the results indicate that participants respond to the instructions by reducing their range widths under certain conditions – in this case, the simple and mild remark that they are free to choose a smaller range was sufficient to induce an effect. Even though this may not automatically lead to overprecision, it certainly reduces the likelihood of participants capturing the true value.

4.3 Additional Analyses

In this section, I provide further insights into the data. First, I analyze which mistakes participants typically make when providing their ranges. As can be seen in appendix 2 (table 9 and figure 4), ranges are rarely located above the true value. In other words, participants typically underestimate the number of dots. About 10 % of ranges overestimate the number of dots. Interestingly, this ratio is relatively stable across all treatment conditions. The main influence of the *helper* condition is therefore to push participants' estimates upwards. While about 70 % of the ranges in the *no helper* condition are too low, only a little less than 30 % of the ranges in the *helper* condition are below the true value. Overall, this leads to an accuracy of about 60 % in the *helper conditions* and about 20 % in the *no helper* conditions. Participants'

accuracy is relatively stable across rounds, with some increase observable for the *no helper* condition (see appendix 4). Note that this slight increase does not interfere with my results, as all treatment conditions, as well as recalibration options and true values, are randomized.

Appendix 3 informs about participants' decision to recalibrate. As noted earlier, participants in the *low cost* condition use the recalibration option more often than participants in the *high cost* condition. Figure 5 shows that the use of the recalibration option also changes whether participants follow the *helper*. People who use recalibration more often choose range widths that are more similar in size to the *helper's* range. Since the cost conditions were randomised across participants, the implied causal chain is that people in the low cost condition more often follow the helper's advice, likely mediated through their increased use of the recalibration option.

Appendix 5 shows how people rounded when entering their ranges. People without a *helper* appear to round their ranges more often. In the no helper condition, about 98 % of input values end in a 5 or 10, while this applies "only" to about 90 % of ranges in the helper condition. About half the people (again with less people in the *helper* condition) seem to round their inputs to 100s. Some, yet few people, follow the helpers' advice literally. Overall, 6.7 % of ranges entered follow the helpers' advice exactly (untabulated).

Finally, I present some answers from the open text fields. I asked participants why they used the helper, why they used recalibration, and I left space for a fully open text field. Overall, I identify three main reasons that people mention why they used the helper:

1. To increase precision: *"I thought it would help me to be more precise"*
2. To shift the range: *"It seemed to provide me with numbers that seemed impossibly large to me, and that is precisely why I decided to trust it. Initially, I thought it might be wrong (1/9 times it is going to provide me with inaccurate results!), but after it kept*

appearing and consistently giving me the large estimates I just assumed I was very bad at determining ranges and decided to trust it instead.”

3. Because it confirmed prior beliefs: *“Mainly the fact that it was in the range of my own estimations.”*

When asked about why participants recalibrated – keep in mind that there was no incentive to recalibrate – two themes emerged:

1. Perhaps unsurprisingly, curiosity: *“I wanted to see if the helper provides with similar estimates after recalibration”*
2. But also, a number of participants voiced the desire for more precise answers: *“I was naturally trying to make my ranges more precise, even though that was not part of the instructions.”*

Finally, the open text field revealed largely that participants were engaged in the study, but also found it challenging. Participants seemed aware of the challenges and the trade-off between accuracy and precision. As one participant put it: *“This experiment was very hard. I tried to always keep a range of 50 so that it my guesses were very precise. However I think i have underestimated each time my guesses. I do think though that the balance between my answer was somewhat decent.”*

Overall, 52 % of participants are still overprecise when asked after the experiment how many answers they believe they had correct. This implies that even the frequentist view cannot fully eliminate overprecision (Gigerenzer et al. [1991]; Langnickel and Zeisberger [2016]). About 13 % of participants are underprecise per this definition (untabulated).

5 Conclusion

I analyse how people balance accuracy and precision when providing range estimates. This is important because people are overprecise: when asked to provide estimates that are correct in 90 % of cases, they regularly are only correct in about half the cases. I confirm this in my experiment.

I find evidence that without any additional information, people seem to provide range estimates in line with *their* preferred level of precision. They thus seem to resort to providing ranges they “feel comfortable with” (Yaniv and Foster [1997]). Furthermore, I find that participants are able to incorporate additional information and adjust their range estimates. They use additional information to increase or decrease precision, depending on whether the information suggests larger or smaller ranges. Additional information also significantly improves accuracy in my setting, although it is not able to mitigate overprecision entirely. With additional information, participants are correct in about 60 % of the cases as opposed to approximately 20 % without additional information. This supports the notion that people do not use additional information to reach the desired level of 90 % accuracy, but to balance accuracy and precision.

In addition to balancing accuracy and precision, I find that participants use additional information to adjust the position of their range estimates. This finding is important, because it supports Moore, Carter et al. [2015] who suggest that precision may not be the only problem in overprecision. It appears that quite often the issue is not the precision of the interval, but rather its position. This highlights that it may not be sufficient to simply ask people to become more accurate or increase their ranges. Rather, people should be encouraged to re-think the position of their range estimates, preferably with the help of outside information that people evidently use.

Finally, I find that even a small change in wording that suggests that it is fine to provide more precise ranges if participants deem appropriate, has a measurable effect on precision. Although this may not directly translate into lower accuracy, my findings show that it is important to be aware that if accuracy is the goal, instructions should avoid any suggestion hinting at a desire for precision.

References

- ALPERT, M., AND H. RAIFFA "A Progress Report on the Training of Probability Assessors." *Judgment under Uncertainty*, Cambridge University Press (1982): 294–305.
- BARRERO, J. M. "The Micro and Macro of Managerial Beliefs." *Journal of Financial Economics* 143.2 (2022): 640–667.
- BAR-YOSEF, S., AND I. VENEZIA "An Experimental Study of Overconfidence in Accounting Numbers Predictions." *International Journal of Economics Sciences* III (2014): 78–89.
- BAZERMAN, M. H., AND D. A. MOORE *Judgment in managerial decision making* (8th ed.) John Wiley & Sons (2012).
- BEN-DAVID, I., J. R. GRAHAM, AND C. R. HARVEY "Managerial Miscalibration." *Quarterly Journal of Economics* 128 (2013): 1547–1584.
- BOLGER, F., AND D. ÖNKAL-ATAY "The Effects of Feedback on Judgmental Interval Predictions." *International Journal of Forecasting* 20 (2004): 29–39.
- CESARINI, D., Ö. SANDEWALL, AND M. JOHANNESSON "Confidence Interval Estimation Tasks and the Economics of Overconfidence." *Journal of Economic Behavior and Organization* 61 (2006): 453–470.
- CRAWFORD, J. D., AND L. STANKOV "Adult Age Differences in the Realism of Confidence Judgments: A Calibration Study Using Tests of Fluid and Crystallized Intelligence." *Learning and Individual Differences* 8 (1996): 83–103.
- DEAVES, R., J. LEI, AND M. SCHRÖDER "Forecaster Overconfidence and Market Survey Performance." *Journal of Behavioral Finance* 20 (2019): 173–194.
- DEAVES, R., E. LÜDERS, AND G. Y. LUO "An Experimental Test of the Impact of Overconfidence and Gender on Trading Activity." *Review of Finance* 13 (2009): 555–575.

- DEAVES, R., E. LÜDERS, AND M. SCHRÖDER "The Dynamics of Overconfidence: Evidence from Stock Market Forecasters." *Journal of Economic Behavior and Organization* 75 (2010): 402–412.
- FEDYK, T., Z. SINGER, AND T. SOUGIANNIS "The Accrual Anomaly: Accrual Originations, Accrual Reversals, and Resolution of Uncertainty*." *Contemporary Accounting Research* 37 (2020): 885–916.
- FELLNER, G., AND S. KRÜGEL "Judgmental Overconfidence: Three Measures, One Bias?" *Journal of Economic Psychology* 33 (2012): 142–154.
- FELLNER-RÖHLING, G., AND S. KRÜGEL "Judgmental Overconfidence and Trading Activity." *Journal of Economic Behavior and Organization* 107 (2014): 827–842.
- GIGERENZER, G., U. HOFFRAGE, AND H. KLEINBÜLTING "Probabilistic Mental Models: A Brunswikian Theory of Confidence." In *Psychological Review* 98.4 (1991): 506–528.
- GLASER, M., T. LANGER, AND M. WEBER "True Overconfidence in Interval Estimates: Evidence Based on a New Measure of Miscalibration." *Journal of Behavioral Decision Making* 26 (2013): 405–417.
- GLASER, M., AND M. WEBER "Overconfidence and Trading Volume." *GENEVA Risk and Insurance Review* 32 (2007): 1–36.
- GRICE, H. P. "Logic and conversation." *Speech acts*. Brill, 1975. 41–58.
- HARAN, U, D. A. MOORE, AND C. K. MOREWEDGE. "A simple remedy for overprecision in judgment." *Judgment and Decision Making* 5.7 (2010): 467–476.
- HILARY, G., AND C. HSU "Endogenous Overconfidence in Managerial Forecasts." *Journal of Accounting and Economics* 51 (2011): 300–313.

- JUSLIN, P., A. WINMAN, AND P. HANSSON "The Naïve Intuitive Statistician: A Naïve Sampling Model of Intuitive Confidence Intervals." *Psychological Review* 114 (2007): 678–703.
- JUSLIN, P., A. WINMAN, AND H. OLSSON "Naive Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect." *Psychological Review* 107 (2000): 384–396.
- KAESLER, M., M. WELSH, AND C. SEMMLER. "Predicting overprecision in range estimation." *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (2016): 502–507.
- KLAYMAN, J., J. B. SOLL, C. GONZÁ LEZ-VALLEJO, S. BARLAS, P. JUSLIN, E. SHAFIR, D. BUDESCU, R. HOGARTH, C. HSEE, J. E. RUSSO, W. FERRELL, AND T. CONNOLLY "Overconfidence: It Depends on How, What, and Whom You Ask." In *Organizational Behavior and Human Decision Processes* 79.3 (1999): 216–247.
- KORIAT, A., S. LICHTENSTEIN, AND B. FISCHHOFF "Reasons for Confidence." In *Journal of Experimental Psychology: Human Learning and Memory* 6.2 (1980): 107–118.
- LANGNICKEL, F., AND S. ZEISBERGER "Do We Measure Overconfidence? A Closer Look at the Interval Production Task." *Journal of Economic Behavior and Organization* 128 (2016): 121–133.
- LICHTENSTEIN, S., AND B. FISCHHOFF "Do Those Who Know More Also Know More about How Much They Know?" *Organizational Behavior and Human Performance* 20 (1977): 159–183.
- LØHRE, E., AND K. H. TEIGEN "Probabilities Associated with Precise and Vague Forecasts." *Journal of Behavioral Decision Making* 30 (2017): 1014–1026.

- LUNDEBERG, M. A., P. W. FOX, AND J. PUNČOHAŘ "Highly Confident, but Wrong: Gender Differences and Similarities in Confidence Judgments." *Journal of Educational Psychology* 86 (1994): 1–20.
- MCBRIDE, M. F., F. FIDLER, AND M. A. BURGMAN "Evaluating the Accuracy and Calibration of Expert Predictions under Uncertainty: Predicting the Outcomes of Ecological Research." *Diversity and Distributions* 18 (2012): 782–794.
- MCKENZIE, C. R. M., M. J. LIERSCH, AND I. YANIV "Overconfidence in Interval Estimates: What Does Expertise Buy You?" *Organizational Behavior and Human Decision Processes* 107 (2008): 179–191.
- MOORE, D. A., A. B. CARTER, AND H. H. J. YANG "Wide of the Mark: Evidence on the Underlying Causes of Overprecision in Judgment." *Organizational Behavior and Human Decision Processes* 131 (2015): 110–120.
- MOORE, D. A., AND P. J. HEALY "The Trouble With Overconfidence." *Psychological Review* 115 (2008): 502–517.
- MOORE, D. A., AND D. SCHATZ "The Three Faces of Overconfidence." *Social and Personality Psychology Compass* 11.8 (2017): e12331.
- MOORE, D. A., S. A. SWIFT, A. MINSTER, B. MELLERS, L. UNGAR, P. TETLOCK, H. H. J. YANG, AND E. R. TENNEY "Confidence Calibration in a Multiyear Geopolitical Forecasting Competition." *Management Science* 63 (2017): 3552–3565.
- MOORE, D. A., E. R. TENNEY, AND U. HARAN "Overprecision in Judgment." In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making* (1st ed.) John Wiley & Sons, 2015.

- O'CONNOR, M., W. REMUS, AND K. GRIGGS "The Asymmetry of Judgemental Confidence Intervals in Time Series Forecasting." *International Journal of Forecasting* 17 (2001): 623–633.
- PALAN, S., AND C. SCHITTER "Prolific.Ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17 (2018): 22–27.
- PEER, E., L. BRANDIMARTE, S. SAMAT, AND A. ACQUISTI "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70 (2017): 153–163.
- PLOUS, S. "A Comparison of Strategies for Reducing Interval Overconfidence in Group Judgments." *Journal of Applied Psychology* 80 (1995): 443–454.
- REMUS, W., M. O'CONNOR, AND K. GRIGGS "Does Reliable Information Improve the Accuracy of Judgmental Forecasts?" *International Journal of Forecasting* 11.2 (1995): 285–293.
- REN, Y., AND R. CROSON "Overconfidence in Newsvendor Orders: An Experimental Study." *Management Science* 59 (2013): 2502–2517.
- RUSSO, J. E., AND P. H. SCHOEMAKER "Managing Overconfidence." *Sloan Management Review* 33 (1992): 7–17.
- SOLL, J. B., AND J. KLAYMAN "Overconfidence in Interval Estimates." In *Journal of Experimental Psychology: Learning Memory and Cognition* 30.2 (2004): 299–314.
- SPEIRS-BRIDGE, A., F. FIDLER, M. MCBRIDE, L. FLANDER, G. CUMMING, AND M. BURGMAN "Reducing Overconfidence in the Interval Judgments of Experts." *Risk Analysis* 30 (2010): 512–523.
- TEIGEN, K. H., AND M. JØRGENSEN "When 90% Confidence Intervals Are 50% Certain: On the Credibility of Credible Intervals." *Applied Cognitive Psychology* 19 (2005): 455–475.

YANIV, I., AND D. P. FOSTER "Graininess of Judgment under Uncertainty: An Accuracy-Informativeness Trade-Off." *Journal of Experimental Psychology: General* 124 (1995): 424–432.

YANIV, I., AND D. P. FOSTER "Precision and Accuracy of Judgmental Estimation." In *Journal of Behavioral Decision Making* 10.1 (1997): 21–32.

Appendix

Appendix 1: Treatment Conditions and Variable Definitions

Treatment conditions:

I use a 2x3 between-subjects experimental design. The treatment arms and the number of participant responses (all attention checks correct) are shown below:

	No helper	Low cost	High cost
High precision	29	27	32
Low precision	32	31	29

In the first part of the experiment, participants had to provide a range that covers the number of dots on a picture. They saw 10 pictures and had to guess at least 9 ranges correctly.

In the second part of the experiment, participants had to judge 25 ranges according to their usefulness in that the given ranges balance accuracy and precision well (see below for a definition of accuracy and precision).

Follow this link to the experiment: https://trr266.wiwi.hu-berlin.de/shiny/finding_range_estimates/

Variables:

Randomised: Treatment conditions

High cost condition (helper condition)	Participants in this condition had a “little helper”, providing them with suggested ranges estimates that cover the correct value in exactly 9 out of 10 cases. In half the cases, participants could “recalibrate” their helper so it would suggest a smaller range. Recalibration takes <u>30 seconds</u> .
Low cost condition (helper condition)	Participants in this condition had a “little helper”, providing them with suggested ranges estimates that cover the correct value in exactly 9 out of 10 cases. In half the cases, participants could “recalibrate” their helper so it would suggest a smaller range. Recalibration takes <u>10 seconds</u> .
No helper condition	Participants in this condition had no “little helper”, i.e. no information about suggested range estimates.
High precision condition	Participants in this condition had the following remark in their experimental material: “In the end, selecting a range is entirely up to you: If you feel like a range is too small, then it may make sense to choose a smaller range.”
Low precision condition	Participants in this condition had no additional remark in their experimental design.

Non-randomised: Dependent variables

Accuracy	A binary variable indicating whether the range provided by participants covered the true value (1) or not (0).
Precision	The width of a participant's range: (<i>upper end of range</i> – <i>lower end of range</i>).
Range evaluation	A participant's evaluation of the usefulness of a given range (second half of the experiment). Participants were supposed to assess the statement that the given range is useful in that it balances accuracy and precision well, on a scale of 1 (strongly disagree) to 7 (strongly agree).

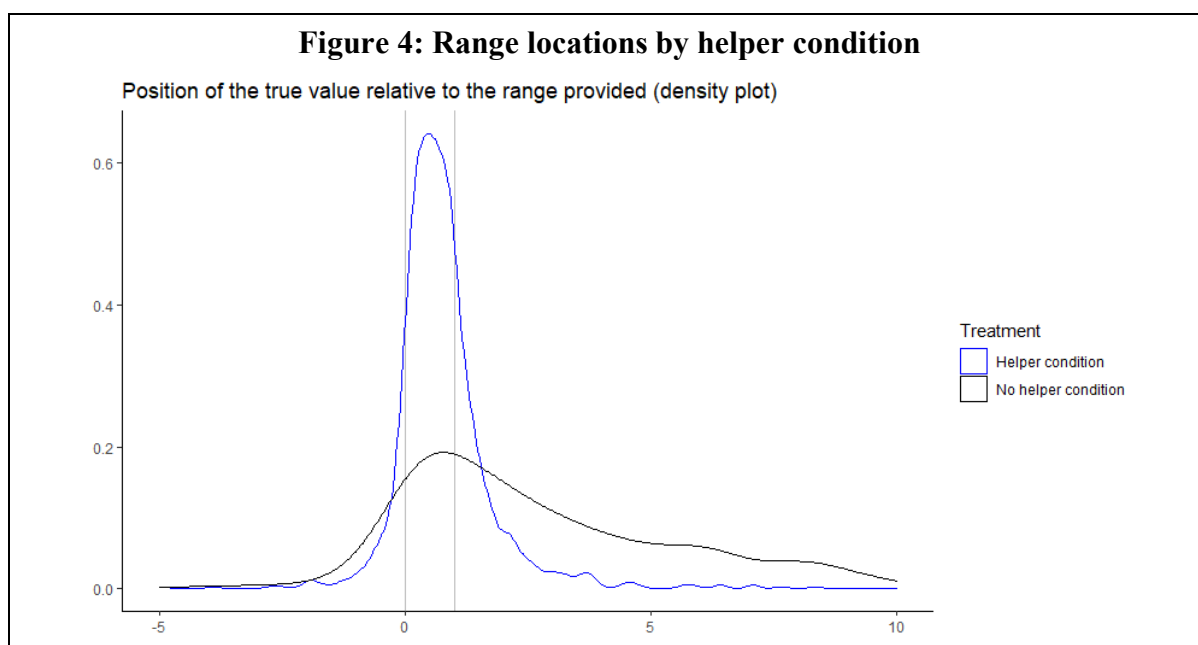
Non-randomised: Independent variables

Age	The age of the participant in years.
Attention check 1 failed	Dummy variable equal to 1 if a participant failed the first attention check (failed to state that the first half of the experiment was about accuracy alone).
Attention check 2 failed	Dummy variable equal to 1 if a participant failed the second attention check (failed to state that the second half of the experiment was about the balance of accuracy and precision).
Confidence group	Ex-post self-assessment of each participant estimating how many of their ranges provided in the first half of the experiment they believe to cover the true number of dots. Groups are: 0-2, 3-5, 6-8, 9-10.
Duration	Time the participant spent on the experiment in minutes.
Female	Dummy variable equal to 1 if the participant identifies as female.
Helper expectation	Ex-post assessment of the participant how much they agree to the statement that it was expected from them to use the range that the helper provided from 1 (strongly disagree) to 7 (strongly agree). (Only in the helper conditions.)
Math	Ex-post assessment of the participant how much they like math on a scale from 1 (strongly dislike) to 7 (strongly like).
Recalibration expectation	Ex-post assessment of the participant how much they agree to the statement that it was expected from them to use the recalibration option on a scale from 1 (strongly disagree) to 7 (strongly agree). (Only available in the helper conditions.)
Recalibration used	Binary variable equal to 1 if the participant used the option to recalibrate their little helper. (Only available in the helper conditions.)
Recalibration used often	Binary variable equal to 1 if the participant used the option to recalibrate their little helper more often than the median participant (3 or more times out of 5). (Only available in the helper conditions.)
Range proposed	Width of the range that was proposed to the participant by the little helper or by the recalibrated little helper if participants chose to recalibrate. (Only available in the helper conditions.)
Student	Binary variable equal to 1 if the participant is a student.

Appendix 2: Range Locations

Table 9: Range locations by treatment

Treatment condition	Range location (%)			Log(precision)		n
	Too low	Fitting	Too high	Mean	Median	
High cost x high precision	25.39	61.44	13.17	5.24	5.30	319
High cost x low precision	27.24	63.10	9.66	5.07	5.06	290
Low cost x high precision	36.30	54.07	9.63	5.00	5.01	270
Low cost x low precision	24.27	64.72	11.00	5.15	5.19	309
No helper x high precision	72.66	18.34	9.00	4.75	4.61	289
No helper x low precision	67.40	22.26	10.34	5.01	4.61	319

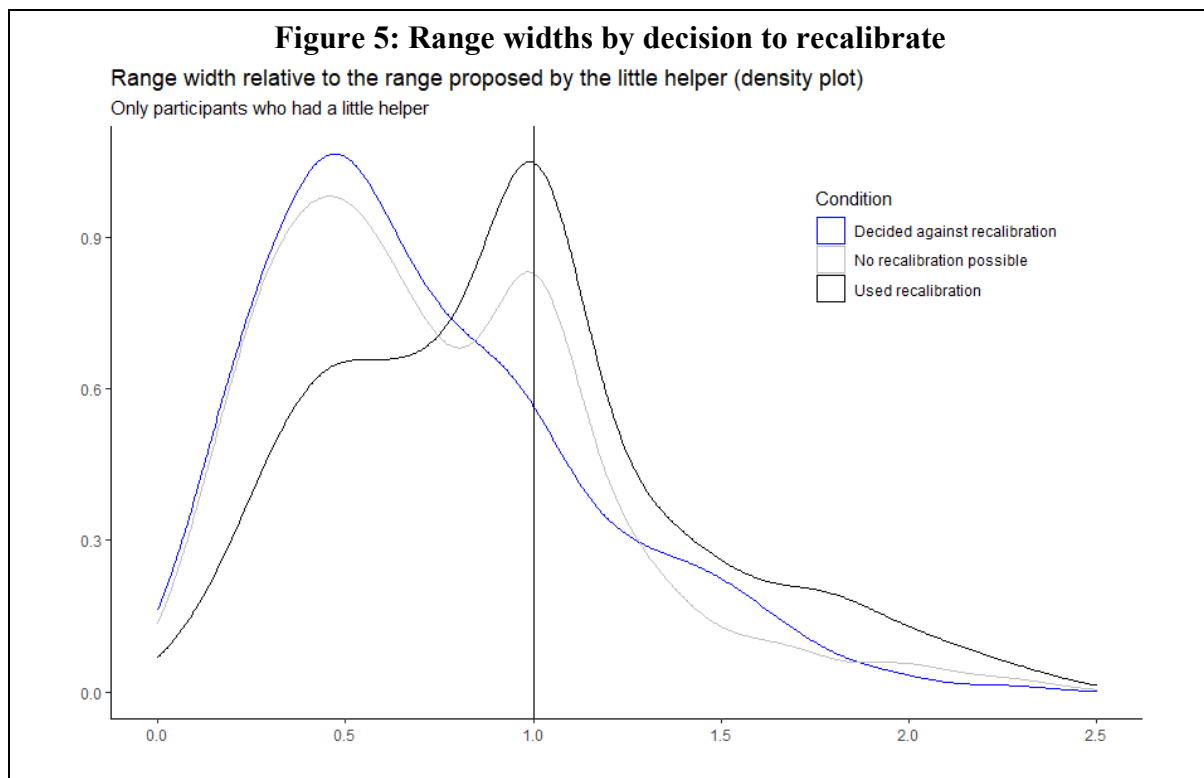


Notes: Table 9 and figure 4 give information on the position of ranges provided by participants. Table 9 shows whether the respective ranges are located below or above the correct value, or whether they cover it (“fitting”). It also shows the log of *precision*, i.e., the range width as provided by participants. Descriptions of treatment conditions and variable definitions can be found in appendix 1. Figure 4 shows a density plot of the position of the true value relative to the range provided by participants, split by whether the participants had a little helper or not. Values on the x axis between 0 and 1 indicate that the range covered the true value. Values below 0 indicate that the true value was lower than the range provided (i.e., lower than the lower end of the range) and values above 1 indicate that the true value was larger than the range provided (i.e., larger than the upper end of the range).

Appendix 3: Recalibration Decisions

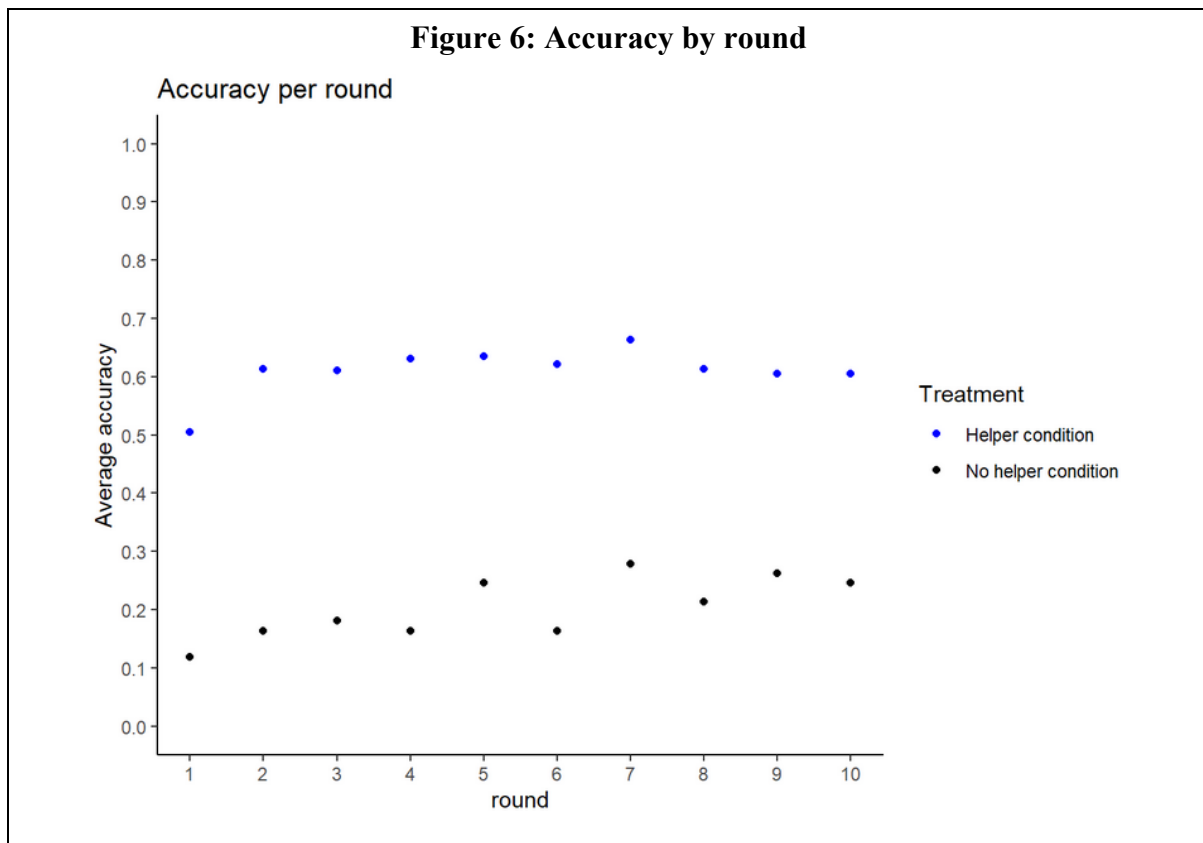
Table 10: Recalibration usage

<i>Number of times recalibration was used</i>	<i>Absolute</i>		<i>Percentages</i>	
	<i>Low cost con- dition</i>	<i>High cost condition</i>	<i>Low cost con- dition</i>	<i>High cost condition</i>
0	5	7	8.62	11.48
1	7	17	12.07	27.87
2	15	18	25.86	29.51
3	9	5	15.52	8.20
4	6	3	10.34	4.92
5	16	11	27.59	18.03



Notes: Table 10 and figure 5 give information on recalibration decisions. All participants who had a little helper had the option to recalibrate it in 50% of all cases in order to receive smaller range suggestions. See appendix 1 for treatment conditions and variable definitions. Table 10 shows the absolute numbers and percentages of participants' decisions to recalibrate, split by the high and low cost conditions ($n = 119$, all participants who had a helper). Figure 5 shows the range widths relative to the proposed range width of the helper, depending on whether participants decided to use recalibration or not ($n = 1188$, all ranges provided by participants who had a helper).

Appendix 4: Accuracy by Round



Notes: Figure 6 shows the average accuracy per round with accuracy being defined as a binary variable equal to 1 if the range provided by participants covered the true value and 0 if it did not. Each participant provided 10 range estimates. The blue dots represent the average accuracy in the helper treatment condition and the black dots in the no helper condition. See appendix 1 for treatment conditions and variable definitions. The number of observations is 1,796, representing all ranges provided by participants after data cleaning (see table 1).

