



PHD

Automated Assistance and Perceptions of Trust and Confidence: Experiments in the Domain of Grammar and Spelling Checking

Zijlstra, Melle

Award date:
2022

Awarding institution:
University of Bath

[Link to publication](#)

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC-ND 4.0) Licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

Take down policy

If you consider content within Bath's Research Portal to be in breach of UK law, please contact: openaccess@bath.ac.uk with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

Automated Assistance and
Perceptions of Trust and
Confidence:
*Experiments in the Domain of
Grammar
and Spelling Checking*

submitted by

Melle Zijlstra

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Computer Science

May 2022

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with the author. A copy of this thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that they must not copy it or use material from it except as permitted by law or with the consent of the author.

This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation with effect from
. (date)

Signed on behalf of the Faculty of Science

Summary

Although interaction with mundane automated aids like spelling and grammar checkers is commonplace, it is a surprisingly little studied subject, of which not all characteristics are equally well understood. In a series of experimental studies, we demonstrate how a novel experimental paradigm based on Signal Detection Theory can be used to study a cognitive task, in which different aspects of performance, trust, and confidence of users interacting with an imperfect automated writing aid are tested. Five closely related experiments are reported, in which participants make a series of judgments of which of a pair of similar sentences is better, or whether a single sentence is correct or not. Our experimental hypotheses all derive from the overarching hypothesis that participants will be able to interpret and make use of an automated aid's suggestions and the aid's own estimation of the likelihood of its suggestions being correct.

The first, and overriding contribution of this thesis is to begin an experimental exploration of personal beliefs in relation to performance under uncertainty, and with support from an imperfect automated aid in the domain of text writing and editing, and in particular spelling and grammar checking. Especially the measure of bias, as the propensity to accept automated advice, is an essential measure for our studies, and arguably a major novel contribution of the thesis.

The experiments show that trust in similar systems has less of an effect on participants' performance than we anticipated on basis of the literature. This is also true of perceived self-efficacy, although our findings suggest it may play a more important role if the advice from the system is weak and users must be more reliant on their own skills.

We demonstrate that improving the reliability of the aid's advice positively affects users' performance, trust in the aid, and confidence in their own responses, but also that a highly reliable automated aid still gets underused. Throughout the five experiments, we confirmed the *above average effect*, people's assumption that their own ability is on average higher than that of others, as well as the *overconfidence effect*, an overestimation of performance if measured as probabilities of success during a task, but less so if measured as an estimate of success-frequency post-task.

Another novel contribution of this series of experiments is the finding that users can recognise how well a system is doing, even if they do not receive any feedback on the system's performance. Users of a more reliable system proved to be more willing to accept the aid's advice, which suggests effects of the *reliability* and *strength* of the advice, the latter of which is represented by

the system's communicated likelihood estimation. Without receiving feedback about their own performance, users also show they have an awareness of their own performance, which is demonstrated by a higher level of self-reported confidence in correct responses than in incorrect ones.

We believe our research successfully demonstrates opportunities and limitations of using an experimental paradigm based on Signal Detection Theory to explore various aspects of performance, trust, and confidence this domain. We think that our findings will be valuable for future research as well as for the design of automated aids, and that the methods and analyses developed could usefully be transferred to assisted cognitive tasks in other domains.

Acknowledgements

'His [the ant's] horizons are very close, so that he deals with each obstacle as he comes to it; he probes for ways around or over it, without much thought for future obstacles. It is easy to trap him into deep detours.'

– Herbert Simon (1996)

The parable of the Ant on the Beach, in Sciences of the Artificial

Like that of the ant, humans' behaviour is largely dependent on their environment, Simon argues. During my PhD, I have felt incredibly fortunate to have landed in an environment defined by patience and support. First and foremost, I would like to thank my supervisors for their incredible endurance and fantastic guidance during what is best described as a drawn-out PhD process. Firstly, Professor Stephen Payne, and then in order of appearance, Professor Linda Newnes, Dr Simon Jones, and Dr James Laird, thank you so much for keeping the faith and reassuring me all would be well in the end.

I would also like to thank my colleagues in the Department of Mechanical Engineering for being understanding and flexible with my working hours, the University of Bath for offering me a scholarship, and Student Services and the Doctoral College for accommodating my needs. A special thank you goes to all those support staff who make being on campus such a pleasurable experience, especially those who rarely get a mention, such as the wonderful cleaners, porters, security and hospitality staff, and others who work tirelessly and get so little in return.

Lastly, Geertje m'n liefje, probing for ways around and over obstacles together has always been a joy. Believe me, I love our detours too, even if I sometimes protest at the prospect of having to climb yet another pile of rocks, just to have a peek over yet another summit (yes you, Benbulbin!). I admire your endless enthusiasm and energy to keep pushing on, let's see what's around the next corner!

Table of contents

SUMMARY	2
ACKNOWLEDGEMENTS	4
TABLE OF CONTENTS	5
TABLE OF FIGURES	7
LIST OF TABLES	8
INTRODUCTION	9
CHAPTER 1 – SUBJECT AREA BACKGROUND AND LITERATURE REVIEW	14
1.1 Introduction	14
1.2 Experimental domain – decision making under uncertainty with imperfect and uncertain advice	14
1.3 Experimental approach – modelling and testing collaborative complex decision making under uncertainty	33
1.4 How word-processors deal with specific writing error types relevant to our research	45
1.5 Summary	46
CHAPTER 2 – RESEARCH APPROACH AND METHODOLOGY	48
2.1 Introduction	48
2.2 Research approach and survey design	49
2.3 Analysis and reporting of results	54
2.4 Analysis and reporting of results: hypotheses and analyses	64
2.5 Discussion of research approach and methodology	65
2.6 Ethics and data management	68
CHAPTER 3 – STYLUS 1: DEVELOPING A PARADIGM TO TEST EFFECTS OF TRUST AND PERCEIVED SELF-EFFICACY ON PERFORMANCE AND CONFIDENCE	69
3.1 Introduction	69
3.2 Method	70
3.3 Analysis strategy	76
3.4 Results	77
3.5 Summary of results	89
3.6 Conclusions and discussion	90
CHAPTER 4 – STYLUS 2: TESTING EFFECTS OF TRUST AND PERCEIVED SELF-EFFICACY WITH AN IMPROVED EXPERIMENT	91
4.1 Introduction	91
4.2 Method	91
4.3 Analysis strategy	93
4.4 Results	94
4.5 Conclusions and discussion	105

CHAPTER 5 – <i>STYLUS 3: INTRODUCING AN AID THAT COMMUNICATES ITS UNCERTAINTY</i>	109
5.1 Introduction	109
5.2 Method	112
5.3 Analysis strategy	116
5.4 Results	117
5.5 Conclusion and discussion	127
CHAPTER 6 – <i>STYLUS 4: INCREASING THE RELIABILITY OF THE AID AND THE STRENGTH ITS ADVICE</i>	129
6.1 Introduction	129
6.2 Method	129
6.3 Analysis strategy	132
6.4 Results	133
6.5 Comparing S3 and S4	140
6.6 Conclusions and discussion	141
CHAPTER 7 – <i>STYLUS 5: TESTING MODELS OF AIDED INTERACTION</i>	144
7.1 Introduction	144
7.2 Method	147
7.3 Analysis strategy	152
7.4 Results	157
7.5 Conclusions and discussion	170
CHAPTER 8 – <i>CONCLUSIONS AND DISCUSSION</i>	174
8.1 Introduction	174
8.2 Overview and discussion of findings	175
8.3 Models of aided performance	178
8.4 Discussion of method	180
8.5 Some implications of findings, and recommendations for further research and future systems development	183
8.6 Concluding remarks	185
BIBLIOGRAPHY	187
APPENDICES	200
Appendix A3 – Stylus 1 data overview	201
Appendix A4 – Stylus 2 data overview	208
Appendix A5 – Stylus 3 data overview	215
Appendix A6 – Stylus 5 data overview	219
Appendix A7 – Stylus 5 data overview	223
Appendix B5 – Stylus 3 item distribution	230
Appendix B6 – Stylus 4 item distribution	231
Appendix B7 – Stylus 5 item distribution	232
Appendix C5 – Stylus 3 analyses with non-parametric measures	237
Appendix C7 – Stylus 5 ANOVA table	238
Appendix D – Hypotheses overview	240
Appendix E1 – University of Bath Department of Computer Science ethics check list	245
Appendix E2 – Stylus Qualtrics consent screen	248
Appendix E3 – Stylus Qualtrics debrief screen	249

Table of figures

FIGURE 1.1 – DECISION MAKING DIAGRAM BASED ON EDWARDS 1954	16
FIGURE 1.2 – PERCEIVED SELF-EFFICACY VS. CONFIDENCE	27
FIGURE 1.3 – PERCEIVED SELF-EFFICACY, SINGLE-EVENT CONFIDENCE, AND FREQUENCY CONFIDENCE	28
FIGURE 1.4 – SPELLING AND GRAMMAR CHECKING CUSTOMISATION SETTINGS IN MS WORD 2021.	42
FIGURE 3.1 – S1 TRIAL INTERFACE EXAMPLE SCREENSHOT	75
FIGURE 3.2 – S1 PERCENTAGE CORRECT RESPONSES, MEAN AND STANDARD ERROR PER GROUP	79
FIGURE 3.3A – S1 PARAMETRIC SENSITIVITY (D'), MEAN AND STANDARD ERROR PER GROUP	80
FIGURE 3.3B – S1 NON-PARAMETRIC SENSITIVITY (A'), MEAN AND STANDARD ERROR PER GROUP	81
FIGURE 3.4A – S1 PARAMETRIC BIAS (C), MEAN AND STANDARD ERROR PER GROUP	81
FIGURE 3.4B – S1 NON-PARAMETRIC BIAS (B''), MEAN AND STANDARD ERROR PER GROUP	82
FIGURE 3.5 – S1 PERCENTAGE H, M, FA AND CR CONFIDENCE, MEAN AND STANDARD ERROR PER GROUP	85
FIGURE 3.6 – S1 COMPARISON OF PERCENTAGE TRIAL CONFIDENCE, PERCENTAGE CORRECT RESPONSES, AND OVERALL ESTIMATED PERCENTAGE CORRECT RESPONSES, MEAN AND STANDARD ERROR PER GROUP	87
FIGURE 4.1A – S2 PARAMETRIC SENSITIVITY (D'), MEAN AND STANDARD ERROR PER GROUP	96
FIGURE 4.1B – S2 NON-PARAMETRIC SENSITIVITY (A'), MEAN AND STANDARD ERROR PER GROUP	96
FIGURE 4.2A – S2 PARAMETRIC BIAS (C), MEAN AND STANDARD ERROR PER GROUP	97
FIGURE 4.2B – S2 NON-PARAMETRIC BIAS (B''), MEAN AND STANDARD ERROR PER GROUP	97
FIGURE 4.3 – S2 PERCENTAGE H, M, FA AND CR CONFIDENCE, MEAN AND STANDARD ERROR PER GROUP	101
FIGURE 4.4 – S2 PERCENTAGE AVERAGE TRIAL CONFIDENCE, PERCENTAGE CORRECT RESPONSES, AND OVERALL ESTIMATED PERCENTAGE CORRECT RESPONSES, MEAN AND STANDARD ERROR PER GROUP	103
FIGURE 5.1 – S3 TRIAL EXAMPLE SCREENSHOT	111
FIGURE 5.2 – S3 PARAMETRIC SENSITIVITY (D'), MEAN AND STANDARD ERROR PER CONDITION	119
FIGURE 5.3 – S3 PERCENTAGE H, M, FA AND CR CONFIDENCE, MEAN AND STANDARD ERROR PER CONDITION	122
FIGURE 5.4 – S3 PERCENTAGE AVERAGE TRIAL CONFIDENCE, PERCENTAGE CORRECT RESPONSES, AND OVERALL ESTIMATED PERCENTAGE CORRECT RESPONSES, MEAN AND STANDARD ERROR PER CONDITION	125
FIGURE 6.1 – S4 PERCENTAGE H, M, FA AND CR CONFIDENCE, MEAN AND STANDARD ERROR PER CONDITION	135
FIGURE 6.2 – S4 PERCENTAGE AVERAGE TRIAL CONFIDENCE, PERCENTAGE CORRECT RESPONSES, AND OVERALL ESTIMATED PERCENTAGE CORRECT RESPONSES, MEAN AND STANDARD ERROR PER CONDITION	139
FIGURE 7.1A – S5 TRIAL EXAMPLE SCREENSHOT; STYLUS INDICATES NO ERROR	146
FIGURE 7.1B – S5 TRIAL EXAMPLE SCREENSHOT; STYLUS INDICATES SUPPOSED ERROR	146
FIGURE 7.2 – S5 $D'_{\text{PARTICIPANT}}$ MEAN AND STANDARD ERROR, D'_{STYLUS} , AND INTERACTION MODELS D'_{TEAM} PREDICTIONS PER GROUP. CF = COIN FLIP; PM = PROBABILITY MATCHING; OW = OPTIMAL WEIGHTING; UW = UNIFORM WEIGHTING.	159
FIGURE 7.3 – S5 YES/ NO BIAS ($C_{Y/N}$), MEAN AND STANDARD ERROR PER GROUP	161
FIGURE 7.4 – S5 EFFECT OF STYLUS ADVICE ON PERCENTAGE H, M, FA AND CR CONFIDENCE, MEAN AND STANDARD ERROR PER GROUP	164
FIGURE 7.5 – S5 PERCENTAGE AVERAGE TRIAL CONFIDENCE, PERCENTAGE CORRECT RESPONSES, AND OVERALL ESTIMATED PERCENTAGE CORRECT RESPONSES (= POST-TASK FREQUENCY CONFIDENCE MEASURE), MEAN AND STANDARD ERROR PER GROUP (NOT CORRECTED FOR $\pm SD * 1.5$)	168

List of tables

TABLE 2.1 – STYLUS STUDY DESIGNS	53
TABLE 2.2 – SIGNAL DETECTION MATRIX	56
TABLE 2.3 – STYLUS SIGNAL DETECTION MATRIX	57
TABLE 2.4 – STYLUS SIGNAL DETECTION EXAMPLE MATRIX	63
TABLE 3.1 – S1 NUMBER OF RESPONSES PER CATEGORY PER GROUP	80
TABLE 3.2 – S1 MEAN CONFIDENCE PERCENTAGE PER CATEGORY PER GROUP	84
TABLE 4.1 – S2 NUMBER OF RESPONSES PER CATEGORY PER GROUP	95
TABLE 4.2 – S2 MEAN CONFIDENCE PERCENTAGE PER CATEGORY PER GROUP	100
TABLE 5.1 – S3 STYLUS LIKELIHOOD ESTIMATION DISTRIBUTION	116
TABLE 5.2 – S3 MEAN NUMBER OF RESPONSES PER CATEGORY PER CONDITION	118
5.4.2.1 PROPORTION OF CORRECT RESPONSES BETWEEN CONDITIONS	118
TABLE 5.3 – S3 MEAN CONFIDENCE PERCENTAGE PER CATEGORY PER CONDITION	122
TABLE 6.1 – S4 ITEM DISTRIBUTION OVER STYLUS RELIABILITY CONDITIONS	132
TABLE 6.2 – S4 MEAN NUMBER OF RESPONSES PER CATEGORY PER CONDITION	133
TABLE 7.1 – S5 NUMBER OF TRIALS PER CONDITION, PER GROUP	152
TABLE 7.2 – S5 H, M, FA, CR DISTRIBUTION OVER CONDITIONS PER GROUP	153
TABLE 7.3 – S5 H, M, FA, CR DISTRIBUTION OVER SENTENCE CORRECT AND INCORRECT CONDITIONS FOR THE PURPOSE OF D'_{STYLUS} AND C_{STYLUS} PER GROUP	155
TABLE 7.4 – S5 H, M, FA, CR DISTRIBUTION FOR THE PURPOSE OF D'_{STYLUS} AND C_{STYLUS} PER GROUP	155
TABLE 7.5 – S5 CONFIDENCE ANOVA FACTORS AND LEVELS	156
TABLE 7.6 – S5 MEAN ABSOLUTE PERFORMANCE PER CATEGORY PER GROUP	158
TABLE 7.7 – S5 MEAN PERFORMANCE RATES PER CATEGORY PER GROUP	158
TABLE 7.8 – S5 MEAN CONFIDENCE PER CATEGORY PER GROUP	163

Introduction

Although terms like "Interaction Design", "User-centred design" and "Human-centred design" are ubiquitous in among others the fields of Sociotechnical Systems Design, Ergonomics and Human-Computer Interaction, human interactions under uncertainty with imperfect (semi) automated systems are still often ill understood, both in theory and in practice. Initially we set out to explore practice, the design of user interface design variations of automated decision aids, but early in our research we learned that the theory of the fundamental relation between users making uncertain judgements and advice from automated aids should be addressed first to get a better understanding of how interface design might be improved.

The general aim of this thesis is to examine how human decision makers' decisions under uncertainty and their confidence in these decisions might be affected by an automated system's advice, especially when that advice itself is somewhat uncertain or imperfect. We also show how Signal Detection Theory (SDT; Macmillan and Creelman 2005) can be used as a method to analyse performance and confidence in this type of cognitive task, and especially how the construct of bias can be appropriated to measure users' willingness to accept or reject advice.

For our research we have developed an experimental paradigm, based on the familiar concept of suggestions for spelling and grammar alternatives in word processors, and the novel factor of the system offering its own estimation of the likelihood of its suggestions being correct. We use this paradigm to explore the influence of participants' trust in similar systems and their perception of their own efficacy on their performance when working with our system, and on the confidence they have in their decisions in conjunction with the system's advice. We regard user confidence, and in particular the calibration of user confidence with performance, to be a key metric for the usability of advice systems; our paradigm allows this idea to be explored.

One innovation in our research, is that we measure self-reported confidence in the domain of interaction with grammar and spelling checkers in three distinct ways: prior to a task (as perceived self-efficacy), after a single event, and after a complete task (a series of events). We observe that although users acknowledge the system, they are overconfident, and underutilise the help from the automation.

The first two chapters of this thesis are a literature review and an overview of our research approach and methodology. The remaining chapters are a chronological account of our experimental research work, followed by a round-

up of the conclusions we have reached at the end of each chapter, and a discussion with recommendations for future research and potential for implementation of our findings in user interface designs.

We start Chapter 1, *Subject area background and literature review* with a discussion of key literature around human decision making under uncertainty, such as Edwards' model of riskless and risky choice, and the division between risk and uncertainty (Edwards 1954). This leads to a review of literature relating to uncertain decision making in relation to adaptive and adaptable automation (e.g., Parasuraman and Wickens 2008, Kidwell, Calhoun, Ruff, and Parasuraman 2012), automated decision-making aids (e.g., Woods 1985, Robinson and Sorkin 1985, Woods, Johannesen, and Potter 1991), and interaction with Artificial Intelligence (AI; e.g., Doran, Schulz, and Besold 2017, Wang and Yin 2021). Often observed behaviours in human interaction with assistive technology in the literature, are underutilisation and overreliance on automated aids (e.g, Parasuraman and Riley 1997).

Important notions we discuss are among others trust (e.g., Muir 1987, Lee and Moray 1992, Lee and Moray 1994), perceived self-efficacy (e.g., Bandura 1997, Bandura 2006, Carroll and Reese 2003), confidence (e.g., Allwood and Montgomery 1987, Stankov, Kleitman and Jackson 2015), users' perception of system reliability (e.g., Dzindolet, Pierce, Beck and Dawe 1999 and 2002, Wiegmann 2002, Rice, and McCarley 2011) and related personal beliefs that play a role in human interaction with automated aids. The interplay between trust, perceived self-efficacy, and confidence is widely recognised as an important factor in predicting users' acceptance of, and reliance on automation (e.g., Lee and Moray 1994, Moray, Hiskes, Lee and Muir 1994, Wiczorek and Meyer 2019), but there is little agreement on the exact mechanisms of this relationship. There are also controversies in the literature around different measures of confidence and overconfidence (e.g., Kahneman and Tversky 1973, Gigerenzer, Hoffrage, and Kleinbölting 1991, Moore and Healy 2008), and in this light we discuss, in the following chapter, how we test these factors in our experiments.

In Chapter 2, *Research approach and methodology*, the operational side of our experimental research is explained. We discuss the design of our experiments, the two different SDT models that we use in our experiments, i.e., Two Alternative Forced Choice (2AFC) and Yes/No (Y/N) (Macmillan and Creelman 2005), and how we test factors such as trust, perceived self-efficacy, confidence, sensitivity, and bias. We also explain the methodology we used to test four models of interaction with automated aids, based on a comparison of seven models by Bartlett and McCarley's (2017), which we discussed in Chapter 1.

In Chapters 3 and 4, which describe our first experimental studies that we name Stylus 1 and Stylus 2, we discuss the influence of participants' prior trust in spelling and grammar checking systems, and their perceived spelling and grammar self-efficacy on their performance and confidence in an experimental

decision-making task. Participants had to choose which one of two sentences was better, the "Original" sentence written by a human, or an alternative purportedly being a correction from an automated system called "Stylus".

In Chapters 5 and 6, describing Stylus 3 and 4, the experimental task is similar, but we shift our focus to how users' performance and confidence are influenced by systems that display their own judgement of the likelihood of their suggestions being correct, when such systems perform at different levels.

In Chapter 7, our last experimental chapter that describes Stylus 5, we acknowledge some limitations of our experimental set-up and modify the design of the cognitive task so that it is more in line with perceptual experiments in the literature that use the same analysis framework. We also test four different statistical models of aided interaction.

The research contributes to knowledge of human interaction with complex sociotechnical systems under uncertainty, and to the usefulness of certain novel research methods in this domain. Our key findings show that users can indeed benefit from advice in judgment tasks of this sort, even when their performance is rather good, and the aid's performance is imperfect. We find that prior perceived self-efficacy and trust in spelling and grammar suggestions in general have a weak effect on users' performance and confidence in an experimental task, but that users can benefit from, and acquire trust in, a system that gives some useful information alongside a reliable representation of its own reliability. We also demonstrate that an adaptation of the Signal Detection model is a viable tool to analyse users' performance and confidence in this type of cognitive decision-making task.

Having introduced the subject, we now briefly anticipate the main findings and contributions of this thesis.

Summary of the main observations from the literature tested in our research

- Confirmation of the *Above Average Effect* (Dunning, Meyerowitz, and Holzberg 1989).
 - *Evidenced among others by confirmation of S1-H1; S2-H1*
- Confirmation of the *Overconfidence Effect*, and how it can reduce or disappear by comparing single-event probabilities and post-task frequencies (Gigerenzer, Hoffrage, and Kleinbölting 1991; Gigerenzer 1994; Kahneman and Tversky 1996).
 - *Evidenced among others by confirmation of S1-H5; S2-H5; S2-H6; S3-H7; S3-H9; S4-H9*
- Confirmation that users can benefit from advice from imperfect automated aids (Wickens and Dixon 2007) – Higher aid reliability positively affects users' performance.
 - *Evidenced among others by confirmation of S3-H1; S4-H1*
- Perceived self-efficacy and trust in similar systems, both measured prior to the task, had less of an effect on participants' performance than

hypothesised on basis of the literature (Lee and Moray 1994; Moray et al. 1994; Wiczorek and Meyer 2019).

- *Evidenced among others by rejection of S1-H3; S1-H4; S2-H3; S2-H4; S3-H3; S4-H10; S4-H11 and confirmation of S5-H5*
- None of four statistical interaction models from the literature (Bartlett and McCarley 2017, 2019) that we tested with S5 data, described or predicted participant–aid team sensitivity in our experiment very well. When the aid’s reliability was high (90%) all models overestimated team sensitivity, when it was just above the reliability threshold for usefulness (70%; suggested by Wickens and Dixon 2007) the Optimal Weighting and Uniform Weighting models overestimated, and the Coin Flip and Probability Matching models underestimated team sensitivity.

Summary of the main novel findings in this thesis

- Improving the reliability of the aid’s advice positively affects users’ performance, acceptance of the aid’s advice, trust in the aid, and the confidence they have in their own responses.
 - *Evidenced among others by confirmation of S3-H1; S3-H2; S3-H6; S4-H1; S4-H6, and a comparison between S3 and S4 (performance only)*
- Users can, to some extent, recognise and acknowledge the system’s level of performance, even without receiving feedback, as in S1 and S2. Significant differences between groups in bias towards following Stylus’ advice, trust in Stylus, and rating of believability of Stylus’ suggestions being created by an automated system, suggest effects of the reliability and strength of the system’s advice.
 - *Evidenced among others by confirmation of S1-H7; S2-H7; S3-H2*
- Users can recognise their own level of performance (metacognition), even without receiving feedback, as demonstrated by higher confidence in correct responses than in incorrect ones.
 - *Evidenced among others by confirmation of S3-H8; S5-H4*
- Even highly reliable aids are underused, most acutely if their performance is better than that of participants.
 - *Evidenced among others by performance in S4-C94 and S5-G90*

Summary of the main general contributions this thesis makes

- An exploration of various forms of personal beliefs of confidence and trust in knowledge-dependent cognitive linguistic judgement and decision-making tasks.
- An exploration of the potential of an automated text-editing aid that shares a statistically reliable estimate of the likeliness of its suggestions being correct with users.