Appendix 5: Last Digits

Table 11 : Last digits

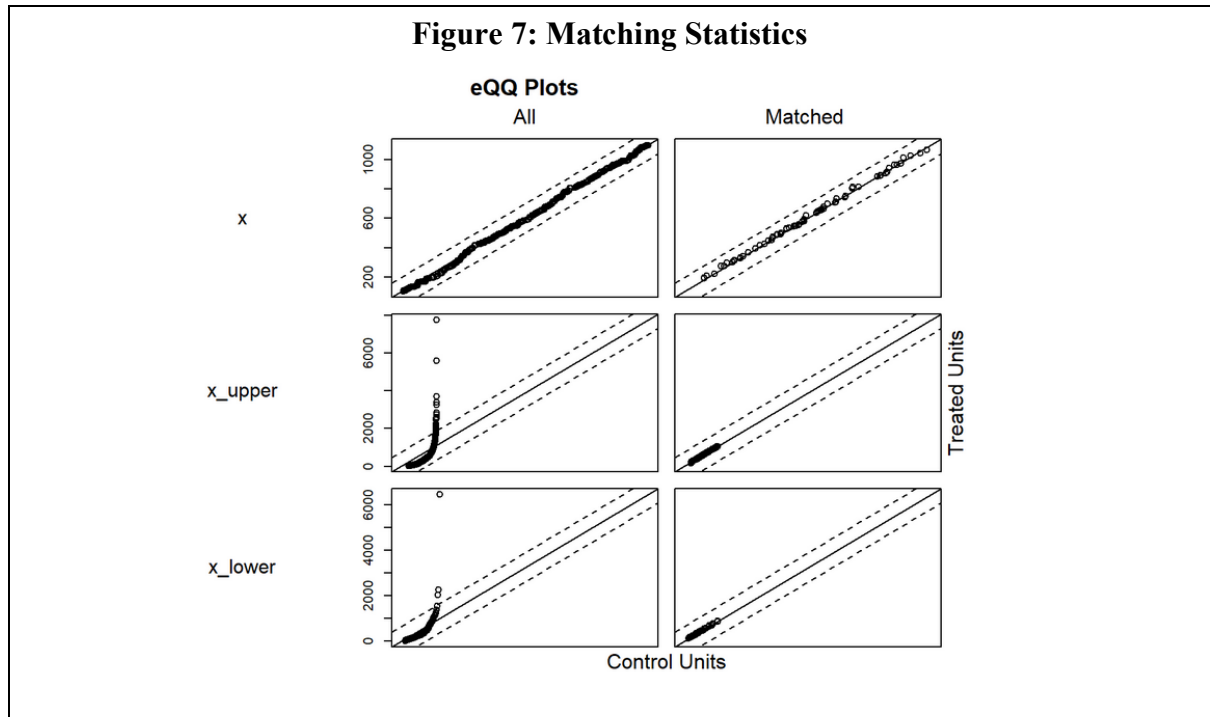
<i>Ends in...</i>	Lower range end		Upper range end		Range width	
	<i>No helper</i>	<i>Helper</i>	<i>No helper</i>	<i>Helper</i>	<i>No helper</i>	<i>Helper</i>
<i>5 or 10</i>	0.984	0.894	0.984	0.897	0.975	0.865
<i>10</i>	0.965	0.852	0.964	0.857	0.946	0.808
<i>50 or 100</i>	0.757	0.662	0.809	0.683	0.720	0.583
<i>100</i>	0.546	0.449	0.607	0.484	0.484	0.380

Notes: Table 11 shows whether the lower end of the range, the upper end of the range, or the width of the range (i.e., upper end minus lower end of the range) end in a number divisible by 5, 10, 50, or 100. The columns are split by whether participants are in the helper or no helper conditions. Treatment conditions and variable definitions can be found in appendix 1. The number of observations is 1,796, representing all ranges provided by participants after data cleaning (see table 1).

Appendix 6: Coarsened Exact Matching

Table 12: Matching statistics

Panel A: Summary of balance for all data							
	<i>Means treated</i>	<i>Means control</i>	<i>Std. mean diff.</i>	<i>Var. ratio</i>	<i>eCDF mean</i>	<i>eCDF max</i>	
X_{upper}	635.9671	743.8731	-0.1329	12.5081	0.2494	0.4395	
X_{lower}	346.1842	448.2378	-0.2146	4.5497	0.1663	0.3177	
X	588.0822	593.5785	-0.0180	1.0409	0.0097	0.0312	
<i>Value within range</i>	0.1941	0.5000	-0.7735		0.3059	0.3059	
Panel B: Summary of balance for matched data							
	<i>Means treated</i>	<i>Means control</i>	<i>Std. mean diff.</i>	<i>Var. ratio</i>	<i>eCDF mean</i>	<i>eCDF max</i>	<i>Std. Pair Dist.</i>
X_{upper}	597.3857	596.7714	0.0008	1.0368	0.0080	0.0571	0.0215
X_{lower}	360.0286	356.3571	0.0077	0.9977	0.0117	0.0857	0.0307
X	609.3286	603.8143	0.0180	0.9745	0.0085	0.0429	0.0524
<i>Value within range</i>	0.3714	0.3714	0.0000		0.0000	0.0000	0.0000
Panel C: Percent balance improvement							
	<i>Std. mean diff.</i>	<i>Var. ratio</i>	<i>eCDF mean</i>	<i>eCDF max</i>			
X_{upper}	99.4	98.6	96.8	87.0			
X_{lower}	96.4	99.8	92.9	73.0			
X	-0.3	35.6	11.6	-37.3			
<i>Value within range</i>	100.0		100.0	100.0			
Panel D: Sample sizes							
	<i>Control</i>	<i>Treated</i>					
<i>All</i>	3592	304					
<i>Matched</i>	70	70					
<i>Unmatched</i>	3522	234					
<i>Discarded</i>	0	0					



Notes: Table 12 and figure 7 show matching statistics for the coarsened exact matching used in table 5, columns 1 and 4 (statistics on the other columns are available upon request). Matching is done via coarsened exact matching with cut-points being determined by the Scott algorithm. Matching variables are the upper end of the range, the lower end of the range, the true value, and whether the range covers the true value. Matching on whether the range covers the true value needs to be exact, since coarsened exact matching may lead to (rare) cases where one range

just covers the true value, while the other just does not, and the question of whether the range covers the true value has a sizeable impact on participant judgments (see, for instance, table 4). The matching quality appears to be high with eCDF (empirical Cumulative Distribution Function) and standard mean differences being close to zero. 70 out of 304 observations can be matched. Matching emphasizes finding close matches at the expense of power.

Appendix 7: Experimental Material

Link to the experiment: https://trr266.wiwi.hu-berlin.de/shiny/finding_range_estimates/

Ethical Approval by the acting ethics committee of the School of Business and Economics of Humboldt-Universität zu Berlin was granted on September 29, 2021.

Introduction and task:

Finding range estimates

Let's get to the topic: A range estimate is an interval that covers an unknown value with a certain probability. People estimate ranges regularly and thus, a real-life example is quickly found. Consider the following scene from a supermarket:

<< I only have cash on me - how much, do you think, do we already have in our shopping cart? >>

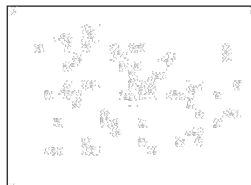
<< I would say between 60 and 70 dollars' worth. >>

<< Are you sure? >>

<< Well, 90% sure. >>

Examples like this are commonplace. Now, the goal of this experiment is to find out more about how people come up with these range estimates.

For this, I developed a game for you to play. You will get to see 10 different pictures with dots on them. Just like this one, except larger:



Each picture will disappear after 10 seconds, so there is most likely not enough time to count the dots. Your task is to estimate a range that covers the correct number of dots on each picture. The aim is to be correct in (at least) 9 out of the 10 cases. In other words: In at least 9 out of the 10 cases, your range should indeed cover the number of dots displayed. In the example above, the number of dots is 50, so any range covering that value would be correct.

It may be difficult to quickly assess the number of dots on each picture. This is why I provide you with some help. In each round, you will have a "little helper." The little helper will give you a suggestion for a range estimate. It is programmed so that it is correct in 9 out of 10 cases. Feel free to use this as guidance in case you struggle to find a range you feel comfortable with.

In some instances, it will be possible to re-calibrate the "little helper" before it shows a range. Re-calibration can be helpful because it usually results in a smaller range that is still correct in 9 out of 10 cases. However, this re-calibration will take 10 seconds to complete. Feel free to use it whenever you like.

In the end, selecting a range is entirely up to you: if you feel like a range is too large, then it may make sense to choose a smaller range.

If you are ready, simply press the next-button below. The next-button on the following pages will always appear after you gave a range estimate. You cannot navigate backwards. Dropping out of the experiment is possible at any point in time, but will result in not being compensated.

At the very end of the experiment, I will tell you whether you were successful in that your estimated range did cover the correct number of dots in at least 9 out of the 10 cases. Please note that your payment does not depend on your success. It is always 3 GBP.

Finally: Enjoy and thanks again for participating!

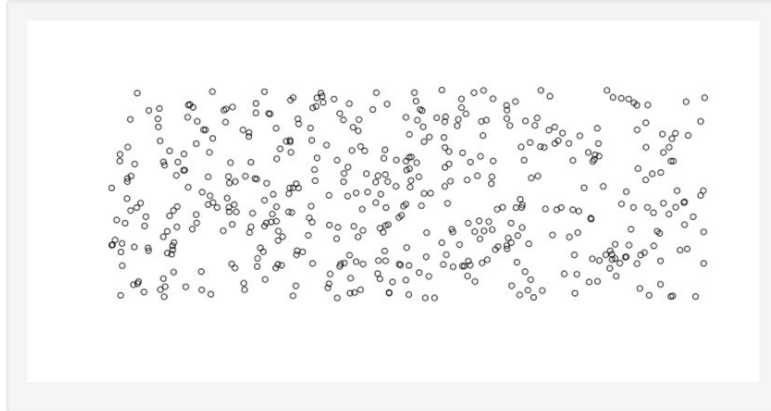
First part of the experiment:

Finding range estimates

After pressing the "Begin" button, an image with dots will appear for 10 seconds. It is your task to estimate a range that covers the number of dots. You will see 10 such images in this experiment. Make sure that the ranges you give cover the correct number of dots at least 9 out of these 10 times.

There are 10 images in total. This is number 4.

Begin



In case you have difficulties with coming up with a range, a "little helper" is there to support you. This time, it is possible to re-calibrate your "little helper," which usually results in a smaller range that it proposes to you. If you would like to re-calibrate your little helper, please press the respective button below. If you decide against re-calibration, it will still show you the larger range. Please note that the re-calibration procedure will take 10 seconds to complete. The little helper proposes a range that covers the true number of dots in 9 out of 10 cases.

Re-calibrate now

Do not re-calibrate

Your little helper shows that it estimates between 445 and 796 dots.

Please provide the lower end of your range estimate here.

0

Please provide the upper end of your range estimate here.

0

Note: The correct number of dots on the picture is 483.

Second part of the experiment:

Finding range estimates

When people come up with range estimates, there is always a trade-off between precision (i.e. giving small, informative ranges), and accuracy (i.e. giving ranges that cover the correct value). Large ranges may be less precise, but have a higher chance of covering the correct number: on average, they are more accurate. Smaller ranges may be more precise, but have a lower chance of covering the correct number: on average, they are less accurate.

The balance between precision and accuracy makes ranges more or less useful. For five pages altogether, I want you to assess the usefulness of different ranges. Each page shows you five different ranges that were estimated to cover the number of dots in a picture. You will see that they differ in precision and accuracy. The correct number of dots is given to you. You can also see it as a red "x" in the graphs below. Now, please imagine that five different people gave you the five different estimates below. Then, please state how much you agree with the following statement:

The given range estimate is useful in that it balances accuracy and precision well.

The correct number of dots was ... 662 - (see the red x in the graphs).

Estimate Number	Range (Estimated)	Agreement
1	199 to 629	Agree
2	1 to 1	Strongly disagree
3	582 to 1068	Agree
4	324 to 664	Somewhat agree
5	742 to 843	Agree

Tables

Table 1: Sample selection

Participants	Ranges provided	Ranges evaluated	
432	4,320	10,800	
-6	-60	-150	Lower limit is above upper limit more than once
-11	-110	-275	Upper limit is smaller than 51 more than once
	-8	-4	Lower limit is larger than upper limit
	-1		Upper limit is larger than 10,000 (typo)
	-1		Lower limit is smaller than 25 (typo)
		-59	No answer provided
415	4,140	10,312	
-200	-1,995	-4,950	Failed attention check 1
-35	-349	-461	Failed attention check 2
180	1,796	4,489	Final Sample

Notes: Table 1 shows the sample selection. Data was collected between October 12 and October 14, 2021 via Prolific.co. Participants were required to speak English fluently in order to participate. Data cleaning included removing participants that made implausible guesses (e.g. the lower end of the range that they provided was higher than the larger end of the range; note that I specifically did not restrict such selections in order to sort out inattentive participants) and removing participants because they answered the questions to the attention checks incorrectly. The attention checks asked participants what their task in the first and in the second half of the experiment was (see section 3.3 for more details). The final sample comprises 180 participants. *Participants* is the number of participants that joined and completed the experiment. *Ranges provided* is the number of range estimates that the participants made (10 each). *Ranges evaluated* is the number of ranges that participants judged in part 2 of the experiment (25 each).

Table 2: Descriptive Statistics

Panel A: Summary Statistics								
	n	mean	sd	min	P25	median	P75	max
<i>accuracy</i>	180	0.47	0.32	0	0.2	0.5	0.7	1
<i>confidence</i>	171	2.78	0.73	1	2	3	3	4
<i>duration (in minutes)</i>	180	17.77	7.01	7.2	12.8	15.8	21.2	42.3
<i>female</i>	180	0.47						
<i>student</i>	170	0.58						
<i>age</i>	179	26.72	7.48	18	22	25	30	59
<i>math</i>	172	4.80	1.55	1	4	5	6	7
<i>expectation to recalibrate</i>	117	4.37	1.79	1	3	5	6	7
<i>recalibration used</i>	119	0.51	0.33	0	0.2	0.4	0.8	1
<i>expectation to use the little helper</i>	118	5.47	1.34	1	5	6	6	7
<i>range proposed</i>	1188	272.7	117.9	99	165	262	371	499
<i>precision (range estimated) with helper</i>	1188	209.2	142.6	10	100	180	300	1000
<i>precision (range estimated) w/o helper</i>	608	293.9	602.8	5	50	100	250	9000

Panel B: Correlation Table

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)
<i>(a) accuracy</i>		.15	-.12	.13	.00	.10	-.01	.31	.16	.08	.22
<i>(b) precision (range estimated)</i>	.38		.06	-.04	.00	.10	.08	.07	.06	-.06	.09
<i>(c) range proposed</i>	-.12	.16		-.37	.03	-.03	.01	-.07	-.08	-.09	-.06
<i>(d) recalibration</i>	.13	-.06	-.38		.05	.03	-.04	.05	.17	.04	.08
<i>(e) female</i>	.00	.05	.03	.05		.05	-.13	-.10	-.03	-.14	.03
<i>(f) age</i>	.10	.16	-.03	.04	.05		-.06	.07	-.04	.00	.05
<i>(g) math</i>	-.02	.06	.02	-.05	-.13	-.07		.14	-.12	-.03	.07
<i>(h) confidence</i>	.32	.30	-.08	.06	-.08	.06	.14		-.02	.08	.20
<i>(i) expectation to recalibrate</i>	.16	.06	-.08	.17	-.04	-.02	-.13	-.03		.40	.14
<i>(j) expectation to use the little helper</i>	.04	-.06	-.08	.03	-.13	-.05	.03	.06	.37		.20
<i>(k) duration (in minutes)</i>	.24	.20	-.07	.08	.07	.03	.07	.22	.09	.13	

Notes: Table 2 presents descriptive statistics based on 180 participants of the clean sample. The number of observations can be lower than 180 because participants had the option not to answer demographic questions. **Panel A** shows summary statistics. *Accuracy* is defined as a binary variable equal to 1 if the range covered the true value and 0 if the range did not cover the true value. In this table, accuracy is aggregated on the participant level since the aim for participants was to meet 90 % correct guesses. *Precision* is the range width provided by participants, split by treatment conditions with and without the helper. Every participant provided 10 such ranges, resulting in the number of observations as shown (less removed cases as explained in table 1). The remaining variable descriptions can be found in appendix 1. **Panel B** shows the correlation matrix for the variables presented in panel A. It presents Pearson correlation coefficients above and Spearman correlation coefficients below the diagonal.

Table 3: Two-sided t-tests on accuracy and precision

Panel A: Accuracy				
<i>Treatment</i>	n	Control Mean	Treatment Mean	p-value (two-sided t-test)
<i>High precision condition</i>	1,796	0.49	0.45	0.06
<i>Helper condition</i>	1,796	0.20	0.61	< 0.01
<i>High cost condition</i>	1,188	0.60	0.62	0.38
Panel B: Log(precision)				
<i>Treatment</i>	n	Control Mean	Treatment Mean	p-value (two-sided t-test)
<i>High precision condition</i>	1,796	5.07	5.00	0.06
<i>Helper condition</i>	1,796	4.88	5.12	< 0.01
<i>High cost condition</i>	1,188	5.08	5.16	0.06
Panel C: Recalibration and costs				
<i>Helper: Log(precision)</i>		Recalibration available	Recalibration not available	p-value (two-sided t-test)
	High cost	5.43	5.62	< 0.01
	Low cost	5.33	5.65	< 0.01
	p-value (two-sided t-test)	0.01	0.32	
<i>Participants: Log(precision)</i>		Recalibration available	Recalibration not available	p-value (two-sided t-test)
	High cost	5.10	5.21	0.06
	Low cost	5.02	5.14	0.04
	p-value (two-sided t-test)	0.15	0.20	
<i>Participants: Accuracy</i>		Recalibration available	Recalibration not available	p-value (two-sided t-test)
	High cost	0.63	0.61	0.59
	Low cost	0.61	0.58	0.53
	p-value (two-sided t-test)	0.57	0.50	

Notes: Table 3 presents the results of two-sided t-tests for control and treatment conditions across all treatment arms. **Panel A** and **Panel B** test the influence of the *high precision condition*, the *helper condition*, and the *high cost condition* on *accuracy* and *precision*, respectively. The first two conditions entail all 1,796 observations, while the *high vs. low cost* condition is only available for participants who see a *helper* ($n = 1,188$). Panel C shows how *accuracy* and *precision* differ within the *helper* condition, depending on the cost condition and whether

recalibration was possible. The first set shows the precision provided by the helper, while the last two sets show precision and accuracy by participants.

Table 4: Range Judgment

	<i>judgment</i>	<i>judgment</i>	<i>judgment</i>	<i>judgment</i>	<i>judgment</i>
(Intercept)	3.148*** (0.430)	3.886*** (0.162)	2.842*** (0.182)	2.892*** (0.201)	3.042*** (0.440)
<i>accuracy</i>	6.019*** (0.716)				5.689*** (0.711)
<i>log(precision)</i>	-0.141* (0.081)				-0.158* (0.081)
<i>accuracy x precision</i>	-0.544*** (0.135)				-0.500*** (0.134)
<i>high precision condition</i>		-0.069 (0.218)		-0.105 (0.187)	0.086 (0.149)
<i>helper condition</i>			1.383*** (0.246)	1.382*** (0.245)	0.200 (0.201)
<i>high cost condition</i>			0.285 (0.217)	0.291 (0.213)	0.180 (0.162)
<i>Clustered SE</i>	Part.	Part.	Part.	Part.	Part.
<i>Num.Obs.</i>	897	897	897	897	897
<i>R2</i>	0.477	0.000	0.107	0.108	0.481
<i>R2 Adj.</i>	0.475	-0.001	0.105	0.105	0.478
<i>AIC</i>	3416.9	3993.7	3894.1	3895.5	3414.8
<i>BIC</i>	3436.1	4003.3	3908.5	3914.7	3448.4
<i>Log.Lik.</i>	-1704.471	-1994.872	-1944.051	-1943.774	-1700.401

Notes: Table 4 presents OLS regression results for associations between estimate characteristics provided by participants and participants' evaluation of their own ranges estimates. Participant *judgments* are provided on a Likert scale from 1 to 7, assessing whether the given range balances accuracy and precision well. Participants were not aware that they were judging their own (transformed) ranges. See section 3.2 for a more detailed explanation of the procedure. *Accuracy* is a binary variable equal to 1 if the range covered the true value, and 0 if the range did not cover the true value. *Precision* is the width of the range. Standard errors are clustered on the participant level. Each participant evaluated 25 ranges, with 5 of those being provided by themselves. The maximum number of observations is thus $180 * 5 = 900$. Three observations needed to be removed during data cleaning (see table 1). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Inherent Preferences for Range Calibration

	<i>judgment</i>	<i>judgment</i>	<i>judgment</i>	<i>judgment</i>	<i>judgment</i>	<i>judgment</i>
<i>(Intercept)</i>	4.749 (2.941)	2.907* (1.560)	4.751** (1.522)	3.121*** (0.577)	2.410*** (0.373)	2.646*** (0.340)
<i>participant range</i>	0.392 (0.274)	0.381** (0.179)	0.423** (0.173)	0.403 (0.275)	0.377** (0.178)	0.429** (0.175)
<i>accuracy</i>	6.279 (3.795)	3.711 (2.530)	0.561 (2.667)	2.806*** (0.354)	3.114*** (0.215)	3.148*** (0.210)
<i>precision</i>	-0.436 (0.536)	-0.121 (0.283)	-0.438 (0.282)			
<i>accuracy x precision</i>	-0.619 (0.692)	-0.122 (0.463)	0.463 (0.487)			
<i>lower end of range</i>				0.005** (0.002)	0.002 (0.001)	0.001 (0.001)
<i>upper end of range</i>				-0.003** (0.002)	-0.002 (0.001)	-0.002 (0.001)
<i>true value</i>				-0.001 (0.001)	0.000 (0.001)	0.000 (0.001)
<i>Clustered SE</i>	subclass (match)	subclass (match)	subclass (match)	subclass (match)	subclass (match)	subclass (match)
<i>Num.Obs.</i>	140	272	314	140	272	314
<i>R2</i>	0.458	0.482	0.478	0.467	0.487	0.477
<i>R2 Adj.</i>	0.442	0.474	0.471	0.447	0.477	0.468
<i>AIC</i>	533.1	1023.6	1192.9	532.8	1023.0	1195.6
<i>BIC</i>	547.9	1041.6	1211.6	550.5	1044.6	1218.1
<i>Log.Lik.</i>	-261.572	-506.805	-591.433	-260.419	-505.489	-591.782

Notes: Table 5 presents OLS regression results for associations between participants' judgments of range estimates and the range estimates' origins, i.e. whether the participant unknowingly provided the (transformed) range or not. Regressions are based on matched data. Participants' own ranges were taken from the treatment condition without the little helper and matched with other participants' assessments of ranges that they did not originally provide. I use coarsened exact matching on the lower end of the range, the upper end of the range, and the correct value the range is supposed to cover (see control variables for the last three regressions). The algorithm furthermore searches for exact matches as to whether the range covers the true value. Therefore, each match contains two judgments on a range similar in position and width, with one range provided and unknowingly judged by one of the participants, and the other computer-generated range judged by a different participant. Information on the matching quality can be found in appendix 6. *Participant range* is a binary variable with 1 indicating that the participant provided the range and 0 indicating they did not. Participant judgments are provided on a Likert scale from 1 to 7, assessing whether the given range balances accuracy and precision well. Participants were not aware that they were judging their own (transformed) ranges. See section 4.2 for a more detailed explanation of the procedure. Standard errors are clustered on the match level. The first and fourth regressions include participants who answered both attention checks correctly, while the second and fifth regressions include participants who answered the second attention check correctly. The third and sixth regressions include participants irrespective of whether they answered the attention checks correctly. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Accuracy and Precision (Regression Results)

	<i>accuracy</i>				<i>log(precision)</i>			
<i>(Intercept)</i>	-0.022 (0.135)	-1.362*** (0.196)	-1.247*** (0.219)	-2.291*** (0.556)	5.08*** (0.08)	4.88*** (0.12)	4.92*** (0.14)	4.08*** (0.27)
<i>high precision condition</i>	-0.179 (0.188)		-0.251 (0.173)	-0.253 (0.179)	-0.08 (0.10)		-0.09 (0.10)	-0.10 (0.10)
<i>high cost condition</i>		0.104 (0.193)	0.119 (0.192)	0.140 (0.203)		0.08 (0.09)	0.08 (0.09)	0.07 (0.10)
<i>helper condition</i>		1.757*** (0.236)	1.761*** (0.234)	1.892*** (0.250)		0.20 (0.14)	0.20 (0.14)	0.22 (0.14)
<i>age</i>				0.030+ (0.015)				0.02** (0.01)
<i>math</i>				0.015 (0.056)				0.06* (0.03)
<i>female</i>				0.171 (0.190)				0.12 (0.11)
<i>Clustered SE</i>	Part.	Part.	Part.	Part.	Part.	Part.	Part.	Part.
<i>Num.Obs.</i>	1796	1796	1796	1706	1796	1796	1796	1706
<i>Pseudo R2</i>	0.001	0.113	0.116	0.137				
<i>R2</i>					0.002	0.02	0.02	0.06
<i>R2 Adj.</i>					0.001	0.02	0.02	0.06
<i>AIC</i>	2484.8	2209.0	2205.0	2050.9	4651.3	4625.1	4622.8	4351.9
<i>BIC</i>	2495.8	2225.5	2227.0	2089.0	4662.3	4641.6	4644.8	4390.0

Notes: Table 6 presents results for tests related to hypotheses 2 and 3. *Accuracy* is defined as a binary variable equal to 1 if the range covered the true value and 0 if the range did not cover the true value. *Precision* is defined as the width of the range that participants provided. *Accuracy*-related regressions are logit regressions with standard errors clustered at the participant-level. *Precision*-related regressions are OLS regressions with standard errors clustered at the participant level. The number of participants is 180 with 1,796 ranges being estimated. Models with full controls have less observations since participants had the option not to answer demographic survey questions. Standard errors are shown in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Accuracy and Precision, High and Low True Values

Panel A: Accuracy								
	accuracy							
	true value < 375				true value >= 375			
(Intercept)	0.208 (0.158)	-1.021*** (0.212)	-0.887*** (0.224)	-1.506** (0.598)	-0.099 (0.148)	-1.491*** (0.239)	-1.378*** (0.268)	-2.595*** (0.645)
High precision condition	-0.208 (0.240)		-0.279 (0.231)	-0.119 (0.235)	-0.177 (0.207)		-0.252 (0.196)	-0.317 (0.204)
high cost condition		0.305 (0.275)	0.331 (0.275)	0.412 (0.273)		0.031 (0.213)	0.043 (0.212)	0.033 (0.230)
helper condition		1.482*** (0.282)	1.479*** (0.278)	1.561*** (0.292)		1.863*** (0.277)	1.870*** (0.275)	2.045*** (0.296)
age				0.015 (0.015)				0.036** (0.018)
math				0.017 (0.072)				0.009 (0.062)
female				0.044 (0.239)				0.206 (0.219)
Clustered SE	Part.	Part.	Part.	Part.	Part.	Part.	Part.	Part.
Num.Obs.	468	468	468	440	1328	1328	1328	1266
R2 Pseudo	0.002	0.099	0.102	0.120	0.001	0.119	0.122	0.147
AIC	650.3	589.5	589.5	548.3	1831.2	1617.3	1614.9	1501.6
BIC	658.6	601.9	606.1	576.9	1841.6	1632.9	1635.7	1537.6
Log.Lik.	-323.149	-291.734	-290.761	-267.171	-913.606	-805.640	-803.445	-743.819

Panel B: Precision								
	log(precision)							
	true value < 375				true value >= 375			
(Intercept)	4.814*** (0.076)	4.331*** (0.128)	4.438*** (0.131)	3.981*** (0.315)	5.168*** (0.087)	5.070*** (0.134)	5.083*** (0.151)	4.091*** (0.286)
high precision condition	-0.202* (0.118)		-0.214** (0.107)	-0.239** (0.112)	-0.021 (0.114)		-0.026 (0.114)	-0.044 (0.113)
high cost condition		0.090 (0.105)	0.110 (0.105)	0.106 (0.110)		0.078 (0.103)	0.079 (0.103)	0.058 (0.105)
helper condition		0.515*** (0.152)	0.507*** (0.147)	0.512*** (0.153)		0.094 (0.150)	0.094 (0.150)	0.136 (0.148)
age				0.008 (0.007)				0.021*** (0.008)
math				0.040 (0.035)				0.073* (0.038)
female				0.093 (0.114)				0.147 (0.117)
Clustered SE	Part.	Part.	Part.	Part.	Part.	Part.	Part.	Part.
Num.Obs.	468	468	468	440	1328	1328	1328	1266
R2	0.016	0.109	0.126	0.147	0.000	0.006	0.007	0.059
R2 Adj.	0.014	0.105	0.121	0.135	-0.001	0.005	0.004	0.054
AIC	1122.2	1077.7	1070.4	1011.7	3440.1	3433.7	3435.4	3232.6
BIC	1130.5	1090.2	1087.0	1040.3	3450.5	3449.3	3456.2	3268.6
Log.Lik.	-559.114	-535.872	-531.199	-498.838	-1718.040	-1713.846	-1713.697	-1609.281

Notes: Table 7 presents results for tests related to hypotheses 2 and 3, split by the correct value being smaller or larger/equal to 375. *Accuracy* is defined as a binary variable equal to 1 if the range covered the true value and 0 if the range did not cover the true value. *Precision* is defined as the width of the range that participants provided. *Accuracy*-related regressions are logit regressions on the range level. See appendix 1 for all other variable definitions. Standard errors are clustered at the participant level. The total number of ranges to be estimated is 1,796, with 468 of those being provided for a true value smaller than 375 and 1328 ranges being provided for a true value

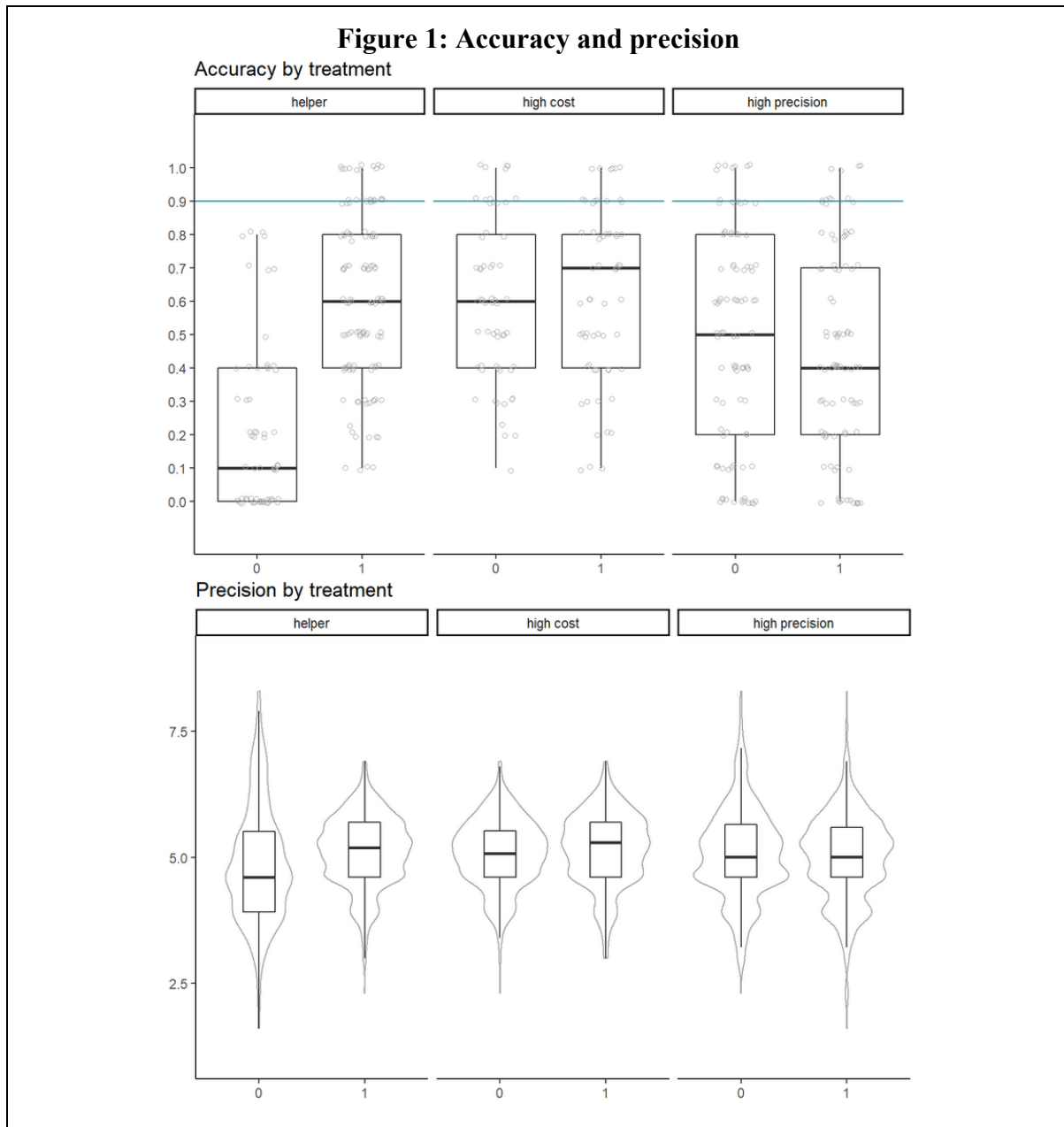
larger or equal to 375. Models with full controls have less observations since participants had the option not to answer demographic survey questions. Standard errors are shown in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Changes in precision and accuracy

	<i>true value < 375</i>		<i>true value ≥ 375</i>	
	Log(precision)	Accuracy	Log(precision)	Accuracy
<i>(Intercept)</i>	4.885*** (0.103)	0.651*** (0.051)	5.249*** (0.067)	0.557*** (0.042)
<i>helper</i>	0.772*** (0.113)	0.233*** (0.061)	0.400*** (0.072)	0.340*** (0.046)
<i>High cost</i>	0.087 (0.133)	0.011 (0.084)	0.055 (0.103)	0.036 (0.059)
<i>Smaller option</i>	-0.091 (0.095)	-0.089 (0.075)	-0.160*** (0.045)	0.067* (0.040)
<i>helper × high cost</i>	-0.092 (0.149)	0.033 (0.094)	-0.102 (0.109)	-0.019 (0.064)
<i>helper × smaller option</i>	-0.176 (0.123)	0.111 (0.091)	-0.181*** (0.061)	-0.057 (0.050)
<i>High cost × smaller option</i>	0.020 (0.124)	0.128 (0.112)	0.038 (0.064)	-0.054 (0.059)
<i>helper × high cost × smaller option</i>	-0.039 (0.168)	-0.186 (0.132)	0.150* (0.084)	0.007 (0.073)
<i>Clustered SE</i>	Part.	Part.	Part.	Part.
<i>Num.Obs.</i>	634	634	1742	1742
<i>R2</i>	0.267		0.090	
<i>R2 Adj.</i>	0.259		0.087	
<i>R2 Pseudo</i>		0.096		0.113
<i>AIC</i>	1060.4	641.7	3081.4	1827.3
<i>BIC</i>	1096.0	677.3	3125.1	1871.0
<i>Log.Lik.</i>	-522.199	-312.841	-1532.678	-905.664

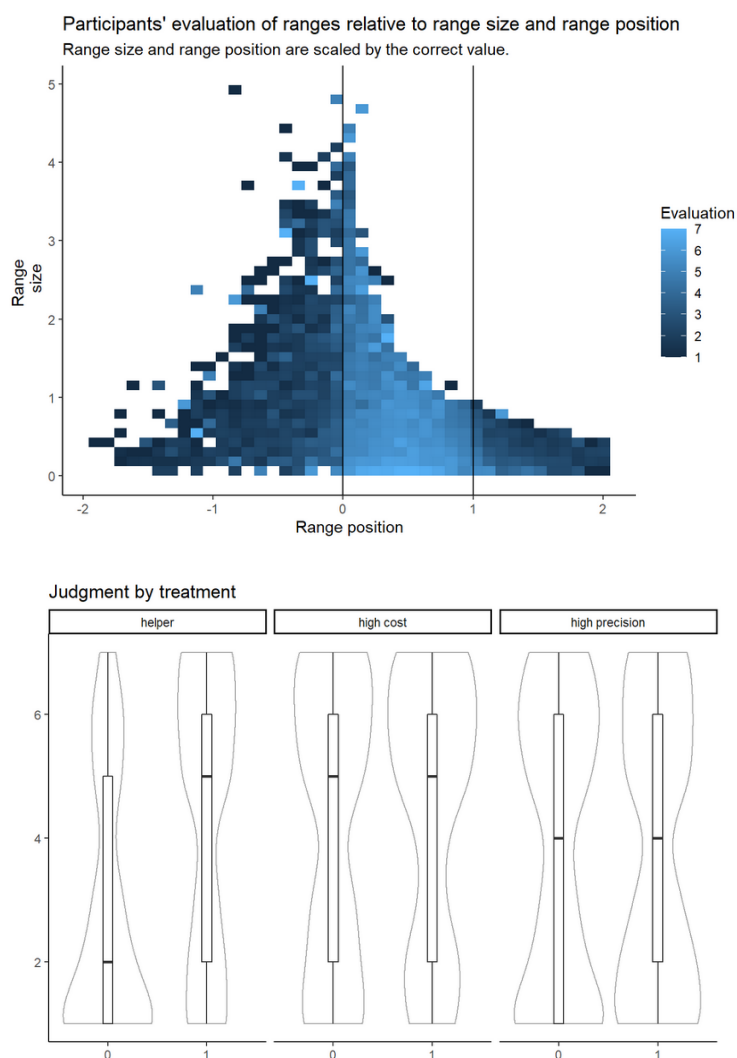
Notes: Table 8 presents results related to hypothesis 2. *Log(precision)* in the first and third regressions is the natural logarithm of the range width provided by the helper or the participant to the same underlying true values. The dependent variable in the second and fourth regressions is *accuracy*, defined as a binary variable equal to 1 if the range covers the true value. *Helper* in table 8 is a binary variable equal to one if the range was provided by the helper. *High cost* is a binary variable equal to one for estimates provided in the high cost condition. *Smaller option* is a binary variable equal to one for estimates provided when recalibration was possible (not when it was used). The number of observations is the sum of all participants in the helper conditions (1,188) plus the helper estimates for the same underlying values (also 1,188, in sum 2,376) – split by low true values ($x < 375$) and high true values ($x \geq 375$). * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figures

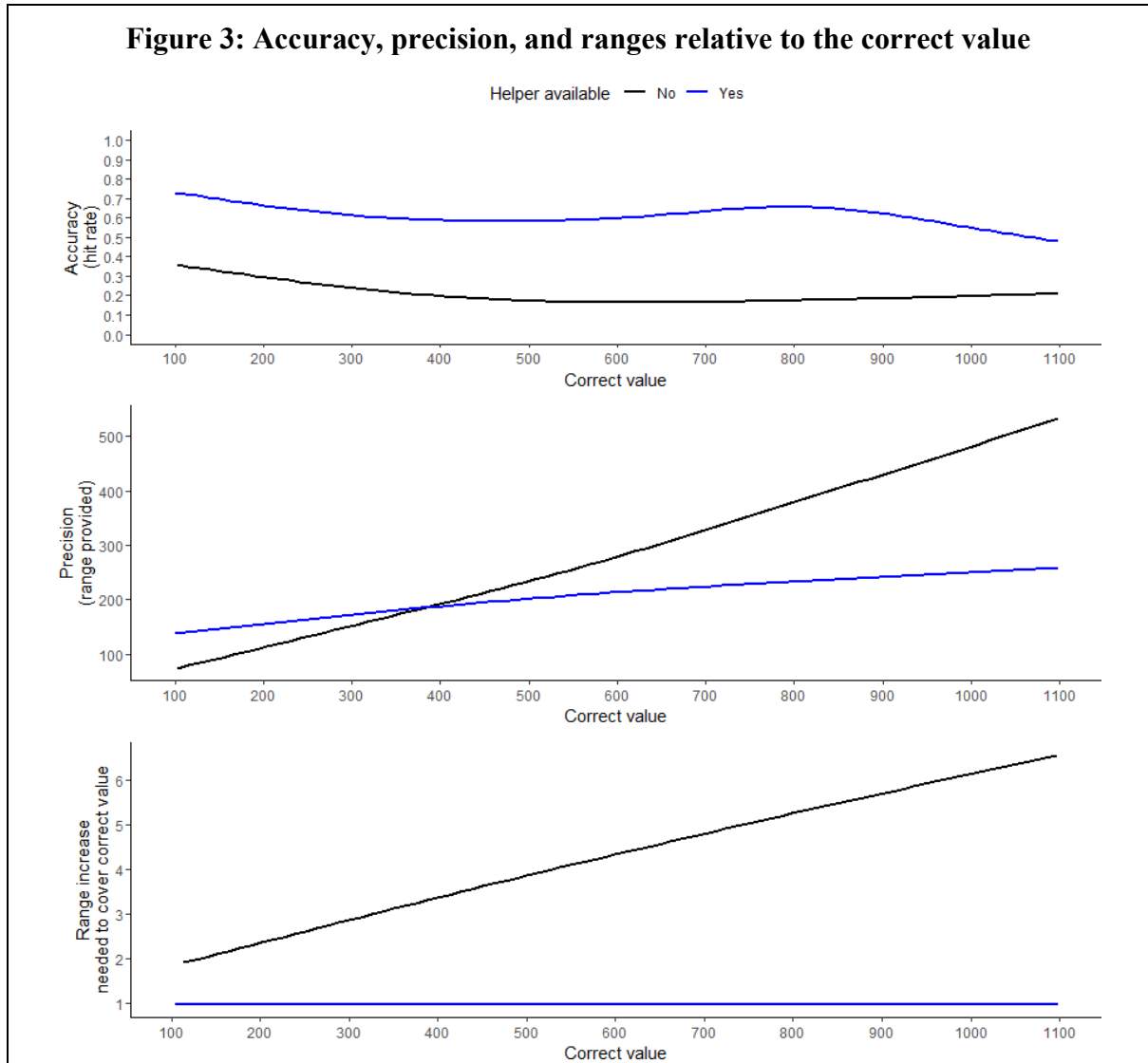


Notes: Figure 1 presents results for tests related to hypotheses 2 and 3. *Accuracy* is defined as the percentage of estimates covering the correct value. *Precision* is defined as the natural log of the width of the range that participants provided. Each graph contains information on the three treatment conditions: *helper* denotes whether participants had access to the little helper (1 indicates yes, 0 indicates no, number of participants = 180), *high cost* denotes whether participants were in the high cost condition (1 indicates yes, 0 indicates no, number of participants = 119), and *high precision* denotes whether participants were in the high precision condition (1 indicates yes, 0 indicates no, number of participants = 180). The turquoise line in the top graph indicates an accuracy of 90 %, the aim that each participant had.

Figure 2: Participant judgments



Notes: Figure 2 presents participants' evaluation of range width relative to range position (first graph). The range width on the y axis is scaled by the true value. The range position is determined by the following equation: $range\ position = (true\ value - lower\ end\ of\ range) / (upper\ end\ of\ range - lower\ end\ of\ range)$. Values on the x axis between 0 and 1 contain ranges that cover the true value (e.g. a true value of 100 with a range from 50 to 150 would result in a range position of $(100 - 50) / (150 - 50) = 0.5$). Values smaller than 0 indicate that the true value is smaller than the range provided (e.g., a true value of 100 with a range from 150 to 250 has a range position of -0.5), and values larger than 1 indicate that the true value is larger than the range provided (e.g., a true value of 100 with a range from 0 to 50 has a range position of 2). The colour shading indicates participants' judgments of the given estimates with brighter blues indicating better evaluations on a Likert scale from 1 to 7, assessing whether the given range balances accuracy and precision well. The second graph shows differences in range judgment depending on which treatment condition the participants were in: *helper* denotes whether participants had access to the little helper (1 indicates yes, 0 indicates no, number of participants = 180), *high cost* denotes whether participants were in the high cost condition (1 indicates yes, 0 indicates no, number of participants = 119), and *high precision* denotes whether participants were in the high precision condition (1 indicates yes, 0 indicates no, number of participants = 180).



Notes: Figure 3 presents first, the *accuracy* (defined as a binary variable equal to 1 if the range covered the true value and 0 if the range did not cover the true value) and second, the *precision* (the width of the range that participants provided) relative to the true value. The third graph presents only observations from participants whose range did not cover the true value. It shows the range increase that would be necessary in order to cover the true value, relative to the true value on the x axis. Each graph shows smoothed regression lines based on generalised additive models, with the black line for participants who did not have the helper and the blue line for participants who had the helper. The number of observations is 1,796 in the first and second graph, and 947 in the third graph.

Translation and Retail Investor Perception

Rico Chaskel

Humboldt University of Berlin

Tom Fischer

Ernst & Young

Abstract

We examine whether the translation of financial disclosures influences retail investors' perception of a firm as a potential investment. We assess three different textual characteristics through which the effect of translation could work: readability, tone, and precision. Our novel research design is based on a survey experiment with real firm disclosures provided by firms in German and English and employs genuine retail investors. We find that German disclosures are perceived easier to read than their English counterparts (both by English and German native speakers), but do not differ in the perception of tone and precision. Surprisingly, better readability does not result in higher attractiveness of the firm as an investment. Our evidence implies that this may be caused by participants finding it relatively difficult to assess readability and precision in the real company disclosures. Finally, we analyse how well retail investors' perceptions correlate with textual measures commonly employed in the accounting literature. We find reasonably high correlations for tone, but not for readability and precision.

1 Introduction

We analyze how the reporting language of firm disclosures influences retail investors' assessment of the attractiveness of a firm as an investment. We examine three different channels through which the reporting language may lead to a different assessment: through perceived readability, perceived tone, and perceived precision of the underlying text. Furthermore, we provide evidence on the ease with which people can assess the three textual characteristics. Finally, we examine how well retail investors' perceptions of textual characteristics match with commonly applied textual measures for readability, tone, and precision.¹

Language barriers can be powerful as they appear to deter investments in firms, even for institutional investors (Cuypers et al. [2015]; Lundholm et al. [2018]). As a response, firms with a high degree of internationalisation, a larger need for external financing, and larger language barriers between their native language and English more often translate their financial documents to English (Jeanjean et al. [2010]). Firms that adopt English as a (second) reporting language then benefit from lower information asymmetry and higher foreign ownership (Jeanjean et al. [2015]). As such, translations play an important role in capital markets.

Research on translated financial documents however is scarce, yet important. After all, it informs us whether the market is a “level playing field” for all investors – ultimately, only if translations do not significantly alter the (perceived) mix of information could international investors come to similar conclusions.² Evidence on the interplay between the translation of accounting narratives and perception of textual characteristics by retail investors informs us

¹ We mainly focus on the Flesch Reading Ease score for readability (Flesch [1948]) and the word lists developed by Loughran and McDonald [2011]. We also use alternative measures for readability which lead to qualitatively similar results.

² Following Nida (1964), we understand high translation quality (also: adequate translations) as evoking the same reaction among readers of the source text and the target text.

about potential market consequences as well as it helps future (experimental) researchers to decide whether translated disclosures are useful for assessing textual characteristics of firm disclosures.

We apply a novel research method in order to determine whether retail investors perceive textual characteristics of firms' original disclosures differently to their translated counterparts and whether this results in differences in the attractiveness of the issuing firm as an investment. We design a survey experiment using a random sample of forecast reports extracted from German firms' 2019 half-year reports.³ Key to our analysis is that these reports are issued in two languages (i.e., German and English) and do not differ in information content.⁴ We acquire survey participants who have investment experience and are native in either German or English, and fluent in the other language. Each participant reads six different, randomly assigned forecast reports out of a sample of 61 firms in either German or English. We ensure that no participant receives the same report in both languages. All participants answer ten different questions per report, starting with the overall perceived attractiveness of the firm's shares as an investment. The remaining questions capture participants' perceptions of readability-, tone-, and precision-related textual characteristics. Since each report (in either language) is read by a large number of survey participants and each participant reads six different reports, we can use participant- and firm-fixed effects. We are thus able to draw conclusions independent of underlying firm characteristics (e.g. profitability or visibility) and participant attributes (e.g. the belief that firms report overly positive). In summary, the design

³ Forecast reports have the advantage that they are reasonably short (i.e. "digestible" for survey participants), but they also likely contain new and thus potentially price-influencing information at the time of publication. Furthermore, as opposed to many other sections of firms' half-year or annual reports, forecast reports are typically non-standardized and thus offer sufficient variation for our analysis.

⁴ We manually compared the German version to its English translation for each report included in the sample and did not detect any differences in content.

allows us to investigate whether retail investors perceive the textual characteristics of real company disclosures differently depending on their disclosure language.

Focusing on German firms has two main advantages: first, firms listed in the Prime Standard segment of Frankfurt Stock Exchange are obliged to prepare their annual and half-year reports in German and English.⁵ Second, our analyses on textual characteristics require us to employ real retail investors who are fluent in two languages (see below for more details). Germany offers a setting where both the legal requirements, as well as the expected sample size seemed promising enough. We focus on retail investors since the importance of properly translated accounting disclosures has risen for retail investors as barriers to invest in foreign stocks have been reduced over the past decades. This implies a growing demand for firm information provided in the English language – the lingua franca of business (Blenkinsopp and Pajouh [2010]) – and underscores the importance of well-translated accounting disclosures. We leave the question whether our findings extend to professional investors to future research.

Our findings indicate that the German disclosures are perceived to be significantly more readable than their English counterparts – both by German and English native speakers. We do not find statistically significant differences between German and English disclosures for tone and precision. Interestingly, the better readability of German disclosures does not result in German firms being rated as more attractive as an investment. We find that this is likely caused by readability *not* having a significant effect on investment attractiveness in our setting. We believe this to be caused by participants finding it relatively difficult to judge readability and precision, leading to an overall null-result of the effect of readability on investment attractiveness: The survey responses to readability- and precision-related questions have 40-50 % higher standard deviations than responses to the tone-related question. This is an

⁵ Para. 51-52 Börsenordnung der Frankfurter Wertpapierbörse (Exchange Rules for the Frankfurter Wertpapierbörse).

interesting starting point for future research that could examine how insecurity about readability and precision attenuate previously documented effects of readability and precision on investment attractiveness.

Finally, we horserace our participants' responses with the Flesch Reading Ease score (proxy for readability, Flesch [1948]) as well as tone and precision measured via the wordlists developed by Loughran and McDonald [2011] in their 2014 version. Both lead to similar results: Tone is identified as being significantly positively associated with investment attractiveness, while readability and precision are insignificant. Still, further analyses reveal very low correlations between our participants' responses and the textual measures. Solely tone seems to capture a similar underlying construct. We thus encourage researchers to be cautious when applying textual measures (especially for short texts) and to cross-validate their findings using sufficiently different methods (e.g. Loughran and McDonald [2014], [2016]; Siano and Wysocki [2021]).

2 Research Questions and Hypotheses

2.1 Translation and Perception of Textual Characteristics

English has emerged as the unequivocal lingua franca for firm communication (Blenkinsopp and Pajouh [2010]; Jeanjean et al. [2010]), which is perhaps partially responsible for the low number of studies interested in translations of firm disclosures. Nevertheless, translations are important: Sonney [2009] finds that analysts who specialize in a country outperform those who specialize in a sector and names a better understanding of the (native) language as one possible reason. Jeanjean et al. [2010] show that larger language barriers between a firm's native language and English, as well as a relative lower importance of the native language are associated with a firm's decision to translate their annual report into

English. Jeanjean et al. [2015] show that adopting English as a reporting language leads to an increase in foreign ownership and Lundholm et al. [2018] provide evidence that language differences are at least partially responsible for an underweighting of foreign stocks by institutional investors, even if the foreign firm issues English annual reports.

As important as having translated disclosures may be, the translation of accounting disclosure reaches beyond the pure mechanical transition of information between languages (Evans [2018]). Even direct translations of firm publications may evoke different interpretations among readers. Pan et al. [2015] let Chinese readers make judgments on whether to consolidate financial statements of two firms. The information on the relationship between two companies is the same, but differs in whether the English or the Chinese conception of the word “control” was used. Despite the same underlying information, judgments were inconsistent depending on the language. Douppnik and Richter [2003] examine the effect of language culture and translation on the interpretation of verbal uncertainty expressions. Their results suggest that German and American certified accountants interpret the same set of uncertainty expressions differently. Moreover, the translation of extreme probability expressions from English to German results in significantly different interpretations⁶.

In conclusion, it is not just important to compare a translated text with its source document, but also to assess how translations are understood by their readers. Research on the topic is scarce and often inconclusive. Only few studies look at how actual firm disclosures are translated and understood. A first step to assess translation quality is to analyze its informational content. Campbell et al. [2005] compare German and (British) English

⁶ Some differences in perception of translated documents also lie in the eye of the reader: “Language translation is not a simple technical, but a socio-cultural, subjective and ideological process” (Evans, 2018, p. 1844). Linguistic relativism suggests that specific characteristics of our native language shape the manner how we interpret the world, information in general, and thus also accounting disclosures (Sapir [1985]; Whorf [1956]). Hence, information recipients from different origins might perceive and value the same set of accounting information differently (Zeff [2007]). This issue is beyond the scope of our paper. Nevertheless, we account for it in our research design explained in section 3.

environmental narratives issued by German firms and find that those do not seem to intentionally discriminate against foreign investors by disclosing less information or highlighting certain information differently in their English translations. Overall, the English translations can be assumed to be accurate.

Going beyond content, Courtis and Hassan [1973] examine the reading ease of Hong Kong and Malaysian firm disclosures and their English counterparts. They provide evidence that the native version of the accounting information is easier to read than the English version. This suggests that foreign investors, with no command of the original language, face higher information processing costs than domestic investors with sound knowledge of the respective original language. Comparing the same language, but different providers, Schroeder et al. [1991] investigate the financial reporting of Japanese firms at the New York Stock Exchange. They find that Japanese firms use similar vocabulary as their U.S. counterparts, but differ in terms of sentence structure, grammatical structure, and readability measures. Similarly, Campbell et al. [2005] find that in a sample of German and British firms, translation is mostly done on the sentence level and thus may lack customs that the target language would usually require. This implies that investors relying on the translated version of financial disclosures may face higher information costs than investors with proficient skills in the original language. Lundholm et al. [2014] on the other hand, find that foreign firms' accounting disclosures are easier to read than the disclosures of their U.S. counterparts.

Even though the literature on how source text and translation of company disclosures are perceived by readers is still small, it is nonetheless important. After all, low quality translations or large differences in understanding of the same underlying information may lead to differences in information processing costs or even behavioral outcomes, and thus may lead to frictions in the market. We therefore examine how (retail) investors perceive company disclosures that are the same in content, but different in language. Our focus is on prospective

investment behavior as well as three different textual characteristics usually associated with investment decisions (see section 2.2 for a discussion): readability, tone, and precision. Considering the inherent difficulty in providing adequate translations, as well as the fact that translations may be perceived differently than intended, we expect that the translated versions of the same underlying information will evoke different perceptions among readers than the native version. As discussed earlier, our focus is on retail investors. Thus, our first (non-directional) hypothesis is:

H1: Retail investors perceive the textual characteristics of native and translated versions of accounting disclosures differently.

2.2 Perception of Textual Characteristics and Investment Decisions

Presuming that textual characteristics of native and translated versions are perceived differently, we examine how this is reflected in the attractiveness of an investment to readers of either language version. The effect of translation on investment attractiveness can work through a number of channels. We examine three different textual characteristics that are documented to have a positive association with the attractiveness of an investment: readability, tone, and precision. In the following, we will briefly discuss each of them.

Readability: Both experimental (Asay et al. [2017]; Rennekamp [2012]) and archival papers (Lawrence [2013]; Miller [2010]) show that readability and complexity⁷ of a given text influence investment decisions. Researchers propose different theories as to why retail investors may react adversely to less readable disclosures. Miller [2010] for instance suggests that investors may abstain from investing in a firm with complex disclosures, because acquiring

⁷ The distinction between readability and complexity of a given text is not clear-cut in the literature. We therefore use the terms synonymously.

the relevant information may be too costly.⁸ Rennekamp [2012] suggests “processing fluency” as another possible explanation. Building upon psychology research (most notably Oppenheimer [2006]; as well as Shah and Oppenheimer [2007]), she explains that simply the perceived ease of processing the given information may act as a heuristic for viewing the given information as more reliable and the messenger as more credible.⁹

In line with these theoretical arguments, Miller [2010] finds that complex disclosures decrease the trading volume of retail investors and Lawrence [2013] discovers that especially smaller investors seem to invest more in stocks of firms with more readable disclosures. Asay et al. [2017] find that less readable disclosures may lead people to rely more on outside information (i.e. not issued by the firm) since they feel less comfortable assessing the firm. Apparently, people who do not seek such outside information value firms that issue less readable disclosures lower.

In summary, we expect to observe a positive relationship between the readability of firm disclosures and the attractiveness of a firm as an investment.

Tone: Tone (or sentiment¹⁰) has been identified as another major factor driving investor judgment and decision making. Just as for readability, a different tone does not imply that the information content of the text has been changed or is distorted; the text is simply differently

⁸ Lo et al. [2017] show that managers seem to strategically hide information in complex disclosures. More specifically, it appears that managers try to conceal earnings management via issuing more complex disclosures. As Li [2008] points out, these findings are in line with the “incomplete revelation hypothesis” stated by Bloomfield [2002]. It demonstrates that managers have an incentive to obfuscate bad news. In an experimental setting, Asay et al. [2018] show that what is largely perceived as obfuscation of bad news, may actually be the opposite: Managers make good news easier to read, while not trying to hide bad news. While this is an important observation, it does not influence our predictions coming from an investor’s perspective.

⁹ While Rennekamp [2012] finds evidence in line with her predictions, Tan et al. [2015] find that processing fluency is not significantly associated with investors’ judgments in a setting where the given information appears inconsistent. In a setting with consistent messages, readability overall seems to matter less for participants’ judgments.

¹⁰ The literature often uses the terms “sentiment”, “valence”, and “tone” interchangeably. For our study, we will generally use the term “tone” to identify the directional presentation (positive vs. negative) of the underlying information.

worded.¹¹ The phenomenon falls within a large array of framing effects that show consistent results across disciplines. As for tone, the general notion is that positively phrased statements trigger heuristics that lead to a higher perceived attractiveness of the item in question (Levin et al. [1998]). Within an accounting setting, this reaction may not be unwarranted, assuming that management intends to provide a signal about the firm's future performance: Davis et al. [2012] analyse the tone of earnings press releases and find that a positive tone is associated with positive future return on assets, as well as a positive market reaction (i.e. positive abnormal returns around the earnings release). Henry [2008] makes a similar observation, also noting that the tone of earnings press releases has a significant influence on abnormal returns around the time of publication.

However, a positive tone may not always be a credible signal. It appears that managers whose compensation is more strongly linked to the firm's stock price development also use positive tone more aggressively. However, the stronger these incentives are for managers, the more investors seem to see through their strategy and discount their firm valuations accordingly (Arslan-Ayaydin et al. [2016]). Huang, Teoh et al. [2014] also observe that firms may use positive tone strategically. Overall, they find a positive association between the abnormal tone of an earnings press release and the cumulative stock-price returns around the announcement. However, they further discover that abnormally positive tone is associated with lower operating cash flows and earnings in the mid-term future.

From an investor's perspective, evidence suggests that not all investors process tone in a similar way, with professional investors and short sellers being better able to detect unwarranted positive tone (Baginski et al. [2018]; Blau et al. [2015]). The resulting disagreement among investors also appears to increase for more positive disclosures (i.e., the

¹¹ A good example would be "the glass is half full" vs. "the glass is half empty." Both phrases contain the exact same information, with the former being expressed positively and the latter negatively.

more positive a disclosure is presented, the higher the disagreement between professional and non-professional investors). Since our setting is designed around retail investors, we expect results in line with a heuristics-based processing of the information at hand and thus expect that positive tone is positively associated with the attractiveness of a firm as an investment.

Precision: As our third textual characteristic, we draw the attention to (perceived) precision¹² in firms' disclosures. Even though this aspect is less prominent in the literature than readability and tone, it has the potential to provide similarly interesting conclusions. One notable paper in the experimental literature is Elliott et al. [2015], who examine how more concrete language influences investment decisions. They build their argument around construal theory, stating that more concrete language reduces the perceived distance to the firm and thereby makes it more attractive as an investment. Furthermore, more concrete language should make people more comfortable in assessing a firm since it makes it easier to grasp potential outcomes. In their experimental setting, the authors present participants with (fake) IPO prospectuses. They do not manipulate the content, but rather just highlight different sections (concrete vs. abstract) of the report. They find that the condition which highlights the concrete sections on average increases the perception of a firm as an attractive investment.¹³

Demers et al. [2014] offer "cheap talk" as an explanation for why investors may react adversely to less precise information. Less precise information would be considered less credible, thus leading investors to view the information as less reliable. They observe that less precise language in quarterly earnings announcements is related to more negative investor reactions. Interestingly though, they also find that more abstract language is associated with

¹² It is important to note that we are not referring to numerical precision, such as point vs. range estimates (e.g. Baginski et al. [2018]). We are referring to perceived precision in disclosure language. For this, the literature offers several synonyms, including "straight talk" and "concreteness". In the spirit of Elliott et al. [2015], we define precision as information being more specific, yet without any difference in verifiability. Elliott et al. [2020] offer an example by stating that "Dave hurts Bill" is more abstract than "Dave pushes Bill."