Summary of the main methodological contributions this thesis makes

- Introduction of a novel experimental paradigm based on Signal Detection Theory to study different aspects of *performance* and *confidence* in a cognitive decision-making task.

- Introduction of the measure of bias, as propensity to accept automated advice, in an aided text-editing task.
- Introduction of an integrated framework for comparing three distinct measures of confidence in the domain of aided linguistic judgement:
 - prior to a task: as *perceived self-efficacy*
 - after a single event: as a probability of being correct
 - after a complete task (a series of events): as an estimated frequency of correct responses

Chapter 1 – *Subject area background and literature review*

1.1 Introduction

In the Introduction, we presented our overarching goal of better understanding human decision making under uncertainty when aided by imperfect automated aids that are themselves uncertain. We established that our experimental domain is that of linguistic judgement with the help of automated spelling and grammar checking aids, while our experimental approach is to present a series of cognitive judgement tasks, for which some aspects of performance, confidence and trust will be measured. We also analyse some aspects of performance by means of perceptual Signal Detection Theory (SDT) methods. In this chapter we discuss this conjunction from the perspective of the literature.

We start discussing the general experimental domain with a broad introduction to human decision making under uncertainty aided by advice from automated systems, followed by a discussion of the effects of ability, trust and perceived self-efficacy, and the perception of confidence users have in their own decisions. We also look at how metacognition may be used to assess joint human-automation decision making. Subsequently, the literature related to our experimental approach and models of aided interaction will be discussed, focussing on applications of methodology similar to ours. Lastly, we discuss research in the specific domain of interaction with text editing aids, and we take a brief look at typical categories of language errors relevant for our experimental studies and how word processors deal with them.

Some technical aspects of our methodology that are touched on in this chapter will be discussed in more detail in the following chapter, *Research approach and methodology*.

1.2 Experimental domain – decision making under uncertainty with imperfect and uncertain advice

In the first section of this literature review we concentrate on decision making and task allocation processes where humans must vet automated systems'

judgements and advice, in order to accept or override them, and how perceptions of self-efficacy and of trust in the system might affect this vetting process. The focus is on core concepts, while experimental research and results are mentioned where relevant.

1.2.1 Human decision making under uncertainty

1.2.1.1 Decision making under uncertainty

Decision making processes are part and parcel of humans' daily lives, with complexity ranging from reversible routine decisions (selecting a coffee cup from the cupboard in the morning), to potentially life-changing irreversible ones (diving off an unknown cliff). While decisions at the low-risk end of the scale, where all options are known and the consequences of the choice can be overseen (*riskless choice* (Edwards 1954)), can usually be made without any assistance, all but the most reckless humans will probably do research or seek advice before making a decision of the latter type. This type of decision making, which Edwards (1954) calls *risky choice*, hinges on two factors: risk (others, like Weisberg (2014), speak of *doubt*), and uncertainty (or *ambiguity* (Weisberg 2014)). Risk is a projection of the future that can be described with a probability, i.e., the likelihood that something will happen. Uncertainty on the other hand is a strength of belief, and has no generally accepted probability attached (Edwards 1954). These useful, if slightly confusingly named, distinctions between riskless choice and risky choice, and risk and uncertainty (see Figure 1.1) is widely used in academic domains like Philosophy, Mathematics, Economy and Finance (see e.g., Keynes 1921, Knight 1921, Mousavi and Gigerenzer 2014, 2017). However, it has to be noted that the notions of *risk* and *uncertainty* are sometimes used inversely, and that some scholars use them more or less interchangeably (e.g., Spiegelhalter and Riesch 2011), or as synonyms of respectively *aleatory* and *epistemic* uncertainty (Agarwal, Renaud, Preston and Padmanabhan 2004).

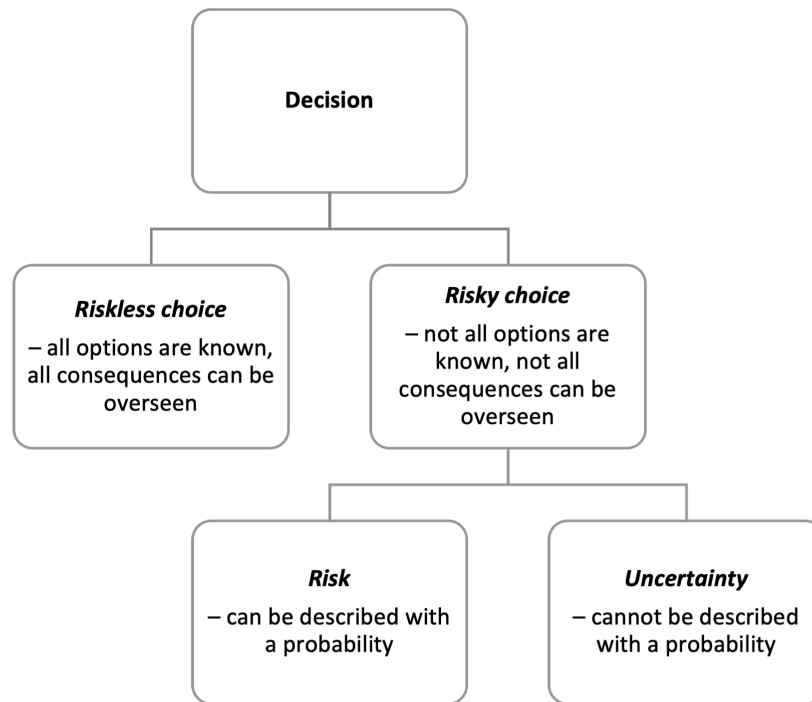


Figure 1.1 – Decision making diagram based on Edwards 1954

1.2.1.2 Automation and uncertainty

Edwards introduced the distinction between risk and uncertainty in the domain of Psychology in 1954, which coincided with a rising interest in human decision making in the 1950s, when decisions in business were increasingly made by, or in conjunction with, automated systems. Lee and See (2004) broadly describe automation as 'technology that actively selects data, transforms information, makes decisions or controls processes', and Parasuraman, Sheridan, and Wickens (2000) propose four function application classes for automation: information acquisition, information analysis, decision and action selection, and action implementation.

Automation is a continuum, that ranges from *lights out* systems that are fully automated and are supposed to require no human supervision (*decision making systems*), to sociotechnical systems that are designed for humans and technology to work together, ideally enhancing each other's qualities. Parasuraman (2000) presents a model for different types of automation, that can then each be indexed on a 10-level classification based on the work of Sheridan (1992), that itself builds upon Sheridan and Verplank's five-level classification (1978). Although decision making systems are automated to an extent that the user usually cannot influence decisions, they will still need periodic maintenance and occasional intervention (Brann, Thurman, and Mitchell 1996). Often parts of complex automated systems are subject to a division in which either the system (partially) allocates tasks to users, or users (partially) allocate tasks to the automation (*joint human-automation decision making*), or systems advise users in their decision making, for example by detecting state changes and anomalies and suggesting potential

improvements, e.g., suggestions for improvements in word processing (*decision support systems, decision aids, automated support systems*, e.g., Woods 1985, Trentesaux, Moray, and Tahon 1998, Skitka, Mosier, and Burdick 1999, Parasuraman, Sheridan, and Wickens 2000, Solomon 2014, Chavaillaz, Wastell, and Sauer 2016). The division of active and passive control and different levels of delegation of tasks between users and automation can be authorised by the user (*adaptable automation*) or by the system (*adaptive automation*) (Parasuraman and Wickens 2008, Kidwell et al. 2012). For illustrative examples of adaptable levels of automation in a rail traffic control context, see Balfe, Wilson, Sharpless and Clarke 2012.

Although the mathematical quantification of uncertainty in automation is an important factor in the performance and believability of (semi) automated decision aids, and indeed of Artificial Intelligence (AI) systems in general, the technical aspects of how uncertainty is mathematically modelled are not within the scope of our research. However, we note that the literature in this domain (see e.g., Agarwal et al. 2004) Roy and Oberkampff 2011, Smith 2013, Sullivan 2015, Ghanem, Higdon, and Owhadi 2017, Begoli, Bhattacharya, and Kusnezov 2019) is much more developed than the literature around human interaction with automated systems that communicate their own uncertainty. We aim for our current research to positively contribute to strengthening the latter.

1.2.2 Decision making by humans aided by judgements and recommendations from automated systems

1.2.2.1 Automated decision aids and their effect on decision making

As early as the mid 1980s, Woods (1985) pointed out that an increase in control automation has meant the emphasis on the role of humans has shifted from perceptual motor skills to cognitive skills of a supervisory and managerial nature. This has accelerated not only research into human operators, but also into joint human-automation systems (or similar terms such as *human-machine system*, e.g., Woods 1985, *human-intelligent system cooperation* e.g., Woods, Johannesen, and Potter 1991, and *person-machine system*, e.g., Robinson and Sorkin 1985) as integrated cognitive structures. The collaborative human-system decision making process is akin to interpersonal co-operative decision making, where judgements from different group members are combined to achieve the highest possible level of group sensitivity (Sorkin and Dai call this the *Ideal Group* (1994)).

Rice and McCarley (2011) argue that the level of performance of joint human-automation systems depends on the interplay of three factors: the human operator's performance level independently of recommendations from the system, the reliability of the system's recommendations, which we mentioned above, and the operator's reliance on the system's recommendations. Gadala, Strigini, and Ayton (2021) argue that the effectiveness of binary automated aids varies with users' ability and experience and the difficulty of a task, and

Meyer and Kuchar (2021) present the optimal effectivity of user interaction with binary aids (see below, *Binary cues vs. evidence sharing decision aids and strength of advice*) as a function of the combined sensitivities of the user and the aid. In the following sections we will dive deeper into the user aspects of this relationship by discussing users' ability and performance level, and their beliefs about their own and automated systems' ability and performance.

1.2.2.2 Ability and performance level independent of system advice

Our experimental research aims to better understand the potential benefits of the help of imperfect and uncertain advice in human decision making (*advisory interactions*, Woods, Johannesen, and Potter 1991), in interaction with text editing aids. Although our studies use participants with a potentially wide range of proficiency in spelling and grammar, we will not discuss the formation of users' ability here in detail, as although it is an important factor contributing to the performance of joint human-automation systems as pointed out above (see Rice and McCarley 2011), it is not one that can be addressed during a single-occasion task. We discuss users' beliefs about their ability in the paragraphs about perceived self-efficacy.

Participants' performance level during the task on the other hand, is influenced by among others a combination of their ability, and by factors that can change and thus be measured during a task. We discuss the ones most relevant to our research, such as humans' beliefs about their ability and their trust in systems, in the following sections.

1.2.2.3 Diagnosis vs. advice

Decision aids can perform two discretely different judgement functions, often in tandem: to *alert*, and to *advise* (or *recommend*). The purpose of the alert function is to indicate a situational change that may require action on behalf of the user, the recommendation function serves to advise on actions or alternatives (Parasuraman and Manzey 2010). An example of an alert is a red wavy underline under a potentially misspelt word in MS Word or a similar text editor, and an example of a recommendation is the list of alternative spellings the system provides when the underlined word is clicked on or hovered over.

1.2.2.4 Binary cues vs. evidence sharing decision aids and strength of advice

Alerts can be *binary* (system judgement meeting or missing a pre-set threshold, see e.g., Bartlett and McCarley 2017 and Meyer and Kuchar 2021), or *graded* (system judgement falls within a bandwidth). The *strength* of advice from graded systems may vary according to their level of uncertainty, or in other words, their own judgement of the likelihood of their judgement being correct. Graded systems, or *Likelihood Alarm Displays* (LADs, Sorkin, Kantowitz, and Kantowitz 1988) employ a form of evidence sharing, which can be direct, e.g., by accompanying judgements with *likelihood estimates* (or *alarm validity information*, Manzey, Gérard, and Wiczorek 2014), or indirect, e.g., by discretising levels of alarms into several distinct alarm categories (Bartlett and McCarley 2017), e.g., "traffic lights". Referring to the aforementioned two systems, Wiczorek (2017) speaks of *Binary Alarm*

Systems (BASs) and graded systems *Likelihood Alarm Systems* (LASs), and describes the latter as effectively binary systems with an added additional threshold.

1.2.2.5 Reliability of judgements and advice: reliability thresholds, and reliability perception

Just as the *strength* of a system's advice may vary, so does the *reliability* (also known as *accuracy*) of its judgements. The objective level of a system's reliability is measured by its overall sensitivity (i.e., its performance level), which can be estimated based on prior information (e.g., historical performance), generalised from similar systems, or calculated post-hoc (Johnson, Cavanagh, Spooner and Samet 1973).

Even if systems are imperfect and a system displays a certain level of false positives and/ or false negatives, several studies have demonstrated that users might still benefit from them (see Wickens and Dixon 2007 for a comparison of 20 experimental studies that suggests as much). We discuss system reliability thresholds in some more detail in the second part of this literature review, and *perception* of system reliability, which affects trust in a system (Wiegmann, Rich, and Zhang 2001), after we have introduced the concept of trust.

1.2.3 Trust in automation and explainable Artificial Intelligence

1.2.3.1 Trust between humans and systems analogous with interpersonal trust

Trust is a widely studied concept in a diverse array of domains, ranging from studies of interpersonal trust in Psychology (e.g., Rotter 1980) and Management (e.g., Mayer 1995), to research of trust between humans and AI systems in overlapping disciplines in the Social Sciences (e.g., Nowotny 2021 and Schoenherr 2022), Management (e.g., Glikson and Woolley 2020) and Computer Science (e.g., Rosenfeld 2021). Muir (1987) pioneered a model of trust in human-machine relationships based on the sociologist Barber's (1983) concept of the phenomenon of trust between humans. Muir adapted Barber's idea that trust is a multi-faceted concept, for example in a marriage, in politics or in business, which is based on the expectation that the other will demonstrate technically competent performance, and that they will act in good faith (i.e., that they will let common interest prevail above their own interest). Trust, in Muir's words, is '*[...] the expectation (E), held by a member (i) of a system, of persistence (P) of the natural (n) and moral social (m) orders, and of technically competent performance (TCP), and of fiduciary responsibility (FR), from a member (j) of the system, and is related to, but not necessarily isomorphic with, objective measures of these qualities.*'

Trust is widely accepted as an important factor in the interaction between humans and systems (see e.g., Muir 1987, Lee and Moray 1992, Lee and Moray 1994, Moray et al. 1994, Moray and Inagaki 1999, Bisantz and Seong 2001, Dzindolet, Peterson, Pomranky, Pierce and Beck 2003, Chavallaz,

Wastell, and Sauer 2016). If a decision aid is not trusted by a human user, they may reject it, or use alternative means that are potentially less time and energy efficient. No level of sophistication of the system can compensate for this dismissal. In their review article *Humans and automation: Use, Misuse, Disuse, Abuse* (which' key importance is itself reviewed by Lee (2008)), Parasuraman and Riley (1997) call this type of under-reliant behaviour of ignoring most, even correct, recommendations *disuse*, and among others Parasuraman (2000), Parasuraman and Manzey (2010), Wickens and Dixon (2007), and Prinzel (2002) talk of *complacency*. A factor that potentially leads to humans distrusting and as an effect potentially disusing a system, is a too-low False Alarm (FA) threshold, or a system alerting users at an inappropriately high system alert rate (Lee and See 2004), which we discuss later in more detail in section 1.3.3.1 *Reliability of judgements and advice: reliability and alerting thresholds*). This *cry wolf* effect is widely described in the literature (e.g., Wickens, Rice, Keller, Hutchins, Hughes and Clayton 2009 and Breznitz 2013), and Madhavan, Wiegmann, and Lacson (2006) found that it has a particularly detrimental effect on users' trust and reliance if tasks are perceived as "easy". Specifically relevant for our research, Gadala, Strigini, and Ayton (2021) describe the effect in relation to the use of spell-checkers, which we discuss in section 1.3.4 of this chapter.

The inverse problem, overreliance or *misuse* (Parasuraman and Riley 1997), is equally problematic. When users trust a system more than warranted by its performance, they may neglect errors the system has missed, or act on FAs, which in turn may lead to incorrect decisions. Referring to the same phenomenon, Swets (1992) likens it to the *engineering fail-safe approach*, Parasuraman and Manzey (2010) refer to it as *automation bias*, and Rice and McCarley (2011) call it *response bias*.

Parasuraman and Manzey (2010) observe that this type of bias, where users' critical analysis is replaced with blind trust in the automated system, may lead to two types of error: *error of omission* and *commission error*. An error of omission is a situation where a user fails to act because they have not received an alert from the aid (*Miss*, Macmillan and Creelman 2005), e.g., a text editor fails to underline a misspelt word and the user ignores the error. This may even be the case if the error is obvious, but the user decides the system "must know best". The second type, commission error, is the inverse, e.g., a word editor underlines a word that is obviously correct, but the user still acts on the system's alarm and accepts its recommendation (*False Alarm*, Macmillan and Creelman 2005).

Lee and See (2004) point out that reliance on automation is not just a binary process, but rather a more graded one due to the often complex nature of automation. To avoid a situation where users over or under-estimate the decision aid's capabilities, their trust must be calibrated to the aid (see e.g., Muir 1987, Lee and Moray 1992, Parasuraman and Riley 1997, McGuirl and Sarter 2006). Tomsett, Preece, Braines, Cerutti, Chakraborty, Srivastava, Pearson and Kaplan 2020 argue that to build users' trust in a system, they

need an adequate mental model (Payne 2003) of the system's knowledge that includes an idea of potential gaps in this knowledge. The project described in further chapters of this thesis utilises a simple method for calibration where users receive uncertain, yet honest advice from aids.

It must be noted that trust is not thought to be the only moderator in the use of decision aids. Factors like workload can also play a key role in disuse and misuse (Parasuraman and Riley 1997), for example because they might lead to a form of effort reduction (Davis and Tuttle 2013).

1.2.3.2 Prior trust vs. trust developed during a task

Other than expectation, which affects initial trust and serves as a predictor of adoption of new technology (Li, Hess, and Valacich (2008) indicate four construct levels: trusting base factors, trusting beliefs, trusting attitudes and subjective norm, and trusting intention), it is widely argued that experience plays an equally important part in the development of trust in a system (Hutton and Klein 1999, Moray, Lootsteen, and Pajak 1986) and that it can influence decision making behaviour over time, both in the short and the long term. Development of the level of trust users have in a system is not necessarily gradual over time, but depends on different factors and processes at different stages of the relationship (Li, Hess, and Valacich 2008).

Much of the trust literature focusses on the development of initial trust, or on trust development based on experience, while in our research we expect there to be a blend of a well-developed level of initial trust (prior to the task, our participants report a reasonably high level of trust in systems comparable with the one used in the experiments), and a gradually developing level of trust based on experience with the system at hand. Participants interact with a system that has some properties that are familiar, and some that are new. While initial trust formation in a system is general and depends chiefly on external factors, and long-term trust is influenced by experience, there are indications that the level of initial trust also might affect long term trust (see Manchon, Bueno, and Navarro 2021 for an example of this phenomenon in the context of highly automated driving).

1.2.3.3 System errors affecting users' trust

Trust in a system developed during its use is affected by users' perception of its performance, which is partly guided by the errors users see a system make (De Vries, Midden, and Bouwhuis 2003, Dzindolet et al. 2003). Where the task can be analysed by Signal Detection Theory, errors made by aids can be subdivided into system False Alarms (FA) and system Misses (M) (not to be confused with respondents' FAs and Ms). A FA occurs when the alarm is not justified by the event, whereas a M happens when a system misses an opportunity for raising the alarm. Several studies have found that FAs and Ms affect users' perception of the reliability of the system differently. In two experiments, Rice and McCarley (2011) tested the benefits of imperfect aids in a security screening task to see whether systems with a high FA rate would

affect users' performance more negatively than M-prone aids that were equally reliable, i.e., performed as well.

In the first experiment, participants performed a "baggage search" task on coloured x-ray images of passenger bags with everyday objects on a monitor, where presence of a knife had to be detected. The researchers created 180 pairs of images, each pair being identical, apart from the absence or presence of the knife, which if present was presented at a randomly selected angle in a random location. Participants were shown a randomly presented series of 180 images with a 50% signal rate, i.e., 90 images that contained the knife, and 90 that did not. Participants were randomly assigned to either a group that received assistance from an automated system that gave its own judgement of the knife being present or absent, or to a group that had to complete the task unaided. Participants in the aided group received help from a system that was 100%, 95%, 80% or 65% reliable, which was communicated with them prior to the task. The system's judgements were presented as a text message in each trial before the test image was shown. When the automation misjudged the images, it either missed the knife, or raised a FA, never both during the experiment for a single participant, and this was briefed to participants as well. This means that there were 8 conditions in total of which 7 were automation-aided; 12 participants were randomly assigned to each of the conditions. Participants were given feedback on their performance after each trial ("Correct!" or "Incorrect" text message).

Perhaps surprisingly, they found that although the automated aids at all levels of reliability improved participants' performance, FA-prone aids led to poorer user performance than M-prone aids with the same reliability. What is more important in the light of our research though, is that even though they were told that M-prone systems would never raise a FA, participants did not agree with all the system's target-present judgements, and vice versa for FA-prone systems. An ANOVA revealed a bias towards FA-prone aids over M-prone ones, no effect was found for automation reliability, nor for interaction between bias and reliability.

Even though presented at the same rate, the findings of the first study suggest that, as (Rice and McCarley 2011) note, 'automation FAs may be more likely than misses to be noticed, remembered, and/ or heavily weighted in the operator's judgements.' In their second study, attention was turned towards participants' perception of the aid and how this is affected by how the aid's advice is framed, based on the suggestion that M and FA-prone aids fundamentally differ in cognitive salience. Framing (Tversky and Kahneman 1989), or the way a decision problem is described, affects the way users treat the advice of automated aids. By comparing conditions in which the automation performed at a stable reliability (65% in this study), but errors are framed as neutral messages instead of incorrect diagnoses, an attempt was made to influence participants' behaviour. One of the main findings was that automation errors framed as neutral messages rather than errors reduced the effect of the aid's errors on users' performance. For users aided by an M-prone

aid, neutral framing eradicated the consequences of the aid's misjudgements on their compliance, while FA-prone aids' errors framed as neutral messages statistically significantly increased users' sensitivity, d' . The authors note that the gains made in ignoring the aid's misjudgements through reframing the aid's errors, comes at the cost of a reduction in agreement rates when the aid's judgement is correct in users assisted by FA-prone aids. Their overall conclusion is that potential positive effects of neutral framing likely depend on the reliability and response bias of the aid, and the trade-off between costs and benefits of FA-prone vs. M-prone systems. While the aids in our experimental studies make both FA and M-type errors and we have not experimented with negative and neutral framing of system errors, Rice and McCarley's research (2011) shows that users' perception of an aid might affect their performance. We will discuss our findings on participants' performance and self-reported confidence related to correct judgements vs. errors, and FAs vs. Ms, which suggest that users have an awareness of their own performance in the light of that of an aid.

1.2.3.4 Reliability perception as a factor affecting trust

As users' default approach to automation in general relies on the heuristic of automation being "perfect" (see e.g., Dzindolet, Pierce, Beck and Dawe 1999, 2002 and Wiegmann 2002), it is crucial that if aids are imperfect, users' perception of the automation is matched accordingly to counterbalance this positivity bias (NB: Hutchinson, Strickland, Farrell and Loft (2022) found that although in their experiments users' perception of reliability of automation systematically undercut actual reliability, acceptance of the automation was much closer to the actual reliability). A result of the over-expectancy of an aid's reliability, is that system errors are judged unjustifiably harshly, which leads to mistrust and underutilisation (Wickens, Helton, Hollands, Banbury 2021). We argue that these forms of mental processing by users affect their trust in a system, which in turn might affect performance, as discussed earlier. Dzindolet et al. (2003) observed on basis of three experimental studies that when users see a system make errors, their trust in the system diminishes, even if the system is highly reliable. A way of solving this, they argue, is to give users insights as to *why* systems make errors. In our experiments we cannot explain to users why the aid makes errors due to the uncertain nature of its advice, but we can explain the extent to which its advice is accurate, and the strength of the advice. We believe this will increase transparency, and we test whether this might in return positively affect trust and performance.

Rice and McCarley's experiments (2011) focussed on the overall performance of the system in terms of types of errors (FA vs. M), and reliability (expressed as a percentage correct performance). Guznov, Lyons, Nelson and Woolley (2016) studied users' interaction with an automated aid in an improvised explosive devices (IEDs) identification task, with a focus on error types and the *severity* of errors. Interestingly, they report several findings from the literature (they cite Dixon and Wickens 2006, Sanchez 2006 and Geels-Blair, Rice, and Schwark 2013) that FAs affect trust and reliance behaviour more negatively than misses, which, they presume, is perhaps because FAs are

more noticeable than Ms. This ties in with Rice and McCarley's finding (2011) that system FAs have a greater negative effect on user's performance than Ms. Since error severity is not a relevant concept in our research, we will not discuss Guznov et al.'s studies (2016) in any greater detail, but it bears mentioning that, after considering the limitations of their experiments, their conclusions are that 'there may not necessarily be a positive correlation between trustworthiness, subjective trust, and trust outcome reliance behavior' and that 'human-machine systems should be tailored to account for automation errors'. Gadala, Strigini, and Ayton (2021) draw similar conclusions, but they note that inexperienced users and specialists may be affected differently.