¹³ Elliott et al. [2020] re-visit the question of concrete language and investment decisions. They confirm their earlier findings that more concrete language is a possible tool to reduce psychological distance to a given firm and can thus help in reducing investors' home bias.

higher future return volatility. The authors thus argue that not all imprecise information is cheap talk, but at least sometimes a rather credible signal about the firms' own insecurity about their future.

Dzielinski et al. [2021] look at managers' statements during conference calls and find that the market reacts more strongly to the news conveyed in the calls if the manager uses less vague (or "clearer") language. Furthermore, they find that markets react stronger to CEOs with clearer language. Pan et al. [2018] use transcripts of quarterly earnings calls. They too find that more precise language is significantly associated with a more positive reaction from investors (measured as abnormal returns around the event date).

In summary, we build upon the above arguments and prior findings and expect that precision is positively associated with the attractiveness of a firm as an investment. We therefore test the following three hypotheses:

H2a: *Disclosures that are perceived as more readable lead retail investors to view the issuing firms more attractive as an investment.*

H2b: *Disclosures that are perceived as more positive in tone lead retail investors to view the issuing firms more attractive as an investment.*

H2c: *Disclosures that are perceived as more precise lead retail investors to view the issuing firm more attractive as an investment.*

Please note that our study is methodologically different from the studies cited above. To the best of our knowledge, we are the first to let people assess textual characteristics of a large number of real (and unmanipulated) company disclosures, while being able to control entirely for the underlying firm fundamentals (see section 3).

2.3 Textual Characteristics: Perception and Standard Measures

Finally, we are interested in how well commonly used textual measures compare with perceptions provided by retail investors. Textual measures are widely used tools to gather textual characteristics from company disclosures. Again, we focus on the three common textual characteristics: readability, tone, and precision.

Recent examples in accounting that examine **readability** typically include more than one textual measure. Among the most common measures for readability are the (modified) FOG index and the Flesch Reading Ease Score (see Du and Yu [2021]; Hasan [2020]; Kim et al. [2019] for recent examples). Albeit slightly different in execution, many of the applied readability measures rely on some combination of word-/sentence-length and the number of syllables. Even though these measures face criticism (e.g. Loughran and McDonald [2014], [2016]) and different approaches such as using file size (e.g. Hasan [2020]; Loughran and McDonald [2014] or the BOG index (Bonsall et al. [2017]) have been employed, the above measures are still widely used.

Tone in accounting research is often measured via word lists. Tone is then defined as positive words minus negative words, divided by either the total number of words in the text or the sum of positive and negative words (e.g., D'Augusta and DeAngelis [2020]; Druz et al. [2020]; Lee and Park [2019]). The workhorse in the literature is the word list developed by Loughran and McDonald [2011]. Just as for readability, criticism has been voiced and other approaches have been suggested (e.g. the BERT model using machine learning, see Siano and Wysocki [2021], and Naïve Bayes approaches, see Huang, Zang et al. [2014]). Nevertheless, the Loughran and McDonald [2011] word list (also in its later versions) remains one of the most widely applied tone measures in accounting research.

It is somewhat more difficult to assign one standard measure that is used to determine **precision** of a given text, not least because the definition of precision sometimes varies across

accounting studies. As noted above, we follow Elliott et al. [2015] and define precision as information being more specific, yet without any difference in verifiability. We choose the Loughran and McDonald [2011] uncertainty word list as a proxy for precision, following the creators' assessment that the "Fin-Unc list includes words denoting uncertainty, with emphasis on the general notion of imprecision rather than exclusively focusing on risk" (Loughran and McDonald [2011]). The list is applied in Demers et al. [2014] and Dzielinski et al. [2021].¹⁴

In the following, we will focus on the Flesch Reading Ease index for readability, and the word lists developed by Loughran and McDonald [2011] for tone and precision. Despite wide application, evidence on how well these measures correlate with real investor perceptions is scarce (a notable exception in the accounting literature is Bonsall et al. [2017]). We therefore add to the literature by showing how well these workhorses in the accounting literature correlate with real retail investor perception for real company disclosures.

3 Research Approach

3.1 Study Design

Language-related accounting literature usually suffers from a necessary trade-off between internal and external validity. Archival studies are often subject to various endogeneity concerns and carefully executed experiments are often restricted to designs which do not leave much room for the variability of disclosures found in practice. We aim to overcome some of these limitations and employ a method offering external validity through the use of real company disclosures, while keeping internal validity high.

¹⁴ The applications differ slightly in what they proxy for, however we believe the related contents to be sufficiently close to our notion of precision.

We use the fact that publicly listed German firms reporting in accordance with the demanding requirements of the Prime Standard¹⁵ are obliged to issue certain company reports in both their native language and English. Therefore, we have two texts with the same informational content, yet potentially different nuances in textual characteristics. The idea is to let both language versions be read by multilanguage retail investors. This way, we can gauge investors' perceptions of and reactions to the disclosure without having any differences in the underlying information content or firm characteristics. Participants' perceptions are captured via 10 different survey questions.

The basis for our investigation is forecast reports included in the half-year reports of German firms which are reporting in accordance with Prime Standard requirements. For each firm, we manually extract the forecast section from the 2019 half-year report. We focus on half-year reports since the provided information is usually shorter than in annual reports. Hence, we believe the information is easier to digest for our survey participants. Obviously, it is not clear whether forecast reports are overall important to retail investors (Loughran [2018]). However, forecast reports have the advantage that all publicly listed firms must incorporate forward-looking information in their half-year reports (§115, para 4, Wertpapierhandelsgesetz; *German Securities Trading Law*). Furthermore, they are one of the few sections of company reports that are future-oriented and thus contain new (price-relevant) information. Finally, the forecast section is usually relatively short: the average forecast report in our sample is 390 words or about 1 page long. This ensures that we can ask participants to read multiple forecast reports while minimising the risk of them growing tired of reading them. We draw a random sample of 61 firms out of the population of 250 Prime Standard firms disclosing separate forecast reports in their half-year reports.¹⁶ We make sure that each language version of the

¹⁵ The Prime Standard is part of the regulated market in Germany and its stock exchange segment with the highest transparency requirements.

¹⁶ The sample size is determined by our budget restrictions with respect to the payment of the participants and the minimum of the required responses per report.

same forecast report contains the same factual information by reading and comparing each of them. Please see Appendices 1 and 2 for sample forecast reports.

[Table 1]

In order to collect the characteristics of each report, we acquire participants who are fluent in both German and English. This offers a key advantage: If our participants were only fluent in one language (i.e., Germans would read German reports etc.), then their cultural background alone could be responsible for finding different effects (Sapir [1985]; Whorf [1956]). By using bilingual participants, we can ensure that each report is read by both English and German native speakers, thus allowing us to control for cultural-related confounders.¹⁷ We ensure that participants do not receive the English and German language version of the same firm. Overall, each participant reads six different, randomly selected forecast reports in either German or English. In turn, each forecast report is read by many participants (see section 4 for further details). This set-up allows us to use participant- and firm-fixed effects in our analyses. We are therefore able to rule out many alternative explanations for our findings that are commonly voiced in the archival literature (e.g. the effect of information content, see Li [2008] and Tan et al. [2014]). However, since we use real company forecast reports, we are able to keep external validity high and our reports free of experimenters' manipulation that may unwillingly drive results.

Our strict sample requirements led us to acquire participants via Prolific.co. Prolific is an online service specialised in providing participants for scientific studies. It allows researchers to pre-screen participants based on a long list of demographic and other questions. In terms of quality, Prolific seems to be on par, if not better than participants acquired via Amazon MTurk (for a comparison of platforms see Peer et al. [2017] and Palan and Schitter [2018]). In order

¹⁷ We subsume Germany, Austria and Switzerland under domestic investors since these countries have a similar language culture. The native language of the Swiss participants in our sample is German.

to gain a sample of retail investors, we further narrowed down our criteria to people who had prior investments in shares. We estimated that each session would take participants about an hour and we paid 9 GBP (about 12 U.S. dollars or 10 Euros at the time) for each completed survey. We acknowledge that this amount is rather at the upper end of participant payments commonly offered. However, besides ethical reasons,¹⁸ we saw that our prospective maximum sample was already relatively small and the task demanding. Hence, our payment was also meant as an incentive to induce a sufficiently high motivation to participate and also complete the long survey. Participants were aware that, just as for any task on Prolific, we were able to exclude them if their answers did not meet quality expectations. Overall, we had to exclude 21 participants, most of which made differing statements about their native languages in Prolific and in our survey (see Table 1).

We invited all pre-screened Prolific users to a survey called “Characteristics of Forecast Reports.” Before joining, they were informed that they had to read six different forecast reports and answer 10 content-related questions plus an open-ended “comment” field per report. Apart from that, the invitation text informed participants that they were pre-screened with regard to their information on Prolific and that the forecast reports would be matched to the languages they are fluent in. We did not inform participants that the reports would only be in German and English to avoid hypothesis guessing. The university granted us ethical approval on November 3rd, 2020, and we ran the study from November 3rd to November 9th, 2020. We provide the survey instructions in Appendix 3.

¹⁸ We decided to pay participants no less than the minimum wage in Germany, which at the time was EUR 9.35.

3.2 Hypotheses Testing

3.2.1 Translation

Our first hypothesis focuses on the impact of translation on the perception of textual characteristics and the evaluation of a firm's attractiveness as an investment. We use the following three survey questions to assess readability, tone, and precision:

For **readability**, we refrain from directly asking participants how readable a certain text is. As prevalent as the term may be in the accounting literature, we believe it to be rather ambiguous to people outside the field. Therefore, we take inspiration from Rennekamp [2012] and capture readability by phrasing our question rather close along the definition of processing fluency as stated by Alter and Oppenheimer [2009, p.219]: Processing fluency is “the subjective experience of ease with which people process information.” We pose the following question:

Q - readability: *How easy do you find it to extract relevant information from the text for making an investment decision?*

Participant responses are captured via a 7-point Likert scale ranging from “very easy” to “very difficult” (see Appendix 4 for a definition of all survey questions).

Our question for **tone** directly asks how participants view the tone of the disclosure, ranging from “very positive” to “very negative”. We include both directions in our question since we did not expect all disclosures to be perceived positively. Also, this choice keeps our design closer aligned with common textual measures that also allow for negative tone. We pose the following question:

Q - tone: *How would you rate the overall tone of this text?*

As our last textual characteristic, we measure the perceived **precision** of the text. We decided to ask participants for their impression of precision, rather than concreteness. Even

though both terms capture similar underlying concepts, we choose precision since it emphasizes detail and exactness. Concreteness on the other hand is often understood as being grounded in facts (consider the term “concrete evidence” as opposed to “precise evidence”). Since the former highlights an angle we do not want to emphasize in the assessment (i.e., the element of truth), we choose precision as our wording. We pose the following question:

Q - *precision*: *How precise is the information presented in this text?*

Again, we offer participants a 7-point Likert scale for this question, ranging from “very vague” to “very precise.”

Finally, we are interested in examining potential consequences for investor behavior due to differences in translation quality. As such, the (experimental) literature offers a variety of dependent variables ranging from firm valuation (Asay et al. [2017]) to earnings judgments (Tan et al. [2014]). In our setting, we want to incorporate the fact that accounting disclosures do not solely, but primarily, aim to provide outsiders with firm information that enable them to value the firm as a(n) (potential) investment. We therefore ask participants for the attractiveness of the respective firm as a share investment. We view investment attractiveness as a necessary antecedent for actual investment decisions. Our question capturing the dependent variable is thus similar to Elliott et al. [2015]:

Q - *attractive_investment*: *After reading this text, how attractive are this firm's shares as an investment to you?*

Answers are provided on a 7-point Likert scale ranging from “very attractive” to “very unattractive.”

We first analyze differences in perceptions via simple t-tests. Additionally, we employ OLS regressions that allow us to use firm- and participant-fixed effects that control for unwanted confounders in case our randomization did not work as intended. In these

regressions, we include *foreign_language* as our independent variable of interest to examine whether the translation of accounting disclosures influences retail investors' assessment of investment attractiveness or textual characteristics. *foreign_language* is a binary variable which is coded as 1 for the English translation and 0 for the German version of the firm disclosure. Furthermore, we include an interaction term between *foreign_language* and the respective participants' origin (*foreign_origin*). This allows us to assess whether the effects hold for all participants irrespective of whether they are German or English native speakers. Our regressions including fixed effects look as follows. β_1 is our main coefficient of interest:

$$1) \text{ dependent variable}_{f,p} = \beta_1 \cdot \text{reporting_language}_{f,p} + \beta_2 \cdot \text{reporting_language}_{f,p} * \text{foreign_origin}_p + \text{firm } FE_f + \text{participant } FE_p + \varepsilon_{f,p}$$

Note that the participant-fixed effects subsume the main effect of *foreign_origin*. The firm-fixed effects on the other hand do not subsume the reporting language, as each firm issues two reports.

3.2.2 Perception of Textual Characteristics and Investment Decisions

In this section, we are interested in the relationship between our three textual characteristics of interest (*readability*, *tone*, and *precision*) and *attractive_investment*. These variables are defined as above. We face two concerns that need addressing in our empirical design: First, our setting does not incorporate an experimental manipulation of the textual characteristics, which is why we include additional survey questions as controls. Second, each participant provides both the independent and dependent variables, making our design prone to common method bias. We deal with these issues in the following ways.

Readability: Two concurrent thoughts may predominantly appear when participants assess the readability of a given text. First, they may evaluate a text relative to what they believe is an appropriate difficulty. For instance, while they may deem it difficult to extract information from the given text, they may view it relatively easy given that forecast reports stem from a rather technical financial document. In order to separate this effect, we control for participants' impressions about how much education would be necessary to understand the text. Second, we acknowledge that texts may feel unnecessarily complicated. As this could be viewed as obfuscation and thus a negative sign by participants, we control for whether they believe the text to be unnecessarily difficult. Our control questions are as follows:

Q – *text_complexity*: *What level of education do you think is required to fully understand the given text?*

Q - *information_difficulty*: *The firm has made this text intentionally difficult to understand.*

Tone: Han and Tan [2010] suspect that (perceived) credibility may make a difference in assessing disclosure tone. Tan et al. [2014] provide such evidence. In an experimental setting, they show that at least for more sophisticated investors, the credibility of an earnings release appears to make a difference in earnings judgments. The less information is perceived as credible, the lower the earnings judgment that participants make. We thus decide to control for perceived credibility. However, credibility as a term may be prone to confusing and confounding interpretations, considering that forecast reports stem from audited financial statements. Hence, we ask participants for perceived objectiveness of the given text. We deem this a fitting construct as it is interpretable and sufficiently specific in the given context. We gather participants' impressions by assessing how much they (dis-)agree with the following statement:

Q - *objectiveness*: *The firm presents the information in this text objectively.*

We furthermore consider that the attractiveness of an investment may also be influenced by whether investors believe that the disclosure tone may be a signal to investors and thus drive share prices (see section 2). Participants would therefore not react to positive tone, but to their expectations of what follows from a positive tone. We therefore include the following control:

Q – *tone_shareprice*: *How do you think will the overall tone of the presented information influence the share price of the firm?*

Precision: In section 2, we note that imprecise language may be a signal about the firms' own insecurity about its future (Demers et al. [2014]). We aim to extract the residual effect of precision after controlling for that by asking participants whether they believe the firm to be sure about its own future. We ask participants to evaluate the following statement based on a 7-point Likert scale ranging from “strongly agree” to “strongly disagree”:

Q - *firm_certainty*: *The firm is certain about its future development.*

Finally, we also control for how certain the participants themselves feel about the future of the firm. Controlling for this is particularly important in our setting; looking at forecast reports ex post. All participants will at least know major economic developments after the first half-year of 2019, and thus feel more or less sure about a certain firm's development independent of the information provided in the report (e.g. an airline during the COVID-19 pandemic). Even though we specifically asked participants to ignore such information, we cannot rule out that it may have influenced their ratings. *individual_certainty* allows us a possibility to control for this:

Q – *individual_certainty*: If someone asked you about the future development of the firm, how sure would you be about your answer?

Controlling for common method bias: Common method bias (CMB; alternatively “common method variance”) is a particular problem in our setting where participants self-

report both independent and dependent variables (Conway and Lance [2010]). The issue is of concern for our hypotheses H2a-c, where all our main variables of interest stem from the same source. It is of no concern for our analyses related to H1 since the reporting language is randomly allocated across participants and thus exogenous. Conway and Lance [2010] point out that the fear of inflated correlations due to CMB is often unwarranted. Yet, we acknowledge that it has the potential to influence our results.

Overall, we believe that the possibility of CMB was a necessary cost in our setting. We are ultimately interested in persons' perceptions of and reactions to the company disclosures. This implies that no other source except real people could be an adequate proxy for perception. In this case, one potent remedy is to have different people providing the independent and dependent variables (Tehseen et al. [2017]). Unfortunately, this is not possible in our setting as the number of bilingual survey participants was already very small. Nevertheless, we aim to reduce CMB in our setting by taking advantage of the fact that each report is read by multiple participants. Assuming that each participant's judgment of a textual characteristic will be a noisy estimate of the true underlying textual characteristic, we choose to replace the independent variables on the participant level with the median answer per report.¹⁹ We run all regressions related to H2a-c in a similar fashion. We use *attractive_investment* as the dependent variable and the median textual characteristic per report (i.e., *readability*, *tone*, and *precision*) as independent variables. We also use the median response per report for all other survey-related control variables. Finally, we use firm- and participant-fixed effects that address

¹⁹ An alternative approach is to split the sample (randomly) into providers of the independent variable and taking the averages across samples to obtain measures free of common method bias (e.g., Antonakis and House [2014]). We applied and repeated this procedure 1,000 times with random allocation of participants each time to estimate the effects. The results are qualitatively similar to the median approach with *tone* being significantly associated with *attractive_investment* on the 5% level in more than 99% of regressions, while *readability* and *precision* are significant in less than 1% of regressions (using firm fixed effects). Note that this approach severely reduces the sample size to 122, as we need to collapse data on the report level. The resulting lack of power and limited opportunity to present alternative specifications in a meaningful way (e.g. using control variables and fixed effects) led us to present the qualitatively similar results from the median approach.

potential endogeneity concerns such as the firms' profitability and participants' origins. Our regressions thus look like the following in their strictest specifications:

$$2) \text{ attractive investment}_{f,p} = \beta \cdot \widetilde{\text{textual characteristics}} + \widetilde{\text{controls}}_f + \text{firm FE}_f + \text{participant FE}_p + \varepsilon_{f,p}$$

3.2.3 Textual Measures

We first restrict our sample to English reports only since the textual measures we apply were developed for English documents. We apply the Flesch Reading Ease index to proxy for *readability*²⁰ and use the word lists developed by Loughran and McDonald [2011] in their 2014 version to capture tone and precision. *Tone* is defined as the number of positive words minus the number of negative words divided by the sum of positive and negative words. Results remain unchanged if the denominator is the total number of words. *Precision* is defined as the number of uncertainty-related words divided by the total number of words, times minus one (i.e., lower values imply less precise texts). We present correlations, graphical evidence, and an application in our regression framework in section 4.

4 Results

4.1 Descriptive Statistics

Table 2 Panel A provides descriptive statistics on characteristics of firms and their respective forecast reports included in our survey sample and the underlying population of Prime Standard firms. We observe no significant differences between our sample and the Prime Standard. We thus believe our results to be applicable to the wider group of publicly listed firms. The median firm in our survey sample has a market capitalization of 590 million Euros and return on assets of 3.4 %. 26 % of the sample firms are audited by a Big4 auditor. Overall,

²⁰ *Flesch score* = $206.835 - 1.015 * \left(\frac{\text{total number of words}}{\text{total number of sentences}} \right) - 84.6 * \left(\frac{\text{total number of syllables}}{\text{total number of words}} \right)$, see Flesch [1948]. We also use the Flesch-Kincaid score and the Gunning Fog index and come to similar conclusions (Gunning [1952]; Kincaid et al. [1975]).

the English reports in our survey sample contain significantly more words than the German reports (55 words difference in means, $p < 0.01$). The same holds true for the Prime Standard population (59 words difference in means, $p < 0.01$). This is in line with findings from prior literature (Campbell et al. [2005]). Our manual comparison of both report versions ensures that this difference is driven by differing language characteristics rather than differences in the information disclosed.

The median (English) report has a Flesch score of 34.3 (college level; mean = 33.23, sd = 9.55), is positive in tone (mean = 0.54, sd = 0.30) and contains three to four precision-related words (mean per 100 words = 0.008, sd = 0.006). Overall, the Flesch score of our forecast reports appears to be relatively high, but overall in line with prior studies (e.g., Courtis and Hassan [1973]; Li [2008]; Richards and van Staden [2015]; Schroeder et al. [1991]). The tone appears to be slightly more positive than in prior literature (e.g. for management forecasts: Baginski et al. [2018]; for MD&A: Lee and Park [2019]; for 10-Ks: Loughran and McDonald [2011], and for earnings press releases: Huang, Teoh et al. [2014]). Precision is slightly lower than in prior literature (Loughran and McDonald [2011]).

[Table 2]

Panel B of Table 2 presents descriptive statistics of our participants' demographics. The median participant in our survey is male, between 31 years and 40 years old, and has a Bachelor's degree or an equivalent education. With 77 %, the large majority of participants state to have read a financial report before. Furthermore, 87 % of our participants have directly invested in shares before.²¹ The median investor trades up to once per quarter and invests

²¹ Despite our explicit participant requirement to have invested in shares at the Prolific database, 13 % of our respondents stated to have never invested in shares before. However, since the majority of these participants had either read financial reports before or stated afterwards that they have indirectly invested in shares before (e.g. through funds), we chose to keep these observations in our sample. Nonetheless, in order to ensure that "non-investors" do not bias our results, we repeated our main analyses without these observations and did not find significant differences in comparison to the total sample.

between EUR 101 and EUR 500 per trade. In summary, the results suggest that the characteristics of our group of participants are suitable for research on retail investors. However, we also see that some core demographic variables (e.g. *education* and *age_group*) differ significantly between foreign and domestic investors. This highlights the importance of participant fixed effects in our sample since it may bias our results due to the link between cultural imprint and text perception. In additional questions that we use as control variables in specifications without participant-fixed effects, we find that participants tend to trust firm disclosures but also believe that firms present themselves overly positive in their disclosures. Participants' impressions are mixed as to whether company managers overall care about retail investors and whether managers on average are competent.

Panel C of Table 2 presents descriptive statistics of participants' responses to our survey questions presented in Appendix 4. The mean response in our sample is centered around the "neutral" response (0). This is not entirely surprising since any other result would have suggested a strong overall bias in firms' disclosures. Furthermore, we are confident that this observation is not driven by participants simply choosing the "neutral" option out of convenience. Both the summary statistics as well as the histograms shown for each survey question indicate an active use of the whole Likert scale. Overall, we observe that participants on average perceive the tone of the firm disclosures as slightly positive while slightly disagreeing to the statement that the firms made their disclosures harder to read on purpose. This matches participants' general beliefs about firms' overly positive reporting and a general trust in firm disclosures (see again Table 2, Panel B). This correlation does not automatically imply that participants are necessarily biased when assessing the firms' disclosures. Still, it highlights the need for participant-fixed effects (and appropriate controls, respectively) in our setting.

Table 3 presents results of univariate tests for differences between report languages (see Appendix 6 for a correlation table). Panel A shows differences in means. We find that solely the perceived *readability* of German reports is significantly easier ($p < 0.01$) than that of their English translations. All other differences are not significant on conventional levels.

[Table 3]

Panel B presents differences in standard deviations per language. We present standard deviations since differences in text evaluation may not just manifest in average differences, but also in the variance of survey responses. If, for example, German reports were easier to assess than English reports, we would expect German reports to have a lower standard deviation per survey question. Our results indicate that German and English reports are remarkably close. All differences are insignificant across languages – including for *readability*. However, differences between questions can be substantial. For instance, the standard deviation for *readability* is about 50 % larger than the standard deviation for *tone*, implying higher disagreement between participants.

4.2 The Effect of Translation on the Perception of Textual Characteristics

In this section, we provide evidence on hypothesis H1. As noted in the previous section, simple t-tests only reveal significant differences in participant perceptions for readability, with German reports being easier to read. Table 4 supports these findings.

[Table 4]

Table 4 provides OLS-regression results on the impact of reporting language on retail investors' perceptions of *readability*, *tone*, and *precision* and the assessment of *attractive_investment* of the respective firm. We use participant- and firm-fixed effects to control for any bias left after randomization. Our analyses confirm the descriptive findings

that only *readability* is perceived differently between reporting languages with German reports being easier to read. Importantly, this holds for both German and non-German native speakers.

We fail to find a significant relationship between reporting language and *attractive_investment* – despite better readability usually being thought of being positively related to an investment being more attractive. We analyze potential causes for that in the following subsection.

4.3 Perception of Textual Characteristics and Investment Decisions

Table 5 presents our regressions covering hypotheses H2a-c. We first present results without controls and then add participant- and firm-fixed effects as well as our survey-related control variables developed in section 3. All regressions include all three main textual characteristics as independent variables. The results with full controls are similar if only one textual characteristic is included at a time.

[Table 5]

We observe that solely *tone* is significantly associated with *attractive_investment* in all specifications. The significant effect of *precision* on *attractive_investment* becomes indistinguishable from zero after including firm fixed effects. *Readability* is not significant in any specification. This finding may help to explain why differences in *readability* between German and English reports do not seem to translate into differences in *attractive_investment* between the two languages. At the same time, we observe that without controlling for common method bias (i.e. using both independent and dependent variables from the same participant), all three textual characteristics are significantly positively correlated with *attractive_investment*, albeit *readability* only at the 10 % level (not tabulated). This indicates that within each participant's evaluation, we observe significant correlations between

independent and dependent variables, but this does not translate into an overall observable effect of *readability* and *precision* on *attractive_investment*. Solely *tone* seems to be positively associated with *attractive_investment* after controlling for firm- and participant-related factors, survey-related controls, and common method bias.

This result is not in line with our hypotheses H2a and H2c that expected *readability* and *precision* to be positively associated with *attractive_investment*. Solely H2b is supported. For readability and precision our findings therefore also do not confirm prior literature. We revisit the distribution of our participants' responses in order to more closely examine these findings.

Table 2 Panel C and Figure 1 present information on the standard deviation of survey responses. Larger standard deviations indicate higher disagreement between participants which we interpret as the textual characteristic being more difficult to assess. For example, if it was easy for participants to judge a given text as highly precise, we would expect a lower standard deviation than for a text where it is more difficult to assess precision. In terms of results, we first refer to our earlier findings indicating that *tone* has a substantially lower standard deviation than the other textual characteristics and thus seems to be easier to assess than the other characteristics. This is supported by Figure 1: In Figure 1, we present the differences between the average response to English and German reports (i.e., English minus German) for each textual characteristic (black dots), sorted by the resulting differences on the firm-level.²² The red dots indicate the differences between the average response to English and German reports for *attractive_investment*. Since the differences do not have a standard deviation themselves, we slightly adapt our indicator for difficulty. The whiskers therefore present differences between English and German reports if each report was judged up to one standard deviation

²² Appendix 1 and Appendix 2 present the “best” and “worst” translations, meaning the reports with the lowest and highest average of absolute mean differences across all textual measures.

higher or lower.²³ Intuitively, larger whiskers again indicate larger disagreement between participants and thus the textual characteristic being more difficult to assess.

[Figure 1]

The figure supports the notion that *readability* and *precision* seem to be inherently more difficult to assess than *tone*. The whiskers for readability and precision are on average more than 30 % larger than for tone. At the same time, we observe a stronger correlation between mean differences for *attractive_investment* and *tone* ($r = 0.55$) than for *attractive_investment* and *readability* (0.34), respectively *precision* ($r = 0.40$). Taken together, this indicates that the insignificant association between *precision* and *readability* with *attractive_investment* in the earlier regressions may be driven by the fact that participants find it substantially more difficult to assess *readability* and *precision* in a given text. *Tone*, on the other hand, seems to be comparably easy to assess. We believe this to be an interesting finding as it indicates that while readability and precision may generally be associated with the perceived attractiveness of an investment, readers may find it difficult to assess them in real firm disclosures, leading to assessments that are too noisy to lead to effects significantly different from zero.

4.4 Textual Characteristics: Perception and Standard Measures

Finally, we present descriptive evidence on the association between our participants' judgments and textual characteristics measured via common textual measures. Table 6 and Figure 2 present our results.

In table 6, we horserace our textual characteristics with the textual measures from the literature. We use the median response to *attractive_investment* as the dependent variable and control for firm-related confounders. We are not able to use firm fixed effects since we can

²³ The endpoints of the whiskers thus represent the English report being judged one standard deviation higher and the German report being judged one standard deviation lower; and vice versa.

only take the English version of each report. Again, we only find tone to be significantly associated with *attractive_investment*. This finding holds for both our participants' responses and the textual measures.

[Table 6]

However, we caution to interpret these results as meaning that the textual measures always lead to the same results as participant responses would. First, keep in mind that our participants' responses are far from being unanimous assessments of each textual characteristic. Second, the correlation between the textual measures and the median response in our setting is low. Figure 2 plots standardized median responses to our survey questions capturing textual characteristics against standardized measures from the literature. It shows that solely for tone, the correlation between participants' responses and the textual measure is reasonably high ($r = 0.55$). For precision ($r = -0.28$) and readability ($r = 0.06$) the correlation is substantially lower.²⁴ We therefore believe the similar results for readability and precision to be rather driven by coincidence, rather than conceptual similarity.²⁵

[Figure 2]

Obviously, these results do not indicate that textual measures per se are a good or bad proxy for textual characteristics. For instance, it may well be that retail investors fail to understand the intended tone in some documents (Loughran [2018]). Nevertheless, it shows that these measures need to be applied with caution, especially in shorter texts. For instance, we have a notable fraction of firm disclosures with only very few (or no) words indicating precision. This is similar for our tone measures. In such small documents, noise in positive,

²⁴ The correlation for precision is still low if we use *firm_certainty* from our control survey questions as a different plausible proxy to compare it ($r = 0.09$).