Wiegmann, Rich, and Zhang (2001) argue that quite how the relation between varying levels of system reliability and users' trust and reliance on automated aids that vary in reliability works exactly, is ill-understood and under-studied. They examined the concept of subjective confidence (which we discuss later in this chapter) in relation to perception of reliability, and objective measures of performance. Their main finding was that although users noticed an aid's different reliability levels, objective performance measures were related to subjective measures of confidence and reliability estimates, and their calibration was not perfect. From this, they concluded that there is a need to 'distinguish between automation trust as a psychological construct that can be assessed only through subjective measures, and automation reliance that can only be defined in terms of performance data.' We believe that this distinction between the subjective measure of automation trust on the one hand, and the objective measure of automation reliance on the other is important, because there is often a discrepancy between what users say, and what they do in interaction with automation (see e.g., Wickens et al. 2021).

Wiegmann (2002) notes that despite substantial individual differences in automation-adoption strategies, with some users following the automation blindly and others seemingly employing a probability matching strategy (see section 1.3.2.5 of this chapter), users' subjective reliability estimates of the automated aids were always lower than their actual agreement with the aid. To test this observation in the domain of spelling and grammar checking, in our studies we ask participants after the task how often they thought they had agreed with the aid.

1.2.3.5 Explainable Artificial Intelligence as a means to calibrate trust in users

Although many interactive systems are "opaque", in the sense that they do not let users know how they arrive at judgements and make decisions (see e.g., Norman 2002 and 2011), developments in human interaction with Artificial Intelligence (AI) are currently trending towards explainable or transparent, rather than "black box" AI systems (see e.g., Wang and Yin 2021). This development puts trust between humans and system in a new perspective. In Explainable AI (XAI) or Transparent AI (TAI), it is not so much the *effects* of a system that are to be trusted by the user, but the relationship becomes a much more active one based on trust in the *underlying mechanisms*, i.e., the

system's algorithms or rationale. Doran, Schulz, and Besold (2017) discuss four notions of increasing levels of system transparency: opaque systems (users have no insight into underlying mechanisms), interpretable systems (algorithms can be mathematically analysed by users), comprehensible systems (output is accompanied by interpretable symbols like words, visualisations etc.), and explainable systems (completely comprehensible explanations). They argue that the more explainable a system is, the more trustworthy users will find its algorithms, and the more trust they will have in the system operating accurately. A system that communicates an estimation of the likelihood of its judgements and/ or suggestions being correct, such as the systems in three of our experiments, can be classified as a comprehensible system because it shares an interpretation of its evidence.

We have to note here that we assume that the information provided by systems to increase transparency is *relevant* to the decisions to be made by users. Hall, Ariss, and Todorov (2007) demonstrate the effect of providing irrelevant information, by giving half of the participants in a basketball game prediction experiment useful statistical information that helps them predict results, and the other half with (irrelevant) team names as well. The latter group was more confident because they believed their knowledge helped them improve their predictions, whereas in reality their performance dropped because their reliance on the useful statistical cues decreased.

1.2.4 Metacognition, perceived self-efficacy, and confidence judgements

1.2.4.1 Using metacognition to assess joint human-automation decision making

Metacognition is *cognition about cognition* (Cox 2005): a human's ability to recognise their own successful cognitive processing (Fleming and Lau 2014), or their ability to effectively assess the accuracy of their own inferences (Garcia-Retamero, Cokely, and Hoffrage 2015). Schraw and Moshman (1995) observe a distinction between 1) metacognitive knowledge (which ties in with Fleming and Lau's definition (2014)), and 2) metacognitive control processes (to which Garcia-Retamero, Cokely, and Hoffrage's definition (2015) conforms). In our project, we have researched both categories in the form of 1) participants' beliefs about their abilities prior to an experimental task (perceived self-efficacy), and 2) their confidence during and after the task.

There is a broad confidence literature in the domain of Cognitive Psychology, that largely focusses on overconfidence and its causes and effects (see e.g., Garcia-Retamero, Cokely, and Hoffrage 2015). Because our research is concerned with gaining insights into the calibration of confidence in general rather than overconfidence specifically, we review literature from a range of academic disciplines.

Stankov, Kleitman, and Jackson (2015) observed two distinctive types of individual differences in confidence assessments in contemporary studies in

the literature: 1) personal beliefs in ability to accomplish tasks (i.e., perceived self-efficacy), which we discuss in the following section, and 2) post-hoc judgements of accuracy, or likelihood of success, which we discuss thereafter. Stankov, Kleitman, and Jackson (2015) argue that the first class describes self-beliefs without the necessity of evidence, whereas in the latter, judgements are made directly following a cognitive or behavioural act, which renders them *online*. Stankov, Kleitman, and Jackson (2015) observe that research of the two classes of measures have evolved independently and that not many empirical studies link them. Elaborating this link is one thing we aimed to do in our research by measuring participants' beliefs of self-efficacy prior to the task, judgement of the likelihood of success during the task, and confidence of overall performance after the task.

1.2.4.2 Personal beliefs of competence and confidence

If trust in automation, which we discussed earlier, is likely to increase the intended use of the automation, perhaps "trust" in one's own unaided performance is likely to diminish any such reliance. What follows is a section about personal beliefs of competence (*perceived self-efficacy*) and performance (*confidence*).

1.2.4.3 Avoiding potential confusion about the terms perceived self-efficacy, self-confidence and confidence

Although the terms *perceived self-efficacy*, *self-confidence* and *confidence* are often used somewhat interchangeably in Human Factors and HCI literature, we think this is potentially confusing because all three can relate to both the pre-trial condition (single event), pre-task condition (series of trials), and a participant's state during, or resulting from, the performance of a task. To avoid any potential confusion, we therefore consistently in this thesis use the term *perceived self-efficacy* when we talk about a human's personal pre-task or pre-trial beliefs about their own ability in relation to the task, and the term *confidence* when we discuss people's personal beliefs about the accuracy of their judgements (Allwood and Montgomery 1987, Stankov, Kleitman, and Jackson 2015). We do not use the phrase *self-confidence* at all. To warrant consistency, we follow this logic even when referring to literature that uses these terms differently or interchangeably but discusses the same concepts. Figure 1.2 shows a schematic of how we use these terms.

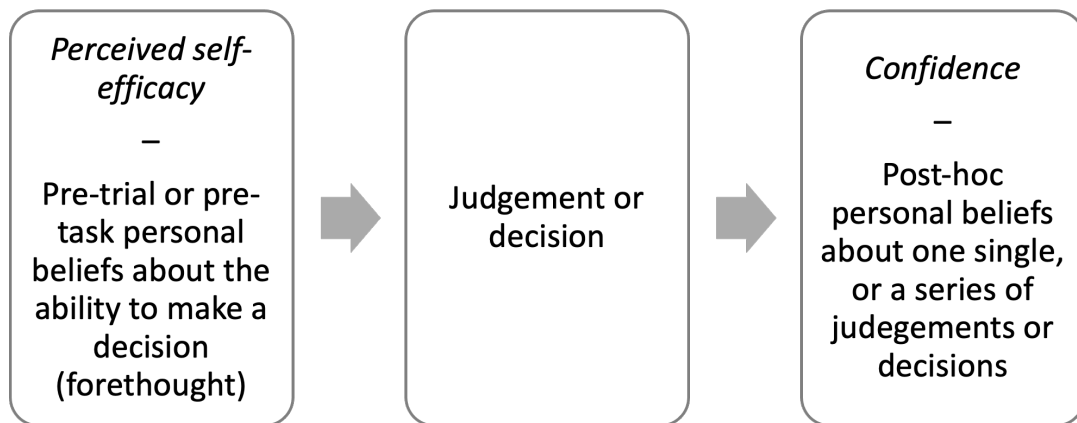


Figure 1.2 – *Perceived self-efficacy vs. confidence*

1.2.4.4 Perceived self-efficacy – personal beliefs about ability

Bandura (1997) coined the phrase *perceived self-efficacy* to describe 'a function of domain-specific beliefs about personal capacities' (Carroll and Reese 2003). In other words, as a way of describing a person's own estimation of their ability to "get a job done satisfactorily". Bandura (1997) describes perceived self-efficacy as a belief system about one's own ability based on four information sources:

- '[...] enactive mastery experiences that serve as indicators of capability' – I.e., successfully completing or failing tasks.
- '[...] vicarious experiences that alter efficacy beliefs through transmission of competencies and comparison with the attainment of others' – I.e., others (peers, role-models etc.) successfully completing or failing tasks.
- '[...] verbal persuasion and allied types of social influences that one possesses certain capabilities' – I.e., positive or negative feedback from others, including systems.
- '[...] psychological and affective states from which people partly judge their capableness, strength, and vulnerability to dysfunction.' – I.e., physical and emotional wellbeing, e.g., mood, stress, and pain.

Bandura notes that 'Any given influence, depending on its form, may operate through one or more of these sources of efficacy information' (1997). This effect of potential partial influence is important to note, because although we believe that perceived self-efficacy plays an important role in how participants interact with our experimental set-up, e.g., social comparison factors are not currently tested in our design (although we initially asked several pre and post-task social comparison questions; for social comparison in relation to a text-editing task, see Figueredo and Varnhagen 2005).

While (*self-*) *efficacy* refers to a person's actual ability, *perceived self-efficacy* refers to the judgement that person makes of their own *specific* proficiency level, so it does not denote their actual skill level, or their general global

knowledge level. As a result, there is no single measure of perceived self-efficacy (Bandura 2006). Actual self-efficacy can be built over the years, whereas perceived self-efficacy is much more dynamic and can even change instantly, for example when a person gets stuck in a process or successfully completes, or fails, a task (*enactive mastery experiences*, see above and e.g., Bandura 1997 and Bandura 2006). Although perceived self-efficacy may serve as a proxy for ability in some cases, it cannot be treated as such by default. However, Bandura (1984) explains that perceived self-efficacy is a reliable predictor not only of people's causal attribution of successes and errors, but also of actual performance.

1.2.4.5 Confidence: evaluating decisions

As pointed out earlier in this section, confidence is a person's post-hoc degree of belief about the level of their judgement or decision (post-trial (single event)) or performance (post-task (series of trials)). We will now discuss the two subclasses of confidence, a) single-event judgements of confidence, measured after each trial in a task, and b) judgements of overall confidence, or frequencies, measured after completion of a sequence of trials (Gigerenzer 1994, Gigerenzer, Hoffrage, and Kleinbölting 1991, Gigerenzer 1991). Figure 1.3 shows a schematic of how we use these terms.

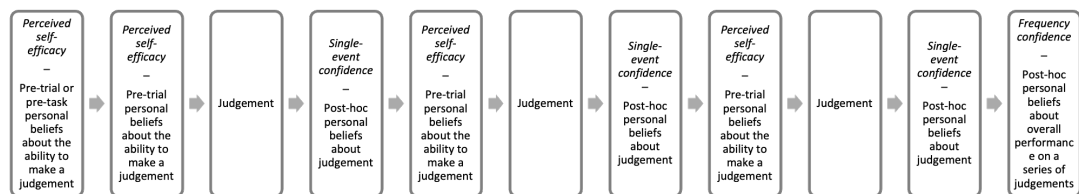


Figure 1.3 – Perceived self-efficacy, single-event confidence, and frequency confidence

1.2.4.6 Overconfidence and confidence calibration

Following on from the breakdown of confidence evaluations into single event and frequency measures, we identify a similar division of *overconfidence* measures in the literature: 1) pre-task, or forecasting, reports that are essentially measures of overassessment of perceived self-efficacy (see e.g., Dunning 2011), and 2) post-hoc measures of overconfidence that represent overestimations of performance (see e.g., Moore and Healy 2008). We subdivide the latter into 2a) post trial, or *online*, and 2b) post-task, or *post experiment*, appraisals.

Classic examples of 1), pre-task "overconfidence", or perceived self-efficacy, measures, are drivers, on average, thinking they are better drivers than average (DeJoy 1989, Svenson 1981, Wohleber and Matthews 2014), medical doctors assuming they make fewer diagnostic errors than their peers (Graber and Berner 2008), or people's willingness to pay a premium to bet on familiar areas (Heath and Tversky 1991). This *above average effect* (Dunning, Meyerowitz, and Holzberg 1989) or *illusory superiority effect* is a bias of

overestimation of one's own knowledge or skills in relation to others (Hoorens 1993).

The well-known Dunning-Kruger effect specifies this bias by describing how people with low ability in a specific domain tend to overestimate their skills and knowledge, whereas the inverse is true for those with high ability. The Dunning-Kruger effect hinges on the "double burden" of unawareness of the lack of competence, combined with unawareness of this deficiency, or as Dunning (2011) describes it, '*meta-ignorance (or ignorance of ignorance)*'. (NB: the equivalent phrase "unknown unknowns" that is also used in this context in the literature (see e.g., Dunning 2011) is perhaps too tainted by U.S. Defence Secretary Donald Rumsfeld's use in 2002 in the run-up to the Iraq war (Rumsfeld 2002) for it to be carried forward in the context of self-awareness of competence in human-system interaction.) In our experimental set-up we have no measures of participants' actual ability available, but we assume that their reported perceived self-efficacy relates to ability, albeit in a complex way moderated by the effects noted above. We only report perceived self-efficacy and its relation to performance and the use of automation.

The discrepancy between perceived and actual ability is usually measured by comparing self-assed performance with objective frequencies. We have to note that the concept of the Dunning-Kruger effect is not free of controversy, and it is viewed as a mere statistical artifact by some (e.g., Gignac and Zajenkowski 2020). Without wading too deeply into this statistical debate, we believe that in the context of our experimental research it is at least useful to compare perceived self-efficacy ratings with performance as a means of establishing to what extent perceived self-efficacy might predict or contribute to performance, and to reliance on information.

As outlined above, in our current research we use the term *confidence* to indicate a post-hoc self-reported estimation of *performance*, as opposed to the measures of perceived self-efficacy we described above. We follow this logic for overconfidence measures. Moore and Healy (2008) identify three types of performance overconfidence: '(a) overestimation of one's actual performance, (b) overplacement of one's performance relative to others, and (c) excessive precision in one's beliefs'. In the context of our experimental research, only the first type, which we will discuss in the next paragraph in more detail, is relevant. The second is a measure of social comparison that is outside the scope of this research, and the third relates to *precision* (e.g., number of decimal places in estimation of value), while our research deals with *reliability*, or *accuracy* (closeness of estimation to true value) of self-reported confidence measures.

As indicated earlier, we subdivide Moore and Healy's '*overestimation of one's actual performance*' (2008) into two subclasses: 2a) post trial (probabilistic) and 2b) post-task (frequency) measures of overconfidence. Classic examples of 2a) are studies where participants are asked trivia questions followed by a measurement of the confidence they have in their response, where they report

a mean confidence that is systematically higher than their mean performance warrants. Questions like 'Which city has more inhabitants? a) Hyderabad, b) Islamabad – How confident are you that your answer is correct? [50% – 100%]', are often used to demonstrate the overconfidence effect (Gigerenzer, Hertwig, Hoffrage and Sedlmeier 2008). Although the correctness of the response to which the probability is assigned might be checked later, these probabilities are spontaneous and subjective, and thus never "true" or "false" (Lichtenstein, Fischhoff, and Phillips 1977). Despite Lichtenstein, Fischhoff, and Phillips's assertion that individual probabilities themselves cannot be invalidated (1977), they suggest that a participant's level of calibration can be checked by comparing the means of their confidence ratings with actual performance measures. A judge (participant), they argue, is well-calibrated if 'over the long run, for all propositions assigned a given probability, the proportion that is true is equal to the probability assigned.'

In post-task frequency measures of overconfidence (2b in the previous paragraph), confidence is measured by asking study participants after completing a task (a series of trials in a study) how many of their responses they think were correct (Gigerenzer et al. 2008). Gigerenzer, Hoffrage, and Kleinbölting (1991) seek to frame overconfidence as a fabricated problem. Their work is largely a critique of social psychology in general, and Kahneman and Tversky's work on heuristics and biases (see e.g., Kahneman and Tversky 1973 and Kahneman, Slovic, Slovic and Tversky 1982) in particular (for a detailed analysis of the arguments of these two "camps" and the third one of the "Thurstonians", see Ayton and McClelland 1997). Their definition of overconfidence as 'mean confidence is higher than percentage of answers correct' makes it possible to compare participants' confidence over a series of trials in an experimental study with their confidence after completion of a task, our class 2b). By thus framing confidence as a frequency effect and comparing it with a series of probabilities that suggest a bias, Gigerenzer (1994) demonstrates that the overconfidence effect can completely disappear if confidence is measured differently. He argues that people are much better at gauging their performance after completing a task than after individual trials, which he takes to serve as evidence that the overconfidence effect is illusory. By their own account (Kahneman and Tversky 1996), Kahneman and Tversky largely agree with Gigerenzer. In earlier work, they dubbed probabilistic measures *inside view*, and frequentist ones *outside view*.

Gigerenzer et al. (2008), in what is largely a recount of a fierce debate with Kahneman and Tversky from the 1990s, claim that although overconfidence has shown to be systematic (i.e., certain levels of overconfidence match set levels of performance) and robust if tested this way, it is an effect of a combination of three factors. The first is the type of questions used by the researchers (which are usually not random), the second the sampling technique they employ (the selection of questions is usually not random either, and thirdly the phenomenon of regression to the mean (also see Gigerenzer, Hoffrage, and Kleinbölting 1991 for discussion of these three phenomena) . Rather than a personality trait that should be treated as a "bias", they argue, it

is a research artifact, that can be made to appear, disappear, or reverse depending on how the research is carried out. Asking participants in experimental studies to estimate a frequency of correct answers, our type 2b), results in reasonably adequate approximations of performance.

We described earlier how we intend to measure perceived self-efficacy and performance, which connect forecast and actual ability, in our experimental studies. We also compare the means of participants' subjective self-reported confidence after each trial with their post-task estimation of frequency to test Gigerenzer's hypothesis that the overconfidence effect disappears post-task, and that the description of the calibration of participants' confidence is merely a result of the way it is defined (1994).

1.2.4.7 Trust vs. confidence, or the interplay between trust and confidence?

Although the interplay between users' trust in a system and their confidence is widely acknowledged as an important factor in predicting operators' task allocation strategies in joint human-automation systems general (e.g., Lee and Moray 1994, Moray, Hiskes, Lee and Muir 1994, Wiczorek and Meyer 2019), there is no universal agreement on the exact nature of this relationship, and it is not always fully clear if perceived self-efficacy or confidence is measured. Below, we discuss two studies that focus on the relationship between humans' trust in an automated aid, and their self-reported confidence in their performance.

To examine the relationship between trust and performance confidence in complex process control, Lee and Moray (1994) performed a series of lab-based microworld- type experiments with a simulated semi-automatic pasteurisation plant. Microworlds simulate the complexity of real-world (work) environments, but at the same time allow for a high level of control of the experimental conditions (Brehmer and Dörner 1993, Rigas, Carling, and Brehmer 2002, Lee and Moray 1994). In the orange juice pasteurisation plant experiments, participants had to control a mock industrial process, thereby relying on automated control algorithms, the option to switch to controlling some of the processes manually, or a combination of the two. The system state could be monitored real-time as a graphical plant diagram on a monitor, and occasional warning messages were provided by the system as well. The information provided to participants was enough to run the plant, but not complete. Each of three sub-systems had its own control algorithm, and no matter what control mode (automatic or manual) was selected, there was always a delay in responses from the system. This delay varied for the different sub-systems. In addition to controlling the process, participants had to frequently log process data. This sub-task was intended to mimic real-world task load conditions and encourage the use of automation. By fully relying on the automation participants could achieve a good result, by fully relying on manual control the maximum attainable outcome would be lower, and by employing a combination of automated and manual control maximum efficiency could be reached. Participants received extensive briefings and training before starting the experiments proper, and during the experiments

they had to balance the rivalling goals of safety and performance. All participants were non-professionals, so the development of their trust and confidence could be monitored. After running the plant for a while, participants were confronted with faults in either the manual or the automated system, that could be sidestepped by participants by changing their control allocation strategy. Participants were asked to rate their trust in the automation and their confidence, both in the overall system and in the individual sub-systems. With their experimental set-up, Lee and Moray (1994) found that trust in combination with confidence predicted operators' strategy for choosing between automated and manual control: 'In general, automation is used when trust exceeds self-confidence and manual control when the opposite is true.' In other words: trust and confidence are communicating vessels.

Wiczorek and Meyer (2019) followed up on this presumed relationship between trust and confidence, describing it as commonly projected as depending on the weighting of both values. Our projection of their hypothesis is, again, that of trust and confidence as communicating vessels: if confidence is higher than trust users will rely on their own ability, whereas if it is lower, they will rely on the system. Important factors affecting trust from the literature, they write, are the earlier discussed issues of the commonly employed fail-safe approach (high number of FAs to minimise number of Ms), and of the negative effect of FAs on users' trust in automation, even if a system is in fact highly reliable. They argue that a higher level of confidence should decrease the level of system reliance, whereas higher trust should increase it. This hypothesis, among others, was tested in an experimental study in which randomly assigned participants had to carry out a signal detection task aided by either a high or low sensitive aid. The highly sensitive aid outperforms human sensitivity, whereas the low sensitivity one underperforms in comparison with the sensitivity level of the study participants. The task was a simple selection task where participants had to identify "intact" and "faulty" products by clicking a "sort out" or "pass" button on a screen, based on the length of a bar on the monitor. The automated decision aid provided an onscreen cue, in the form of either a red box with the message "sort out", or a green "pass" one. Performance, self-reported trust, and confidence were recorded, and the performance measures d' (sensitivity) and c (bias) were calculated with the usual SDT analyses (see Chapter 2 of this thesis), as well as measures of *reliance* (action on presence of alarms) and *compliance* (lack of action in absence of alarms). The most important findings in the light of our current project, were that there was a miscalibration between confidence and trust that may be explained as resulting from overconfidence or of undertrust, and that if two (human or automated) decision makers of different sensitivity levels collaborate, sensitivity of the joint human-machine system remains below that of the better one. Or in other words, if this effect generalises: highly sensitive automated aids will tend to get underused, whereas low sensitivity ones will be over-relied on.

1.3 Experimental approach – modelling and testing collaborative complex decision making under uncertainty

In this section of our review of the literature, we discuss several conceptual models of humans interacting with automated aids under uncertainty, followed by experimental models of human decision making and of complex joint human-automation collaboration. Although most of the models rely on statistical decision theory, the focus in this chapter is on their concepts and where relevant mathematical formulae will be discussed in the following chapter, *Research approach and methodology*. What all the models we discuss have in common, is that they are described as forms of *signal detection systems* (Sorkin and Woods 1985).

In the literature the terms *strategy* and *model* are often used interchangeably (e.g., in Duncan-Reid and McCarley 2021). In this thesis we use the term *strategy* to describe users' (plans of) action, and *model* to describe optimal strategies.

1.3.1 Modelling uncertainty in people and in automated systems

In joint human-automation systems there are two agents: a human (or more specifically, e.g., *user* (e.g., Gadala, Strigini, and Ayton 2021), *operator* (e.g., Moray, Sanderson, and Vicente 1992 and Rovira, McGarry, and Parasuraman 2007), *detector* (e.g., Pollack and Madans 1964 and Robinson and Sorkin 1985), *supervisory controller* (e.g., Moray and Inagaki 1999)), and an automated decision aid (or, e.g., *automated monitor* (Sorkin and Woods 1985), *alerted monitor* (Sorkin and Woods 1985), *computer alerting tool* (e.g., Gadala, Strigini, and Ayton 2021)). Together the two agents form a system that is supposed to collaboratively outperform the individual agents, but in reality often performs even worse than the worst of the two would on their own (Parasuraman and Riley 1997, Parasuraman 2000, Gadala, Strigini, and Ayton 2021), and also worse than most models suggest (Bartlett and McCarley 2017). These interesting paradoxes are widely acknowledged, but quite how this works is a more disputed subject.

1.3.2 Testing joint human-automation system behaviour: experimental models of collaborative decision making

We observed earlier that users' interaction with decision aids is often not optimal, which is shown by performance being lower than what can be statistically expected (Bartlett and McCarley 2017, 2021, 2019). Bartlett and McCarley (2017) tested the performance of joint human-automation systems in two experimental studies against seven different statistical models of collaborative decision-making strategies from the literature, which all have

different levels of efficiency. Some of the models have an origin in binary task (e.g, "yes/ no") group decision making literature (see e.g., Davis 1992 for a discussion on aggregate interpersonal decision making in general, and Sorkin, Hays, and West (2001) for a specific discussion about Group Signal Detection Theory).

We follow Bartlett and McCarley's lead by describing the seven models they mention, followed by a discussion of their experiments that test the models. We only describe the concepts of the models here, for formulae see Bartlett and McCarley (2017), and where relevant the chapter *Research approach and methodology* from this thesis. As context, it is worth reiterating that the paradigm task being considered in the models is perceptual signal detection. For example, to use the task of Bartlett and McCarley's own studies (2017), a stimulus might show a large number of blue and orange dots, and the participant is asked to judge whether the proportion of blue or orange is the larger. (One of the inequalities is arbitrarily chosen to be 'signal', the other 'noise'.) Or, to give a further example, a large array of letters might be presented, and the task would be to note the presence (or absence) of the letter Y (signal).

1.3.2.1 Contingent Criterion (CC) model

Robinson and Sorkin's pioneering CC model (1985) has long been the yardstick for models of joint human-system decision making. The model postulates that in an aided binary signal detection task, judgement depends on the user and the aid operating consecutively (Elvers and Elrif (1997) describe this conditional nature as in effect 'two cascaded signal detection systems'). In other words, the user has to make two judgements, one on the raw data, and one on the aid's judgement (see e.g., Scott-Sharoni, Yamani, Kneeland, Long, Chen and Houpt 2021 for perceptual separability of both sources).

The automation continuously monitors and evaluates processes, and in case of specific conditions alerts the user about perceived anomalies. This model assumes that the user considers the aid's judgement first, before analysing the available information, and either agreeing or disagreeing with the system's verdict. If deemed required, they can then follow up with further action. The system's alerts are binary (alarm or no alarm), and it does not present any background information on its judgements. The alert can either be presence of a signal, or the absence thereof, based on either meeting or missing a pre-set criterion. It is assumed that if the system gives a positive cue (signal present), the user will respond more liberally than if the cue is negative (no signal). This way the user's judgement will build upon that of the system, supposedly resulting in a performance improvement compared to that of either the automation or the human alone.

Robinson and Sorkin (1985) tested their model in two experiments. In the first one, four participants had to detect a tonal signal amid brief noise bursts in eight blocks of 100 trials, and they were occasionally helped by a simulated

alarm system. Performance of this simulated system was carefully matched to participants' level. The conclusion was that the combination of the human participant and the alarm system did indeed perform better than either alone. In the second experiment, three participants performed two tasks simultaneously. The first task, which served as a distraction from the second, was a scrolling letter detection task, the second was a diagnostic decision task aided by a simulated automated detector. Both tasks were presented on a video display terminal. In the second task participants had to recognise signals in number arrays on the monitor. They occasionally heard an audio signal, indicating a signal detected by the automation. The conclusion was again that the combination of the automated aid and the human performed at least better than the poorer of the two, and usually outperformed the best judge as well.

Although the main premise of the article, human-automation system performance depends on human behaviour as well as on a system's alarm threshold, is widely accepted (see e.g., Gadala, Strigini, and Ayton 2021), Robinson and Sorkin's model (1985) has attracted criticism as well since its inception. One of the main criticisms is that the model describes an ideal scenario, and does not necessarily predict real-world practice (e.g., Bartlett and McCarley 2021). Another critique is that the model is overly rigid, and that it assumes participants always consider a system's judgement before making their own judgement (see for a discussion on the influence of experience level e.g., Gadala, Strigini, and Ayton 2021).

We argue that this model is not very useful to describe user-aid interaction in our experiments, because one aspect of this model is that the aid signalling the need to make a decision serves as the first of two decisions made by the user – which is not true of our experiments, where the need for the user to make a decision is there on every trial.

1.3.2.2 Best Decides (BD) model

The BD model was initially presented to describe decision making between pairs of humans (see e.g., Bahrami, Olsen, Latham, Roepstorff, Rees and Frith 2010), assuming each of both actors would be able judge their own sensitivity in relation to the other. If applied to human-aid relations, the model assumes the user recognises their own sensitivity in relation to that of the aid. If they think they are more sensitive than the aid they will ignore it and make their own judgement, if they think the aid knows best, they will follow it by default. This simple model is essentially one of relative hierarchy of sensitivity of user and aid, with the basic premise that the best judgement automatically prevails. Bartlett and McCarley (2017) note that this simpler model makes less efficient use of the combination of the user's and the aid's judgements than the CC model, and that it produces lower levels of user-automation sensitivity. They also note that observed real-world observation of automation-aided performance often still undercuts the predictions of the BD model.

1.3.2.3 Yes/Yes (Y/Y) and No/No (N/N) models

Under the Y/Y and N/N models (Pollack and Madans 1964), which are either very conservative or very liberal, users and systems collaboratively make parallel judgements. This means that the Y/Y model assumes a "signal is present" (H) judgement is reached if both the user and the systems judge there is a signal. If only one of both judges "signal" (N/Y or Y/N), this counts as a M if in reality a signal is present. The inverse is true for the N/N model, which decides on CR (N/N) or FA (N/Y or Y/N) rate. Bartlett and McCarley (2017) note also these models are inefficient, and like the BD model, overestimate sensitivity in automation aided performance.

1.3.2.4 Coin Flip (CF) model

The CF model is another adaptation from Bahrami et al. (2010) from a human decision making to a joint human-automation decision-making context. Like the Y/Y and N/N models, it assumes that a valid decision is based on agreement between the user and the aid. However, where disagreement leads to respectively M or FA in the aforementioned models, the CF model assumes that the disagreement is resolved with a random selection of an answer (signal vs. no signal) with equal probabilities. Bartlett and McCarley (2017) suggest this model 'might provide a more plausible and better-fitting process model of human-automation performance' than that of Pollack and Madans (1964), because it reflects the inefficiency of the combination of user and aid that was observed to be problematic in the reliability of the earlier discussed models.

1.3.2.5 Probability Matching (PM) model

The element of randomness is also present in the PM model (coined by Humphreys (1939) in relation to a conditioning experiment), which is conceptually closely related to the CF model. The difference though, is that although the PM model assumes agreements between user and aid are treated the same as in the CF model, disagreements are solved randomly, but with a probability that matches the aid's overall reliability, rather than 50/50 by default. Although some studies find that a high number of participants use a PM strategy (e.g., 90% in Bliss, Gilson, and Deaton 1995), PM as a strategy is non-optimal because it is not a maximising strategy (Fantino and Esfandiari 2002, Wiegmann 2002), and it is therefore sometimes viewed as a decision-making error (Koehler and James 2009). Bartlett and McCarley (2017) argue that although suboptimal, it is still more efficient than CF, assuming the aid's overall reliability is higher than that of the user.

1.3.2.6 Optimal Weighting (OW)

The aforementioned strategies all apply to aids that provide signal strength weightings transformed into binary cues. An alternative approach whereby aids provide a more refined type of feedback, e.g., as graded or tiered cues, seems less well researched in the context of automation-aided human decision making. The OW model (or, in sensory tasks in an interpersonal context, Weighted Confidence Sharing (WCS) model, first introduced by Bahrami et al. (2010)) assumes that an aid sharing (a representation of) its evidence with the user offers the best-possible joint user-aid performance,

under the condition of both the user's and the aid's judgements being normally-distributed (Bartlett and McCarley 2017). The model presumes the user and the aid assess stimuli independently for the presence or absence of a signal. The aid presents its estimation of signal strength, i.e., the likelihood of a signal being present, to the user, who then averages their own likelihood estimation with that of the aid, weighting both their own and the aid's estimates by the agent's average sensitivity.

1.3.2.7 Uniform Weighting (UW)

The UW model (in a group decision making context, see: Sorkin, Hays, and West 2001) is near-identical to the OW model, but the user's and the aid's likelihood estimations are treated equally, i.e., they are unweighted when averaged by the user.

Bartlett and McCarley (2017) note that in a comparison between the first five models on one hand and the UW and OW models on the other, there is a suggestion that users might benefit from evidence assessments from aids that are shared directly, rather than transformed into binary cues. However, they also note that at the time of writing, there did not seem to be much evidence of this having been tested empirically. Wiczorek (2017) cites Wickens and Colcombe (2007) as the only experiment that tests how users might benefit from different threshold settings in likelihood alarm systems (LASs). Their own research focusses mainly on this interaction in the light of concurrent task load and operational safety in a task similar to that of Lee and Moray's orange juice pasteurisation plant experiments (1994), which is not directly relevant to our current research because our experiments deal with a single task in isolation. Five years later, we are still not aware of any further evidence of empirical exploration of the potential benefits of likelihood sharing systems.

Bartlett and McCarley's tests of the seven models

Bartlett and McCarley (2017) tested the seven models described above with three perceptual experiments, of which only the first two are discussed here, as the third one introduced feedback as a new variable, which is not relevant in the context of our current research. In the first experiment, participants had to choose which colour predominated in each of 300 blue and orange random dot images on a monitor. The cover story was that of geologists having to sort samples of a fictional mineral, and the probability of the dominant colour was 0.52, that of the contrasting colour 0.48. On some trials participants received help from a 93% accurate (or $d' = 3$) automated decision aid, the verdict of which was expressed as a numeric value on the screen before the start of each trial, so as to satisfy the order in the CC model. Performance feedback ("Correct!" or "Incorrect!") was given to participants after each trial. Each block of trials consisted of 50 unaided and then 50 aided practise trials, followed by a block of 100 each experimental trials, with stimuli presented randomly.

To test model performance, d' -scores from the automation aided conditions were compared with the models' predicted scores based on participants' unaided sensitivity. It was found that although the aid improved performance,

the OW, UW, CC, NN and BD model overestimated participants' automation-aided sensitivity, which was underestimated by the CF model. No great differences were observed with the PM model, which can either mean that participants adopted a probability matching strategy, or that they employed a different strategy that mimics the PM model's sensitivity. A bias towards the aid's judgements was observed in all models, but the magnitude differed, and did not match the models' predictions. One of the most striking observations, and one that is highly relevant for our research and certainly requires further investigation in the future, is the suggestion that participants seem to hardly have made use of the aid's graded evidence values, which is demonstrated by performance substantially below the predictions of the OW and UW models. In order to test this suggestion, the second experiment replicated the first one, but with binary instead of graded signal strength ratings from the aid. The hypothesis was that if participants did not benefit from the graded signal strength ratings, their performance should match that of the first experiment, which was indeed confirmed. After a high-level model comparison, Bartlett and McCarley (2017) conclude that although the PM model performs best when it comes to predicting participants' automation-aided sensitivity, it does not seem reasonable to assume they have employed a PM strategy, or any of the other strategies for that matter. It seems the references to *model* and *strategy* are used somewhat interchangeably to predict and to retrospectively describe user's actions here, which is potentially confusing, especially because of the observed discrepancies between strategies (a term we use to describe users' (plans of) action), and models (which we use to describe optimal strategies). We will not repeat Bartlett and McCarley's general post-test discussion (2017) here, but instead extend our support for their suggestion that further research is needed to determine whether the observed underuse of the aid's graded evidence judgements is caused by the instructions, the presentation format of the aid's information, or perhaps by an entirely different factor. We think understanding the exact causes is important not only for future research, but for future systems design as well, particularly because positive effects of likelihood information sharing systems have been reported in the past, e.g., by Sorkin, Kantowitz, and Kantowitz in relation to workload (1988).

Although Bartlett and McCarley (2017) test between the seven collaboration models by generating precise quantitative predictions of sensitivity, this approach is harder to pursue in our domain, where judgments are *conceptual* and *specific*. Also, pragmatically, items of varying and known difficulty for experimental cognitive study are much harder to generate than those for perceptual experiments. It can be noted that their seven models fall into two classes, and distinguishing between these two classes is itself a worthwhile empirical aim. The first three models (BD, YY/NN, and CF) pay no attention to the reliability of the automated aid, so already seem unlikely to properly respect what we know about the role of trust. The last four, quite similar, models are all sensitive to reliability. One aim of our research is to test whether, and to what extent, assisted conceptual judgments are influenced by aid reliability. In chapter 7 we test the CF, PM, OW and UW models with our S5 data.

1.3.3 Factors affecting interaction models

Bartlett and McCarley (2017) showed that none of the seven models they describe do a very good job of either predicting or describing user interaction with a sub-optimal aid, regardless of the aid providing binary or graded judgements. One problem that besets all these models we suspect, is that interaction between users and aids often depends on multiple interaction effects, among others between users' ability and experience, systems' performance and alarm threshold, task difficulty, type of task, and environmental factors such as workload (task difficulty + concurrent task load), aspects of which are noted by Gadala, Strigini, and Ayton (2021). Meyer and Kuchar (2021) also note that aids may provide benefits other than just signal detection, such as maintaining situational awareness, and that studying interaction with an aid in isolation has its intrinsic dangers, an observation most models cannot account for either. It is for example possible, they say, that an aid is a hindrance despite being very sensitive, or conversely, advantageous despite being not very sensitive at all. That being said, we do not discuss workload and other environmental factors here in any more detail as they are not tested in the current research; for a discussion of multiple task workload see Wickens (2020), for an overview of the literature on the combination of workload and automation reliability see Wickens and Dixon (2007), and for a discussion of four experimental studies on this relationship see Manzey, Gérard, and Wiczorek (2014). We discuss system reliability and thresholds, which we use as variables in our experiments, below.

As discussed, interaction between humans and automated aids is usually less straightforward and more contingent than many models predict, and as such those models describe ideal scenarios (*normative models*, Douer and Meyer 2019), and as we have seen in Bartlett and McCarley (2017), often fail their task as straightforward descriptive or predictive tools because in reality interaction with aids is non-optimal (Parasuraman and Riley 1997). Bearing in mind statistician George Box' maxim that '*all models are wrong but some are useful*' (first alluded to in Box 1976, and found in its current form in Box and Draper 1987), we believe that although there currently is no single overarching model that can describe the interactions in our research in a satisfactory way, there is potential to develop one that is useful. Therefore, one of our aims is to develop models to describe the findings of our experiments that consider degrees of belief of confidence and trust as factors as well.

1.3.3.1 Reliability of judgements and advice: reliability and alerting thresholds

Although there seems to be a broad agreement in the literature that automated aids are rarely used in an optimal way, there is less agreement on the exact causes of this mismatch. As discussed in section 1.2.3.1, contributing factors are among others the *cry wolf* effect, or an overrepresentation of false alarms that leads to diminished trust, and user perception of a system's reliability that does not match its actual performance, resulting in disuse.

A system's reliability (number of valid alarms minus number of false alarms) depends on its alarm threshold, or how alarms are defined. The alarm threshold is defined by the compromise between the highest possible sensitivity to targets (true positives), and the lowest possible number of false alarms (false positives), but a lower threshold will generally lead to more of both (Meyer and Sheridan 2017). However, this relationship is not a given, and just as the division of active and passive control and different levels of delegation of tasks between users and automation we discussed in section 1.2.1.2, it can be *static* and predetermined by the system's designers, or *dynamic* (Meyer and Bitan 2002) and automatically adjusted by the system based on users' performance (*adaptive automation*), or sometimes directly controlled by users (*adaptable automation*) (Parasuraman and Wickens 2008, Kidwell et al. 2012). The main purpose of optimising the alarm threshold, is calibrating users' trust in a system to moderate optimal use. This can be done by matching a system's sensitivity to the user's ability before they start a task, or by dynamically altering it during a task (Gadala, Strigini, and Ayton 2021).

Meyer and Sheridan (2017) state that the (system developers') premise that "users should always be in charge" is naive, because users simply do not always have all the information to identify a system's optimal settings. However, they say, many organisations misuse the idea of user involvement to (legally) offload responsibility for system failures onto users: in case something goes wrong, the user is (partially) responsible for the failure because they have incorrectly altered the settings. On the other hand, user control over system settings can empower users and give them a sense of autonomy. As a rule of thumb, Meyer and Sheridan (2017) say that users should be able to adjust settings if they have information that a system does not have that might affect decisions, when certain criteria are met, or to tailor a system to different users' needs. In cases where there is only one source of information, such as the level of CO in a carbon-monoxide alarm (CO is tasteless, odourless, and invisible, so humans cannot detect it before it is too late), users can generally not correctly set thresholds. If alarm information can be matched with other available information, such as a warning that an email may be malicious (the user can check the warning against among others the content and the sender), greater freedom to adjust settings may be justified. As Meyer and Sheridan's study (2017) deals with alarms as users' only information source, we will not discuss it here in more detail as in our experiments users can rely on their experience and skills too. Gadala, Strigini, and Ayton (2021) test a scenario in which customised thresholds in a spell-checking task are pre-set; we discuss their study in section 1.3.4.

1.3.3.2 Reliability thresholds in our experimental studies

In two of our studies (Stylus 3 and 4) where participants have to choose which of two sentences is better and where the aid gives tailored advice per trial, the concepts of reliability (performance) and alarm thresholds overlap, because the reliability of alarms is statistically matched with the likelihood of an alarm being correct as presented to participants. In other words, if the system tells a participant the likelihood of its alarm being correct is 75%, its performance will

be 75% correct in all instances where it does so. This means that reliability and alerting thresholds are not set separately. In our study where participants must choose if a sentence is correct and the aid's overall performance is briefed before the task (S5), the system's errors are equally balanced between false positives (a sentence is correct, and the system's judgement that the sentence is incorrect is bad) and false negatives (the sentence is correct, and the system's judgement that the sentence is incorrect is bad). In other words, if the number of false positives is high, the number of false negatives will be high too, and vice versa. Even though this means that the system's performance is balanced, as we have seen earlier, false negatives may impact trust and therefore performance differently than false positives.

1.3.4 Research in the domain of user interaction with language checkers

Text editing software like Microsoft Word or Apple pages, as well as internet browsers and email clients, have the option to provide users with instant, dynamic, feedback on their writing, or retrospective feedback when editing text. Although admittedly in text writing and editing the stakes may not be as high as, say, in supervisory control in a nuclear powerplant, its use is so widespread that it affects the behaviour of almost everyone who uses text editing software, an internet browser, or an email client. This feedback consists of two components: *error detection*, and *error correction* (compare: Parasuraman and Manzey 2010). The system can indicate e.g., perceived typos and homophone errors, punctuation errors such as double spaces and apostrophe errors, as well as, but to a lesser extent, writing style errors, such as passive voice and colloquialisms. Perceived errors are usually indicated by highlighting a word or phrase with a red wavy underline in simple systems, while more sophisticated ones highlight errors in a variety of ways for different types of errors, and/ or give detailed suggestions for potential alternatives, for example with a drop-down list with replacement options.

Spelling and grammar checkers can fundamentally differ in nature from some of the aforementioned supervisory control systems, in the sense that they can provide the user with direct feedback on work of their own making, rather than on an external process they are monitoring. If, and how exactly, this difference affects interaction with an automated aid is outside the scope of this thesis, and we are not aware of any literature that has attempted to describe or test potential effects.

In the case of our first two experiments (S1 and S2), participants are asked to compare sentences 'written by human writers' with suggestions for improvements 'from an automated editing tool called Stylus'. In the following two studies (S3 and S4), participants are asked to imagine they have 'just typed the "original sentence" into a word processor, like Microsoft Word or Apple Pages, and see suggested improvements 'from a writing style checker called Stylus'. Similarly, in S5, participants are asked to imagine they have 'just typed the sentence you see in each question into a word processor, like

Microsoft Word or Apple Pages. In some sentences you will see a word highlighted yellow, these are suggestions of errors from Stylus, an imaginary text editing aid based on artificial intelligence technology'.

Contemporary language checking software is so powerful that users can be provided with real-time feedback, in contrast to spelling and grammar checking being a separate post-hoc operation when computers were much less powerful when these systems were first introduced (Galletta, Durcikova, Everard and Jones 2005). Users are free to allocate the software to automatically amend all perceived errors (full-automation, error detection and correction), choose to accept or reject suggestions on a case-by-case basis (joint human-automation system, automated detection with manual acceptance of suggestion for correction), or switch suggestions off altogether (no system feedback). In more refined systems, the nature of the advice and the types of perceived errors that are highlighted can be customised by the user as well, figure 1.4 shows the options in MS Word 2021. However, no matter how sophisticated the system, the feedback is uncertain by nature because a system cannot "know" the writer's intention.

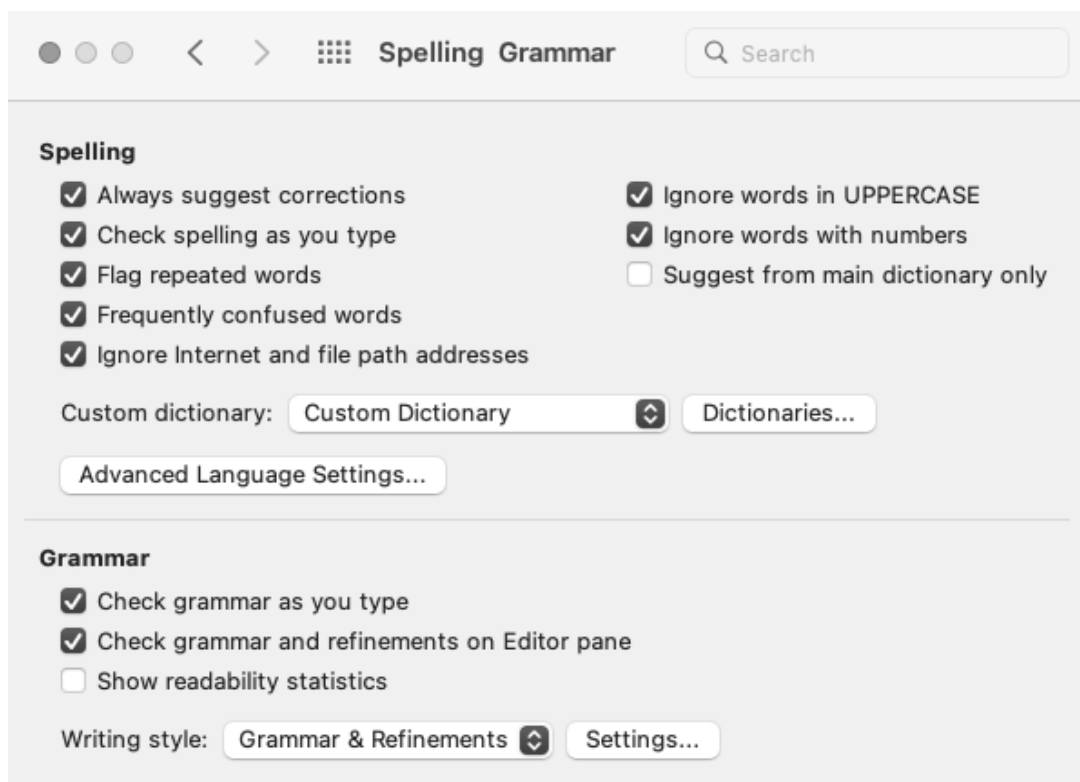


Figure 1.4 – Spelling and grammar checking customisation settings in MS Word 2021.

Although there is a vast body of literature on how spell-checking software packages work or might be improved in technical terms (see e.g., Birn 2000, Kukich 1992, and Major 2010), or on the effect of the use of spell-checkers on e.g., readers' expectations (e.g., Figueredo and Varnhagen 2005), second language learners' spelling proficiency (e.g., Lin 2017), users with learning

disabilities (e.g., MacArthur, Graham, Hayes and DeLaPaz 1996), or dyslexia (e.g., Pedler 2001 and Quattrini Li, Sbattella, and Tedesco 2013), we found a surprisingly limited literature on user interaction with language (spelling, grammar etc.) checkers in general first-language users. Many of these articles include unsubstantiated claims or assumptions about users' interaction with spell-checkers (e.g., 'Aside from the absolute faith many students put into checkers and the hostility other writers feel toward computer assistance', Major 2010). To our knowledge, very few attempts have been made to experimentally test writers' and editors' interactions with language checkers systematically.

One example of such experimental research is Galletta et al. (2005), which uses a form of signal detection theory without explicitly mentioning the term. 65 participants, classed as either "high" or "low verbals" based on test scores from their school records were asked to edit a business letter, using Word. For one half of the participants the language checking software was switched on, the other half received no help. In the text there were five instances each of genuine errors, false errors, and errors that were missed. Performance was measured by counting the remaining number of errors of each type that were left after participants had completed the task. Potential new errors introduced by participants were apparently not measured, and it is unclear how it is possible that false errors were presented to participants that worked without the checker, as they would not see any system alerts at all. The main findings were an effect of experience level and of presence of the automated checker, and a general overreliance on the checker in high verbals. This latter finding is interesting as it contradicts other findings of users with low experience falling for unreliable software, and it shows that *misuse* (Parasuraman and Riley 1997) is not simply an effect of lack of experience.

Galletta et al. (2005) conclude that interaction with spell checkers is often sub-optimal due to overreliance, and that warning labels may be a solution to this phenomenon. Although they almost immediately neutralise this proposal with the caveat (without evidence or argumentation) that the likelihood of effectivity is probably low, the main premise of their article seems to be that of providing a warning to users. From a usability perspective the hope that instructions and manuals in general, and warning labels in particular, might influence behaviour, is often frowned upon. In the usability literature the consensus seems to be that the need for such measures should be omitted with better design solutions (see e.g., Norman 2002, 2011). However, in almost two decades since Galletta et al. (2005) proposed the use of warning labels for text editing aids, no better alternative seems to have presented itself. Our research aims to contribute to bridging this gap by integrating tools to improve users' awareness of the level of imperfection of the aid, without the need to affix warning labels.

Based on Galletta et al.'s findings (2005), Gadala, Strigini, and Ayton (2021) test adjustable alerting thresholds in a spell-checking task. It should be noted here that this study is a preprint, not yet published, which appeared online after

the experimental studies in this thesis were complete. Analogous to our considerations when selecting a domain, they mention that some of the benefits of this application are the ability to run explorative studies at low cost, yet with a high level of experimental freedom. In addition to that, they argue, observations made in one type of task or domain are to an extent applicable to entirely different domains. This is because the tasks, at a higher level, are all a form of pattern detection and response. Gadala, Strigini, and Ayton also mention the empirical argument of findings from experimental research in very different domains corroborating each other (they compare Galletta et al.'s research on interaction with text editors (2005) and Alberdi, Strigini, and Ayton (2010), who tested behavioural patterns in mammography screening).