²⁵ Results are qualitatively similar if we replace the Flesch Reading Ease score (Flesch [1948]) with the Flesch-Kincaid score or the Gunning Fog index (Gunning [1952]; Kincaid et al. [1975]).

negative, and precision-related words can have large consequences on the calculated measures. This supports the notion of better proxies, especially in smaller documents (Siano and Wysocki [2021]).

5 Conclusion

Our findings provide valuable insights on how translation (quality) influences retail investors' perceptions of textual characteristics and a firm's attractiveness as an investment. We employ a novel research design by randomly presenting real company disclosures issued in two different languages to (potential) retail investors. The retail investors then complete a survey about their perceptions of the textual characteristics of the presented disclosures as well as the investment attractiveness of the respective firm. This design allows us to assess perception differences across languages, as well as the relationship between textual characteristics and investment attractiveness while holding participant and firm characteristics constant.

We find that the German versions of firm disclosures are significantly easier to read than their English translations. This finding holds for German and English native speakers alike. However, this does not lead to German reports being assessed as more attractive as an investment. We find that after controlling for participant- and firm-fixed effects as well as for common method bias, only tone (but not readability and precision) seems to be positively associated with the attractiveness of a firm as an investment. Overall, participants' assessments of each textual characteristic are very heterogeneous, potentially contributing to the lack of support for a relationship between readability and precision with investment attractiveness. We encourage further research to explore how this difficulty in assessing textual characteristics is linked to an attenuation in former documented relationships between textual characteristics and investment attractiveness.

Finally, we show that the correlation between our participants' assessment of tone seems to capture a similar underlying construct as tone measured via the Loughran and McDonald [2011] word list. The correlation with precision is smaller, and there is virtually no correlation between our participants' assessment of readability and the Flesch Reading Ease score (we obtain similar results for the Flesch Kincaid score and the Gunning Fog index). Even though these findings do not necessarily extend to other settings, we encourage researchers to not just rely on one standard method to capture textual characteristics. We acknowledge that this insight is already widely applied, especially in readability research (e.g. Du and Yu [2021]; Hasan [2020]; Kim et al. [2019]).

One limitation of our study is that we cannot directly control for participant-firm related factors (e.g. people letting their knowledge about subsequent firm performance influence their choices). However, we believe this risk to be low due to our experimental randomization and use of median assessments in our regressions.

References

- ALTER, A. L., AND D. M. OPPENHEIMER "Uniting the Tribes of Fluency to Form a Metacognitive Nation." *Personality and Social Psychology Review* 13 (2009): 219–235.
- ANTONAKIS, J., AND R. J. HOUSE "Instrumental Leadership: Measurement and Extension of Transformational-Transactional Leadership Theory." *Leadership Quarterly* 25 (2014): 746–771.
- ARSLAN-AYAYDIN, Ö., K. BOUDT, AND J. THEWISSEN "Managers Set the Tone: Equity Incentives and the Tone of Earnings Press Releases." *Journal of Banking and Finance* 72 (2016): S132–S147.
- ASAY, H. S., W. B. ELLIOTT, AND K. RENNEKAMP "Disclosure Readability and the Sensitivity of Investors' Valuation Judgments to Outside Information." *The Accounting Review* 92 (2017): 1–25.
- ASAY, H. S., R. LIBBY, AND K. RENNEKAMP "Firm Performance, Reporting Goals, and Language Choices in Narrative Disclosures." *Journal of Accounting and Economics* 65 (2018): 380–398.
- BAGINSKI, S. P., E. DEMERS, A. KAUSAR, AND Y. J. YU "Linguistic Tone and the Small Trader." *Accounting, Organizations and Society* 68–69 (2018): 21–37.
- BLAU, B. M., J. R. DELISLE, AND S. M. K. PRICE "Do Sophisticated Investors Interpret Earnings Conference Call Tone Differently than Investors at Large? Evidence from Short Sales." *Journal of Corporate Finance* 31 (2015): 203–219.
- BLENKINSOPP, J., AND M. S. PAJOUH "Lost in Translation? Culture, Language and the Role of the Translator in International Business." *Critical Perspectives on International Business* 6 (2010): 38–52.

- BLOOMFIELD, R. J. "The "Incomplete Revelation" and Financial Reporting." *Accounting Horizons* 16 (2002): 233–243.
- BONSALL, S. B., A. J. LEONE, B. P. MILLER, AND K. RENNEKAMP "A Plain English Measure of Financial Reporting Readability." *Journal of Accounting and Economics* 63 (2017): 329–357.
- CAMPBELL, D., A. C. BECK, AND P. SHRIVES "A Note on Comparative Language Interrogation for Content Analysis: The Example of English vs. German." *British Accounting Review* 37 (2005): 339–350.
- CONWAY, J. M., AND C. E. LANCE "What Reviewers Should Expect from Authors Regarding Common Method Bias in Organizational Research." *Journal of Business and Psychology* 25 (2010): 325–334.
- COURTIS, J. K., AND S. HASSAN "Reading Ease of Bilingual Annual Reports." *The Journal of Business Communication* 39 (1973): 394–413.
- CUYPERS, I. R. P., G. ERTUG, AND J.-F. HENNART "The Effects of Linguistic Distance and Lingua Franca Proficiency on the Stake Taken by Acquirers in Cross-Border Acquisitions." In *Source: Journal of International Business Studies* 46.4 (2015): 429–442.
- D'AUGUSTA, C., AND M. D. DEANGELIS "Tone Concavity around Expected Earnings." *The Accounting Review* 95 (2020): 133–164.
- DAVIS, A. K., J. M. PIGER, AND L. M. SEDOR "Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language." *Contemporary Accounting Research* 29 (2012): 845–868.
- DEMERS, E., AND C. VEGA. "The impact of credibility on the pricing of managerial textual content." Available at SSRN 1153450 (2014).

- DOUPNIK, T. S., AND M. RICHTER "Interpretation of Uncertainty Expressions: A Cross-National Study." *Accounting, Organizations and Society* 28 (2003): 15–35.
- DRUZ, M., I. PETZEV, A. F. WAGNER, AND R. J. ZECKHAUSER "When Managers Change Their Tone, Analysts and Investors Change Their Tune." *Financial Analysts Journal* 76 (2020): 47–69.
- DU, S., AND K. YU "Do Corporate Social Responsibility Reports Convey Value Relevant Information? Evidence from Report Readability and Tone." *Journal of Business Ethics* 172 (2021): 253–274.
- DZIELIŃSKI, M., A. F. WAGNER, AND R. J. ZECKHAUSER. "CEO Clarity." RWP17-017, Swiss Finance Institute Research Paper 17–13. Available at SSRN: 2965108 (2021).
- ELLIOTT, W. B., K. M. RENNEKAMP, AND B. J. WHITE "Does Concrete Language in Disclosures Increase Willingness to Invest?" *Review of Accounting Studies* 20 (2015): 839–865.
- ELLIOTT, W. B., K. M. RENNEKAMP, AND B. J. WHITE "Can Concrete Language Help to Mitigate the Home Bias in Equity Investing? An Extension of Elliott, Rennekamp, and White (2015)." *Journal of Financial Reporting* 5 (2020): 51–64.
- EVANS, L. "Language, translation and accounting: towards a critical research agenda." *Accounting, Auditing & Accountability Journal* 31.7 (2018): 1844–1873.
- FLESCH, R. "A New Readability Yardstick." *Journal of Applied Psychology* 32 (1948): 221.
- GUNNING, R. *The Technique of Clear Writing* (1952).
- HAN, J., AND H. T. TAN "Investors' Reactions to Management Earnings Guidance: The Joint Effect of Investment Position, News Valence, and Guidance Form." *Journal of Accounting Research* 48 (2010): 123–146.

- HASAN, M. M. "Readability of Narrative Disclosures in 10-K Reports: Does Managerial Ability Matter?" *European Accounting Review* 29 (2020): 147–168.
- HENRY, E. "Are Investors Influenced by How Earnings Press Releases Are Written?" *Journal of Business Communication* 45 (2008): 363–407.
- HUANG, A. H., A. Y. ZANG, AND R. ZHENG "Evidence on the Information Content of Text in Analyst Reports." *The Accounting Review* 89 (2014): 2151–2180.
- HUANG, X., S. H. TEOH, AND Y. ZHANG "Tone Management." *The Accounting Review* 89 (2014): 1083–1113.
- JEANJEAN, T., C. LESAGE, AND H. STOLOWY "Why Do You Speak English (in Your Annual Report)?" *The International Journal of Accounting* 45 (2010): 200–223.
- JEANJEAN, T., H. STOLOWY, M. ERKENS, AND T. L. YOHAN "International Evidence on the Impact of Adopting English as an External Reporting Language." *Journal of International Business Studies* 46 (2015): 180–205.
- KIM, C., K. WANG, AND L. ZHANG "Readability of 10-K Reports and Stock Price Crash Risk." *Contemporary Accounting Research* 36 (2019): 1184–1216.
- KINCAID, J. P., R. P. JR. FISHBURNE, R. L. ROGERS, AND B. S. CHISSOM "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. " Naval Technical Training Command Millington TN Research Branch (1975).
- LAWRENCE, A. "Individual Investors and Financial Disclosure." *Journal of Accounting and Economics* 56 (2013): 130–147.

- LEE, J., AND J. PARK "The Impact of Audit Committee Financial Expertise on Management Discussion and Analysis (MD&A) Tone." *European Accounting Review* 28 (2019): 129–150.
- LEVIN, I. P., S. L. SCHNEIDER, AND G. J. GAETH "All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects." *Organizational Behavior and Human Decision Processes* 76 (1998): 149–188.
- LI, F. "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 221–247.
- LO, K., F. RAMOS, AND R. ROGO "Earnings Management and Annual Report Readability." *Journal of Accounting and Economics* 63 (2017): 1–25.
- LOUGHRAN, T. "Linguistic Tone and the Small Trader: Measurement Issues, Regulatory Implications, and Directions for Future Research." *Accounting, Organizations and Society* 68–69 (2018): 38–41.
- LOUGHRAN, T., AND B. McDONALD "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66 (2011): 35–65.
- LOUGHRAN, T., AND B. McDONALD "Measuring Readability in Financial Disclosures." *Journal of Finance* 69 (2014): 1643–1671.
- LOUGHRAN, T., AND B. McDONALD "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research* 54 (2016): 1187–1230.
- LUNDHOLM, R. J., R. ROGO, AND J. L. ZHANG "Restoring the Tower of Babel: How Foreign Firms Communicate with U.S. Investors." *The Accounting Review* 89 (2014): 1453–1485.
- LUNDHOLM, R., N. RAHMAN, AND R. ROGO "The Foreign Investor Bias and Its Linguistic Origins." *Management Science* 64 (2018): 4433–4450.

- MILLER, B. P. "The Effects of Reporting Complexity on Small and Large Investor Trading." *Accounting Review* 85 (2010): 2107–2143.
- NIDA, E. A. *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive (1964).
- OPPENHEIMER, D. M. "Consequences of Erudite Vernacular Utilized Irrespective of Necessity: Problems with Using Long Words Needlessly." *Applied Cognitive Psychology* 20 (2006): 139–156.
- PALAN, S., AND C. SCHITTER "Prolific.Ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17 (2018): 22–27.
- PAN, L., G. MCNAMARA, J. J. LEE, J. (JOHN) HALEBLIAN, AND C. E. DEVERS "Give It to Us Straight (Most of the Time): Top Managers' Use of Concrete Language and Its Effect on Investor Reactions." *Strategic Management Journal* 39 (2018): 2204–2225.
- PAN, P., C. PATEL, AND R. MALA "Questioning the Uncritical Application of Translation and Back-Translation Methodology in Accounting: Evidence from China." *Corporate Ownership and Control* 12 (2015): 479–491.
- PEER, E., L. BRANDIMARTE, S. SAMAT, AND A. ACQUISTI "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research." *Journal of Experimental Social Psychology* 70 (2017): 153–163.
- RENNEKAMP, K. "Processing Fluency and Investors' Reactions to Disclosure Readability." *Journal of Accounting Research* 50 (2012): 1319–1354.
- RICHARDS, G., AND C. VAN STADEN "The Readability Impact of International Financial Reporting Standards." *Pacific Accounting Review* 27 (2015): 282–303.

- SAPIR, E. *Culture, language and personality: Selected essays* (Vol. 342) Univ of California Press (1985).
- SCHROEDER, N., R. AGGARWAL, AND C. GIBSON "Financial Reporting by Japanese Firms on the NYSE: An Analysis of Linguistic Content." *MIR: Management International Review* 31 (1991): 233–251.
- SHAH, A. K., AND D. M. OPPENHEIMER "Easy Does It: The Role of Fluency in Cue Weighting." *Judgment and Decision Making* 2 (2007): 371–379.
- SIANO, F., AND P. WYSOCKI "Transfer Learning and Textual Analysis of Accounting Disclosures: Applying Big Data Methods to Small(Er) Datasets." *Accounting Horizons* 35 (2021): 217–244.
- SONNEY, F. "Financial Analysts' Performance: Sector versus Country Specialization." *Review of Financial Studies* 22 (2009): 2087–2131.
- TAN, H. T., E. Y. WANG, AND B. ZHOU "When the Use of Positive Language Backfires: The Joint Effect of Tone, Readability, and Investor Sophistication on Earnings Judgments." *Journal of Accounting Research* 52 (2014): 273–302.
- TAN, H. T., E. Y. WANG, AND B. ZHOU "How Does Readability Influence Investors' Judgments? Consistency of Benchmark Performance Matters." *The Accounting Review* 90 (2015): 371–393.
- TEHSEEN, S., T. RAMAYAH, AND S. SAJILAN "Testing and Controlling for Common Method Variance: A Review of Available Methods." *Journal of Management Sciences* 4 (2017): 142–168.
- WHORF, B. L. "Language, thought, and reality: selected writings of....(Edited by John B. Carroll.)." (1956).

ZEFF, S. A. "Some Obstacles to Global Financial Reporting Comparability and Convergence at a High Level of Quality." *British Accounting Review* 39 (2007): 290–302.

Appendix

Appendix 1: Forecast Report - Deutsche Konsum Reit-AG (ISIN: DE000A14KRD3)

Taking absolute differences between the German and English reports from participants' judgments, Deutsche Konsum-REIT AG provided the forecast report with the smallest average of absolute differences (i.e., they provided the "best" translation on average). They ranked on similarity:

- readability: 2nd out of 61 (German version more readable)
- tone: 18th out of 61 (English version more positive)
- precision: 12th out of 61 (German version more precise)

German Ausblick und Prognose

Profitabilitätssteigerungen durch weiteres Portfoliowachstum

Die DKR konnte im ersten Halbjahr 2018/2019 bereits fast das Ankaufovolumen des gesamten vorherigen Geschäftsjahres erreichen und findet weiterhin attraktive Objekte, die den Investitionskriterien entsprechen. Insofern soll das Wachstumstempo weiterhin hoch bleiben und das Portfolio werterhöhend vergrößert werden. Zudem arbeitet die Gesellschaft intensiv an der Revitalisierung einzelner erworbener Immobilien, um hier stille Reserven zu realisieren. Dabei steht die Erzielung einer nachhaltig attraktiven Rendite im Vordergrund. Zum weiteren Aufbau des Immobilienportfolios wird die DKR die auf der ordentlichen Hauptversammlung beschlossenen Kapitalmaßnahmen einsetzen.

Die Fremdkapitalaufnahme zur Finanzierung des Portfolioaufbaus soll dabei im Rahmen des angestrebten Ziel-LTV von rund 50% begrenzt bleiben. Durch die mittlerweile stark verbesserte Bonität der DKR und die im Februar 2019 erfolgte Anhebung des Unternehmensratings werden sich bei zukünftigen Darlehensaufnahmen weiter leicht verbesserte Zinskonditionen ergeben, die den FFO entlasten

English Outlook and forecast

Profitability increases through further portfolio growth

In the first half year of 2018/2019, DKR was able to achieve almost the purchase volume of the entire previous financial year and continues to find attractive properties that meet the investment criteria. In this respect, the pace of growth should remain high and the portfolio should be increased in value. In addition, the Company is working intensively on the revitalisation of individual properties acquired in order to realise hidden reserves. The focus is on achieving a sustainably attractive return.

For the further development of the real estate portfolio, DKR will make moderate use of the capital decided on at the ordinary general meeting.

Borrowing to finance the portfolio build-up should remain limited within the target LTV of around 50%. Due to DKR's meanwhile much improved credit rating and the increase of the corporate rating in February 2019, future loan borrowing will continue to result in slightly improved interest rates, which will relieve the burden on FFO and further increase corporate profitability.

Earnings forecast confirmed

und die Unternehmensprofitabilität weiter steigern werden.

Ergebnisprognose bestätigt

Aufgrund des erwartungsgemäß guten Halbjahresergebnisses sowie den noch im zweiten Halbjahr folgenden Nutzen-/Lastenübergängen der erst erworbenen Immobilien bestätigen wir unsere Prognose und erwarten einen FFO zwischen EUR 26 Mio. und EUR 29 Mio. im Geschäftsjahr 2018/2019. Weiterhin bestätigen wir unsere Erwartung, eine FFO Run Rate zum 30. September 2019 von EUR 33 Mio. zu erzielen.

Based on the good half-year results as expected and the subsequent transfer of benefits and encumbrances in the second half of the year of the recently acquired real estate, we confirm our forecast and expect an FFO of between EUR 26 and 29 million in the 2018/2019 financial year. Furthermore, we confirm our expectation to achieve an FFO run rate of EUR 33 million as of 30 September 2019.

Appendix 2: Forecast Report - Delivery Hero (ISIN: DE000A2E4K43)

Taking absolute differences between the German and English reports from participants' judgments, Delivery Hero provided the forecast report with the largest average of absolute differences (i.e., they provided the “worst” translation on average). They ranked on similarity:

- readability: 61st out of 61 (German version more readable)
- tone: 43rd out of 61 (English version more positive)
- precision: 49th out of 61 (German version more precise)

German Ausblick 2019

Die Erwartungen für das globale Wachstum in 2019 und 2020 sind weiterhin stabil bzw. leicht positiv. Darüber hinaus erwartet Delivery Hero, von den strukturellen Trends im Umgang mit Technologie, Logistik und Lifestyle weiter zu profitieren.

Getrieben von der positiven Geschäftsentwicklung hat Delivery Hero seine Umsatzprognose für das Gesamtjahr 2019 am 19. Juni 2019 auf eine Spanne zwischen € 1,3 Mrd. und € 1,4 Mrd. angehoben. Vor dem Hintergrund des sich weiter fortsetzenden positiven Momentums mit einem höheren Level an Kundenakquisitionen, Bestellungen und Umsatzerlösen, erwartet die Gruppe für das Gesamtjahr 2019 Umsatzerlöse am oberen Ende der zuvor veröffentlichten Spanne.

Vor dem Hintergrund des beschleunigten Wachstums kündigte Delivery Hero im Juni 2019 an, für die zweite Jahreshälfte 2019 opportunistisch zusätzlich bis zu € 100 Mio. zu investieren, solange die Renditen attraktiv bleiben. Entsprechend wurde das erwartete negative adjusted EBITDA für das gesamte Jahr 2019 verglichen zur Prognose im Geschäftsbericht 2018 erhöht. Es wird ein adjusted EBITDA zwischen minus € 370 Mio. und minus € 420 Mio. erwartet. Es wird nach wie vor davon ausgegangen, dass das MENA

English Outlook 2019

Global growth expectation for 2019 and 2020 continues to be stable and slightly positive. Furthermore, Delivery Hero expects to further benefit from structural trends in the use of technology, logistics and lifestyle.

As a result of the strong business performance Delivery Hero raised its revenue guidance to between € 1.3 and € 1.4 billion for the Full Year 2019 on June 19, 2019. Given the continued positive momentum with higher levels of new customer acquisitions, orders and revenues, the Company expects to achieve full year revenues in line with the top end of the previously announced guidance range.

Based on the investment returns observed, in June 2019, Delivery Hero announced to opportunistically invest up to additional € 100 million in the second half of 2019 if returns remain attractive. Accordingly, the expected adjusted negative EBITDA for full year 2019 was raised compared with the outlook as disclosed in the annual report. It is expected to be between negative € 370 million and negative € 420 million. The MENA segment is still expected to contribute a positive adjusted EBITDA of € 70 million based on the expectation of significant operating profits from the strong underlying segment performance. One-of effects in MENA are not expected to be carried forward in H2

Segment ein positives adjusted EBITDA von € 70 Mio. beitragen wird, ausgehend von der Erwartung signifikanter operativer Gewinne aus der starken zugrunde liegenden Segmententwicklung. Einmaleffekte im Segment MENA über das erste Halbjahr hinaus werden nicht erwartet. Für Europa wird weiterhin erwartet, dass das Segment in der zweiten Jahreshälfte 2019 auf Basis des adjusted EBITDA den Breakeven erreichen wird.

Aufgrund der vergleichsweise kurzen Historie des Konzerns und der Tatsache, dass Delivery Hero in einem relativ neuen Markt agiert, ist eine Prognose der Ergebnisentwicklung mit erheblichen Unsicherheiten behaftet. Neben den Faktoren, die von Delivery Hero beeinflusst werden können, ist das adjusted EBITDA auch von Faktoren abhängig, die nicht beeinflusst werden können. Wenn der Konzern beispielsweise gezwungen wäre, seine Position gegen neue Wettbewerber in bestimmten Märkten zu verteidigen oder auf Umsatzeinbußen zu reagieren, müssen möglicherweise Maßnahmen ergriffen werden, die zuvor nicht geplant waren (z. B. steigende Marketingaufwendungen), die sich negativ auf das adjusted EBITDA auswirken und zu erheblichen Abweichungen von den geschätzten Ergebnissen führen können.

Die Annahmen zur wirtschaftlichen Entwicklung des Marktes und der Branche beruhen auf Einschätzungen, die das Management der Delivery Hero Gruppe nach derzeitigem Kenntnisstand für realistisch hält. Diese Schätzungen sind jedoch mit Unsicherheiten behaftet und bergen das unvermeidbare Risiko, dass die Prognosen weder in Richtung noch in Bezug auf das Ausmaß eintreten. Die Prognose für den Prognosezeitraum basiert auf der Zusammensetzung des Konzerns zum Zeitpunkt der Erstellung des Konzernzwischenlageberichts.

2019. For Europe expectation remains that the segment will reach breakeven on an adjusted EBITDA level during the second half of 2019.

Due to the comparatively short history of the Group and the fact that Delivery Hero is operating in a relatively new market, any forecast on the earnings trend is subject to considerable uncertainty. Besides factors that can be impacted by Delivery Hero, adjusted EBITDA is also contingent on factors that cannot be influenced. For example, if the Group were forced to defend its position against new competitors in specific markets or to react to revenue downturns, then measures which may not have been scheduled previously may have to be implemented (e.g. increasing marketing expenditure) which can negatively affect adjusted EBITDA and trigger considerable deviations from the estimated results.

The assumptions on the economic development of the market and the industry are based on assessments which the management of the Delivery Hero group considers realistic in line with currently available information. However, these estimates are subject to uncertainty and bring with them the unavoidable risk that the forecasts do not occur, either in terms of direction or in relation to extent. The forecast for the forecast period is based on the composition of the Group at the time the interim group management report was prepared.

Appendix 3: Survey Instructions

Dear participant,

Thank you very much for participating in our survey titled "Characteristics of Forecast Reports." We are [REDACTED]

Purpose and design of our study

This study examines characteristics of forecast reports issued by publicly listed firms. This is important since even though projections for the future may play an important role for investors and other stakeholders of the firm, there are still significant gaps in our knowledge about firms' forecast reports. Closing these gaps helps us understanding whether the market is truly a "level playing field" for all investors.

In the following, you are asked to evaluate certain characteristics of six different forecast reports. Please note that our sample covers different languages. We use the information you provided on Prolific to only give you forecast reports in languages that you are fluent in. You were selected for this study because of your previous investment in shares.

Please evaluate the reports based on the provided information and ignore all other information you may have about the companies' development after 2019 (e.g. COVID 19 crisis).

Compensation

The compensation for participating in this study is 9 GBP. We estimate the time to complete the survey to be approximately one hour, depending on the length of your reports. Each report is between 50 and 1000 words long (up to approximately two pages). We randomly draw six reports for you. Please read them carefully, even the longer ones.

Your participation in this study is completely voluntary and you can withdraw at any time. Please note however, that you will only get compensated when you complete the survey. You can take the survey only once and you are not able to edit your responses.

Personal information

We collect your Prolific ID for payment purposes. Besides that, the survey does not collect any identifying personal information. We will delete your Prolific ID after payment.

Further questions

If you have any further questions, please contact us via [REDACTED]

(Identifying author information redacted)

Appendix 4: Variable Definitions for Survey Questionnaire

Variable	Question / Statement	Likert Scale Endpoints (coded from 3 to -3)
Part 1: Investment Attractiveness		
<i>attractive_investment</i>	<i>After reading this text, how attractive are this firm's shares as an investment to you?</i>	Very attractive (3) – very unattractive (-3)
Part 2: Textual Characteristics		
<i>readability</i>	<i>How easy do you find it to extract relevant information from the text for making an investment decision?</i>	Very easy (3) – very difficult (-3)
<i>tone</i>	<i>How would you rate the overall tone of this text?</i>	Very positive (3) – very negative (-3)
<i>precision</i>	<i>How precise is the information presented in this text?</i>	Very precise (3) – very vague (-3)
Part 3: Control Questions / Statements		
<i>information_difficulty</i>	<i>Please evaluate the following statement: The firm has made this text intentionally difficult to understand.</i>	Strongly disagree (3) – strongly agree (-3)
<i>objectiveness</i>	<i>Please evaluate the following statement: The firm presents the information in this text objectively.</i>	Strongly agree (3) – strongly disagree (-3)
<i>firm_certainty</i>	<i>Please evaluate the following statement: The firm is certain about its future development.</i>	Strongly agree (3) – strongly disagree (-3)
<i>text_complexity</i> ²⁶	<i>What level of education do you think is required to fully understand the given text?</i>	No formal education (3) – Doctor (PhD) or equivalent (-2)
<i>tone_shareprice</i>	<i>How do you think will the overall tone of the presented information influence the share price of the firm?</i>	Strong positive influence (3) – strong negative influence (-3)
<i>individual_certainty</i>	<i>If someone asked you about the future development of the firm, how sure would you be about your answer?</i>	Very certain (3) – very uncertain (-3)

²⁶ For *text_complexity*, the Likert Scale only had 6 points: no formal education, primary school education, secondary school education, Bachelor of equivalent, Master or equivalent, Doctor (PhD) or equivalent.

Appendix 5: Variable Definitions for Demographic Questions

Variable	Definition
<i>foreign_origin</i>	Binary variable, which is coded as 0 if participant's current state of residence is Germany, Austria or Switzerland, and 1 otherwise.
<i>education</i>	Participants education categorized by 7-point Likert Scale from "No formal education" (0) to "Doctor (PhD) or equivalent" (6).
<i>female</i>	Binary variable, which is coded as 1 if participants are female, and 0 if the participants are male.
<i>read_financial_reports</i>	Binary variable, which is coded as 1 if participant's have read financial reports before, and 0 otherwise.
<i>never_invested</i>	Binary variable, which is coded as 1 if participant's have never directly or indirectly invested in shares before, and 0 otherwise.
<i>pension_investment</i>	Binary variable, which is coded as 1 if participant's have invested in shares through a pension plan before, and 0 otherwise.
<i>shares_investment</i>	Binary variable, which is coded as 1 if participant's have actively purchased shares before, and 0 otherwise.
<i>other_investment</i>	Binary variable, which is coded as 1 if participant's have actively purchased other investments (e.g. funds, bonds, options) before, and 0 otherwise.
<i>age_group</i>	<i>age_group</i> is coded based on age groups: 1: Participant's age < 21. 2: Participant's age between 21 and 30 years. 3: Participant's age between 31 and 40 years. 4: Participant's age between 41 and 50 years. 5: Participant's age between 51 and 60 years. 6: Participant's age > 61 years.
<i>trading</i>	<i>trading</i> is coded based on the reply how often participants trade shares: 0: Participant does not trade. 1: Participant trades up to once per year. 2: Participant trades up to once per quarter. 3: Participant trades up to once per month. 4: Participant trades up to once per week. 5: Participant trades multiple times per week.
<i>trading_amount</i>	<i>trading_amount</i> is coded based on the reply for how much money participants on average trade: 0: Participant does not trade. 1: Participant trades between 1 EUR and 100 EUR per trade. 2: Participant trades between 101 EUR and 500 EUR per trade. 3: Participant trades between 501 EUR and 1,000 EUR per trade. 4: Participant trades between 1,001 EUR and 5,000 EUR per trade. 5: Participant trades between 5,001 EUR and 20,000 EUR per trade. 6: Participant trades more than 20,000 EUR per trade.
7-point Likert Scale from "strongly disagree" (0) to "strongly agree" (6)	
<i>no_trust</i>	"I do not trust publications issued by companies."
<i>investor_care</i>	"Companies overall care about investors, including small retail investors."
<i>overly_positive</i>	"Most companies try to portray themselves in an overly positive manner."
<i>poor_management</i>	"Often, company management does not know what they are doing."

Appendix 6 : Correlation Matrix

Panel A: Participant-related control variables																				
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)	(m)	(n)	(o)	(p)	(q)	(r)	(s)	(t)
(a) <i>attractive_investment</i>		.36	.68	.27	-.01	.10	.01	.08	.02	.03	-.02	.02	.00	-.05	.01	.02	-.05	.12	.00	-.04
(b) <i>readability</i>	.38		.32	.42	-.10	.09	-.07	.03	-.03	.06	-.07	-.06	.08	.02	.08	.12	-.10	.17	-.02	-.07
(c) <i>tone</i>	.68	.34		.10	-.03	.06	-.01	.04	.00	.01	-.01	.02	.00	-.02	.00	.00	-.04	.12	.04	-.06
(d) <i>precision</i>	.28	.41	.10		-.01	.04	-.06	.00	-.04	-.06	.01	-.10	.00	-.02	-.01	.04	-.09	.14	-.03	-.03
(e) <i>foreign_language</i>	-.01	-.10	-.04	.00		.03	.01	.03	.07	-.03	-.04	.05	.00	.02	-.01	.02	.02	.00	.02	-.05
(f) <i>foreign_origin</i>	.11	.10	.06	.04	.03		.28	.36	.22	.03	-.03	.24	.07	-.14	-.02	.05	.01	.09	.09	.02
(g) <i>education</i>	.02	-.06	.01	-.06	.01	.29		.27	.29	.07	-.10	.22	.07	-.01	.08	.16	-.03	.07	.04	.02
(h) <i>female</i>	.08	.03	.05	.00	.03	.36	.28		.14	-.07	.08	.15	-.05	-.18	-.24	-.05	-.03	.13	.14	-.12
(i) <i>age_group</i>	.02	-.03	.00	-.03	.07	.21	.31	.15		.19	-.09	.27	.05	.03	-.09	.14	-.02	.01	-.06	.03
(j) <i>read_financial_reports</i>	.03	.06	.03	-.05	-.03	.03	.08	-.07	.19		-.22	.11	.17	.20	.24	.25	.02	-.08	-.03	.09
(k) <i>never_invested</i>	-.02	-.06	-.02	.01	-.04	-.03	-.12	.08	-.11	-.22		-.18	-.59	-.42	-.52	-.51	.05	-.05	.01	.04
(l) <i>pensions_invested</i>	.01	-.05	.02	-.09	.05	.24	.21	.15	.27	.11	-.18		.00	.07	-.01	.11	.04	-.02	.03	.02
(m) <i>shares_invested</i>	.01	.08	.00	.00	.00	.07	.08	-.05	.06	.17	-.59	.00		.15	.53	.45	-.07	.08	-.06	-.03
(n) <i>other_investments</i>	-.05	.02	-.02	.00	.02	-.14	.01	-.18	.05	.20	-.42	.07	.15		.33	.37	-.01	-.02	.03	-.02
(o) <i>trading_frequency</i>	.01	.08	.00	.00	-.01	-.05	.09	-.24	-.08	.24	-.52	-.02	.53	.32		.53	-.15	.07	.01	-.06
(p) <i>trading_amount</i>	.02	.11	.00	.05	.02	.03	.19	-.04	.17	.24	-.50	.11	.44	.37	.49		-.16	.16	.07	.00
(q) <i>no_trust</i>	-.05	-.10	-.04	-.09	.02	.01	-.03	-.02	.00	.02	.05	.04	-.07	-.01	-.16	-.17		-.34	.22	.34
(r) <i>investor_care</i>	.12	.16	.12	.14	.00	.08	.09	.13	.02	-.08	-.05	-.01	.07	-.03	.06	.15	-.32		-.09	-.30
(s) <i>overly_positive</i>	.01	.01	.07	-.01	.01	.13	.09	.17	-.05	-.02	.01	.06	-.05	.04	-.03	.05	.23	-.07		.21
(t) <i>poor_management</i>	-.04	-.07	-.06	-.03	-.04	.02	-.02	-.11	.01	.08	.04	.01	-.03	-.01	-.07	-.02	.33	-.31	.21	

Panel B: Survey-related control variables												
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	(j)	(k)	(l)
(a) <i>attractive_investment</i>		.36	.68	.27	-.01	.10	.24	.15	.51	.03	.72	.48
(b) <i>readability</i>	.38		.32	.42	-.10	.09	.57	.35	.31	.29	.35	.44
(c) <i>tone</i>	.68	.34		.10	-.03	.06	.28	-.01	.54	.12	.76	.39
(d) <i>precision</i>	.28	.41	.10		-.01	.04	.19	.47	.36	-.05	.24	.46
(e) <i>foreign_language</i>	-.01	-.10	-.04	.00		.03	-.05	.00	-.03	-.05	-.05	.00
(f) <i>foreign_origin</i>	.11	.10	.06	.04	.03		.05	.05	.02	-.17	.14	.06
(g) <i>information_difficulty</i>	.27	.57	.31	.18	-.06	.06		.24	.23	.39	.26	.26
(h) <i>objectiveness</i>	.16	.34	-.01	.47	.00	.06	.24		.23	-.03	.12	.34
(i) <i>firm_certainty</i>	.52	.33	.55	.36	-.03	.03	.25	.24		.10	.56	.61
(j) <i>text_complexity</i>	.03	.30	.13	-.05	-.05	-.16	.39	-.04	.10		.06	.08
(k) <i>tone_shareprice</i>	.72	.37	.74	.25	-.05	.15	.29	.13	.57	.06		.46
(l) <i>individual_certainty</i>	.50	.45	.41	.46	-.01	.06	.27	.34	.61	.09	.48	

Notes:

Appendix 6 presents Spearman correlations above and Pearson correlations below the diagonal for *attractive_investment* as well as the three main textual variables used in our analyses (*readability*, *tone*, *precision*). It furthermore includes *foreign_language*, a binary variable which is coded as 1 for the English translation and 0 for the German version of the firm's disclosure, as well as *foreign_origin* which is coded as 1 if the participant is located in a country outside the DACH region and 0 otherwise. Panel A shows correlations with demographic control variables presented in Appendix 5, while Panel B shows correlations with survey-related questions presented in Appendix 4. Definitions for all variables are provided in Appendix 4 and 5.

Tables

Table 1: Sample Selection

Panel A: Firms				
<i>Prime Standard Firms (as of 01.09.2019)</i>				320
<i>./. Duplicates – (e.g. preferred shares)</i>				12
<i>./. Foreign firms (by headquarter)</i>				31
<i>./. No half-year report available (e.g. insolvency)</i>				4
<i>./. Text not extractable from half-year reports</i>				4
<i>./. No separate forecast reports</i>				19
Final population				250
Random sample used for survey experiment				61
Panel B: Survey responses				
	Participants		Reports	
	<i>German</i>	<i>English</i>	<i>German</i>	<i>English</i>
<i>All responses</i>	106	106	595	648
<i>./. Participants with inconsistent demographic responses between Prolific and our survey</i>	1	18	60	54
<i>./. Participants who spent less than 10 minutes on the survey</i>	0	1	3	3
<i>./. Reports of participants without sufficient command of the German language</i>	0	1	24	0
Final sample	105	86	508	591

Notes:


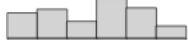








This table presents information on the sample selection process of chosen forecast reports as well as of survey participants. **Panel A:** The determination of the underlying population was based on firms included in the Prime Standard at 30.09.2019. We exclude duplicates (e.g. firms with preferred shares). We also exclude foreign firms since these are not required to disclose a German version of the half-year report. For four firms we were not able to convert the forecast report from PDF to a text file. Lastly, we excluded firms that had no separate forecast report in their half-year report (these firms issued combined reports, mostly joined with risks and opportunities). Out of the 250 remaining firms, we randomly draw 61 for our survey experiment. **Panel B:** Survey data was collected between the 2nd and 16th November 2020 on Prolific.co. Available at maximum were 383 active English native and 705 active German native participants meeting our criteria. Inconsistent responses between Prolific and our survey relate to participants providing a different native language to Prolific than in our survey. Participants without sufficient German abilities are participants who speak German on a level lower than intermediate. There were no participants without sufficient English abilities.

Table 2: Descriptive Statistics

Panel A: Firm and report characteristics														
Firm	Survey sample						Prime Standard						Diff. in means	p-value
	n	mean	sd	min	median	max	n	mean	sd	min	median	max		
market_cap (bEUR)	61	3.93	8.55	0.02	0.59	40.63	227	4.54	11.39	0.00	0.58	93.81	0.84	0.56
return_on_assets (%)	61	1.12	12.28	-67.33	3.40	21.80	239	2.26	10.23	-80.60	3.31	29.20	1.54	0.38
total_assets (bEUR)	61	14.56	62.24	0.01	0.60	458.16	239	24.77	116.33	0.01	0.80	1348.14	13.71	0.28
big4	61	0.26		0.00	0.00	1.00	250	0.29		0.00	0.00	1.00	0.04	0.55
domestic_revenues (%)	61	0.55	0.32	0.00	0.55	1.00	205	0.64	1.32	0.00	0.53	18.97	0.13	0.34
Report														
number_words_GER	61	363.23	221.80	54.00	305.00	836.00	250	401.75	362.38	5.00	296.50	2732.00	50.95	0.21
number_words_ENG	61	418.30	258.87	62.00	344.00	993.00	250	461.59	415.61	5.00	339.50	3073.00	57.27	0.22
readability_measure	61	33.23	9.55	5.34	34.30	54.28	250	31.83	41.72	-581.30	34.51	78.73	-1.86	0.61
tone_measure	61	0.55	0.30	0.00	0.57	1.00	250	0.51	0.31	0.00	0.50	1.00	-0.05	0.28
precision_measure	61	-0.01	0.01	0.00	-0.01	-0.03	250	-0.01	0.01	0.00	-0.01	-0.06	0.00	0.05

Panel B: Participant characteristics																				
	Whole sample						Domestic participants						Foreign participants						Diff in means	p-value
	n	mean	sd	min	median	max	n	mean	sd	min	median	max	n	mean	sd	min	median	max		
education	187	4.23	1.00	2	4	6	88	3.93	0.96	2	4	5	99	4.49	0.97	2	5	6	0.56	<0.01
female	191	0.37		0	0	1	89	0.19		0	0	1	102	0.53		0	1	1	0.34	<0.01
age_group	191	1.93	1.24	0	2	5	89	1.64	1.06	0	1	4	102	2.19	1.34	0	2	5	0.55	<0.01
read_financial_reports	191	0.77		0	1	1	89	0.75	0.43	0	1	1	102	0.78		0	1	1	0.03	0.61
never_invested	191	0.12		0	0	1	89	0.13	0.34	0	0	1	102	0.11		0	0	1	-0.03	0.57
pensions_invested	191	0.26		0	0	1	89	0.15	0.36	0	0	1	102	0.35		0	0	1	0.21	<0.01
shares_invested	191	0.71		0	1	1	89	0.67	0.47	0	1	1	102	0.74		0	1	1	0.06	0.36
other_investments	191	0.66		0	1	1	89	0.73	0.45	0	1	1	102	0.61		0	1	1	-0.12	0.07
trading_frequency	182	1.88	1.35	0	2	5	85	1.91	1.27	0	2	4	97	1.87	1.42	0	2	5	-0.04	0.84
trading_amount	181	2.30	1.61	0	2	6	84	2.21	1.52	0	2	5	97	2.38	1.68	0	2	6	0.17	0.48
no_trust	191	2.46	1.28	0	2	5	89	2.44	1.23	0	2	5	102	2.48	1.33	0	2	5	0.04	0.82

Panel B: Participant characteristics																		
	Whole sample						Domestic participants						Foreign participants					
	n	mean	sd	min	median	max	n	mean	sd	min	median	max	n	mean	sd	min	median	max
<i>investor_care</i>	191	2.95	1.52	0	3	6	89	2.81	1.41	0	3	5	102	3.08	1.61	0	3	6
<i>overly_positive</i>	191	4.68	0.98	0	5	6	89	4.60	0.90	3	5	6	102	4.75	1.04	0	5	6
<i>poor_management</i>	191	2.59	1.42	0	3	6	89	2.57	1.41	0	3	6	102	2.61	1.44	0	3	6

Panel C: Survey responses									
	n	mean	sd	min	P25	median	P75	max	histogram
<i>attractive_investment</i>	1085	0.39	1.60	-3	-1	1	2	3	
<i>readability</i>	1097	0.33	1.65	-3	-1	1	2	3	
<i>tone</i>	1095	0.82	1.39	-3	0	1	2	3	
<i>precision</i>	1020	0.23	1.58	-3	-1	1	2	2	
<i>information_difficulty</i>	1093	0.96	1.59	-3	0	1	2	3	
<i>objectiveness</i>	1091	0.61	1.45	-3	-1	1	2	3	
<i>firm_certainty</i>	1098	0.71	1.48	-3	0	1	2	3	
<i>text_complexity</i>	1095	0.43	0.80	-2	0	0	1	3	
<i>tone_shareprice</i>	1072	0.45	1.32	-3	0	1	1	3	
<i>individual_certainty</i>	1092	0.13	1.53	-3	-1	1	1	3	

Notes:

Panel A provides information for all Prime Standard firms (as of 30th of September 2019) providing forecast reports, once for all Prime Standard firms and our survey sample. Differences in means are tested for statistical significance via t-tests between the survey sample and the remaining Prime Standard firms. Financial information is reported on the firm level as of 31st of December 2018. The *market capitalization (bEUR)* is taken from Dafne, *return on assets (%)*, *total assets (bEUR)*, and the share of revenues generated in the DACH region divided by total revenues (*domestic_revenues (%)*) stem from CapitalIQ. Big4 indicates whether the 2019 half-year report was audited by a big4 auditor and is hand-collected. Panel A further provides information about textual characteristics of the firms' forecast reports. It presents the number of total words in both languages (*number_words_GER* and *number_words_ENG*), as well as summary statistics for the Flesch Reading Ease readability measure (*readability_measure*). *tone_measure* is computed as the number of positive words divided by the sum of positive and negative words and *precision_measure* is computed as the number of uncertainty

Pabel C: Survey responses

expressions divided by the total amount of words, times minus one (both as defined in the *Loughran & McDonald (2014)* word list). **Panel B** shows participant responses to demographic questions asked at the end of the survey. Variable definitions including survey questions are presented in Appendix 5. We show differences between 89 domestic and 103 foreign participants. Note that investor origin and native language do not have a complete overlap. We used native language for screening participants (Table 1) in order to make sure they understand the firm disclosures. From thereon, we split participants by country of residence and not native language as this better distinguishes between domestic and foreign investors. **Panel C** shows participant responses to textual characteristics-related questions of the forecast reports. A description of the underlying questions is presented in Appendix 4. The answers are based on a total of 1,099 observations. The number of observations is usually lower since participants had the option not to give an answer.

Table 3: Differences between German and English Reports

Panel A: Participant judgments of German and English reports														
	German reports						English reports						Diff in means	p-value
	n	mean	sd	min	median	max	n	mean	sd	min	median	max		
<i>attractive_investment</i>	503	0.41	1.61	-3	1	3	582	0.38	1.59	-3	1	3	-0.03	0.74
<i>readability</i>	507	0.50	1.64	-3	1	3	590	0.19	1.64	-3	0	3	-0.32	<0.01
<i>tone</i>	507	0.86	1.43	-3	1	3	588	0.78	1.36	-3	1	3	-0.09	0.30
<i>precision</i>	470	0.23	1.56	-3	1	2	550	0.22	1.59	-3	1	2	-0.02	0.87
<i>information_difficulty</i>	505	1.05	1.62	-3	2	3	588	0.89	1.57	-3	1	3	-0.16	0.09
<i>objectiveness</i>	502	0.61	1.47	-3	1	3	589	0.60	1.43	-3	1	3	0.00	0.97
<i>firm_certainty</i>	507	0.76	1.46	-3	1	3	591	0.67	1.49	-3	1	3	-0.09	0.34
<i>text_complexity</i>	507	0.48	0.80	-2	0	3	588	0.40	0.79	-2	0	3	-0.08	0.10
<i>tone_shareprice</i>	495	0.52	1.32	-3	1	3	577	0.40	1.32	-3	1	3	-0.12	0.14
<i>individual_certainty</i>	507	0.13	1.57	-3	1	3	585	0.13	1.50	-3	1	3	0.00	0.99

Panel B: Standard deviations of participant judgments per report														
	German reports						English reports						Diff in means	p-value
	n	mean	sd	min	median	max	n	mean	sd	min	median	max		
<i>SD attractive_investment</i>	61	1.36	0.45	0.52	1.34	2.21	61	1.33	0.42	0.00	1.35	2.04	-0.03	0.71
<i>SD readability</i>	61	1.51	0.45	0.45	1.66	2.39	61	1.52	0.38	0.58	1.58	2.71	-0.00	0.99
<i>SD tone</i>	61	1.09	0.40	0.33	1.04	1.83	61	1.10	0.31	0.53	1.01	2.15	0.01	0.86
<i>SD precision</i>	61	1.40	0.40	0.52	1.41	2.51	61	1.39	0.40	0.46	1.45	2.12	-0.00	0.95
<i>SD information_difficulty</i>	61	1.41	0.44	0.52	1.52	2.33	61	1.39	0.43	0.52	1.40	2.50	-0.02	0.81
<i>SD objectiveness</i>	61	1.39	0.36	0.58	1.46	2.04	61	1.33	0.35	0.49	1.37	2.02	-0.05	0.39
<i>SD firm_certainty</i>	61	1.35	0.35	0.49	1.38	2.08	61	1.34	0.32	0.53	1.34	1.99	-0.01	0.84
<i>SD text_complexity</i>	61	0.73	0.21	0.44	0.69	1.29	61	0.71	0.23	0.00	0.69	1.41	-0.02	0.62
<i>SD tone_shareprice</i>	61	1.08	0.38	0.00	1.12	1.72	61	1.14	0.31	0.45	1.14	2.30	-0.06	0.31
<i>SD individual_certainty</i>	61	1.52	0.35	0.60	1.52	2.17	61	1.42	0.35	0.00	1.47	2.16	-0.10	0.13

Notes:

This table presents statistics on participant judgments to our ten main survey questions (see Appendix 4 for a description of the underlying questions). **Panel A** is based on the total 1,099 observations, of which 508 (591) are based on German (English) reports. The number of observations is usually lower since participants had the option not to answer a question. Differences in means are tested for statistical significance via t-tests. For **Panel B**, standard deviations to each of our 10 main survey questions (see Appendix 4) are calculated on the report level (i.e. per language per firm). Panel B presents the resulting summary statistics. Again, t-tests are employed to test for differences in means. Note that *text-complexity* only has 6 points on the Likert scale, so the standard deviation is mechanically lower than for the other variables.

Table 4: Effect of Language on the Perception of Textual Characteristics

	Without fixed effects				With fixed effects			
	<i>attractive_investment</i>	<i>readability</i>	<i>tone</i>	<i>precision</i>	<i>attractive_investment</i>	<i>readability</i>	<i>tone</i>	<i>precision</i>
<i>foreign_language</i>	-0.038 (0.140)	-0.407** (0.143)	-0.099 (0.122)	-0.178 (0.141)	-0.019 (0.131)	-0.480*** (0.138)	-0.047 (0.110)	-0.171 (0.136)
<i>foreign_origin</i>	0.336* (0.142)	0.235 (0.145)	0.152 (0.124)	-0.026 (0.145)				
<i>foreign_language x foreign_origin</i>	-0.006 (0.194)	0.152 (0.198)	0.012 (0.169)	0.310 (0.198)	0.033 (0.199)	0.309 (0.209)	-0.020 (0.144)	0.305 (0.200)
<i>Participant FE</i>	No	No	No	No	Yes	Yes	Yes	Yes
<i>Firm FE</i>	No	No	No	No	Yes	Yes	Yes	Yes
<i>Num.Obs.</i>	1085	1097	1095	1020	1085	1097	1095	1020
<i>R2</i>	0.011	0.019	0.004	0.004	0.448	0.416	0.500	0.451
<i>R2 Adj.</i>	0.008	0.016	0.001	0.001	0.281	0.242	0.351	0.271
<i>AIC</i>	4093.0	4194.6	3838.4	3828.4	3956.1	4121.1	3579.6	3716.8

Notes:

Table 4 presents OLS regressions with *attractive_investment* as well as our three main textual characteristics of interest as dependent variables. The last four regressions include firm- and participant-fixed effects. Note that firm-FE are not the same as report-FE as each firm issues their forecast report in two languages. The maximum number of observations is 1,099. It is usually lower since participants had the option not to give an answer. Regressions based on complete observations only ($n = 978$) yield similar results. *foreign_language* has no detectable effect different from zero on the other text-related characteristics presented in Appendix 4. Standard errors are presented in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Textual Characteristics and Investment Attractiveness

	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>
<i>readability</i>	-0.02 (0.05)	0.01 (0.05)	0.10 (0.08)	0.12 (0.08)	0.12 (0.11)
<i>tone</i>	0.71*** (0.04)	0.78*** (0.05)	0.32** (0.13)	0.53*** (0.14)	0.46*** (0.16)
<i>precision</i>	0.12*** (0.04)	0.11** (0.05)	0.06 (0.09)	-0.08 (0.10)	-0.17 (0.11)
<i>information_difficulty</i>					-0.04 (0.11)
<i>objectiveness</i>					-0.02 (0.13)
<i>firm_certainty</i>					0.11 (0.12)
<i>text_complexity</i>					-0.11 (0.19)
<i>tone_shareprice</i>					0.20 (0.15)
<i>individual_certainty</i>					0.27** (0.13)
<i>foreign_language</i>					0.03 (0.10)
<i>Participant FE</i>	No	Yes	No	Yes	Yes
<i>Firm FE</i>	No	No	Yes	Yes	Yes
<i>Num.Obs.</i>	1085	1085	1085	1085	1085
<i>R2</i>	0.206	0.385	0.275	0.458	0.468
<i>R2 Adj.</i>	0.204	0.252	0.231	0.293	0.300
<i>AIC</i>	3854.5	3957.0	3875.4	3939.9	3934.4
<i>F</i>	93.542	2.893	6.161	2.778	2.787

Notes:

Table 5 presents OLS regressions with *attractive_investment* as the dependent variable. The independent variables are based on our survey questions related to textual

characteristics (see Appendix 4). While the dependent variable is taken from each participant, the independent variables have been replaced by the median survey response per report (i.e. per firm, per language). This controls for common method bias. Using regressions without controlling for common method bias shows a positive and significant relationship for *tone*, *precision*, and *readability* with *attractive_investment* ($p < 0.05$ and smaller across all specifications, except with full controls where readability: $p < 0.1$). Regressions II-V include firm-FE and participant-FE in different combinations. Note that firm-FE are not the same as report-FE as each firm issues their forecast report in two languages. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Horserace between Textual Characteristics and Participants' Judgments

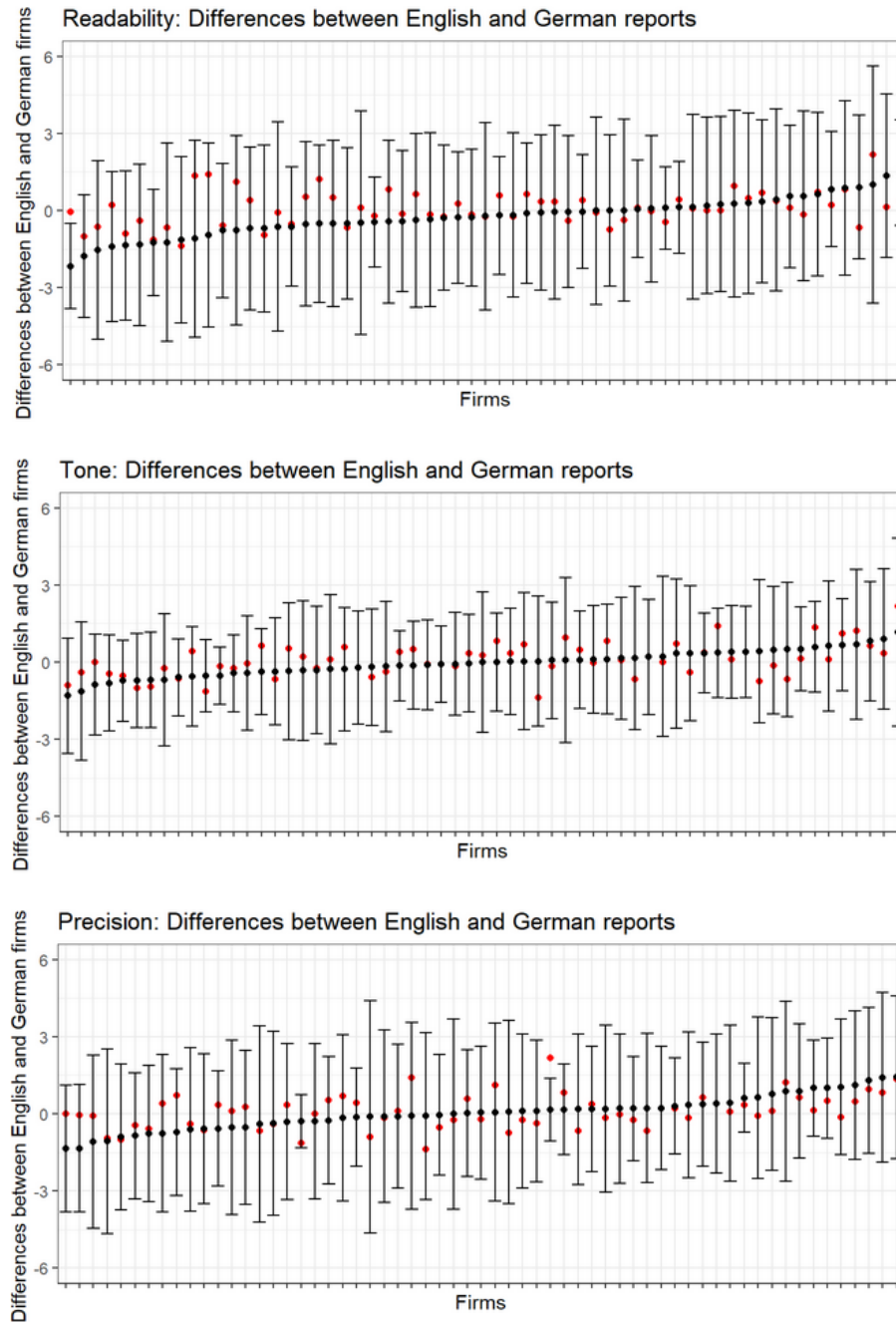
	Textual characteristics				Participant judgment			
	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>	<i>attractive_investment</i>
<i>readability</i>	-0.02 (0.02)			-0.02 (0.01)	0.03 (0.15)			-0.09 (0.11)
<i>tone</i>		1.61*** (0.46)		1.46*** (0.48)		0.87*** (0.10)		0.89*** (0.10)
<i>precision</i>			27.11 (26.42)	13.87 (24.73)			-0.02 (0.14)	0.09 (0.10)
<i>return_on_assets</i>	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	-0.02* (0.01)	0.00 (0.01)	-0.02* (0.01)
<i>log(market_cap)</i>	0.03 (0.10)	0.09 (0.09)	0.06 (0.10)	0.08 (0.09)	0.04 (0.10)	0.06 (0.06)	0.04 (0.10)	0.04 (0.07)
<i>domestic_revenues</i>	0.40 (0.49)	0.49 (0.45)	0.38 (0.50)	0.42 (0.46)	0.43 (0.50)	0.10 (0.32)	0.45 (0.50)	0.11 (0.33)
<i>big4</i>	0.09 (0.38)	0.27 (0.35)	-0.01 (0.38)	0.27 (0.36)	0.03 (0.39)	0.15 (0.25)	0.03 (0.39)	0.11 (0.25)
Num.Obs.	61	61	61	61	61	61	61	61
R2	0.060	0.196	0.035	0.216	0.017	0.590	0.016	0.599
R2 Adj.	-0.026	0.123	-0.053	0.112	-0.073	0.553	-0.073	0.546
AIC	192.6	183.1	194.2	185.5	195.4	141.9	195.4	144.7
BIC	207.4	197.8	209.0	204.5	210.1	156.7	210.1	163.7
Log.Lik.	-89.312	-84.531	-90.113	-83.767	-90.675	-63.969	-90.680	-63.339
F	0.697	2.682	0.394	2.085	0.186	15.849	0.184	11.295

Notes:

Table 6 presents results for OLS regressions using our complete set of 61 sample firms. Only English reports are used since the textual characteristics measures are only applicable for English texts. *attractive_investment* is the median of all participant judgments per firm (again, English reports only). The “textual characteristics” columns present results using the Flesch Reading Ease measure for *readability*, the number of positive words divided by the sum of positive and negative words for *tone*, and the number of uncertainty expression divided by the total number of words times minus one for *precision* (tone and precision words as defined in the Loughran & McDonald (2014) word list). In the “participant judgment” columns, *readability*, *tone*, and *precision* are the median values on the firm level (English reports only) to the survey questions presented in Appendix 4. Results are similar when mean values are used, with the exception of *readability* measured via the Flesch Reading Ease measure being negatively associated with *attractive_investment* on the 10 % significance level. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Figures

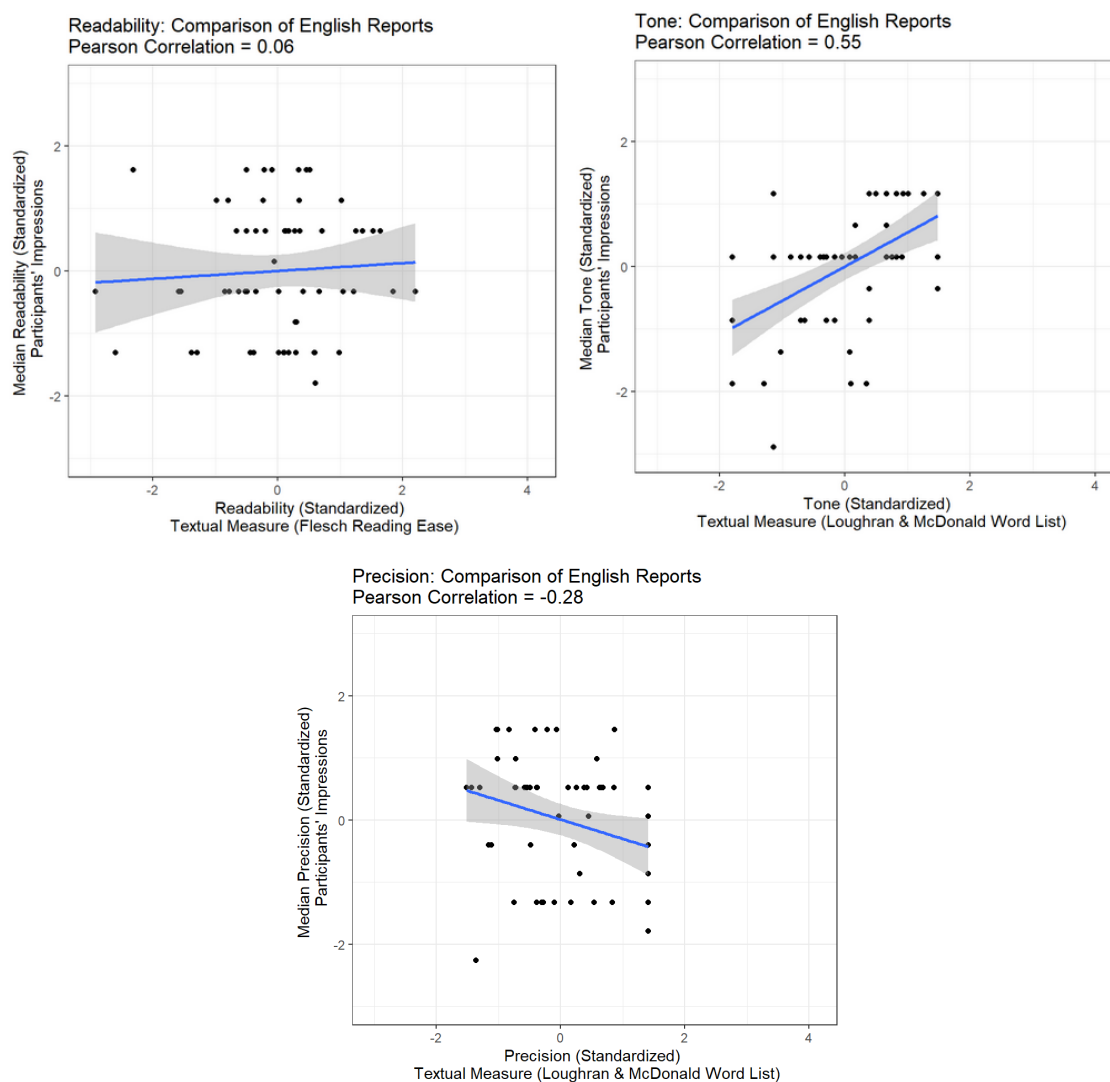
Figure 1: **Differences between English and German Reports**



Notes:

Figure 1 presents differences between the means of participants' judgments of the English and German reports (English minus German; black dots) of each firm, sorted by the size of differences of each textual characteristic. The whiskers represent the differences if either language report was judged one standard deviation higher or lower (i.e. the mean judgment of the English report moved one standard deviation upwards and the mean judgment of the German report one standard deviation downwards, respectively vice versa). The red dots show the differences in means of participants' judgments between English and German reports for *attractive_investment*.