In a within-subjects design, 47 participants interacted with a mock spell-checking aid that highlighted misspelled words in a binary fashion. No suggestions for corrections were made by the system. Participants were briefed that the system could detect non-words and misused words, homophones in particular, but that it was not 100% accurate, hence false negatives and false positives were possible. Participants completed an editing task and a dictation task, and a series of questionnaires. In the text editing task, participants edited two approximately 650-word passages, an easy and a difficult one, under time constraints (to avoid floor and ceiling effects), purportedly with two different aids, each with a different alerting threshold. The threshold, indicating the percentage of correctly highlighted errors, of the first, "sensitive", aid was 100%, that of the second, "specific" one 30%, 67% or 100% depending on error difficulty level. A dictation task, where participants heard a word and had to spell it, was used to assess participants' linguistic sensitivity as defined by Fischer, Shankweiler, and Liberman (1985), and questionnaires were used to measure participants' perception of the aid's behaviour and reliability, trust in both systems, and observed differences and preference for either of the systems. For the editing as well as for the dictation task, the number of errors in participants' output was counted, and two decision types were split out in the editing task: correct detection and correct result. A correct detection means a participant rightly identified an error, correct result meant they were able to meaningfully correct it. The results of the dictation task were crossed with the total count of errors in the editing task, which was interpreted as a different proxy for ability. This cross-check was used to identify errors as a result of lack of knowledge (participant does not know the word or its correct spelling, *knowledge deficit*), or of failure to detect an error (*processing deficit*) (Figueredo and Varnhagen 2004).

Gadala, Strigini, and Ayton (2021, preprint) found limited evidence for their hypothesis that the effect of different alerting thresholds depends on user ability, and more substantial proof for their second hypothesis that this effect is moderated by the interaction between the user's ability and the difficulty of a task, i.e., different thresholds may work better for low vs. high ability and easy vs. difficult tasks. Like Galletta et al. (2005), Gadala, Strigini, and Ayton (2021) also found that although high verbals are better at correcting misspelt words, they are no better than low verbals when it comes to dealing with false

positives (FAs) than low verbals. They conclude that alerting thresholds should be calibrated for optimal performance and reduction of automation bias by empirically measuring specific aids with realistic samples of their users and decision tasks. They stress the importance of interactions, i.e., different users may benefit from different alerting thresholds, and an individual user's performance may benefit from a different alerting threshold depending on the difficulty of a task.

1.4 How word-processors deal with specific writing error types relevant to our research

Since our research is about user interaction with language checkers, it is outside the remit of this literature review to discuss all types of errors these systems can and cannot successfully detect, and how they do so. Two aspects are relevant though, of which the first is the fact that most checkers seem to prioritise minimising the number of false alarms over absolute reliability (Birn 2000, Kukich 1992), and second is the types of errors that are difficult to detect for the combination of the user and the automation. An example of the latter is the potential overlap between spelling and grammar errors, as well as homophone errors as a particular class of errors. Our Stylus 5 study is entirely built on this last category.

1.4.1 Overlap between spelling and punctuation errors.

Major (2010) points out that there is often overlap between spelling and grammar errors, and the way a system deals with them. Apostrophe errors are a poignant example: although words like *wasnt* (instead of *wasn't*), *dont* (instead of *don't*), are technically grammatical errors, they are strings of characters that do not match the system's dictionary and are as such treated as misspelt words, so they will almost always be flagged up by writing aids. Conversely, apostrophe errors like *writer's* vs. *writers'*, where both versions are acceptable spellings although they are grammatical errors if used incorrectly, will often go unnoticed by the automation. In our research we use a mix of spelling and grammar errors, but the technical difference is not relevant, hence we use the collective phrase "writing errors".

1.4.2 Homophone errors

Homophones are words that sound similar, but have a distinctly different meaning, e.g., *flower* vs. *flour* (Parent 2012). Homophone errors, a class of *real-word errors* (Kukich 1992, Rello, Ballesteros, and Bigham 2015), are mistakes where homophonic pairs are confused. Because they are context-dependent, and both variants are in its lexicon as acceptable sequences of letters (Figueredo and Varnhagen 2005), homophones are notoriously difficult to detect for automation (Major 2010), but for humans as well (Riano and Margolin 2018). Homophones can be genuine errors (the writer is convinced

they have spelt the word correctly), as well as typos (e.g., *scnt vs sent*, or *too vs. to*). In some cases, homophones are variants that may not necessarily be technically wrong yet may be perceived as odd in certain fixed phrases (e.g., '*no tool's left in van overnight*' instead of '*no tools left in van overnight*'). Due to their stealth nature, homophones are often not flagged up by the system (omission error), which may contribute to users developing a misplaced high level of trust in the system (see section 1.2.3.4, and e.g., Parasuraman and Riley 1997, Parasuraman and Manzey 2010, Rice and McCarley 2011 and Swets 1992 for this phenomenon in general). Due to their ambiguous nature, homophones are extremely useful in text-editing-based experiments, as demonstrated by among others Figueredo and Varnhagen (2004), Figueredo and Varnhagen (2005) and Gadala, Strigini, and Ayton (2021).

1.5 Summary

As is apparent from our review, most research that directly addresses the performance of joint human-automation performance in simple decision tasks uses perceptual signal detection as a paradigm task. The domain of spelling and grammar checking is clearly quite distinct, in that it relies much more on semantic knowledge and conceptual processes. Yet it surely is an important domain practically, if only because of how widespread the use of automation in writing tasks is. The lack of study of knowledge-dependant tasks in general, and spelling and grammar checking in particular, means, we argue, that many of the findings reviewed above must be tested anew in this domain. Consequently, the studies reported in this thesis seek to test several hypotheses, derived from the above literature and from plausible psychological accounts in parallel. Thus, we will now very briefly summarise the findings reviewed in this chapter that are directly addressed in the following experimental work.

The first, and fundamental Human Factors finding, is that even imperfect decision aids can sometimes improve unaided human performance, and that more reliable aids will be more effective (Wickens and Dixon 2007). This finding is important, and must be established in any new domain.

Effective use of aids is supported, perhaps, by two established factors that influence their use: 1) aids will be less used when users have more perceived self-efficacy or are more confident about individual decisions, and 2) aids will be more used when users trust them more, on the basis of their experience with similar aids, or with the aid itself, or on the basis of some other judgments about the aid (Lee and Moray 1994, Moray et al. 1994, Wiczorek and Meyer 2019).

Finally, the usefulness of imperfect aids is challenged by several biases or dispositions which might undermine their optimal use (Parasuraman and Riley 1997). In particular, in several respects, users are likely to be overconfident (Moore and Healy 2008). Despite its obvious plausible role in the use of automation, and the widespread acceptance of the self-efficacy and trust effects, few empirical studies of aided decision making include confidence measures. Furthermore, we do not know the extent to which varieties of overconfidence might exist in the domain of spelling and grammar. Therefore, our experiments will directly address this question.

Chapter 2 – *Research approach and methodology*

2.1 Introduction

This chapter details our research approach and experimental methodology, and our data analysis strategy. Since our five interrelated studies vary in set-up on crucial points, details of the methods of each individual study, and how they differ from each other, are discussed in the relevant chapters.

2.1.1 Language checkers as an example of interaction with complex automated systems

One mundane example of computer users interacting with a complex automated system, is the use of spelling, grammar and writing style checkers in word processors, email clients and internet browsers, as discussed in section 1.4 of Chapter 1. Such systems typically rely on statistical techniques to parse a user's draft sentences, and to assist users in their writing or editing by highlighting potential errors and offering possible rewrites for these errors.

2.1.2 General research aim: establishing an experimental paradigm

The general aim of our research is to establish an experimental paradigm in which humans and complex automated systems interact, with human operators able to accept or over-ride automation, so that human judgments will depend on how they combine their trust of the automated system with their own perceived self-efficacy.

We also explore a range of effects related to participants' calibration and confidence. We suppose that a well-working automated support system will allow better judgments from users, but also that it may allow better calibrated confidence in judgments: i.e., more valid judgments about the likelihood of uncertain answers being correct. We seek to develop a paradigm in which an automated system that knows its own likelihood of being correct, can meaningfully and usefully communicate this knowledge to the user.

2.1.3 Quantitative vs. qualitative approaches

In the introductory chapter we mentioned our initial desire to study the effects of variations in user interface design on human interaction with automated aids. We also explained that we decided to focus on the theory and fundamental relationships between users making uncertain judgements and advice from aids, because many aspects of these foundations are still ill

understood. Our research focus has consequences for the methodology we opted to employ. Payne and Howes (2013) describe how different types of HCI research can benefit from qualitative and quantitative methods. Research of interface design "in the wild", i.e., in real-world environments, may benefit from qualitative sociological and ethnographic methods, such as observations and interviews, because it relies on strong social (e.g., in teamwork situations) and environmental (e.g., physical work environments) factors, by definition. Notable examples of this type of research can be found in e.g., Suchman 1987, Greenbaum and Kyng 1991, Hutchins 1995, Heath and Luff 2000, and Dourish 2001. We argue that HCI research with a more fundamental orientation, with an intent to test theory-based empirical hypotheses, such as ours, is better served with quantitative methods. Gadala, Strigini, and Ayton (2021) show that there can be analogies in interactions with aids in different types of tasks or domains, i.e., in different contexts. By decontextualising, or at least controlling or demarcating the context to an extent, quantitative methods allow us to observe phenomena of interaction with automated aids that we suppose to be general across contexts and without them being obscured by uncontrolled contextual factors.

An important prerequisite for quantitative research is that factors can be validly and reliably specified. In sections 2.2 and 2.3 we explain for each factor how it was defined and measured.

2.1.4 General research design

Our research consists of a series of five experimental studies, each with between 62 and 140 pre-screened participants (details discussed in relevant chapters). The procedure of each study can be broken down into three clearly distinguishable parts: *Pre-task*, which establishes who our participants are and identifies some of the relevant beliefs they hold, *Task*, in which participants execute an experimental task, and *Post-task*, in which we explore how our participants judge their own and the system's performance.

2.2 Research approach and survey design

Our experimental studies were conducted fully online, with the surveys being created on, and hosted by the survey platform Qualtrics.com, and participants recruited through the recruitment website Prolific.ac. Participants completed the surveys on their own devices in their own time, details of how we selected and pre-screened participants, and how we accepted or rejected their data, are also discussed in each of the experimental chapters of this thesis. Participants' education level was generally high (vocational training/ apprenticeship/ higher education degree), and they all had experience using word processing software such as Microsoft Word or Apple Pages.

2.2.1 Screening participants

For our experiments, we needed a pool of native British English speakers to make sure all participants potentially had the same understanding of the items. Because at the time we used Prolific it was not possible to screen language proficiency any further than “English”, and filters changed a number of times when we used the platform, we applied several different participant screening strategies, details of which are presented in each of the chapters that discuss our experiments.

2.2.2 Sample sizes

The first experiment, Stylus 1, was treated as a pilot study, and is certainly underpowered, but it is reported in full though it generated some useful findings for the development of the research. Thereafter, all experiments had sample sizes determined with the software G*Power to give a power of at least 0.8 for the primary Human Factors hypothesis that a more accurate decision aid would be beneficial to participants’ overall judgment performance if this were tested using a *t*-test for a medium size effect.

2.2.3 Pre-task

Our pre-task questions aimed to establish some of the relevant background characteristics of participants of our studies in terms of:

1. identifying their sociodemographic profile (age, gender and educational level)
2. gauging their experience with word processing software to make sure their task responses are credible
3. measuring their self-reported levels of trust in spelling and grammar suggestions in word processing software packages or internet browsers in general, to test our hypotheses for correlations with their performance
4. assessing their perceived level of self-efficacy in spelling and grammar, to test our hypotheses for how this interacts with trust of automation

The pre-task part of the survey has minor variations between studies because part of the information was obtained through Prolific, rather than from the participants in the survey, in later studies. The differences are detailed in the relevant chapters.

2.2.4 Task, type of research and objectives

The core of the research is done through a series of five experimental studies, each consisting of between 30 and 100 experimental trials, and with either a within-subjects or a between-subjects design, varying depending on the requirements of the individual studies.

During the task, participants assess sentences in the light of an imaginary automated checker ("Stylus"), by comparing an "Original" sentence to one that has been amended by Stylus (S1–4), and determining which is the better, or by judging if sentences, in a percentage of which Stylus flagged up an issue,

are correct (S5). Participants are asked to indicate how confident they are in each of their responses, and overall confidence is measured post-task.

The objective of the experimental task is to obtain measures of performance, strategy and confidence for each participant. Because of the information given on each trial, performance and confidence will be determined not only by the participants' own ability to judge the correctness of English sentences, but also by the trust they have in the Stylus judgement, or recommendation, for each item. For example, in some conditions of the experiments participants could perform quite well without any knowledge of English, by simply accepting Stylus's recommendations.

In order to analyse the separate contributions of participants' English competence and their trust in the automated recommendations, the analytic approach of Signal Detection Theory (SDT) is adapted.

2.2.5 Two Alternative Forced Choice and Yes/No models

SDT is not a singular paradigm, but rather a collection of discrimination methods that use similar analyses (Green and Swets 1966, Stanislaw and Todorov 1999, Macmillan and Creelman 2005, Hautus, Van Hout, and Lee 2009). For our research we used adaptations of the Two Alternative Forced Choice (2AFC) model, which represents comparison and classification tasks, (S1–4) and of the Yes/No (Y/N) model, which can be described as an identification task (Y/N, S5). Table 2.1 shows an overview of the design of our five studies. Confusingly, response categories in a Y/N design (*one-interval design*) can be something other than Y/N, e.g., "true" and "false", "fast" and "slow", "high" and "low" etc., and there can even be more than two response categories, e.g., a 5-point scale. Response categories in 2AFC designs can in some cases be named "Yes" and "No". What sets the type of Y/N design as we use it apart from the 2AFC one (which is 2AFC by design, but perhaps not as a concept as we explain in the following paragraph), is that in the Y/N design there is one single stimulus (in our case one written sentence), whereas there are two stimuli in a 2AFC design (or more in 3>AFC designs; in our case two sentences). Although Y/N designs with more than one stimulus exist, we will not discuss them here because they are outside the scope of our research.

In a 2AFC design, participants are presented with two alternative items in each trial, and it is their task to indicate which one is *signal*, and their choice renders the alternative option *noise*. The terms *signal* and *noise* are rather arbitrary in the context of our design and have been used pragmatically yet consistently in our studies; Table 2.2 shows how we use them.

The reasons for using a 2AFC-like method ('Which one is better, the original sentence or the alternative Stylus suggests') for the first four studies, as opposed to a Y/N design ('Is this sentence correct – Yes/ No', as used in S5), are that 2AFC discourages bias (the only possibility is an arbitrary preference of order of response options, e.g., always selecting the first option for any reason other than merit), and that performance levels (percentage correct) are

high, which allows for measuring sensitivity to smaller stimulus differences than in a Y/N design (Macmillan and Creelman 2005). In particular, in our 2AFC experiments, we eliminate a bias towards assuming that presented sentences are correct, while at the same time we deliberately allow, and analyse, a bias toward accepting automated advice.

A typical perceptual 2AFC task has hundreds up to thousands of trials, whereas the cognitive tasks in S1–4 only have thirty. There are several pragmatic reasons for the comparatively low number of trials in our studies. The first is that it is difficult to create enough items with known and controlled levels of difficulty. Controlling the level of difficulty is vital, because floor and ceiling levels of performance (i.e., near-chance or near-perfect performance) make it impossible to calculate SDT measures. It is important to note in this context that in a "classic" 2AFC SDT paradigm, there is no way to trade off Ms against FAs. However, if a participant in our studies always chooses the Stylus alternative, they will never make a M. Hence, we call our model "2AFC-like" because it confirms with the model, yet less so with the concept. A second reason for the low number of trials, is that responding takes longer in cognitive tasks compared to perceptual ones, which is connected to the third reason, namely that it is difficult to recruit participants on online platforms for studies that last longer than approximately 20-30 minutes.

S5 uses a Y/N design, and 100 trials to offset the potential of bias towards Yes or No being introduced by the set-up. Y/N questions are perhaps easier to understand than classification tasks, and thus response times will likely be shorter than when choosing between alternatives, but they come at the cost of increased bias, which must be compensated by a greater number of trials (Macmillan and Creelman 2005). S5 uses only homophones (word pairs that sound similar, but have a distinctly different meaning), which made it easier to create a large number of difficult enough trials, than creating a set with mixed error-types of known and controlled difficulty. Homophones are also easier to believably swap around as being used correctly vs. incorrectly in sentences, because both alternatives are valid words.

The layout of our studies is as shown in Table 2.1:

<i>Study</i>	<i>Design</i>	<i>Participant task</i>	<i>Number of trials</i>	<i>Number of participants</i>
Stylus 1	Between-subjects	2AFC ('Which sentence is better?')	30	62
Stylus 2	Between-subjects	2AFC	32	120
Stylus 3	Within-subjects	2AFC	32 (+1 dummy)	128
Stylus 4	Within-subjects	2AFC	32 (+1 dummy)	140
Stylus 5	Between-subjects	Y/N ('Is this sentence correct?')	100	114

Table 2.1 – Stylus study designs

It must be noted that S2 is an improved version of S1, and that S4 is a variation of S3. S1 and S2 are between-subjects studies because they compare two groups of participants working with aids with different levels of reliability, S3 and S4 are within-subjects studies because participants receive advice from an aid that itself has two different levels of reliability, and S5 lastly is a between-subjects design again because it compares two groups of participants working with aids with different levels of reliability and a control group.

2.2.6 Post-task

After the experimental task, we asked our participants to reflect on their own performance, and that of the system, by asking them to:

1. estimate how many times they gave the correct answer, and how many times they followed Stylus' advice.
2. estimate how well they did in relation to others
3. judge Stylus' performance.
4. judge the plausibility of the Stylus suggestions being created by an automated system.
5. (S5 only) indicate to what degree they remembered and considered Stylus' level of reliability during the task.

2.2.7 Sliders and scales

In our studies we use two types of scales with sliders. Participants were forced to answer all questions and manipulate the sliders before being able to proceed to the next question.

Pre-task and post-task where we asked participants to rate their own, others' or systems' attributes, we used sliders with 0% – 100% scales. During the experimental task participants were asked to rate their confidence on a 50% ('I guessed') – 100% ('I'm certain') scale; this is a method commonly used in cognitive psychology research (Gigerenzer, Hoffrage, and Kleinbölting 1991). Values below 50% would not make sense, as participants would select the alternative option if their confidence in their response would effectively be negative. Each slider had a pointer set in the middle as a default (resp. at 50% and at 75%), which had to be manipulated before participants could proceed to the next question. If the default was the desired position (50% or 75%), participants could move the pointer and put it back to proceed.

2.3 Analysis and reporting of results

We used Microsoft Excel for initial survey data processing and for the creation of figures. Early explorations of results were done in IBM SPSS, all final analyses have been performed in the open-source statistical software JASP.

2.3.1 Pre-task

For each study we identified the sociodemographic profile of the participants, subdivided into each group in between-subjects studies, by calculating the average and the standard deviation of their age, establishing the gender balance (number female and number male), and the average educational level. Since the latter is reported as categorical data, we describe the results in broad terms, e.g., 'generally relatively high (vocational training/ apprenticeship/ higher education degree)'. Participants' experience with word processing software was treated as a means to potentially reject results. Should a participant have no prior experience with word processing software, on which Stylus' behaviour was modelled, their results would not be credible, and their data would be ignored. We have not identified such a case in any of our five studies though.

Because participants' self-reported levels of trust in spelling and grammar suggestions in checkers in general were measured with aggregated questions, these were tested for reliability (Cronbach α), and then the averages and standard deviations were calculated. The same was done for participants' perceived level of self-efficacy in spelling and grammar, and that of their estimate of average native British English speakers' efficacy.

We also calculated the average and the standard deviation of the time taken to complete the survey from completion time data provided by Qualtrics. If participants completed the task unrealistically quickly, their data was rejected. Where applicable, this is reported for each study in the relevant chapter.

2.3.2 Analysing performance and confidence by adapting SDT calculations

As discussed in Chapter 1, SDT is an approach to the analysis of human performance on perceptual signal detection tasks, where, on any given trial, the human judge is asked to indicate whether a signal is present. In addition to its use on perceptual judgment tasks, this theory has been widely applied to recognition judgments, where, for example, each of a list of words, or faces, must be judged as "Old" (seen during the study phase of the experiment) or "New" (not seen during the experiment). More generally, a stimulus is observed, and the observer's task is to report whether it is *signal* or mere *noise*. Note that the use of the terms *signal* and *noise* is metaphorical in these types of experiments (Abdi 2007).

Considering the recognition example, it is clear why the proportion of "Old"-items that are successfully recognised is not a good measure of performance – a participant could respond "Old" in *every* trial and achieve perfect recognition. In general, a participant's response is taken to be determined by two independent criteria, the evidence they somehow gather from the stimulus, and the threshold they set on this evidence for responding "Old" or "Signal". A participant who sets a very low threshold will achieve high recognition at the expense of making a lot of False Alarms (responding "Old"/"Signal" when the item is in fact "New"/"Noise"). A participant with a higher threshold but the same ability to discriminate will make fewer False Alarms, but will "Miss" more "Old" items, erroneously responding "New" to them. By setting a discretionary threshold, therefore, the human judge can trade-off two types of error, False Alarms vs. Misses. Indeed, there is empirical research that observers can learn to do this optimally (e.g., Kubovy, Rapoport, and Tversky 1971), depending on the relative costs of the two error types, although in many experimental studies all errors are notionally equivalently costly.

By classifying performance on every trial according to a 2x2 matrix, SDT allows separate calculations of a participant's ability to assess stimuli (called *sensitivity*) and their discretionary threshold (called *bias*). In our studies, the advantage of using the SDT analyses is that they give us a measure of sensitivity that is independent of bias, i.e., a measure of ability to determine the correct sentence independent of the willingness to accept Stylus advice. Also, and more specifically, they provide a measure of bias independent of sensitivity, i.e., a measure of willingness to accept Stylus advice independent of whether it is correct or not. Both these measures are independent of the proportion of Stylus-correct trials. This is important, because different H_{rate} and FA_{rate} pairs can result in the same sensitivity d' (*iso-sensitivity*; Macmillan and Creelman 2005), and it is therefore more useful than the easily understood, yet in a way flawed, simple metric *percentage correct* (which we still report for reference and when reliable SDT sensitivity measures cannot be computed).

Thus, SDT analysis uses a 2x2 grid, in which results are classified as Hit, Miss, Correct Rejection, or False Alarm, usually categorised as follows:

	<i>Stimulus = Signal</i>	<i>Stimulus = Noise</i>
<i>Participant responds Signal</i>	Hit (H)	False Alarm (FA)
<i>Participant responds Noise</i>	Miss (M)	Correct rejection (CR)

Table 2.2 – Signal Detection matrix

The classifications in Table 2.2 have the following meanings:

H	True positive	Signal	Correctly identified error
M	False negative	Noise	Missed error
CR	True negative	Signal	Correctly identified lack of error
FA	False positive	Noise	Incorrectly identified error

Before considering the precise mathematical approach of SDT, this grid allows the fundamental logic to be discerned. ‘Hit Rate’ is defined as $H / (H + M)$, ‘False Alarm Rate’ is defined as $FA / (FA + CR)$. By considering Hit Rate - False Alarm Rate one gets a measure of sensitivity independent of bias. By considering Hit Rate + False Alarm Rate one gets a measure of bias independent of sensitivity.

A different approach is adopted in some studies of signal detection – instead of being asked whether a single stimulus is signal or noise (Y/N), participants are given a Two Alternative Forced Choice (2AFC), with one item being signal and one noise in each trial. Such a design effectively nullifies the role of any strategically set threshold and allows measurement of sensitivity in terms of percentage of trials correct. However, our research questions require the keeping-separate of sensitivity and advice taking, which means this approach is not necessarily appropriate. Although the design we use is a 2AFC, it is *not* 2AFC with respect to what we wish to analyse as signal vs. noise, and it allows separate computation of sensitivity and bias, where bias is interpreted to mean willingness to be influenced by the automated advice.

2.3.4 Adapted SDT matrix

In S1–4 the first item in each trial was labelled “Original sentence” and the second item “Stylus suggestion”, and participants had to indicate which one was better (because an error or infelicity was present in the other); our adapted matrix is shown in Table 2.3.

Performance on such a trial, in analogy to SDT trials, combines the participant’s sensitivity to correctness of English sentences with their willingness to accept the advice of an automated system. If Stylus’ suggestions are correct for 80% of the trials, then a participant could achieve 80% correct by simply accepting the Stylus suggestion, even if they had no knowledge of English grammar and spelling.

It should be clear, however, that the logic of SDT analysis can be applied to this situation, to separate sensitivity (ability to identify correct sentences) from bias (willingness to accept automated advice).