Figure 2: Investor Responses and Textual Measures



Notes:

Figure 2 presents the distribution of the median participant perceptions per firm for the 61 firms included in our sample (English reports only), plotted against the results of textual measures for readability, tone, and precision. *Readability* is computed as the result of the Flesh Reading Ease readability measure. *Tone* is computed as the number of positive words divided by the sum of positive and negative words (as defined in the *Loughran & McDonald (2014)* word list). *Precision* is computed as the number of uncertainty expression divided by the total amount of words, times minus to allow higher values to indicate higher precision (words as defined in the *Loughran & McDonald (2014)* word list). Both the participant responses and the textual measures are standardized, i.e. de-measured and divided by the standard deviation. One dot represents one firm. The continuous line presents the regression line. The grey area presents the 95% confidence interval.

Formative Online Assessments and Student Performance

Rico Chaskel

Humboldt University of Berlin

Joachim Gassen

Humboldt University of Berlin

Abstract

This study investigates the effect of formative online assessments on student performance in an entry-level cost accounting class at a German public university. We conduct a randomised experiment that offers students access to online assessments. We exploit the fact that not all students voluntarily join the assessments and randomly assign them to a control condition as well as two treatment conditions: a continuous learning environment in which students solve the online assessments within two weeks, and a flexible learning environment in which students can solve the assessments at any time. Actual test-taking is still voluntary. We show that formative online assessments positively influence student performance for students who do not join the tests voluntarily and who are in the continuous learning environment. We do not find significant effects for all other students. The naïve treatment effect of test-taking on student performance is significantly positive, highlighting the need for randomised studies in order to avoid upward biased results driven by self-selection into treatment.

1 Introduction

Formative online assessments¹ are a cost-effective tool to provide students with additional material to increase their learning outcomes (Peat and Franklin [2002]). Multiple studies across disciplines point towards advantages of formative online assessments and find positive associations between their use and student performance (Sotola and Crede [2021]). The positive effects, however, are not entirely obvious. Conceptually, online assessments are prone to being used inappropriately (e.g. through memorizing questions rather than learning about the content; Brothen and Wambach [2001]) and procrastination may exhaust any positive effects (Häfner et al. [2014]). Methodologically, virtually all studies examining the relationship between online assessments and student performance struggle with unobserved confounders in their analyses (Einig [2013]), potentially leading to severely overstated effects (Angus and Watson [2009]). Even carefully selected control variables may not be able to fully control for unobservable confounders.

We contribute to the literature by creating a randomised experimental setting that allows us to identify the causal intention-to-treat effect (ten Have et al. [2008]) of offering online assessments on student performance. In our design, we are able to overcome ethical concerns that often preclude researchers from engaging in randomised trials in the field (Marriott and Lau [2008]). In our setting, we furthermore differentiate between the use of online assessments offered in a continuous learning environment (i.e. with fixed time intervals for solving the tests), and a flexible learning environment where students can freely choose when they solve the assessment.

¹ According to Black and Wiliam [2009, p. 9], an assessment is formative “to the extent that evidence about student achievement is elicited, interpreted, and used by teachers, learners, or their peers, to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions they would have taken in the absence of the evidence that was elicited”. In contrast, a summative assessment is a simple “judgement which encapsulates all the evidence up to a given point” (Taras [2005, p. 468]), for instance an exam at the end of the year without feedback intentions. Furthermore, we focus on online assessments as part of blended learning rather than pure online learning. In the following, we use the terms “assessment” and “test” synonymously.

At the beginning of the semester, we offer students the possibility to voluntarily sign up for online tests offered throughout the semester. We exploit the fact that some students do not show any interest in the online tests and allocate them to a treatment and control condition. In the treatment condition, students have - despite not signing up for the tests - the opportunity to participate in the online tests. In the control condition, students do not have the opportunity to join the online tests. Students who participate voluntarily as well as students in the treatment condition who do not join voluntarily are randomly assigned to a continuous learning environment and a flexible learning environment. In the continuous learning environment, students are required to solve the formative online tests within a time frame of two weeks. The flexible learning environment does not have a time constraint and all tests stay online until the exam. Students in the two treatment conditions still have the choice not to participate as all tests are voluntary and non-participation is not penalised.

Our results indicate that offering formative online assessments positively influences student performance (measured as their score in the final exam), but only for students who do not join the tests voluntarily and who are in the continuous learning environment. We do not find a significant relationship between offering the online tests and student performance for the other students. Interestingly, the naïve treatment effect of test-taking on exam performance is significantly positive across all specifications, even when we control for student motivation. This highlights the need for randomised studies to explore the true effect of online assessments on student performance.

2 Literature and Hypotheses Development

2.1 Participation in Formative Online Assessments and Student Performance

Eliciting the true effect of formative online assessments is difficult since randomized student trials are largely unfeasible and research suffers from students' self-selection into

treatment (Angus and Watson [2009]; Einig [2013]). Kibble [2007] compares online assessment participation and performance of students with different participation incentives. Without incentives, about half the students joined the two voluntary assessments that were offered. With just a small incentive of 0.5 % of the overall mark per completed quiz, the participation rate increased to nearly 90 %. He further observed that some students began using the quizzes inappropriately, i.e. not in order to gain feedback, but simply to obtain credit points. This indicates that students likely self-select into treatment. Potential drivers of self-selection are factors that drive both voluntary test participation and student performance. Among those are prior (academic) achievements, confidence in one's own skills and abilities, and motivation (Byrne and Flood [2008]; Davidson [2002]). Scholars commonly try to address these confounders with control variables (e.g., Angus and Watson [2009]; Massoudi et al. [2017]). However, as Einig [2013] notes, a truly randomized experiment would be preferable, not least because some confounders are inherently hard to measure (e.g. motivation or confidence in one's own abilities; e.g., Chak and Fung [2015]). Reasons that have kept researchers so far from conducting a randomised trial include ethical concerns (e.g., Marriott and Lau [2008]), potentially confounding communication between treatment groups (e.g., Cleaveland and Larkins [2004]), and practical issues (e.g., Einig [2013]). We aim at extending the literature by establishing a causal relationship between formative online assessments and student performance.

Online assessments offer a variety of advantages such as low administrative costs, quick and easy marking, the potential use of embedded media resources, as well as the provision of timely and frequent feedback (Angus and Watson [2009]; Dufresne et al. [2002]; Einig [2013]; Peat and Franklin [2002]). Most importantly however, the use of online tests throughout the semester may be correlated with improved student performance² across disciplines (e.g.,

² The studies typically measure performance as the students' marks in the final exam. The effect of online tests on performance appears to be comparable to, if not better than, paper-based tests (Bonham et al. [2003]; Dufresne et al. [2002]; Gok [2011]).

Buchanan [2000]; Chak and Fung [2015]; Einig [2013]; Grimstad and Grabe [2004]; Kibble [2007]; Lowry [2005]; Massoudi et al. [2017]; Peat et al. [2005]; Velan et al. [2002]) – although some studies find no or even negative effects (Nagandla et al. [2018]; Sotola and Crede [2021]).

A strong argument for using formative online assessments lies in their ability to provide timely feedback to students and thereby guide their learning (Einig [2013]). Henly [2003] shows that students who access online tests more often on average perform better. Kibble [2007] observes that students who excel in online tests but only access them once (i.e., do not gather feedback), perform worse in the final summative assessment compared to students who access them multiple times. Finally, formative online assessments also provide valuable feedback to instructors who gain a sense of the students' progress and need for repetition (Vos [2000]).

Another advantage of formative online assessments is that they facilitate students to engage in active learning³. A large body of evidence supports the notion that active learning enhances student performance (Prince [2004]). For instance, Riley and Ward [2017] find that individual active learning is associated with better student performance in an accounting information systems setting. Potter and Johnston [2006] use an online learning system designed for active learning in a cost management course and find that students who use the system more frequently have a better exam performance. Active learning further supports deep learning⁴ (Duff and McKinstry [2007]), which is linked to enhanced student performance (Byrne et al. [2002]; Davidson [2002]) and appears to be related to success in the accounting profession (Sharma [1997]). Research shows that instructors can actively trigger deep learning in students.

³ Following Bonwell and Eison [1991], Prince [2004] (p. 223) defines active learning as any instructional method that engages students in the learning process. In short, active learning requires students to do meaningful learning activities and think about what they are doing. However, the author supports the notion of viewing active learning as an approach and not a particular method.

⁴ Hall et al. [2004] (p. 491) define deep learning, as opposed to surface learning, as “a personal commitment to learning and an interest in the subject. The student approaches learning with the intention to understand and seek meaning and, consequently, searches for relationships among the material and interprets knowledge in the light of previous knowledge structures and experiences”.

For instance, Hall et al. [2004] as well as English et al. [2004] observe that the course design can alter students' approaches to learning. Finally, Gikandi et al. [2011] view formative online assessments as a possible component in facilitating deep learning.

Despite strong arguments and a large body of evidence in favour of online assessments, a causal relationship is not obvious because online assessments still may have no or even unwanted effects. First, they are usually unsupervised and their inappropriate use by students may hamper the valuable feedback component of formative assessments (Brothen and Wambach [2001]; Daniel and Broida [2004]; Henly [2003]; Kibble [2007]). Second, as Angus and Watson [2009] point out, online assessments need to be accepted by students in order to be effective. While most studies find high student satisfaction with additional online tests (e.g., Kibble [2007]; Marriott and Lau [2008]; Velan et al. [2002]), satisfaction as well as student performance is influenced by the test design (Brothen and Wambach [2001]; Ricketts and Wilks [2002]). Furthermore, acceptance in terms of participation rates largely seems to decrease over the course of the semester (Henly [2003]; Massoudi et al. [2017]). Finally, largely standardized online tests may fail to facilitate deep learning. Instead, they may encourage surface learning and promote answering strategies corresponding to the type of question, rather than the actual content (Paxton [2000]; Scouller [1998]).

Our study aims to provide causal evidence on whether formative online assessments improve student performance. We phrase the following hypothesis:

H1: Formative online assessments improve student performance.

2.2 Continuous Learning and Student Performance

Students seem to recognise certain advantages of continuously studying throughout the semester (Marriott and Lau [2008]). Still, they often appear to postpone their study efforts and leave large portions of their work for a brief period right before important deadlines (e.g. mid-

semester tests or the final examination; Azorlosa and Renner [2006]; Häfner et al. [2014]; Jar-molowicz et al. [2010]). Scholars point out numerous reasons why this may lead to adverse consequences, including a lack of timely feedback (Einig [2013]), a tendency to adopt a sur-face-approach to learning (Orpen [1998]), a high workload for teaching staff towards the end of the semester (Morris et al. [1978]), ineffective classroom time due to a lack of student prep- aration (Azorlosa and Renner [2006]), and ultimately a worse performance in the final exam (Rotenstein et al. [2009]).

Despite these disadvantages, it is still unclear whether continuous learning leads to en- hanced student performance and whether instructors should take action to foster continuous learning. For instance, while Azorlosa and Renner [2006] find that frequent quizzes during the semester increase student learning time, student performance in the final exam is not enhanced. Morris et al. [1978] randomly assign students to a self-paced and an instructor-paced group. In the latter, students are incentivized by marks to complete practice tests continuously throughout the semester. They do not find any differences in final exam performance between the groups. Einig [2013] lets one group of students have access to online tests for two weeks per test and another one for the whole semester. She finds that restricted access does not lead to signifi- cantly improved performance. Reiser [1984] finds that penalising students for not participating in regular self-instructional units decreases procrastination, but neither a penalty, nor a reward for continuous learning, significantly affects exam performance.

Perrin et al. [2011] argue that adverse consequences from not participating in continuous learning activities may lead to negative reinforcement and thereby fail to improve students' results. They find small-sample evidence (11 subjects), that positive reinforcement to work continuously increases intra-semester test scores (however, the authors do not look at the final examination). Hadsell [2009] finds that continuously offered quizzes in close proximity to the respective lectures moderately increase student performance in the final exam. He shows that

this result is not due to the feedback component of the tests. Finally, Grove and Wasserman [2006] use a natural experiment and find that freshmen students who are incentivized to learn continuously (via marked problem sets) perform significantly better in the final exam.

We add to the literature by conducting a randomized experiment on the influence of continuous learning on student performance. We are able to provide evidence on the difference between students in a flexible learning environment without deadlines, students in a continuous learning environment with deadlines, as well as a control group of students without any additional online tests. This is a significant improvement compared to previous studies (Grove and Wasserman [2006]) that allows us to identify a causal effect. In line with the theoretical arguments provided above, we hypothesize the following:

H2: Continuous learning improves student performance.

3 Research Design

3.1 Experimental Design

Our experiment was conducted in an introductory management accounting class (cost accounting), aimed at undergraduate students in their second semester. We chose this class since it is compulsory for all students pursuing a business degree at the faculty of business and economics and thus promised a sufficiently large participant pool. The summer semester spans over 14 teaching weeks and concludes with two exam periods at the end of the semester. Students who fail the exam during the first period are allowed to repeat it during the second period. Students are also allowed to directly enrol for the second exam period without sitting the first exam. The content of the first exam is different from the second exam.

Self-selection into treatment makes it difficult to find a causal effect between online assessments and exam performance. High-achieving or motivated students might not just opt for

additional training, but might also be intrinsically better equipped for final exams. In order to overcome this concern and elicit a causal relationship, we followed the suggestion by Einig [2013] and conducted our experiment with randomized allocation to control and treatment groups. For ethical reasons we did not exclude any student from participating in the online assessments. Hence, we invited all students to join our online tests on multiple occasions at the start of the semester: in lectures, tutorials, as well as via email announcements (see appendix 3 for the email invitation). Students were told that we would (anonymously) collect data on assessment participation and exam results. They had eight days to register for the online assessments on the university platform Moodle.

As could be expected from previous studies (e.g., Kibble [2007]), not all students would voluntarily participate in the assessments. We exploited the fact that some students opted against participation and randomly assigned two thirds of them to have access to the online tests anyway. These students, as well as the students who voluntarily decided to join the online tests, were then randomly assigned to a continuous learning environment and a flexible learning environment (we used within-bin-randomization to achieve an equal gender distribution).⁵ The remaining third of the non-voluntary student group did not get access to the online assessments and acts as our control group.⁶

We followed Einig [2013] and limited students' test access in the continuous learning environment to two weeks, while students in the flexible learning environment had access to the tests until the end of the semester. Students were invited to each test by a bi-weekly invitation e-mail. While every student in the treatment groups received one invitation email for each of the tests, students in the continuous learning group received an additional reminder email

⁵ Note that actual test participation was still optional. Students were not penalised for not taking the tests, even if they were assigned to a treatment group.

⁶ It is important to note that these students received the assessment questions along with the answers as a PDF document on the Moodle course page two weeks after the respective test went online. Similar to Massoudi et al. [2017], all students therefore had access to the materials before the exam.

towards the end of the test period. Over the course of the semester, we offered five online tests every two weeks with each test containing 9 to 13 questions each. Each student received the same questions in the same order. We provide a sample question in appendix 2.

The online assessments were formative in nature as they provided immediate feedback to students and aimed at improving their future performance (Black and Wiliam [2009]; Marriott and Lau [2008]; Taras [2005]; Wiliam and Black [1996]). Since the university regulations state that marks can only be granted for the final exam, there was no participation incentive. Besides this regulatory reason, we believe the purely voluntary nature of our assessments to be an advantage as it allows us to estimate an effect unconfounded by extrinsic incentives (Kibble [2007]). The students knew that the questions on the final exam would be different from the questions in the online tests since we wanted to avoid having students memorizing questions or studying the question type, rather than the question content (Brothen and Wambach [2001]; Daniel and Broida [2004]; Paxton [2000]).

3.2 Empirical Strategy

In a first step, we are interested in the naïve treatment effect of online learning on exam performance. We include the naïve treatment effect to have a baseline for our findings that is comparable in methodology to prior studies (e.g., Figueroa-Cañas and Sancho-Vinuesa [2021]; Kibble [2007], [2011]; Massoudi et al. [2017]). We use three different variables capturing exposure to the online tests: whether any test was taken over the course of the semester, how many distinct tests were taken (out of five), and how many tests were taken overall (as students could repeat the tests). We explore the relationship between exposure to the online tests and student performance by using a multiple OLS regression framework with the raw exam score as the dependent variable. We use strict marking guidelines to ensure objectivity and consistency in the process (Potter and Johnston [2006]). Each person involved in the marking

process was assigned to one or more questions that they marked for all exams. This ensures that our results are not biased by marker-fixed effects.

We include three different control variables in each regression. First, we control for students sitting the exam twice. Anecdotal evidence tells us that some students choose to fail the first exam in order to gain experience and obtain a general feeling for the exam questions. Since this behaviour may bias our results, we add a dummy control variable equal to one when a student sits the exam twice. Second, we control for students who only sit the exam at the second exam date. We use this control variable since despite our best efforts, we cannot rule out that the second exam date was of slightly different difficulty than the first exam date. We use a dummy variable equal to one if the student only sat the second exam. Finally, we control for gender. After the exams, a member of the chair of accounting with access to the student data but otherwise uninvolved in this experiment matched each student's assessment history to their respective performance in the final exam. Within this procedure, we also obtained information on the participants' gender. In our regression framework, we use a dummy variable equal to one if the student was female. We do not make any predictions of a possible relationship between gender and student performance since results in the literature are largely inconclusive (e.g., Arbaugh [2000]; Byrne and Flood [2008]; Latif and Miles [2020]).

Finally, we examine how controlling for student motivation changes our findings. In order to capture student motivation, we add another dummy variable that is equal to one if a student voluntarily joined the online tests. The regressions including the dummy capturing student motivation look as follows.

$$exam\ score_i = \alpha + \beta_1 test\ taking_i + \beta_2 volunteered_i + \beta_{3,4,5} controls_i + \varepsilon_i$$

Variable definitions are presented in appendix 1.

After estimating the naïve treatment effect, we turn towards our randomised treatment conditions. We first examine whether our treatment conditions change how students opt into test taking. We test this in order to examine whether the type of treatment influences the test exposure of students, as a path through which the different treatment conditions could influence student performance. Test taking is again proxied by the three above-mentioned variables ranging from *any test taken* to the *total number of tests taken*. We split the sample into voluntary and non-voluntary participants and use the same control variables as above. The OLS regressions are as follows:

$$test\ taking_i = \alpha + \beta_1 treatment\ condition_i + \beta_{2,3,4} controls + \varepsilon_i$$

Now, we test our hypotheses by using the randomised treatment design. In order to test our first hypothesis, we look at the effect of being part of the treatment group (i.e., being assigned to the continuous or flexible learning environment) on the exam performance. Note that this is the *intention-to-treat effect* of offering randomly assigned voluntary formative online tests. Random treatment allocation ensures that our results are not biased by unobservable confounders driving both selection into treatment and exam performance (e.g., motivation). However, students may still not participate in the online tests and thus again self-select into the control group. Hence, with estimating the intention-to-treat effect, we look at the potential effects of offering voluntary formative online assessments on student performance. If formative online assessments increase student performance, we would expect a significantly positive relationship between being assigned to a treatment condition and exam performance relative to the control group.

For our second hypothesis, we focus on the effect of being in the continuous learning environment and the flexible learning environment separately. For the second hypothesis, we can use observations from non-voluntary as well as voluntary participants since we are interested

in a comparison between the flexible and continuous learning environment. We employ the following OLS regression framework using the same control variables as above:

$$exam\ score_i = \alpha + \beta_1 treatment\ condition_i + \beta_{2,3,4} controls + \varepsilon_i$$

4 Results

4.1 Descriptive Statistics

In the beginning of the semester, 411 students were enrolled on Moodle. Of those, 150 decided to voluntarily participate in the online tests, while the remaining 261 students decided not to. At 36 %, the participation rate is lower than in comparable studies (e.g., Kibble [2007]). However, students can register for Moodle courses without sitting the final exam. Hence, enrolment on Moodle is not the same as serious course participation. Anecdotal evidence from students tells us that sometimes students simply enrol in a course on Moodle to stay informed.

The 150 students who volunteered for test-taking were equally split into continuous learners and flexible learners. Of the 261 students who did not join voluntarily, 87 (87) were randomly selected for the continuous (flexible) learning group and invited to the online tests. The remaining 87 students acted as control. At the end of the semester, 224 of the initial 411 students sat the final exam. 105 students were voluntary participants, 79 students were non-voluntary participants, and 40 students acted as our control group. Table 1 presents the sample selection.

[Table 1]

Over the course of the semester, 112 out of the final 184 students (61 %) in the treatment condition participated in at least one online test. Participation is considerably lower for students who did not enrol voluntary at the beginning of the semester at 35 %, compared to 80 % of students who joined voluntarily. On average, students who participated took 3.7 distinct tests

and a total of approximately 6 tests. In contrast to prior research, we do not observe a decline in the use of online tests over the course of the semester for the continuous learning environment (Henly [2003]; Massoudi et al. [2017]; results not tabulated). In the flexible learning environment, test-taking slightly increases over the course of the semester and peaks in the week before the first exam. Of the students in our sample, 6 % had to repeat the exam and 17 % sat the exam at the second exam date only. Table 2 presents descriptive statistics.

[Table 2]

We observe a positive correlation of about 0.3 between the number of distinct tests taken and exam performance. It is similar for the other two variables capturing test exposure. Figure 1 provides supporting visual evidence that students who participate in more distinct online tests perform better in the final exam. We furthermore observe a weak positive correlation of 0.2 between joining the online tests voluntarily and exam performance.

[Figure 1]

4.2 Main Tests

We begin by estimating the naïve treatment effect of test exposure on exam performance. Table 3 presents the results. We find that test exposure is positively associated with exam performance with each distinct test increasing exam performance by approximately 1.3 points (out of 60; with 5 distinct tests available). Students who participate in at least one online test on average perform about 4.6 points (confidence interval: 2.28-6.93) better than students who did not participate in any test. The results are similar when controlling for whether students voluntarily joined the online tests.

[Table 3]

We also provide regressions that include all measures of test exposure, i.e., including whether any test was taken, how many distinct tests were taken, and how many tests were taken

in total. The results are insignificant and should be interpreted with caution since the regression suffers from multicollinearity with VIFs between 5 and 10 for the above-mentioned variables (see also table 2 for the correlation coefficients between the variables). Overall, we conclude that test-taking is positively associated with exam performance.

Table 4 presents results on the influence of the different treatment conditions on treatment up-taking. We observe that students in the flexible treatment condition participate in slightly more tests in total than participants in the continuous treatment condition.⁷ This can be explained by a larger number of students in the flexible treatment condition who repeat the tests before the exam. Overall, we do not observe that participants in the flexible treatment condition participate in significantly more distinct tests or that more of these students join the tests. We therefore cannot conclude that a continuous learning environment leads students to engage in more online tests throughout the semester.

[Table 4]

We next turn to the intention-to-treat effect of online tests on student performance. Analyses of variance do not suggest that online tests have a significant influence on exam performance, irrespective of the treatment condition (see appendix 5). Table 5 presents multiple regression results. We find a marginally positive ITT effect from the continuous learning treatment for students who did not voluntarily join the online tests. The effect with 3.4 points difference in the final exam equals approximately one mark in grading. We do not find significant effects for students who join the online tests voluntarily. Overall, we interpret the data such that students who did not voluntarily join the online tests perform slightly better in the final exam, but only when assigned to the continuous learning environment.

[Table 5]

⁷ We also test alternative specifications using a logit regression for *any test taken*, and Poisson regressions for *number of tests taken* and *number of distinct tests taken* as dependent variables. The results are qualitatively similar, with the only exception that the effect of the treatment condition on the *number of tests taken* is also significant for the voluntary participants (with students in the flexible learning environment taking more tests).

4.3 Discussion

Our results indicate that students seem to make a conscious decision about participating in voluntary online assessments. Even with the exam looming, a large fraction of students opt against complementary online tests. This illustrates our concern with non-randomized studies.

The regression results add to the concern. Even with controlling for motivation, we find significantly positive effects of test-taking on exam performance. However, the intention-to-treat effects from the randomised experiment are less clear. We do not find an overall significant effect of online tests on exam performance, but only for the subgroup of people who did not join the tests voluntarily and were in the continuous learning environment. It appears as though students who do the tests are already motivated and write better exams, irrespective of the learning environment. However, students who appear to be less motivated in the sense that they did not enrol for the online tests, may profit from continuous tests. We thus conclude that encouraging students to learn continuously is particularly beneficial for those students who are less motivated to learn. We do not find effects between the continuous and flexible learning environments for students who appear to be motivated to learn.

There are some limitations to our research design. First, all students could access the test questions in a PDF format after the test windows closed for the continuous learning group. We therefore cannot rule out that instead of using the online tests, some students resorted to the PDF version. However, given the participation rate, we do not believe this to be a major problem. In a related fashion, we cannot fully attribute our findings to the circumstance that the tests were administered online. In other words, we cannot speak to the question whether paper-based tests would have led to similar results. However, prior research gives us some comfort that other modes of test facilitation may have resulted in similar outcomes (Bonham et al. [2003]; Gok [2011]).

5 Student Evaluation

Angus and Watson [2009] point out that in order to be successful, online assessments need to be accepted by students in order to avoid negative and counterproductive student experiences. Therefore, we need to rule out that our findings are driven by a low acceptance of the online tests among students.

Students had the chance to (anonymously) fill out a questionnaire regarding their experiences with our online assessments. Overall, students exhibit a very positive attitude towards the assessments. 45 out of 51 students feel like the online tests contributed to a better understanding of the course content. However, just 57 % percent are positive that the tests would help them in the exams (note that the survey was conducted before the exams). Still, no respondent indicated that the tests should not be offered in the future. Overall, our results are in line with a rather positive attitude of students towards the formative online assessments. In summary, we do not believe that a lack of student acceptance influences our results.

We provide the complete results of our semester-end evaluation in appendix 4.

6 Conclusion

We contribute to the debate about formative online assessments by using an experimental design that circumvents the problem of non-randomised treatment allocation. We use the fact that a large number of students at the beginning of the semester did not express interest in joining voluntary online tests and assign them to three different conditions: a continuous learning environment where students have two weeks to complete a given online test, a flexible learning environment with access to the online tests but without the time constraint, and a control environment without access to the online tests. Actual test participation is still voluntary in all treatment conditions. This design allows us to causally identify the intention-to-treat effect of formative online assessments on exam performance. We find that formative online

assessments significantly increase exam performance, but only for students who did not voluntarily join the assessments and who are in the continuous learning environment.

We furthermore illustrate the importance of using a randomised experimental setting by comparing our results with the naïve treatment effect of actual test-taking on exam performance. Even with controlling for students who are comparably motivated, we find a strong statistically significant relationship between test-taking and exam performance with a confidence interval of 2.28 to 6.93 for the naïve treatment effect of test-taking (measured as *any test taken*) on exam performance (measured via points achieved). In comparison, the confidence interval for the overall intention-to-treat effect of being in the treatment condition on exam performance is -1.29 to 5.40 (point estimate: 2.06) and not statistically different from zero. The results indicate that self-selection into treatment is a concern and a student's willingness to participate in online tests and exam performance are likely to be influenced by unobservable confounders. Without appropriately considering this, study results may be significantly upward biased. Hence, we advocate for finding creative yet ethical ways to randomly assign treatment or at a minimum apply an appropriate set of controls (Angus and Watson [2009]).

References

- ANGUS, S. D., AND J. WATSON "Does Regular Online Testing Enhance Student Learning in the Numerical Sciences? Robust Evidence from a Large Data Set." *British Journal of Educational Technology* 40 (2009): 255–272.
- ARBAUGH, J. B. "An Exploratory Study of the Effects of Gender on Student Learning and Class Participation in an Internet-Based MBA Course." *Management Learning* 31 (2000): 503–519.
- AZORLOSA, J. L., AND C. H. RENNER "The Effect of Announced Quizzes on Exam Performance." *Journal of Instructional Psychology* 33 (2006): 278–283.
- BLACK, P., AND D. WILIAM "Developing the Theory of Formative Assessment." *Educational Assessment, Evaluation and Accountability* 21 (2009): 5–31.
- BONHAM, S. W., D. L. DEARDORFF, AND R. J. BEICHNER "Comparison of Student Performance Using Web and Paper-Based Homework in College-Level Physics." *Journal of Research in Science Teaching* 40 (2003): 1050–1071.
- BONWELL, C. C., AND J. A. EISON *Active learning: Creating excitement in the classroom. 1991 ASHE-ERIC higher education reports*. ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183 (1991).
- BROTHEN, T., AND C. WAMBACH "Effective Student Use of Computerized Quizzes." *Teaching of Psychology* 28 (2001): 292–294.
- BUCHANAN, T. "The Efficacy of a World-Wide Web Mediated Formative Assessment." *Journal of Computer Assisted Learning* 16 (2000): 193–200.