Our SD grid looks as follows:

	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	Hit (H)	False Alarm (FA)
<i>Original sentence selected by participant</i>	Miss (M)	Correct rejection (CR)

Table 2.3 – Stylus Signal Detection matrix

The classifications in Table 2.3 have the following meanings:

H	True positive	Correctly identified lack of error
M	False negative	Missed error
CR	True negative	Correctly identified error
FA	False positive	Incorrectly identified error

Note that possible categories in our 2AFC-model depend on the formatting of the trial (Original sentence is correct and Stylus suggestion is incorrect, or vice versa), and that possible categories always come in the same pairs (H and M or CR and FA) depending on formatting, as is the case with SDT trials which are either signal or noise (signal trials will result in H or M, noise trials in FA or CR).

And finally, to complete the analogy with SDT, from above matrix we can see that a participant's sensitivity can be computed by subtracting FA-rate from H-rate, and their willingness to accept advice can be computed by adding FA-rate to H-rate. Sensitivity in this analysis is performance independently of being aided, and consequently has a minor role in our studies, but it allows us to query the validity of self-efficacy judgments. Bias, however, as propensity to accept advice, is an essential measure for our studies, and arguably a major novel contribution of the thesis.

2.3.5 Parametric and non-parametric analysis of sensitivity and bias

Having introduced the logical structure of our approach, by analogy with SDT, and the rationale for the approach, we now turn to the actual mathematics of SDT computations of sensitivity and bias.

Different types of SDT-design require different analysis strategies, and analytical approaches are a contentious topic in the SDT literature (Macmillan and Creelman 2005, Pastore, Crawley, Berens and Skelly 2003, Zhang and Mueller 2005). Traditionally the most commonly used method to analyse participants' sensitivity (ability to distinguish between stimuli) and bias

(tendency to favour one option over the alternative on any basis other than merit), is that of parametric analysis (Macmillan and Creelman 2005), with a non-parametric approach catching up in popularity more recently (Zhang and Mueller 2005). Parametric analysis assumes results have normal distributions with equal variance, whereas a non-parametric approach can seemingly do without these assumptions.

Although the widely used parametric measures of sensitivity (d') and bias (c) seem to work well in case the number of trials in an experiment is high, they do not give sensible answers when H or FA is 0 or 1, which is often the case when the number of trials is low like in most of our studies and any SDT measure might itself be noisy or unreliable as a result. Unfortunately, none of the workarounds mentioned in the literature, one of which we used and discuss below, is perfect. Another disadvantage of d' and c , is that the resulting values are anything but intuitive to interpret. The so-called "non-parametric" measures of sensitivity (A') and bias (B'') are not really non-parametric in the sense that assumptions about the underlying distributions are still made (Pastore et al. 2003) and, worse, they do not produce completely independent measures of sensitivity and bias. However, they have the advantage of being much more intuitive, and having meaningful values when H or FA are 0 or 1. In other words, using parametric or non-parametric measures is a trade-off in which both options carry different risks, of which we are aware, and that we bring up in the discussion of our findings where relevant.

Because the designs of our first four studies are a hybrid in that their appearance is that of 2AFC, but the number of trials is low and the results are closer to those of a Y/N design, it was not clear to us beforehand when any of the above measurement problems might affect our hypothesis testing, so we needed to consider the extent to which any of the effects we wished to test might be affected by inaccuracies or unwanted correlations in our derived measures. Because of the novelty of our paradigm and the uncertainties about the SDT measures, we argued the most reasonable approach would be to test parametric and non-parametric approaches alongside each other and compare the outcomes. In the following paragraphs we discuss the traditional parametric approach and a "non-parametric" alternative.

2.3.6 Measuring sensitivity corrected for bias with parametric measures

Because SDT assumes results have normal distributions with equal variance, we started by testing our data for these conditions. To calculate parametric sensitivity and bias, we need to calculate p -values ($p_H = H / (H + M)$; $p_{FA} = FA / (FA + CR)$) and transform them into z -scores using the NORMSINV function in MS Excel ($z_H = \text{NORMSINV}(p_H)$; $z_{FA} = \text{NORMSINV}(p_{FA})$). p -values (probability values) indicate proportion correct, and z -scores (standard scores) indicate the number of standard deviations the signal distribution is above the noise distribution. NB: NORMSINV is superseded by NORM.S.INV, which is said to be more accurate, in later versions of Excel. We used NORMSINV to allow rearward compatibility, after we checked that both yield the same results for our data.

2.3.7 Correcting 0 Miss and False Alarm rates

As some of our experiments have fairly high performance levels and low numbers of trials in certain categories, some participants had no Ms and/or FAs. M and FA numbers of 0 lead to respectively $p_H = 1$ and $p_{FA} = 1$, which in turn result in a #NUM!-error when calculating z_{FA} and z_H in Excel; this is because for successful application of NORMSINV, p -values values must be between 0 and 1, excluding both. To solve this issue, the following p -formulae for H and FA rates that incorporate a correction, by subtracting 0.5 from the H or FA-rates if they were 0, were used in Excel: H-rate, =IF(M>0,H/(H+M),(H-0.5)/(H+M)); FA-rate, =IF(FA=0,FA/(FA+CR),(FA-0.5)/(FA+CR)).

2.3.8 Parametric sensitivity, d'

$$d' = z_H - z_{FA}$$

Although percentage correct gives a reasonable indication of performance in 2AFC designs as bias is normally of limited concern, a correction for bias is still required. In our 2AFC-like studies, where due to their nature bias is a potential issue as explained in the following paragraph, the dimensionless statistic d' is the measurement of participants' sensitivity. d' is a correction of percentage correct for potential bias in participants to favour the Stylus suggestion over the Original sentence, and it is calculated by subtracting the z-score for FA from the H z-score.

As long as the number of Hits is equal to or greater than the number of False Alarms, d' will be equal to or greater than 0. When $N_H = N_{FA}$, $d' = 0$, i.e., pure chance. This indicates participants cannot discriminate between good and bad Stylus suggestions and their responses are likely the effect of guessing (Bartlett and McCarley (2017) exclude participants with d' -scores below 0.5, because they assume not meeting that lower limit suggests participants have failed to understand or follow the instructions). The higher a participant's absolute d' -score is, the more sensitive they are to discrimination between the correct answer and the incorrect one; theoretically the highest possible d' -score is infinite, but when above corrections are applied to zero M and FA rates, 6.93 is the upper limit. However, when $H = 0.99$ and $FA = 0.1$, the effective upper bound is 4.65. Typical d' -scores are values up to around 2 (Macmillan and Creelman 2005).

2.3.9 Parametric bias, c

$$c = z_H + z_{FA}$$

In a classic SDT 2AFC study, bias would not exist or be very low; the only bias that could possibly exist is that of participants (adopting a strategy of) favouring either the first or the second option. Our first four studies, which have a 2AFC-like design, are different in the sense that both options are labelled ("Original sentence" and "Stylus suggestion"), so we need to measure bias, and correct performance figures per condition because the labels potentially introduce the possibility of bias in participants towards either the Original sentence ("human") or the Stylus suggestion ("automation").

Criterion, c , is an index of bias in participants towards either the original sentence or the Stylus suggestion; it is calculated by adding the z-score for H to z_{FA} . If a participant's c -score is above 0, they favour the Stylus suggestion over the Original sentence, and if it is negative they display bias towards the Original sentence. A c -score of 0 means that a participant displays no bias at all.

As long as the number of False Alarms is equal to the number of Misses, c will be 0. When the number of False Alarms is greater than the number of Misses, the c -score will be positive and indicate bias towards Stylus, and when N_{FA} is lower than N_M , c will be negative, which is indicative of bias towards the Original sentence. Note that the range for c is equal to that of d' , but that the centre is 0.

2.3.10 Two bias measures in our S5 Y/N design

In Stylus 5, where participants must decide whether a sentence is correct by selecting "Yes" or "No", while they receive help from Stylus, there are two types of bias, that require separate analysis. The first type, $c_{Y/N}$, is bias towards responding "Yes" (positive $c_{Y/N}$) or "No" (negative $c_{Y/N}$), the second type, c_{Stylus} is bias towards Stylus' advice (positive c_{Stylus}) or away from it (negative c_{Stylus}). In Chapter 7 we lay out the details of how both were analysed.

2.3.11 Testing interaction models

Four of the interaction models described in the literature review in Chapter 1 will be tested with data from S5 in Chapter 7. The other three models from the literature (Contingent Criterion model, Best Decides model, and Yes/Yes – No/No models) are falsified by general effects in our data, i.e., that tendency to follow Stylus advice is influenced by Stylus reliability (see our classification of these three models in Chapter 1, section 1.3.2, as models that pay no attention to the reliability of the automated aid). More specifically, the concepts of all the models are discussed in Chapter 1, we only print the relevant formulae here. The formulae come from Bartlett and McCarley (2017), although we believe there was an error in the OW model formula, which is corrected in Bartlett and McCarley (2019), which is the version we use here.

d'_{aid} for each model is computed the same way as participants' d' ; The H and FA-values used are the numbers of correct and incorrect Stylus judgements in each condition. d'_{aid} in C70 is 1.05, d'_{aid} in C90 is 2.56.

In Bartlett and McCarley's experiments (2017, 2019), participants completed a combination of unaided and aided trials. Because in our S5 experiment participants in G90 and G70 only encountered aided trials, the operator values used to compute the model predictions are the unaided Control Group (CG) participants' scores.

Coin Flip (CF) model

$$p_{H-CF} = 0.5 * (p_{H-operator} + p_{H-aid})$$

$$p_{FA-CF} = (p_{FA-operator}) * (p_{FA-aid} + (0.5 * p_{FA-operator})) * (1 - p_{FA-aid}) + (0.5 * (1 - p_{FA-operator})) * p_{FA-aid}$$

$$d'_{CF} = Z_{p(H)-CF} - Z_{p(FA)-CF}$$

$$C_{CF} = -0.5 (Z_{p(H)-CF} - Z_{p(FA)-CF})$$

Probability Matching (PM) model

R_{aid} is the aid's average reliability rate

$$p_{H-PM} = R_{aid} * p_{H-aid} + (1 - R_{aid}) * p_{H-operator}$$

$$p_{FA-PM} = R_{aid} * p_{FA-aid} + (1 - R_{aid}) * p_{FA-operator}$$

$$d'_{PM} = Z_{p(H)-PM} - Z_{p(FA)-PM}$$

$$C_{PM} = -0.5 * (Z_{p(H)-PM} - Z_{p(FA)-PM})$$

Optimal Weighting (OW) model

$$d'_{OW} = \text{sqrt}(d'_{operator}{}^2 + d'_{aid}{}^2)$$

Uniform Weighting (UW) model

$$d'_{UW} = (d'_{operator} + d'_{aid}) / \text{sqrt}2$$

When d'_{team} has been computed for each of the models, it can be compared with d' of the participants of each group that received assistance from Stylus with a one-sample t -test.

2.3.12 Measuring sensitivity and bias with non-parametric measures

A non-parametric approach is an alternative to parametric analysis that does not assume a normal distribution of the data and equal variance (although some suggest it is not truly non-parametric in the sense that assumptions about the underlying distributions are still made, and it should therefore be avoided (Pastore et al. 2003)). Introduced by Pollack and Norman in 1964 as a means to analyse the results of recognition tests where no strong underlying assumptions of underlying mechanisms can be made (Pollack and Norman 1964, Zhang and Mueller 2005), its use has broadened over time. The use of non-parametric SDT analyses is not without controversy, and dismissed as a modern fad by some (Pastore et al. 2003). We have to discriminate between the use of non-parametric *measures*, of which the concept of *percentage correct* is perhaps the easiest to understand, and non-parametric *analyses*. We do not attempt any non-parametric analyses, but we do use non-parametric measures in Chapters 3 – 5 of this thesis. The reason for doing so, is that they might be useful because they are thought to largely bypass some of the problems introduced by a relatively small number of trials, such as unequal variance in the data.

The simplest non-parametric measures are observations like “number of Stylus responses” and “confidence in Stylus responses”. But obviously the difference in the number of "Stylus" responses will be different between conditions where Stylus performs at a different level: if Stylus more often judges correctly in one condition than in the other, the number of times a participant agrees with Stylus will almost automatically be higher in the first condition. Although we report these easily interpretable measures throughout

the chapters that discuss our experimental studies, we note that this is typically the kind of measurement artifact that the SDT measures get around. The easiest to interpret non-parametric performance measure we consistently report throughout this thesis, is percentage correct. It may have its own particular issues (Macmillan and Creelman 2005), it is so easy to understand as a way of describing performance that we believe it helps provide conceptual context.

The second category of non-parametric measures are equivalents of the sensitivity measure d' and the bias measure c . Although it has been argued that assumptions about the underlying distributions are still made (Pastore et al. 2003), and, that the bias measure B'' is only superficially related to the sensitivity measure A' (Macmillan and Creelman 2005), which perhaps affects their usefulness, they are widely used, and may be particularly useful in studies with a low number of data points. We have to note here that although we report analyses using A' and B'' in full for Stylus 1 and 2, we report them in an appendix for S3, and deem them impractical or unhelpful in the subsequent studies.

In the same way that there are several different parametric sensitivity and bias measures of which we only use the most common ones, there is a multitude of non-parametric measures. In this thesis we only use A' and B'' , which, again, are the most commonly used of their kind.

2.3.13 Non-parametric sensitivity, A'

$$\text{If } p_H \geq p_{FA}, A' = 0.5 + (((p_H - p_{FA}) * (1 + p_H - p_{FA})) / (4 * p_H * (1 - p_{FA})))$$

$$\text{If } p_H \leq p_{FA}, A' = 0.5 - (((p_{FA} - p_H) * (1 + p_{FA} - p_H)) / (4 * p_{FA} * (1 - p_H)))$$

A' -values near 1 signal good sensitivity, values close to 0.5 indicate chance performance. Note that there are two different formulae, the use of which depends on the relative proportions of Hs and FAs.

2.3.14 Non-parametric bias, B''

$$\text{If } p_H \geq p_{FA}, B'' = (((p_H * (1 - p_H)) - (p_{FA} * (1 - p_{FA}))) / ((p_H * (1 - p_H)) + (p_{FA} * (1 - p_{FA})))) * -1$$

$$\text{If } p_H \leq p_{FA}, B'' = (((p_{FA} * (1 - p_{FA})) - (p_H * (1 - p_H))) / ((p_{FA} * (1 - p_{FA})) + (p_H * (1 - p_H)))) * -1$$

Like c , B'' is a 0-centric measure. As long as the proportion of False Alarms is equal to the proportion of Misses, c will be 0. When the proportion of False Alarms is greater than the proportion of Misses, the B'' -score will be negative, and when p_{FA} is lower than p_M , B'' will be positive. Note that also for B' there are two formulae. As a result of the z-transformations used to arrive at the parametric measures, the equivalents of positive c -scores are negative B'' -scores, and vice versa. Positive and negative scores are entirely a matter of convention, and it is the comparison between two scores arrived at through the same method that is of interest, the positive and negative direction of the data being completely arbitrary. Since bias metrics were inverted between c and B'' in our data, all B'' -results have been multiplied by -1 to arrive at a

symmetrical direction in the parametric and non-parametric representation of the measures, this means that positive B'' , or the proportion of FAs outnumbering the proportion of Ms, indicates bias towards Stylus, and negative B'' , or p_M outnumbering p_{FA} , indicates bias towards the Original Sentence in S1–4.

2.3.15 Example calculation

Table 2.4 shows a matrix with tallies per category for a random S1 participant. We use the data in this table to demonstrate how different sensitivity and bias measures were computed.

<i>ID = R_1gqLibBXSrqcUn0 N trials = 30</i>	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected</i>	H = 14	FA = 6
<i>Original sentence selected</i>	M = 6	CR = 4

Table 2.4 – Stylus Signal Detection example matrix

For above example, this results in:

$$\begin{aligned}
 \text{percentage correct} &= (30 / (14 + 6)) * 100 &&= 60\% \\
 p_H &= 14 / (14 + 6) &&= 0.7 \\
 p_{FA} &= 6 / (4 + 6) &&= 0.6 \\
 z_H &= \text{NORMSINV}(0.7) &&= 0.52 \\
 z_{FA} &= \text{NORMSINV}(0.6) &&= 0.25 \\
 d' &= 0.52 - 0.25 &&= 0.27 \\
 c &= 0.52 + 0.25 &&= 0.78 \\
 A' &= 0.5 + (((0.7 - 0.6) * (1 + 0.7 - 0.6)) && \\
 & / (4 * 0.7 * (1 - 0.6))) &&= 0.60 \\
 B'' &= (((0.7 * (1 - 0.7)) - (0.6 * (1 - 0.6))) && \\
 & / ((0.7 * (1 - 0.7)) + (0.6 * (1 - 0.6)))) * -1 &&= 0.07
 \end{aligned}$$

2.3.16 Post-task

Post-task we calculated participants' performance (number correct calculated by adding up Hits and Correct rejections) and the number of times they followed Stylus' suggestions (Hits + False alarms) and we calculated the average and the standard deviation of that data, and of participants' own estimations of those results. The average and the standard deviation of participants' estimations of how well they did in relation to others were calculated, but these figures are not reported because in hindsight social comparison was outside the remit of this thesis.

Because participants' judgements of their own and of Stylus' performance during the experiment were measured with aggregated questions, the results were tested for internal reliability (Cronbach α) and then their average was reported. Lastly, the average and standard deviation of participants'

judgement of the plausibility of Stylus being an automated system were calculated.

2.4 Analysis and reporting of results: hypotheses and analyses

A series of hypotheses about effects of perceived self-efficacy, trust, performance, and confidence is tested for each of our studies. For ease of reference, we number hypotheses by experiment number and in sequence (S1-H1, etc.), and they are tested once all required data have been reported. An overview of all hypotheses and whether they were confirmed can be found in Appendix D.

Next to the main hypotheses, additional statistical analyses will be undertaken to investigate the relationship between variables. Standard statistical methods were used, e.g., correlations, single and paired samples *t*-tests and AVOVAs. A pretty liberal *p*-value of .05 was used, which is to some extent justified by considerable participant numbers, but opens up the risk of Type 1 errors. To combat this risk, the most important findings are replicated throughout the thesis.

Assumption tests were automatically carried out when doing analyses in JASP. For *t*-tests these were Shapiro-Wilk (normality) and Levene's (equality of variances) tests. For ANOVAs they were Mauchly's sphericity tests, with sphericity corrections (none, Greenhouse-Geisser, Huynh-Feldt), and Levene's (homogeneity) tests. Assumption test results are not reported in the thesis unless assumptions are shown to be violated, in which case this is noted. This happened in only a single test, see Chapter 4, section 4.4.3.1.1.

2.4.1 Missing data in confidence cells

As all participants completed all trials, missing data in confidence cells are not the result of missing responses, but rather the result of the SDT matrix. Therefore, missing values are substituted by the mean of participants' confidence cells that are present. For example: if there was no data in an "FA confidence" cell, the mean of a participant's "H confidence", "M confidence" and "CR confidence" was used instead. Although this method of mean imputation has its drawbacks, such as the potential to introduce bias, it is commonly used (Donders, Van Der Heijden, Stijnen and Moons 2006), and discussing its pros and cons and potential alternative approaches is outside the scope of this thesis.

2.5 Discussion of research approach and methodology

2.5.1 Participants and online platform

We are well aware of the fact that the use of online surveys comes with its own specific advantages and challenges (Evans and Mathur 2005, Nayak and Narayan 2019), and that this also applies to the use of “crowd working” platforms like Prolific (e.g., Lefever, Dal, and Matthíasdóttir 2007, and more specifically discussing Prolific: Lumsden 2018, and Palan and Schitter 2018). Although some of the advantages like ease of recruitment and comparatively low cost are evident, the challenges need to be critically assessed, just as when doing research in a physical lab. A consequence of using Prolific, is that we only have general information on our participants, and that we get anonymised data that we cannot relate to actual people. While Prolific provides a participant ID and personal details like age, gender, nationality, country of birth, and first language that can be cross-referenced with questions in the survey, and which we found to be a near-perfect match in S1 and S2, we do not know whether the information participants have provided is truthful, and if participants are who they claim to be.

Another potential concern with using an online platform is self-selection. At the time of our experiments, Prolific’s total pool amounted to approximately 50,000 participants, of which roughly 22,000 were eligible for the experiment after preselection (screening out non-native British English speakers). Prolific works on the basis of first-come-first-served until the agreed number of responses has been reached, which means there is a risk that the majority of the participants are very active on Prolific and very experienced in answering surveys (“professional survey takers”). Another possible incentive for self-selection, is the amount of the payment participants receive (both average hourly rate and actual payment on completion, which varied between our studies). However, we believe self-selection is a minor concern, because the task in our study is quite specific (e.g., there are no socially preferable responses possible) and there is no direct relation that we are aware of between people’s desired income and their spelling and grammar proficiency.

In regard to the previous paragraphs, a potential advantage of using a crowd working platform like Prolific, is that we had a more demographically varied sample than if we had conducted our study in a lab with only undergraduate students, as would be typical for this type of research.

2.5.2 Gender differences – females somewhat overrepresented in all studies

The gender distribution in S1 was 19 males and 43 females (approximately 1:2.2). At the time of the experiment, Prolific’s pool had approximately 44% male and 56% female members, so we assumed a larger poll would lead to a more accurate representation of this gender distribution. On basis of analysis of our study data we had no reason to assume the observed gender disbalance influenced the findings. However, out of an abundance of caution, we assured an equal gender split in S2 by simultaneously running two identical

versions of the experiment, both with a different gender added as an eligibility filter. Analysis of the S2 data again revealed no obvious gender differences, hence we abandoned this approach for the following three studies.

2.5.3 Method/ environment

Because we could not control participants' environment and we could not observe them, like in a lab, we can only assume they completed the study seriously and concentrated. However, on average participants' performance is consistent throughout the experiments. On the Prolific platform, it is possible to control on which types of platforms participants can participate. For our experiments, we only allowed desktop and laptop computers, but we cannot be fully certain that participants have not used other (mobile) devices, e.g., by using a web browser on a mobile phone instead of a dedicated phone app.

2.5.4 Attention checks

From S3 onward, we introduced a short series of easy practise trials at the start of the task to familiarise participants with the format of the trials. These practise trials also served as an attention check; we would expect all participants who pay attention to complete these trials without fault. In S3 and S4, there was a dummy trial designed so that Stylus' communicated likelihood estimation was completely accurate. This dummy was presented randomly in between the test trials and the result was not used in our analyses, but it served as a second attention check. In Stylus 5, bias towards responding "Yes" or "No" served as a supplementary attention check; participants would not be expected to display any bias because of the equal signal-noise ratio in this study.

2.5.6 Validity of metrics

There is not a single method to measure perceived self-efficacy, nor is there a single metric that describes it (Bandura 2006). We have established participants' *own judgment of capability* by pre-task asking them how they judged their own spelling and grammar skills. The results of these questions were then averaged and internal reliability was checked, thus creating a single metric. Pre-task trust was measured with questions about participants' judgement of the trustworthiness of spelling and grammar checkers in general. Our metrics are based on common examples of reliable measures from the literature, which all use similar questions, and typically 'not at all agree – completely agree'-type Likert-scales (see e.g., McDonnell 1969, Lichtenstein, Fischhoff, and Phillips 1977, Muir 1989, Gigerenzer, Hoffrage, and Kleinbölting 1991, Lee and Moray 1992). The wording of our questions was slightly adapted to fit with the domain, and we used 0–100 scales instead of Likert-scales to warrant consistency with our other measures, such as that of trial confidence.

Participants' measure of confidence, or *judgment of performance*, was self-reported as well. Here we also used the conventional method of asking participants 'how confident are you in your answer?' after each trial, and asking them to estimate in how many trials they thought they responded correctly

after they completed the task (see e.g., Gigerenzer et al. 2008, and section 1.2.4.6 of this thesis). The latter scale is inconsistent with the 0–100-scales we used in the rest of our research, because it is used to test the potential of the overconfidence effect disappearing by comparing post-task frequencies with probabilities during the task (Gigerenzer, Hoffrage, and Kleinbölting 1991; Gigerenzer 1994).

2.5.7 Effect of proportion of signal and noise trials on bias and trust in a system

In our 2AFC-like designs, the proportion of signal and noise trials can only be equal if the aid performs at 50%. This was not desirable in our studies because we wanted to test different levels of performance of the aid, and as a result the distribution is skewed by default in S1–4 (S5 uses a Y/N-design). It is possible that the proportion of Stylus-correct items will affect bias, but in this case not because of a measurement artifact. Rather, such an effect would presumably be because participants are noticing how good the Stylus judgement is or because they believe Stylus' likelihood rating, and then that affects their willingness to accept Stylus advice. We hypothesise that this is an effect that can be expected, rather than an artifact that can be "explained away".

2.5.8 Using traditional SDT terminology in relation to our paradigm

In SDT's original application as a means to discriminate between the presence and absence of radar signals (Peterson, Birdsall, and Fox 1954), the use of the terms *signal* and *noise* had a logical relationship with the observations, and the terms Hit, Miss, False Alarm, and Correct Rejection were undisputed. In our research the designations of *signal* and *noise* are rather arbitrary, as is the case in most contemporary applications of SDT in Human Factors and Social Science research. E.g., in our studies, "signal" might mean that Stylus correctly indicates an error, or it might indicate that Stylus *correctly* indicates an error, or that the Original sentence is correct, depending on among others the design (2AFC-like vs. Y/N). Although we have considered using alternative descriptors, we believe explaining how we use the common SDT terminology in each instance is the least confusing method to describe our findings.

2.5.9 Bloated specifics in aggregated pre and post-task questions

As Newstead (1986) and Kline (2000) point out, there is a danger in aggregated variables, of creating factors that are essentially meaningless because the questions they consist of are too similar. This redundancy leads to a misleadingly high alpha-coefficient because the internal variance is too small. To avoid these tautological factors, or *bloated specifics* (Cattell 1957, Cattell and Kline 1977), and create factors that measure the same construct yet have a sufficient level of internal variance, pre and post-task questions were adjusted between experiments because we observed very high alpha coefficients (>0.95) in the first studies.

2.6 Ethics and data management

Since human subjects were involved, all our studies were designed following the University of Bath Department of Computer Science ethics guidelines (see Appendix E1). A data management plan was created with, and stored by, the online data management software DMPonline. The first two studies were described in an online Ethical Implications of Research Activity (EIRA1) form, checked by the Department Research Ethics Officer and a second reader, and approved by the Head of department. Studies 3 through 5 were, following a newly introduced ethics procedure, also approved by the University of Bath Psychology Research Ethics Committee (PREC 19-285).

2.6.1 Consent

Prior to starting an experiment, participants were asked to consent to taking part in the experiment, their responses being recorded and stored, and their responses being used for publication and research (see Appendix E2).

2.6.2 Debrief

After completing a survey, participants were thanked for their participation, and the objectives of our research were explained. The researchers' contact details were provided as well (see Appendix E3).

Chapter 3 – *Stylus 1: Developing a paradigm to test effects of trust and perceived self-efficacy on performance and confidence*

3.1 Introduction

This chapter details the method and results of Stylus 1 (S1), which is the first in a series of five related experimental studies. S1 explores the idea that accepting/ rejecting suggestions from an automated writing style checker might be influenced by the interplay between general prior trust in systems that give writing advice, and participants' perceived grammar and spelling self-efficacy. This exploration was required in the development of our experimental paradigm as described in section 2.1.2, and to establish to what extent interaction with writing aids aligns with interactions with aids in other domains (e.g., Lee and Moray 1994, Moray, Hiskes, Lee and Muir 1994, Wiczorek and Meyer 2019, and see section 1.2.1).

It is common experience that writing-style checkers, like all decision support systems, are not perfect and their signalling of errors will sometimes be false alarms; their suggestions for rewrites might make matters worse. Therefore, we suggest when users accept or reject a style checker's suggestions, their decision process will depend on exactly the aspects of the interplay of trust in the automation and their perceived self-efficacy that are reviewed in Chapter 1, *Subject area background and literature review*.

S1 was essentially a pilot study with only a limited number of participants, that was used to lay out and refine our methodology. Despite its small sample, and results that are too limited to draw useful conclusions from, we discuss the study in full in this chapter because it gives a good insight in how we developed our methodology. The second study is an improved version of S1, as will be discussed in Chapter 4.

3.2 Method

The objective of S1 was to establish a baseline for our future studies by investigating the fundamental relation between trust and perceived self-efficacy in the use of automation in this domain. To achieve this, we designed a study that consists of a series of pre-task questions, an experimental task and a series of post-task questions. The study was set-up on, and hosted by, the online survey platform Qualtrics. S1 was designed following the University of Bath Department of Computer Science's local ethics guidelines, and a data management plan was filed. The study operates as a pilot study, leading to incremental changes in materials and methods; to maximise the benefits and expose the rationale for changes, the results are reported in full.

3.2.1 Task design, variables and hypotheses

3.2.1.1 Task design

S1 used a between-subjects design, for which participants were randomly split into two equally sized groups (31 participants each). The first group of participants, Group Good (GGood) encountered a version of the experiment in which Stylus, an imaginary language checker, correctly suggests an alternative for twenty sentences with errors ("signal trials"), and incorrectly suggests alternatives for ten correct original sentences ("noise trials"). The second group, Group Bad (GBad) got ten correct suggestions and twenty incorrect ones (i.e., a system that performs very poorly). For GBad, ten suggestions from GGood were inverted (i.e., the "Stylus suggestions" for the first group were now presented as the "Original sentences"). Each participant was presented with 30 trials, each made up of two alternative sentences presented in tandem. Of the first alternative (labelled "Original sentence" in each trial) they were told in an introductory paragraph that it was written by a human writer, of the second that it was a suggestion for improvements from an automated editing tool called Stylus ("Stylus suggestion"). Participants were asked to indicate which sentence was better, the original sentence or the Stylus suggestion.

3.2.1.2 Variables

The individual differences variables we aimed to measure prior to the task were participants' prior trust in automated writing style checkers, and participants' perceived self-efficacy as checkers of grammar and spelling. The independent variable we manipulated during the experiment was the level of correctness of the Stylus recommendations and, lastly, the dependent variables we measured were acceptance of Stylus' recommendations and participants' confidence in their own responses.

3.2.1.3 Hypotheses

The most important hypotheses for this experiment derive from one of the main ideas reviewed in Chapter 1, i.e., that use of an automated aid will be

increased by trust in the aid, and decreased by the user's prior perceived self-efficacy (S1-H3; H4: these hypotheses rely on a SDT-based analysis of bias). The secondary hypotheses relate to the central Human Factors proposition that an aid can improve performance (H2) and to various aspects of confidence and overconfidence that relate to measures of perceived self-efficacy item confidence, and post-hoc confidence (estimated proportion of correct responses).

S1-H1 Participants' prior perceived self-efficacy in the domain of writing will be greater than their estimation of the efficacy of the average British English speaker.

This hypothesis concerns the *Above Average Effect*, as described by among others Dunning, Meyerowitz, and Holzberg 1989 in the domain of writing.

S1-H2 Participants' performance, in terms of percentage correct, will be better in GGood than in GBad.

This hypothesis concerns the very idea that an automated aid might affect participants' performance. If confirmed, this suggests that a more reliable aid might positively affect participants' performance.

S1-H3 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

This hypothesis concerns whether prior trust in automated writing style checkers might transfer to this particular situation, and affect participants' willingness to follow the aid's suggestions. If confirmed, this suggests that the higher participants' level of trust in similar systems, the more likely they are to accept the aid's advice, and vice versa, in line with what was observed earlier by among others Lee and Moray 1994, Moray et al. 1994, and Wiczorek and Meyer 2019 in other domains.

S1-H4 Participants' perceived self-efficacy in the domain of writing will be negatively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

This hypothesis concerns whether perceived self-efficacy in the domain of writing might affect participants' willingness to follow the aid's suggestions. If confirmed, this suggests that the higher participants' level of perceived self-efficacy, the less likely they are to accept the aid's advice, and vice versa, in line with what was observed earlier by among others Lee and Moray 1994, Moray et al. 1994, and Wiczorek and Meyer 2019 in other domains.

S1-H5 Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.

This hypothesis concerns the *Overconfidence Effect*, as described by among others Gigerenzer, Hoffrage, and Kleinbölting 1991, Gigerenzer 1994, and Kahneman and Tversky 1996 in other domains.

S1-H6 The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

This hypothesis concerns a measurement issue with the *Overconfidence Effect* in conjunction with S1-H5 as raised by Gigerenzer, Hoffrage, and Kleinbölting 1991, Gigerenzer 1994, and Kahneman and Tversky 1996.

S1-H7 Participants' trust in Stylus during the experiment (measured post-task) will be higher in GGood than in GBad.

This hypothesis concerns whether the aid's reliability might affect participants' trust as it has developed during the experiment. If confirmed, this suggests that while not given any feedback during the task, users recognise an aid's reliability, and this might in turn affect how much they trust it.

As well as testing these main hypotheses, additional statistical analyses will be undertaken to investigate the relationship between variables.

3.2.2 Participants

62 participants were recruited by listing the experiment on the online survey participant recruitment platform Prolific.ac. The listing specified the topic and the task of the study, the estimated completion time, and the reward. Participants were pre-screened on location (registered as United Kingdom resident), nationality (registered as United Kingdom citizen) and first language (English) to ensure their responses would be credible with regards to their command of British English. This pre-screening was arguably not perfect as it would potentially allow speakers of other variants of English to be part of the sample, but it was as precise as possible within Prolific at the time. Participants were automatically assigned to GGood or GBad by Prolific, according to the order in which they accepted participation in the online experiment.

Of the 62 participants, 19 were male and 43 were female and their age ranged from 19 to 73 ($M = 36.73$, $SD = 12.88$) (Split out per group: GGood, M13 / F18, age $M = 33.48$, $SD = 10.42$; GBad, M6 / F25, age $M = 39.97$, $SD = 14.06$).

3.2.3 Materials

The design of the decision tasks required us to sample correct and incorrect original sentences and rewrites at different levels of writing style and difficulty. We did not have any background data about the kinds of grammatical errors that are most common in our participant sample, nor much intuition about how competent this sample would be at the general task. Therefore, inspiration for the sentences used in the experiment was drawn from common errors observed on the RetroRides internet forum (2018b), readers' comments on the Guardian newspaper website (2018a), and from the Collins Improve your skills series (King 2009a, b, c). Stylus combines the behaviour of a set of real-world style checkers integrated in word processors or internet browsers (e.g., Word, Grammarly, Hemingway, Grammar, Language Tool, Slick Write, Whitesmoke and Espresso) and suggestions from the Collins books. We freely

mixed the suggestions from different sources, so as to gather a wide sample of particular cases. Errors used in the trials included among others punctuation, spelling, homophone and tautology and pleonasm errors. Errors were intuitively assigned to Original and Stylus sentences so they would look "believable". A pilot was run with one participant who was well-versed in British English to check this believability, and several minor changes were made on basis of their feedback.

3.2.3.1 Pre-task perceived self-efficacy

We measured participants' pre-task perceived self-efficacy with the items 'When thinking of how good I am at English grammar, I would class myself as [0; Not very good at all] – [100; Very good]' and 'When thinking of how good I am at English spelling, I would class myself as [0; Not very good at all] – [100; Very good]'. After the internal reliability of the results was checked (Cronbach α) and found to be acceptable ($\geq .70$), the mean score was used to compute participants' level of prior perceived linguistic self-efficacy. We also asked similar questions about participants' perception of average British English speakers' efficacy.

3.2.3.2 Pre-task trust

We measured participants' pre-task trust with the items 'When thinking of the trustworthiness of spelling suggestions in word processing software packages or internet browsers in general, I would class them as [0; Not very trustworthy at all] – [100; Very trustworthy]' and 'When thinking of the trustworthiness of grammar suggestions in word processing software packages or internet browsers in general, I would class them as [0; Not very trustworthy at all] – [100; Very trustworthy]'. After the internal reliability of the results was checked (Cronbach α) and found to be acceptable ($\geq .70$), the mean score was used to compute participants' level of prior trust.

3.2.3.2 Confidence

We measured participants' confidence in their responses at two levels: for single events (per trial) and overall (post-task).

3.2.3.2.1 Response confidence

At each trial participants were asked 'How confident are you of your answer? [50%; I guessed] – [100%; I'm certain]'.

3.2.3.2.2 Post-task estimation of frequency as a measure of long-term confidence

Post-task we asked participants to make an estimation of their own performance as a measurement of their confidence across the entire experiment. They were first asked to estimate how often they chose the Stylus suggestion over the Original sentence ('You have just rated thirty Stylus suggestions for rewrites of original sentences, how often do you estimate you chose the Stylus suggestion over the original sentence? [0-5; 6-10; [...]; 26-30]). They were then asked to rate their own level of performance ('Not all Stylus suggestions were correct; how often do you estimate you chose the correct answer (either the original sentence or the Stylus suggestion)? [0-5; 6-10; [...]; 26-30)'). (NB: In hindsight the 5-point intervals were a poor choice in

the survey design, which was corrected in S2.) Lastly, participants were asked how well they estimated they performed in relation to others who they were told had already completed the survey. This social comparison data, like all other social comparison data from the survey, was eventually not used in our analyses as we deemed it not relevant in regard to our hypotheses.

3.2.3.3 Post-task trust

Participants' retrospective trust in Stylus' suggestions during the task was measured with the single item 'When thinking of the trustworthiness of Stylus' performance during this experiment, I would class it as [0; Not very trustworthy at all] – [100; Very trustworthy]'.

3.2.3.4 Perceived Stylus performance and perceived plausibility of Stylus suggestions being created by an automated system as evidence of engagement with the task

Participants' perception of Stylus' performance during the task was measured with the question 'When thinking of the consistency of Stylus' performance during this experiment, I would class it as [0; Not very consistent at all] – [100; Very consistent]'. They were also asked 'How plausible do you find it that the Stylus suggestions were created by an automated system [0] – [100]?'; we used this as an indication of their level of engagement with the task.

3.2.4 Procedure

Participants conducted the experiment remotely on their own device. After accepting participation in the experiment on the Prolific platform, they were automatically assigned to GGood or GBad by Prolific according to the order in which they accepted participation in the online experiment. Upon completion of the survey, participants were paid an average of £3.20 (based on £12/hour). Participants needed to complete each trial before moving on to the next one and, following Prolific's terms, were only paid if the full survey was completed. The estimated time for completing the survey was 19 minutes (automatically estimated by Qualtrics).

3.2.4.1 Pre-task

The study started with a series of socio-demographic questions and questions about perceived self-efficacy, others' efficacy, and the trustworthiness of writing style checkers.

3.2.4.2 Word processing software use

Participants were asked what word processing software they used (e.g., Microsoft Word and Apache Open Office).

3.2.4.3 Experimental task

The task consisted of thirty trials, presented one after another. The trials were grouped into five randomly ordered blocks, which were each internally

randomised as well. This double form of randomisation was chosen to minimise the chance participants would encounter too many items from the same category (e.g., apostrophe errors, run-on sentences) at one time during the experiment. In each trial, participants were shown two sentences, of which they had to indicate which one they thought was “better”. Of the first alternative (labelled "Original sentence" in each trial) they were told in an introductory paragraph that it was written by a human writer, of the second that it was 'a suggestion for improvements from an automated editing tool called Stylus'.

Participants were also asked to rate their level of confidence in each response with a slider on a 50 (“I guessed”) to 100 (I’m certain) scale. The default for the slider was 75, and in order for them not to miss this rating, participants were forced to manipulate the slider before proceeding to the next question. An example of a single trial interface is shown in Figure 3.1.

No performance feedback was provided to participants during the experiment.

Original sentence: He placed an ad in a local newspaper and lo and behold, after 15 years of searching, found the bike of his dreams.

Stylus suggestion: He placed an ad in a local newspaper and low and behold, after 15 years of searching, found the bike of his dreams.

Which of the above do you think is better:

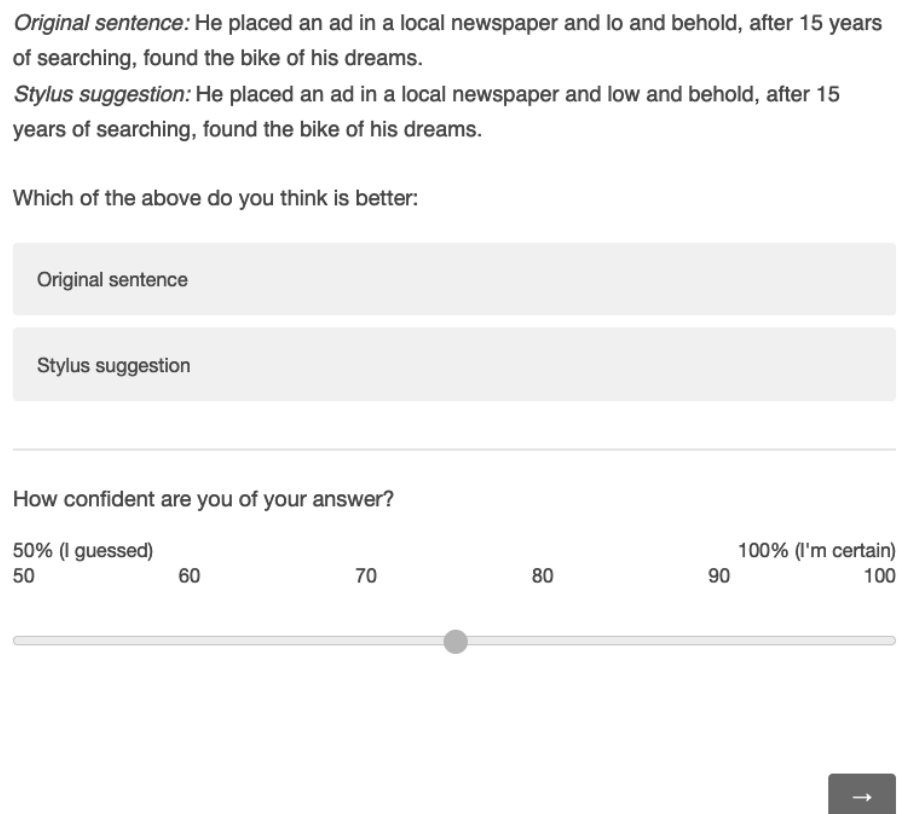
Original sentence

Stylus suggestion

How confident are you of your answer?

50% (I guessed) 100% (I'm certain)

50 60 70 80 90 100



The screenshot shows a trial interface. At the top, two sentences are presented: an 'Original sentence' and a 'Stylus suggestion'. Below these is a question: 'Which of the above do you think is better:'. Two light gray rectangular buttons are shown, one for 'Original sentence' and one for 'Stylus suggestion'. A horizontal line separates this from the next section. Below the line is the question 'How confident are you of your answer?'. A slider scale is shown with labels '50% (I guessed)' at 50 and '100% (I'm certain)' at 100. The slider has major tick marks at 50, 60, 70, 80, 90, and 100. A dark gray circular knob is positioned at the 75 mark. To the right of the slider is a dark gray square button with a white right-pointing arrow.

Figure 3.1 – S1 trial interface example screenshot

3.2.4.4 Post-task

After the sequence of decision tasks, participant's estimations of their own and Stylus' performance were measured. They were also asked how likely they thought it was that the Stylus suggestions were created by an automated system.

3.3 Analysis strategy

3.3.1 Acceptance and rejection of data

Prolific's minimum age for joining the platform is 18, but one participant reported in our pre-task socio-demographic questions that their age was 17. Their response was cross-checked with the data Prolific provided. In the Prolific data their age is 27, so we assumed 17 was a typo and corrected the age in the dataset.

To make sure participants completed the study in one sitting, a maximum duration of 25 minutes was set, the actual completion time was $M = 17.15$ ($SD = 6.24$). For unknown reasons, Prolific allows participants to overrun this limit and in the case of this experiment, seven participants took more time to complete it. Because their results were not statistically significantly different from the average, which suggests time does not have a noticeable effect on performance, their results have been kept.

We observed that four participants "returned" the study (meaning that they voluntarily left the study uncompleted without getting paid). There can be several reasons for this, on which we can only speculate (e.g., task too difficult, other priorities), because we have not asked these participants for their reasons. Partial data of uncompleted tasks was not stored by Qualtrics, hence it has not been used in our analyses. Also, a further three participants "timed out". A "time out" appears when a participant leaves the study inactive for too long (we do not know the specifics of the duration), or it may be a technical fault with the Prolific platform. Two of these three participants contacted us through the Prolific messaging system, telling us they had been shown an error message after completing the study. We could see in Qualtrics that they had indeed completed the study; their results were recorded and used in our analyses on top of the 60 participants originally agreed with Prolific. We made sure these participants got paid as well.

3.3.2 Pre-task questions – socio-demographics; aggregated variables

Prior to the task, participants were asked to rate their own perceived self-efficacy (three questions) in relation to spelling and grammar, their perception of the average British English speaker's efficacy (three questions), and the trustworthiness of writing suggestions in word processing software packages or internet browsers in general (two questions). Because the reported data is derived from aggregated questions, their internal reliability was tested with Cronbach α .

3.3.3 Signal Detection Theory to analyse task data

We used an adapted version of Signal Detection Theory to analyse certain aspects of performance data. An introduction to Signal Detection Theory and our adaptation can be found in Chapter 2, *Research approach and methodology*. Our methodology evolved slightly during our research, in this section we only discuss those aspects of the approach that are relevant to S1, and different to the methods used in the following studies.

3.3.3.1 Post-task questions – interval data

The midpoints of the 5-point intervals were used to analyse participants' estimations of their own performance (number correct and number Stylus). E.g., if a participant selected the 11–15 interval, their response was calculated to be $((11 + 15) / 2) = 13$.

3.3.4 Analysing confidence

Participants' confidence ratings were analysed using a mixed repeated measures ANOVA. The within-subjects variables that were tested were Correctness of the response (i.e., "Correct" vs. "Incorrect") and Type of response (i.e., "Stylus" vs. "Original"). The between-subjects factor was Group (i.e., GGood vs. GBad). Participants' average confidence in each cell of the design was entered into the ANOVA.

3.4 Results

The most important S1 data can be found in tabular form in Appendix A3, including breakdowns of aggregated variables. A table of all hypotheses from this thesis can be found in Appendix D.

3.4.1 Reliability testing

Firstly, the three pre-task aggregated variables (which are the averaged results of three questions each) *prior perceived self-efficacy*, *prior perceived efficacy of others* and *prior trust* were checked for internal consistency to validate the questions did indeed form reliable scales, with Cronbach α (unstandardised) $\geq .70$. Reliability of prior perceived self-efficacy was found to be $\alpha = .96$ and $\alpha = .95$ for GGood and GBad, that of prior perceived efficacy of average British English speakers $\alpha = .95$ and $\alpha = .96$, and that of prior trust $\alpha = .91$ and $\alpha = .93$ respectively. The post-task variable *post-trust* was checked for reliability as well, the Cronbach α scores for GGood and GBad were respectively $\alpha = .87$ and $\alpha = .91$.

3.4.2 Pre-task measures

Because allocation to groups was by order of participation, the two groups were expected to be more or less equal in terms of socio-demographics and pre-task efficacy.

3.4.2.1 Perceived self-efficacy, efficacy of others, and prior trust

Group Good (GGood) reported a perceived self-efficacy of $M = 75.50$ ($SD = 13.89$), an estimation of the efficacy of the average British English speakers' efficacy of $M = 61.76$ ($SD = 20.05$) and a rating of prior trust in style suggestions of $M = 74.29$ ($SD = 18.51$).

Group Bad (GBad) reported a perceived self-efficacy of $M = 69.83$ ($SD = 13.27$), an estimation of the efficacy of the average British English speakers' efficacy of $M = 53.72$ ($SD = 16.99$) and a rating of prior trust in style suggestions of $M = 67.27$ ($SD = 17.71$).

Independent samples t -tests showed that the differences between the groups (prior trust, $t(60) = 1.56$, $p = .13$, $d = 0.393$; prior perceived self-efficacy $t(60) = 1.36$, $p = .18$, $d = 0.344$) and between genders (prior trust, $t(60) = -0.84$, $p = .41$, $d = -0.231$; prior perceived self-efficacy $t(60) = -0.38$, $p = .71$, $d = -0.104$) were not statistically significant.

3.4.2.2 The above average effect

S1-H1 Participants' prior perceived self-efficacy in the domain of writing will be greater than their estimation of the efficacy of the average British English speaker.

In both groups, perceived self-efficacy was statistically significantly higher than perceived efficacy of others, GGood, $t(30) = 3.70$, $p < 0.001$, $d = 0.665$; GBad, $t(30) = 5.45$, $p < 0.001$, $d = 0.978$. Thus, S1-H1 is confirmed, and it replicates the well-known phenomenon of the *above average effect* (Dunning, Meyerowitz, and Holzberg 1989), one important type of over-confidence. This means that S1 participants think, on average, that they are better at spelling and grammar than the average British English speaker.

3.4.3 Performance

Any statistically significant differences between the groups in performance during the task may be explained by the manipulation, but we found none of the comparisons to be statistically significant. GGood took an average of 17.36 minutes ($SD = 6.72$) to complete the study and GBad an average of 16.94 minutes ($SD = 5.82$).

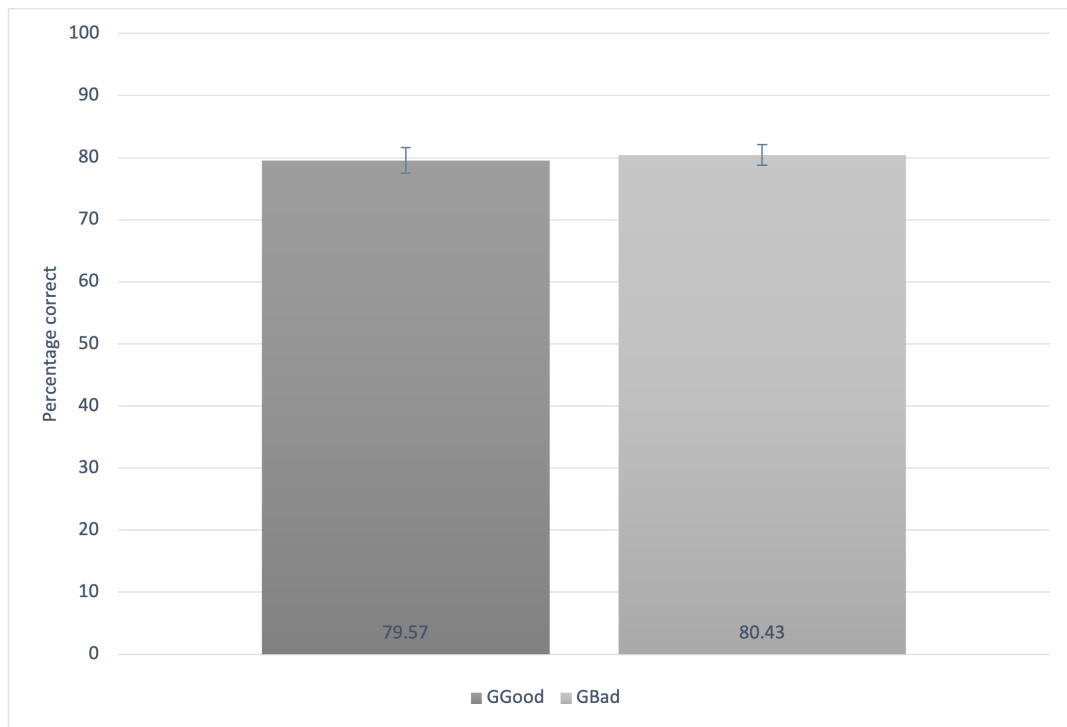


Figure 3.2 – S1 percentage correct responses, mean and standard error per group

S1-H2 Participants' performance, in terms of percentage correct, will be better in GGood than in GBad.