- BYRNE, M., AND B. FLOOD "Examining the Relationships among Background Variables and Academic Performance of First Year Accounting Students at an Irish University." *Journal of Accounting Education* 26 (2008): 202–212.
- BYRNE, M., B. FLOOD, AND P. WILLIS "The Relationship between Learning Approaches and Learning Outcomes: A Study of Irish Accounting Students." *Accounting Education* 11 (2002): 27–42.
- CHAK, S. C., AND H. FUNG "Exploring the Effectiveness of Blended Learning in Cost and Management Accounting: An Empirical Study." In *New Media, Knowledge Practices and Multiliteracies* Springer, Singapore, 2015. 189–203.
- CLEAVELAND, M. C., AND E. R. LARKINS "Web-Based Practice and Feedback Improve Tax Students' Written Communication Skills." *Journal of Accounting Education* 22 (2004): 211–228.
- DANIEL, D. B., AND J. BROIDA "Using Web-Based Quizzing to Improve Exam Performance: Lessons Learned." *Teaching of Psychology* 31 (2004): 207–208.
- DAVIDSON, R. A. "Relationship of Study Approach and Exam Performance." *Journal of Accounting Education* 20 (2002): 29–44.
- DUFF, A., AND S. MCKINSTRY "Students' Approaches to Learning." *Issues in Accounting Education* 22 (2007): 183–214.
- DUFRESNE, R., J. MESTRE, D. M. HART, AND K. A. RATH "The Effect of Web-Based Homework on Test Performance in Large Enrollment Introductory Physics Courses." *Jl. of Computers in Mathematics and Science Teaching* 21 (2002): 229–249.
- EINIG, S. "Supporting Students' Learning: The Use of Formative Online Assessments." *Accounting Education* 22 (2013): 425–444.

- ENGLISH, L., P. LUCKETT, AND R. MLADENOVIC "Encouraging a Deep Approach to Learning through Curriculum Design." *Accounting Education* 13 (2004): 461–488.
- FIGUEROA-CAÑAS, J., AND T. SANCHO-VINUESA "Investigating the Relationship between Optional Quizzes and Final Exam Performance in a Fully Asynchronous Online Calculus Module." *Interactive Learning Environments* 29 (2021): 33–43.
- GIKANDI, J. W., D. MORROW, AND N. E. DAVIS "Online Formative Assessment in Higher Education: A Review of the Literature." *Computers and Education* 57 (2011): 2333–2351.
- GOK, T. "Comparison of Student Performance Using Web- and Paper- Based Homework in Large Enrollment Introductory Physics Courses." *International Journal of Physical Sciences* 6 (2011): 3740–3746.
- GRIMSTAD, K., AND M. GRABE "Are Online Study Questions Beneficial?" *Teaching of Psychology* 31 (2004): 143–146.
- GROVE, W. A., AND T. WASSERMAN "Incentives and Student Learning: A Natural Experiment with Economics Problem Sets." *American Economic Review* 96 (2006): 447–452.
- HADSELL, L. "The Effect of Quiz Timing on Exam Performance." *Journal of Education for Business* 84 (2009): 135–141.
- HÄFNER, A., V. OBERST, AND A. STOCK "Avoiding Procrastination through Time Management: An Experimental Intervention Study." *Educational Studies* 40 (2014): 352–360.
- HALL, M., A. RAMSAY, AND J. RAVEN "Changing the Learning Environment to Promote Deep Learning Approaches in First-Year Accounting Students." *Accounting Education* 13 (2004): 489–505.
- HENLY, D. C. "Use of Web-Based Formative Assessment to Support Student Learning in a Metabolism/Nutrition Unit." *European Journal of Dental Education* 7 (2003): 116–122.

- JARMOLOWICZ, D. P., Y. HAYASHI, AND C. ST. P. PIPKIN "Temporal Patterns of Behavior from the Scheduling of Psychology Quizzes." *Journal of Applied Behavior Analysis* 43 (2010): 297–301.
- KIBBLE, J. "Teaching With Technology Use of Unsupervised Online Quizzes as Formative Assessment in a Medical Physiology Course: Effects of Incentives on Student Participation and Performance." *Adv Physiol Educ* 31 (2007): 253–260.
- KIBBLE, J. D. "Voluntary Participation in Online Formative Quizzes Is a Sensitive Predictor of Student Success." In *American Journal of Physiology - Advances in Physiology Education* 35.1 (2011): 95–96.
- LATIF, E., AND S. MILES "The Impact of Assignments and Quizzes on Exam Grades: A Difference-in-Difference Approach." *Journal of Statistics Education* 28 (2020): 289–294.
- LOWRY, R. "Computer Aided Self Assessment-an Effective Tool." *Chemistry Education Research and Practice* 6 (2005): 198–203.
- MARRIOTT, P., AND A. LAU "The Use of On-Line Summative Assessment in an Undergraduate Financial Accounting Course." *Journal of Accounting Education* 26 (2008): 73–90.
- MASSOUDI, D., S. K. KOH, P. J. HANCOCK, AND L. FUNG "The Effectiveness of Usage of Online Multiple Choice Questions on Student Performance in Introductory Accounting." *Issues in Accounting Education* 32 (2017): 1–17.
- MORRIS, E. K., C. F. SURBER, AND S. W. BIJOU "Self-Pacing Versus Instructor-Pacing: Achievement, Evaluations, and Retention." *Journal of Educational Psychology* 70 (1978): 224–230.
- NAGANDLA, K., S. SULAIHA, AND S. V. NALLIAH "Online Formative Assessments: Exploring Their Educational Value." In *Original Article Journal of Advances in Medical Education & Professionalism* 6.2 (2018): 51–57.

- ORPEN, C. "The Causes and Consequences of Academic Procrastination: A Research Note." *Westminster Studies in Education* 21 (1998): 73–75.
- PAXTON, M. "A Linguistic Perspective on Multiple Choice Questioning." *Assessment and Evaluation in Higher Education* 25 (2000): 109–119.
- PEAT, MARY, AND SUE FRANKLIN. "Supporting student learning: the use of computer-based formative assessment modules." *British Journal of Educational Technology* 33.5 (2002): 515–523.
- PEAT, M., S. FRANKLIN, M. DEVLIN, AND M. CHARLES "Revisiting the Impact of Formative Assessment Opportunities on Student Learning." *Australasian Journal of Educational Technology* 21 (2005): 102–117.
- PERRIN, C. J., N. MILLER, A. T. HABERLIN, J. W. IVY, J. N. MEINDL, AND N. A. NEEF "Measuring and Reducing College Students' Procrastination." *Journal of Applied Behavior Analysis* 44 (2011): 463–474.
- POTTER, B. N., AND C. G. JOHNSTON "The Effect of Interactive On-Line Learning Systems on Student Learning Outcomes in Accounting." *Journal of Accounting Education* 24 (2006): 16–34.
- PRINCE, M. "Does Active Learning Work? A Review of the Research." *Journal of Engineering Education* 93 (2004): 223–231.
- REISER, R. A. "Reducing Student Procrastination in a Personalized System of Instruction Course." *ECTJ* 32 (1984): 41–49.
- RICKETTS, C., AND S. J. WILKS "Improving Student Performance through Computer-Based Assessment: Insights from Recent Research." *Assessment and Evaluation in Higher Education* 27 (2002): 475–479.

- RILEY, J., AND K. WARD "Active Learning, Cooperative Active Learning, and Passive Learning Methods in an Accounting Information Systems Course." *Issues in Accounting Education* 32 (2017): 1–16.
- ROTENSTEIN, A., H. Z. DAVIS, AND L. TATUM "Early Birds versus Just-in-Timers: The Effect of Procrastination on Academic Performance of Accounting Students." *Journal of Accounting Education* 27 (2009): 223–232.
- SCOULLER, K. "The Influence of Assessment Method on Students' Learning Approaches: Multiple Choice Question Examination versus Assignment Essay." *Higher Education* 35.4 (1998): 453–472.
- SHARMA, D. S. "Accounting Students' Learning Conceptions, Approaches to Learning, and the Influence of the Learning–Teaching Context on Approaches to Learning." *International Journal of Phytoremediation* 21 (1997): 125–146.
- SOTOLA, L. K., AND M. CREDE "Regarding Class Quizzes: A Meta-Analytic Synthesis of Studies on the Relationship Between Frequent Low-Stakes Testing and Class Performance." *Educational Psychology Review* 33 (2021): 407–426.
- TARAS, M. "Assessment - Summative and Formative - Some Theoretical Reflections." *British Journal of Educational Studies* 53 (2005): 466–478.
- TEN HAVE, T. R., S. L. T. NORMAND, S. M. MARCUS, C. H. BROWN, P. LAVORI, AND N. DUAN "Intent-to-Treat vs. Non-Intent-to-Treat Analyses under Treatment Non-Adherence in Mental Health Randomized Trials." *Psychiatric Annals* 38 (2008): 772–783.
- VELAN, G. M., R. K. KUMAR, M. DZIEGIELEWSKI, AND D. WAKEFIELD "Web-Based Self-Assessments in Pathology with Questionmark Perception." *Pathology* 34 (2002): 282–284.
- VOS, H. "How to Assess for Improvement of Learning." *European Journal of Engineering Education* 25 (2000): 227–233.

WILLIAM, D., AND P. BLACK "Meanings and Consequences: A Basis for Distinguishing Formative and Summative Functions of Assessment?" *British Educational Research Journal* 22 (1996): 537–548.

Appendix

Appendix 1: Variable Definitions

Mnemonic	Definition
<i>exam points</i>	Highest number of points that the student achieves in the final exam (first or second exam date).
<i>treatment</i>	Indicator variable equal to one if the student is in the treatment group (continuous or flexible).
<i>continuous treatment</i>	Indicator variable equal to one if the student is in the continuous treatment group having time-restricted access (two weeks) to the online tests.
<i>flexible treatment</i>	Indicator variable equal to one if the student is in the flexible treatment group having unrestricted access to the online tests.
<i>any test taken</i>	Indicator variable equal to one if the student completes at least one online test over the course of the semester. Captures test exposure.
<i>number of distinct tests taken</i>	Total number of distinct online tests a student takes over the course of the semester (max. 5). Captures test exposure.
<i>number of tests taken</i>	Total number of online tests a student takes over the course of the semester. The number can be higher than for <i>number of distinct tests taken</i> since students had multiple attempts at one test. Captures test exposure.
<i>volunteered</i> (control variable)	Indicator variable equal to one if the student applies and enrolls voluntarily for the online tests at the beginning of the semester.
<i>sat only second exam</i> (control variable)	Indicator variable equal to one if the student only sits the exam at the second exam date (i.e. has no attempt at the exam during the first exam period).
<i>exam repeat</i> (control variable)	Indicator variable equal to one if the student repeats the final exam.
<i>female</i> (control variable)	Indicator variable equal to one if the student is female.

Appendix 2: Sample Question

The art glazier Claude-Glas OHG produces large illuminated leaded glazing with impressionist glass paintings in the style of Monet. Customers can choose between two options: “Water Lilies” and “The Japanese Bridge”.

The following sales and production data of Claude-Glas OHG are available:

Item	Option 1: “Water Lilies”	Option 2: “The Japanese Bridge”
Output / Year	20	10
Sales Price / Piece	2,500 €	2,000 €
Variable Costs / Piece	1,300 €	1,200 €
Product fixed costs	15,000 €	10,000 €
Company fixed costs	2,000 €	

Please enter your results without a thousands separator.

- a) What profit can be achieved with an optimal short-term production program?

Calculate the contribution margin I of the two options:

Option 1: € / Piece

Option 2: € / Piece

Profit: €

- b) What profit can be achieved with an optimal long-term production program?

Calculate the contribution margin II of the two options:

Option 1: € / Piece

Option 2: - € / Piece

Profit: €

Appendix 3: E-mail Invitation

Dear students, As announced during this morning's lecture, you will have the opportunity to participate in our project “MAME – Management Accounting Made Easy”.

This project is sponsored by the media commission of the Academic Senate. We will offer regular online tests that allow you to deepen your understanding of the topics dealt with in the lectures and tutorials. This is an optional offer that neither replaces the lecture nor the tutorials.

Our goal is to examine whether these additional opportunities for training have a positive influence on your exam success. Since our focus is on the effect of continuous learning on exam success, we will implement two different groups. Group A will be motivated to learn continuously via restricting the tests to be available only for two weeks. Group B on the other hand will not receive this incentive and has permanent access to the tests. The assignment to groups is random.

The online tests are administered by a person from our institute who is not involved in teaching of this course. This way you can be sure that no lecturer can view your personal test results.

How can you join?

You can register for participation by 12:00 noon on Friday, April 28, 2017. To do this, please send an e-mail with the subject "MAME" to (e-mail address) and enter your full name and your university e-mail address so that we can assign you to the respective Moodle courses.

After this registration period, we randomly select additional students and grant them access to the online tests. This way, everyone interested has the opportunity to participate. Independent of the participation, all students receive the answers to the online tests after the end of the respective test window.

We look forward to your registration!

Appendix 4: Student Evaluation

Panel A: Participation in the online tests

	Continuous learning (n = 26)						Flexible learning (n = 25)					
	0	1	2	3	4	5	0	1	2	3	4	5
<i>How many online tests did you attend?</i>	0%	4%	4%	4%	8%	81%	4%	16%	4%	12%	4%	60%

Panel B: Reasons for not participating in the online tests

<i>If you have participated in less than three tests: What was the main reason for not participating?</i>	Continuous learning (number of responses)	Flexible learning (number of responses)
I did not have the time to do the tests.	0	0
Lack of motivation to learn continuously throughout the semester.	1	2
Expectation that the tests avail to nothing.	0	0
Missed the period of time despite the reminders (continuous learning).	1	0
I plan to do the tests during my exam preparation (flexible learning).	0	6

Panel C: Overall impression of the online tests

	Continuous learning (n = 26)					Flexible learning (n = 25)				
	Strongly agree	Rather agree	Indifferent	Rather disagree	Strongly disagree	Strongly agree	Rather agree	Indifferent	Rather disagree	Strongly disagree
The online tests contributed to a better understanding of the content.	38%	50%	8%	4%	0%	44%	44%	12%	0%	0%
The regular online tests motivated me to learn continuously.	42%	23%	23%	8%	4%	24%	20%	16%	28%	12%
Due to my participation in the online tests, I expect a better result in the exam.	27%	27%	27%	15%	4%	24%	36%	28%	12%	0%
I think that the online tests should be continued in the following semesters.	92%	4%	4%	0%	0%	84%	16%	0%	0%	0%

Appendix 5: ANOVA (Intention-to-treat Effect)

Table 7: Analysis of variance (ANOVA)

Panel A: Non-voluntary participants					
Source	Sum of Squares	df	Mean square	F-statistic	p-value
Treatment groups	205	2	102.4	1.41	0.25
Residuals	8434	116	72.7		
Comparison of means					
Treatment groups	Difference	Confidence interval		p-value	
Continuous vs. control	2.95	[-1.67;7.56]		0.29	
Flexible vs. control	0.23	[-4.24;4.71]		0.99	
Continuous vs. flexible	2.71	[-1.85;7.28]		0.34	
Panel B: Voluntary participants					
Source	Sum of Squares	df	Mean square	F-statistic	p-value
Treatment groups	68	1	67.6	0.8	0.37
Residuals	8696	103	84.4		
Comparison of means					
Treatment groups	Difference	Confidence interval		p-value	
Continuous vs. flexible	-1.61	[-5.16;1.95]		0.37	

Notes: Table 7 presents ANOVAs for the comparison of means between treatment conditions with respect to exam performance. The comparison of means uses Tukey Honest Significant Differences. **Panel A** presents results for all 119 observations of students who sat the final exam and did not apply to join the online tests at the beginning of the semester. **Panel B** presents results for all 105 observations of students who joined the online tests voluntarily. This specification has no control condition. Variable definitions can be found in appendix 1.

Tables

Table 1: Sample Selection

	Voluntary test-taking		Non-voluntary test-taking			Σ
	Continuous learning	Flexible learning	Continuous learning	Flexible learning	Control	
Number of students at semester start	75	75	87	87	87	411
.\. students not sitting the final exam	21	24	50	45	47	187
Final sample	54	51	37	42	40	224

Notes: Table 1 shows the sample selection procedure. All students enrolled at the beginning of the semester were invited to join the online tests. Students who joined voluntarily were randomly assigned to the continuous and flexible learning treatment conditions. Students who did not apply to join were randomly assigned to either treatment condition or the control condition. We only use data of students who participate in the final exam.

Table 2: Descriptive Statistics

Panel A: Summary statistics								
	N	Mean	sd	min	P25	median	P75	max
<i>exam points</i>	224	32.01	9.01	2	26.5	32.25	38.5	52.5
<i>any test taken</i>	224	0.50		0	0	0.5	1	1
<i>number of distinct tests taken</i>	112	3.69	1.62	1	2	5	5	5
<i>number of tests taken</i>	112	6.07	4.39	1	2	5	9	19
<i>volunteered</i>	224	0.47		0	0	0	1	1
<i>exam repeat</i>	224	0.06		0	0	0	0	1
<i>female</i>	224	0.58		0	0	1	1	1
<i>sat only second exam</i>	224	0.17		0	0	0	0	1
Panel B: Correlation table								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>(1) exam points</i>		.27	.33	.32	.20	-.16	-.06	.01
<i>(2) any test taken</i>	.28		.85	.70	.56	-.07	-.14	.11
<i>(3) number of distinct tests taken</i>	.32	.94		.89	.49	-.01	-.12	.12
<i>(4) number of tests taken</i>	.33	.93	.98		.40	.00	-.09	.03
<i>(5) volunteered</i>	.20	.56	.53	.53		.02	-.14	.06
<i>(6) exam repeat</i>	-.18	-.07	-.03	-.03	.02		-.12	.07
<i>(7) sat only second exam</i>	-.08	-.14	-.13	-.12	-.14	-.12		-.05
<i>(8) female</i>	.01	.11	.12	.09	.06	.07	-.05	

Notes: Table 2 presents descriptive statistics based on 224 students who participated in the final exam and were assigned to either treatment or control condition at the beginning of the semester. **Panel A** shows summary statistics for all variables used in the regressions (including control variables). **Panel B** shows the correlation matrix for the same variables. It shows Pearson correlation coefficients above and Spearman correlation coefficients below the diagonal. For variable definitions please see appendix 1.

Table 3: Naïve Treatment Effect

	<i>exam points</i>							
<i>Intercept</i>	30.32***	30.31***	30.53***	30.40***	30.05***	30.05***	30.03***	30.14***
	[28.14, 32.50]	[28.27, 32.34]	[28.52, 32.54]	[28.24, 32.56]	[27.80, 32.31]	[27.89, 32.22]	[27.87, 32.20]	[27.91, 32.37]
<i>any test taken</i>	4.60***			-0.76	3.85***			-1.42
	[2.28, 6.93]			[-5.18, 3.65]	[1.05, 6.65]			[-6.05, 3.21]
<i>number of distinct tests taken</i>		1.33***		1.02		1.23***		0.99
		[0.81, 1.85]		[-0.55, 2.60]		[0.64, 1.83]		[-0.59, 2.57]
<i>number of tests taken</i>			0.65***	0.25			0.58***	0.26
			[0.39, 0.91]	[-0.33, 0.84]			[0.30, 0.86]	[-0.32, 0.85]
<i>volunteered</i>					1.34	0.89	1.50	1.27
					[-1.44, 4.12]	[-1.70, 3.47]	[-0.96, 3.96]	[-1.45, 4.00]
<i>Controls</i>	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Num.Obs.</i>	224	224	224	224	224	224	224	224
<i>R2</i>	0.094	0.131	0.127	0.136	0.098	0.133	0.133	0.139
<i>R2 Adj.</i>	0.077	0.116	0.111	0.112	0.077	0.113	0.113	0.111
<i>AIC</i>	1609.6	1600.0	1601.2	1603.0	1610.6	1601.6	1601.8	1604.1
<i>BIC</i>	1630.0	1620.5	1621.7	1630.3	1634.5	1625.5	1625.6	1634.8
<i>Log.Lik.</i>	-798.781	-794.025	-794.615	-793.490	-798.316	-793.791	-793.878	-793.052
<i>F</i>	5.668	8.289	7.958	5.675	4.714	6.706	6.667	4.982

Notes: Table 3 presents OLS regressions to estimate the naïve treatment effect of test-taking on exam performance. The regressions comprise all 224 observations of students who sat the final exam and were assigned to treatment or control conditions at the beginning of the semester. Variable definitions (including control variables) can be found in appendix 1. 95 % Confidence intervals are shown in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Treatment Up-taking

	<i>Non-voluntary participants</i>			<i>Voluntary participants</i>		
	<i>any test taken</i>	<i>number of distinct tests taken</i>	<i>number of tests taken</i>	<i>any test taken</i>	<i>number of distinct tests taken</i>	<i>number of tests taken</i>
<i>(Intercept)</i>	0.38***	1.40***	2.98***	0.77***	2.76***	5.57***
	[0.18, 0.58]	[0.58, 2.23]	[1.29, 4.67]	[0.61, 0.94]	[1.92, 3.60]	[3.72, 7.43]
<i>continuous treatment</i>	-0.05	-0.39	-1.07	-0.07	-0.38	-1.65*
	[-0.26, 0.17]	[-1.28, 0.50]	[-2.90, 0.75]	[-0.23, 0.09]	[-1.20, 0.44]	[-3.45, 0.15]
<i>Controls</i>	Yes	Yes	Yes	Yes	Yes	Yes
<i>Num.Obs.</i>	79	79	79	105	105	105
<i>R2</i>	0.099	0.102	0.071	0.031	0.036	0.043
<i>R2 Adj.</i>	0.050	0.054	0.021	-0.008	-0.002	0.005
<i>AIC</i>	111.5	335.5	448.7	114.3	457.9	623.1
<i>BIC</i>	125.7	349.7	462.9	130.2	473.9	639.0
<i>Log.Lik.</i>	-49.735	-161.760	-218.355	-51.143	-222.968	-305.537
<i>F</i>	2.025	2.109	1.423	0.791	0.946	1.123

Notes: Table 4 presents OLS regressions to estimate the effect of our treatment conditions on student participation in the online tests. The first three regressions comprise all 79 observations of students who sat the final exam, did not apply to join the online tests at the beginning of the semester, and are in the treatment conditions. The last three regressions comprise all 105 observations of students who joined the online tests voluntarily. Variable definitions (including control variables) can be found in appendix 1. 95 % confidence intervals are shown in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

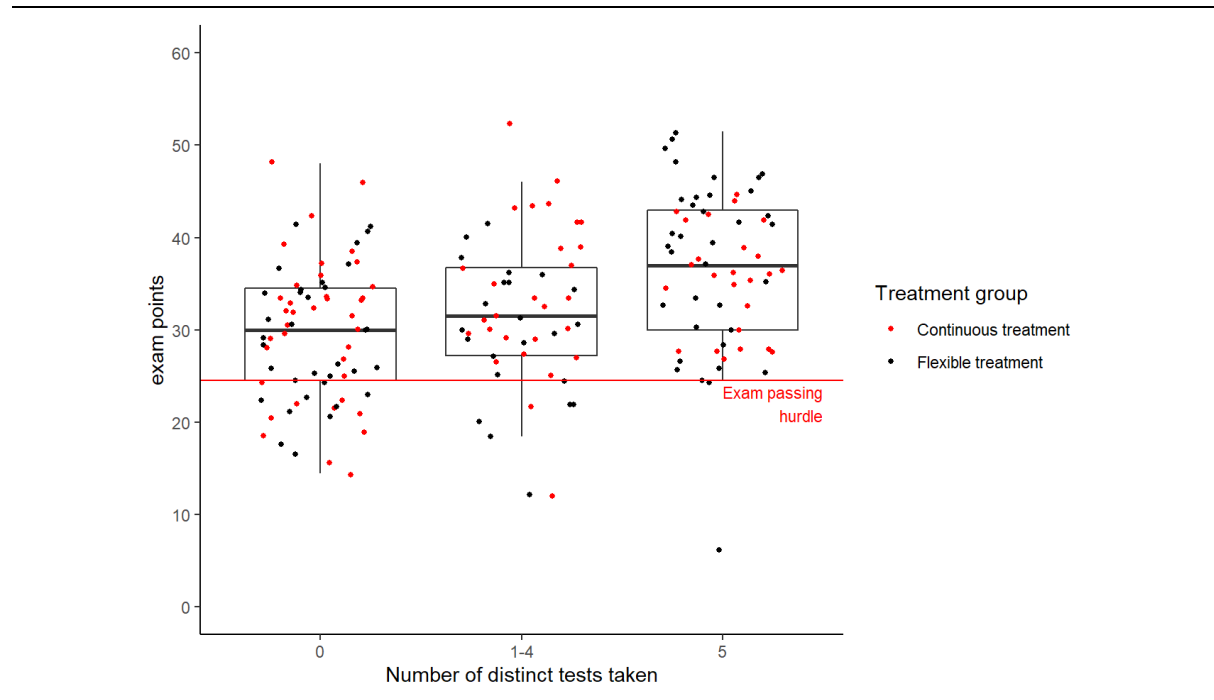
Table 5: Intention-to-treat Effects

	<i>exam points</i>			
<i>treatment</i>	2.06	0.73		
	[-1.29, 5.40]	[-3.10, 4.55]		
<i>continuous treatment</i>	3.42*	2.70	-1.80	
	[-0.43, 7.28]	[-1.12, 6.51]	[-5.44, 1.84]	
<i>flexible treatment</i>	0.73			
	[-3.10, 4.55]			
<i>Controls</i>	Yes	Yes	Yes	Yes
<i>Num.Obs.</i>	119	119	119	105
<i>R2</i>	0.061	0.077	0.077	0.037
<i>R2 Adj.</i>	0.028	0.037	0.037	-0.002
<i>AIC</i>	852.1	852.0	852.0	770.6
<i>BIC</i>	868.7	871.5	871.5	786.5
<i>Log.Lik.</i>	-420.036	-419.011	-419.011	-379.292
<i>F</i>	1.865	1.897	1.897	0.959

Notes: Table 5 presents OLS regressions to estimate the effect of our treatment conditions on exam performance. The first three regressions present intention-to-treat effects and comprise all 119 observations of students who sat the final exam and did not apply to join the online tests at the beginning of the semester. The last regression presents the treatment effects for students who joined the online tests voluntarily (105 observations). This specification has no control condition. Variable definitions (including control variables) can be found in appendix 1. 95 % confidence intervals are shown in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figures

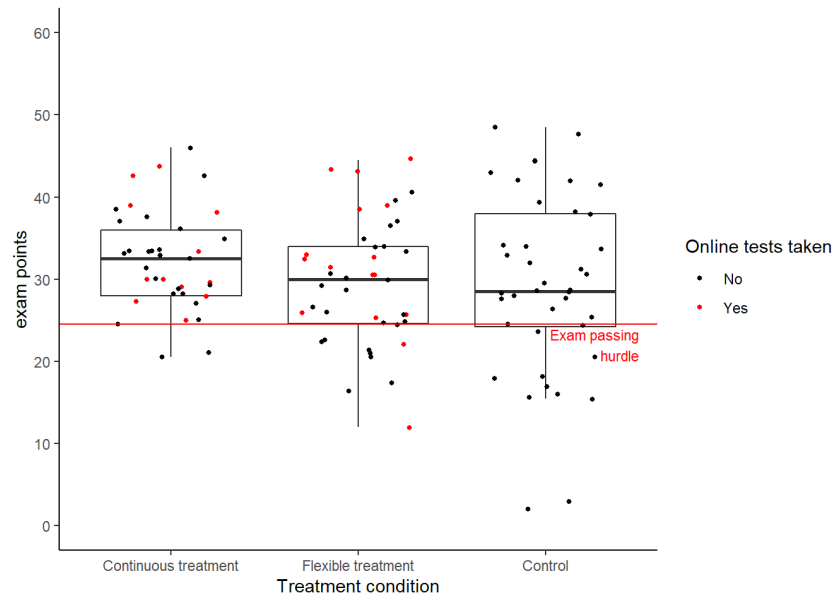
Figure 1: Exam Performance by Test-taking



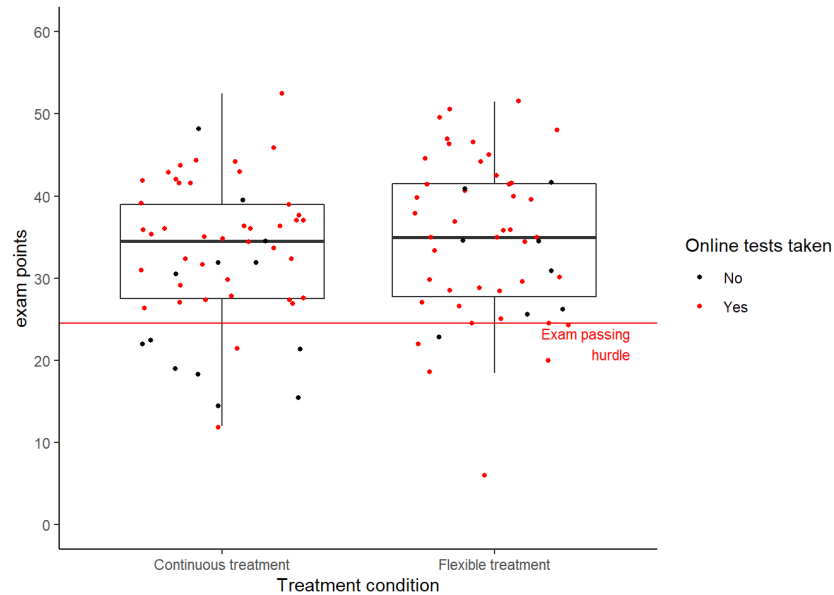
Notes: Figure 1 presents visual evidence on the relationship between treatment up-taking (here defined as the *number of distinct tests taken*) and exam performance. It comprises data of all 182 students who were assigned to the continuous or flexible treatment condition at the beginning of the semester and sat the final exam at the end of the semester. It does not include data from students in the control condition.

Figure 2: Exam Performance by Treatment Condition

Panel A: Non-voluntary participants



Panel B: Voluntary participants



Notes: Figure 2 presents visual evidence on the intention-to-treat effect of being assigned to a treatment condition on exam performance. **Panel A** presents results for all 119 observations of students who sat the final exam and did not apply to join the online tests at the beginning of the semester. **Panel B** presents results for all 105 observations of students who joined the online tests voluntarily. This specification has no control condition.