The average percentage correct for GGood was 79.57 ($SD = 11.38$), while GBad performed $M = 80.43$ ($SD = 9.38$), as shown in Figure 3.2. There was no statistically significant difference between the groups, $t(60) = -0.33$, $p = .746$, $d = -0.083$, i.e., there is no evidence for S1-H2 and thus it is rejected.

Performance can be described according to 2x2 a version of the grid used in Signal Detection Theory (see Table 3.1), to allow bias towards accepting Stylus advice to be separated from ability to distinguish correct sentences. Following SDT, we name the four categories Hit (H), Miss (M), False Alarm (FA), and Correct Rejection (CR)

GGood	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 17.10$ ($SD = 2.12$)	(FA) $M = 3.23$ ($SD = 1.73$)
<i>Original sentence selected by participant</i>	(M) $M = 2.90$ ($SD = 2.12$)	(CR) $M = 6.77$ ($SD = 1.73$)

GBad	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 8.42$ ($SD = 1.06$)	(FA) $M = 4.29$ ($SD = 2.18$)
<i>Original sentence selected by participant</i>	(M) $M = 1.58$ ($SD = 1.06$)	(CR) $M = 15.71$ ($SD = 2.18$)

Table 3.1 – S1 number of responses per category per group

3.4.3.1 Testing sensitivity and bias, parametric vs. non-parametric approach

As discussed in Chapter 2, *Research approach and methodology*, there is a lot of disagreement in the SDT literature about the appropriateness of either parametric or non-parametric measures. We decided initially to use both alongside each other to check how they might affect analysing our hypotheses. Mathematical formulae can be found in Chapter 2 as well. Figures 3.3a and 3.3b show comparisons of the parametric and non-parametric means and standard deviations of sensitivity, and 3.4a and 3.4b show comparisons of bias measures in GGood and GBad.

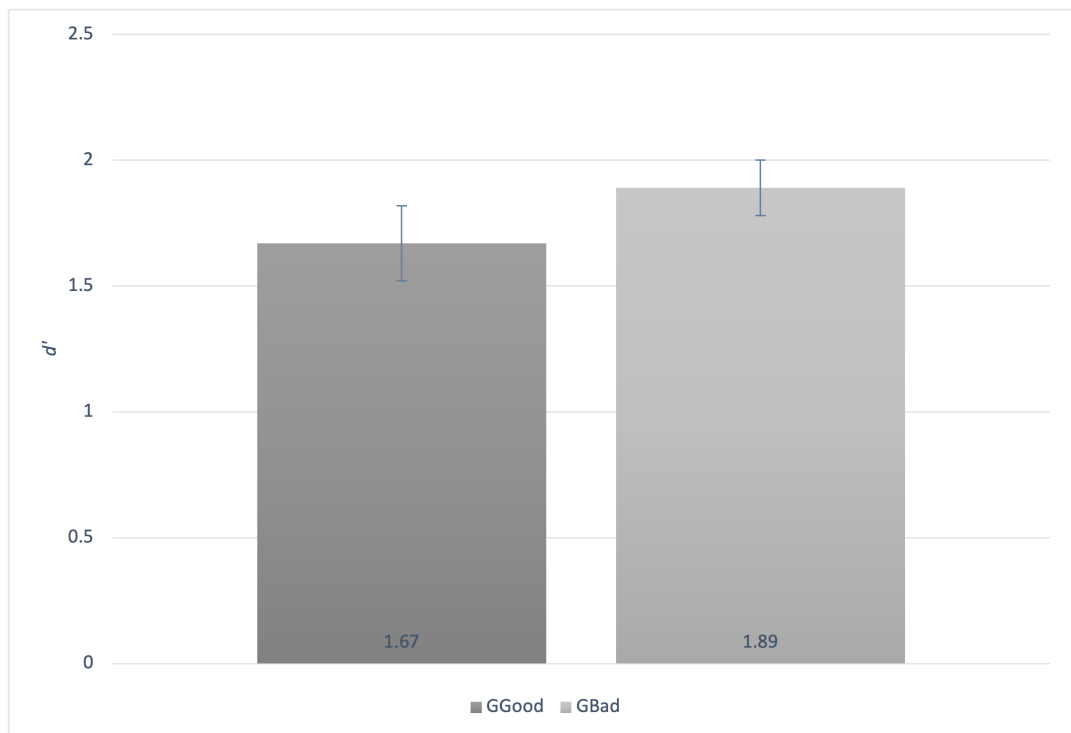


Figure 3.3a – S1 parametric sensitivity (d'), mean and standard error per group

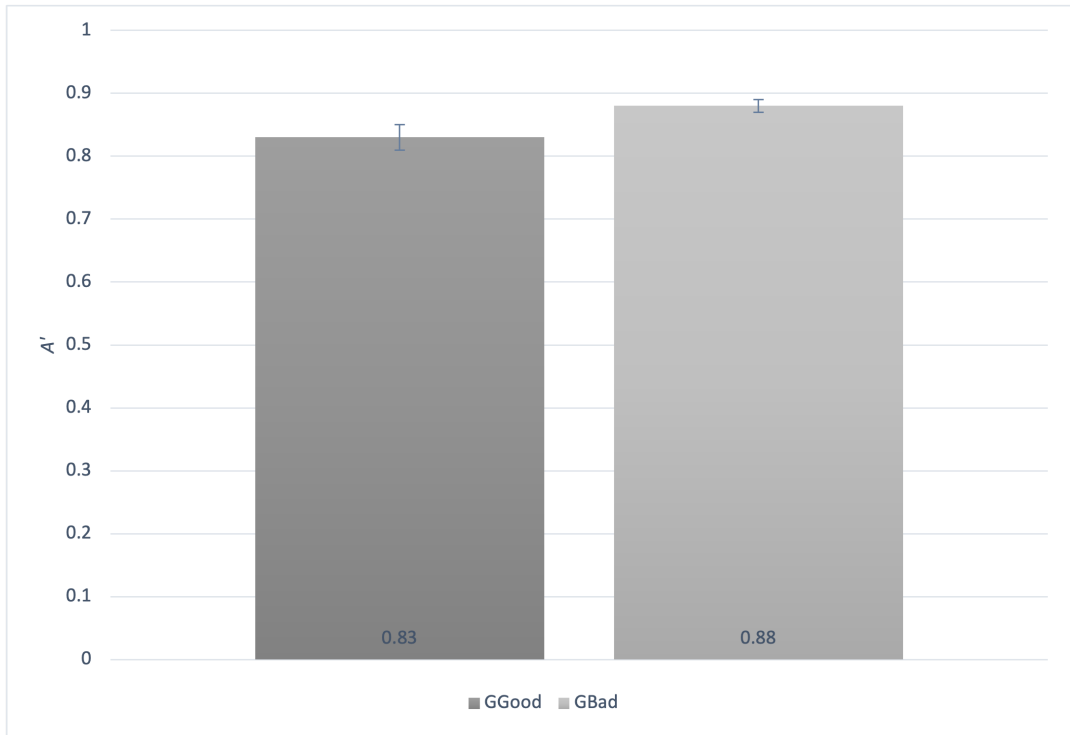


Figure 3.3b – S1 non-parametric sensitivity (A'), mean and standard error per group

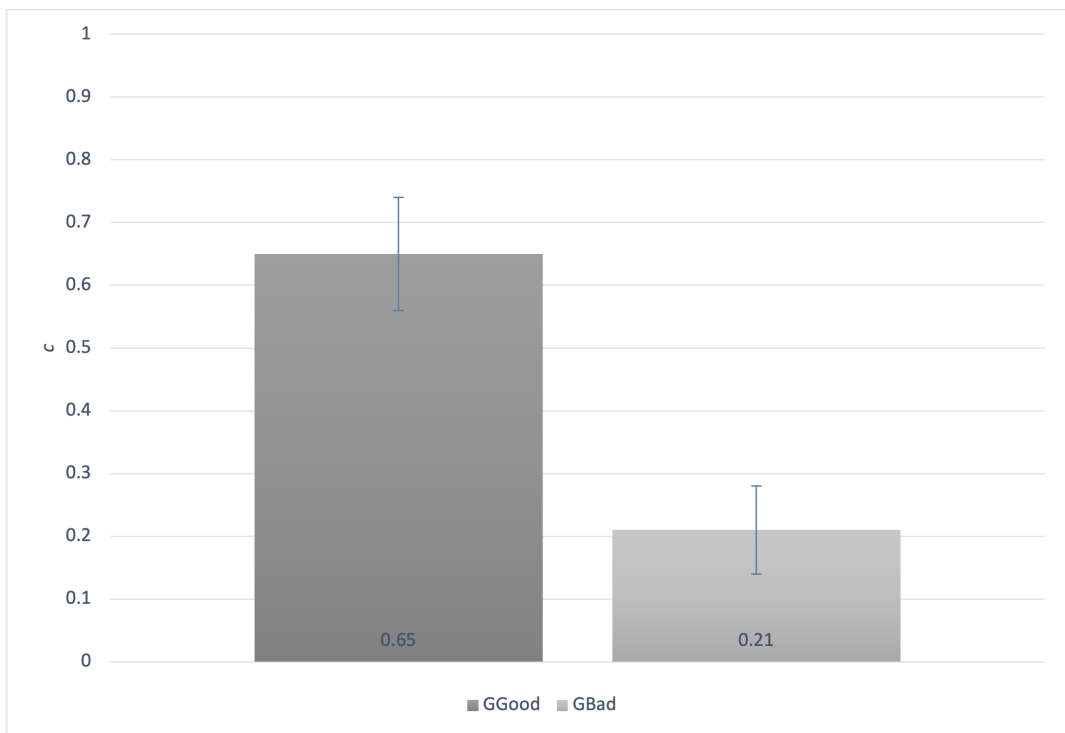


Figure 3.4a – S1 parametric bias (c), mean and standard error per group

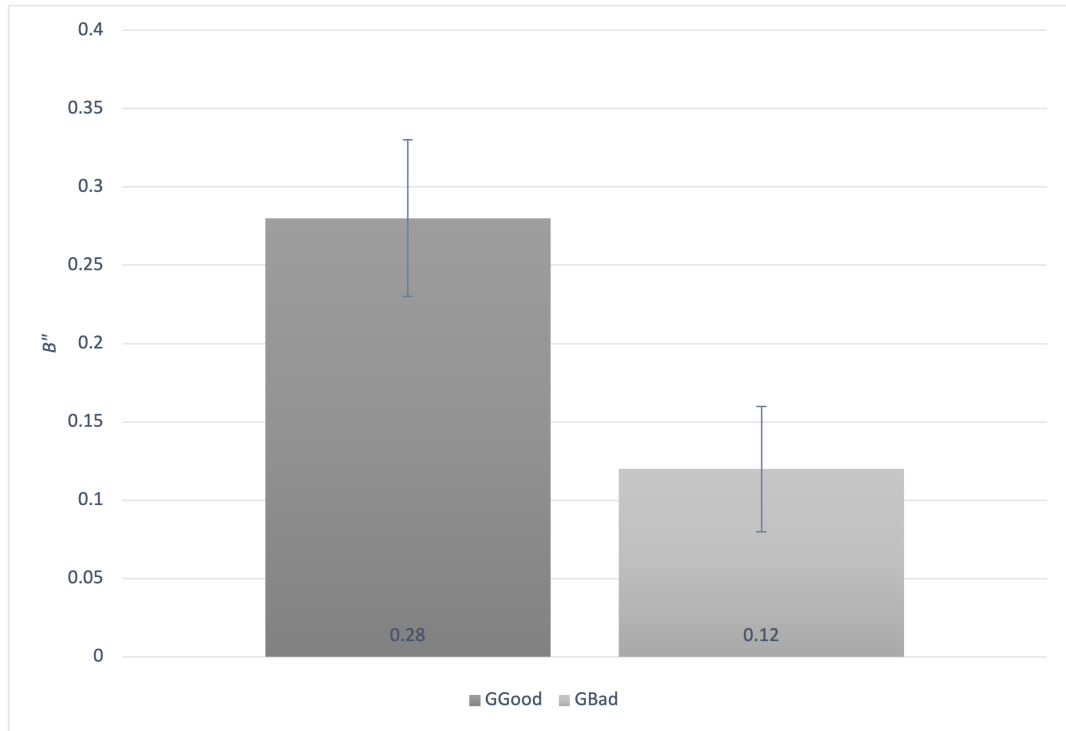


Figure 3.4b – S1 non-parametric bias (B''), mean and standard error per group

3.4.3.1.1 Testing sensitivity and bias with parametric measures

As discussed in Chapter 2, typical d' -scores are values up to 2, with positive scores meaning participants are sensitive to telling signal from noise, and $d' = 0$ meaning participants cannot discriminate between them. The higher the d' -score, the better calibrated participants are in choosing the correct answer. Because d' is not an intuitive metric, it is difficult to tell immediately how well the participants in GGGood and GBad did, and the same is true for c . The higher the d' -score, the better calibrated participants are in choosing the correct answer (i.e., either the Original sentence or the Stylus alternative). The higher their c -score, the more they tend to choose the Stylus sentence, independent of its correctness. GGGood's average d' -score was 1.67 ($SD = 0.84$), that of GBad was $M = 1.89$ ($SD = 0.63$). No difference between the groups would be predicted, and an independent samples t -test revealed no statistically significant difference between the groups, $t(60) = -1.15$, $p = .254$, $d = -0.292$.

GGGood's average c -score was 0.65 ($SD = 0.52$), and GBad scored $M = 0.21$ ($SD = 0.41$). An independent samples t -test showed a statistically significant difference between the groups, which confirmed our prediction that participants in GBad notice Stylus' poor performance and therefore feel less inclined to accept its advice, $t(60) = 3.67$, $p < .001$, $d = 0.932$.

The bias score, c , has a zero point, at which a participant shows no overall preference for choosing Stylus suggestions (independently of their correctness). In both groups the mean of c was greater than zero. One-sample t -tests in both groups compared c with zero, GGGood, $t(30) = 6.99$, $p < .001$, $d = 1.225$; GBad, $t(30) = 2.88$, $p = .007$, $d = 0.517$.

3.4.3.1.2 Testing sensitivity and bias with non-parametric measures

The non-parametric equivalent sensitivity measure A' and bias measure B'' are easier to interpret, although they still need explanation. The closer A' is to 1, the better participants' performance (choosing the correct sentence, either Original or Stylus). Positive B'' indicates a bias towards Stylus, negative B'' a bias towards Original. In S1, a statistically significantly different B'' -score between the groups, with GBad scoring lower than GGood, would indicate that participants in GGood notice the fact that correct Stylus suggestions are more prevalent than Original ones, and vice versa in GBad.

GGood's average A' -score was 0.83 ($SD = 0.13$), that of GBad was $M = 0.88$ ($SD = 0.08$). An independent samples t -test showed no statistically significant difference between the groups, $t(60) = -1.59$, $p = .116$, $d = -0.41$. GGood's average B'' -score was 0.28 ($SD = 0.29$), and GBad scored $M = 0.12$ ($SD = 0.23$). An independent samples t -test showed a statistically significant difference between the groups, which confirmed our prediction that GBad participants notice Stylus' poor performance, $t(60) = -2.38$, $p = .021$, $d = -0.603$.

3.4.3.1.3 Testing the role of pre-task trust and perceived self-efficacy

The bias scores also allow us to test the main hypotheses concerning the role of prior trust and perceived self-efficacy in determining the propensity to accept advice (S1-H3 and S1-H4).

S1-H3 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

The acceptance of Stylus recommendations is the degree of bias participants displayed. We tested S1-H3 with both the parametric bias variable c and the non-parametric variable B'' to understand how they would compare. In GGood there was no statistically significant correlation between prior trust in writing style checkers and c , $r(29) = -.17$, $p = .359$, or B'' , $r(29) = .32$, $p = .08$. Neither were there statistically significant correlations between prior trust in writing style checkers and c , $r(29) = .24$, $p = .204$, and B'' , $r(29) = -.19$, $p = .304$, in GBad. Thus, S1-H3 was rejected.

S1-H4 Participants' perceived self-efficacy in the domain of writing will be negatively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

In GGood there was no statistically significant correlation between perceived self-efficacy and bias toward accepting Stylus suggestions, c , $r(29) = .00$, $p = .999$, or B'' , $r(29) = .03$, $p = .879$. In GBad, there were also no statistically significant correlations between perceived self-efficacy and c , $r(29) = .04$, $p = .828$, or B'' , $r(29) = .04$, $p = .815$. Thus, S1-H4 was rejected.

Additionally, the sensitivity scores allow us to check whether participants' prior perceived self-efficacy, which although as already shown is overestimated, predict their level of performance. In GGood, there was no statistically significant correlation between participants' prior perceived self-efficacy and their sensitivity d' , $r(29) = .12$, $p = .531$, nor between their prior perceived self-

efficacy and A' , $r(29) = .22$, $p = .234$. In GBad there also was no statistically significant correlation between prior perceived self-efficacy and c , $r(29) = .26$, $p = .163$, nor between prior perceived self-efficacy and A' , $r(29) = .22$, $p = .230$.

3.4.4 Confidence analysis

3.4.4.1 Confidence during the task

The average self-reported confidence across the task was 91.36 ($SD = 5.06$) for GGood, and $M = 91.21$ ($SD = 5.44$) for GBad. An independent samples t -test showed no statistically significant difference in average reported confidence between the groups, $t(60) = 0.11$, $p = .91$, $d = 0.029$.

S1-H5 Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.

Across participants, the average confidences can be compared with percentage correct scores to test the standard overconfidence finding for trial-by-trial confidence measures. A paired samples t -test showed a statistically significant difference between confidence and percentage correct in GGood, $t(30) = 5.61$, $p < .001$, $d = 1.008$, as well as in GBad, $t(30) = 5.90$, $p < .001$, $d = 1.059$, which clearly demonstrates that confidence is overall higher than warranted by performance, and thus confirms S1-H5 in line with earlier findings from the literature (e.g., Gigerenzer, Hoffrage, and Kleinbölting 1991).

When we break down the confidence scores for H, M, FA and CR, they are as shown in Table 3.2, also shown in graphical form in Figure 3.5 for ease of comparison.

GGood	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 92.97$ ($SD = 5.37$)	(FA) $M = 88.74$ ($SD = 8.55$)
<i>Original sentence selected by participant</i>	(M) $M = 86.31$ ($SD = 11.13$)	(CR) $M = 90.51$ ($SD = 5.71$)

GBad	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 92.20$ ($SD = 4.99$)	(FA) $M = 86.28$ ($SD = 8.20$)
<i>Original sentence selected by participant</i>	(M) $M = 89.32$ ($SD = 12.69$)	(CR) $M = 91.90$ ($SD = 5.47$)

Table 3.2 – S1 mean confidence percentage per category per group

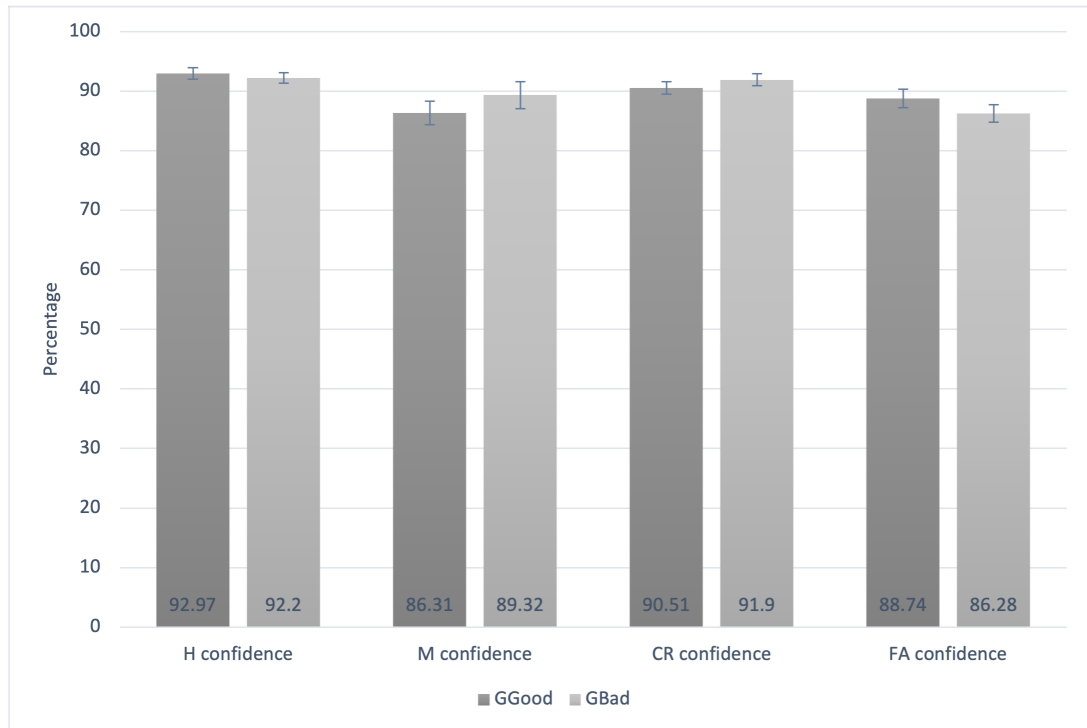


Figure 3.5 – S1 percentage H, M, FA and CR confidence, mean and standard error per group

The confidence data in Table 3.2 were analysed in a 2x2x2 mixed ANOVA, with Group as a between-subjects factor, and Correctness of response (correct v. incorrect) and Response type (Stylus vs. Original) as within-subjects factors.

This ANOVA revealed a statistically significant main effect in confidence of Correctness, $F(1, 60) = 29.10, p < .001, partial \eta^2 = .327$, which suggests that that confidence is at least to some extent meaningfully associated with performance. GGood reported an average confidence in correct responses of 95.25 ($SD = 5.03$) and GBad reported $M = 92.04 (SD = 5.14)$. GGood reported an average confidence in incorrect responses of 87.58 ($SD = 7.40$), and GBad reported $M = 86.98 (SD = 7.61)$. There was no statistically significant effect for Correctness of response x Group, $F(1, 60) = 2.61, p = 1.000, partial \eta^2 = .000$

There was no statistically significant difference in confidence of Type of response (Original vs. Stylus), $F(1, 60) = 0.42, p = .521, partial \eta^2 = 0.007$, although there was a statistically significant effect of Type of response x Group, $F(1, 60) = 4.97, p = 0.030, partial \eta^2 = .076$. The means underlying this interaction effect are as follows: GGood confidence in Original responses $M = 89.24 (SD = 5.19)$ and in Stylus responses $M = 92.38 (SD = 5.52)$; GBad confidence in Original responses $M = 91.69 (SD = 5.76)$ and in Stylus responses $M = 90.46 (SD = 5.39)$. This suggests that only those participants who were using the more reliable aid, were more confident when accepting its advice than when rejecting it.

None of the interaction effects were statistically significant, Correctness of response x Type of response, $F(1, 60) = 0.99$, $p = 0.325$, $partial \eta^2 = .016$; Correctness of response x Type of response x Group, $F(1, 60) = 1.05$, $p = 0.309$, $partial \eta^2 = .017$. There was no statistically significant between-subjects confidence effect, $F(1,60)$, 0.04 , $p = .838$, $partial \eta^2 = .001$.

3.4.5 Post-task measures

3.4.5.1 Post-task confidence

3.4.5.1.1 Post-task estimation of number correct and number Stylus responses

GGood participants' subjective estimation of the number of times they selected the correct answer (either Original or Stylus), $M = 20.73$ ($SD = 6.98$), is lower than, but not statistically significantly correlated with the objective frequency, their real number of correct responses, $M = 23.87$ ($SD = 3.41$), $r(29) = .33$, $p = .069$. GBad's average estimation was 19.60 ($SD = 7.16$), which is statistically significantly correlated with their real number of correct responses, $M = 24.13$ ($SD = 2.81$), $r(29) = .53$, $p = .002$. At least in GBad, and although not significantly significant, to some extent in GGood too, these correlations suggest that participants have some insight into their own performance.

When we compare GGood and GBad's average estimated number of Stylus responses, we note that for GGood there is a statistically significant correlation between the average number of times participants thought they chose the Stylus suggestion, 19.29 ($SD = 4.65$), and their real number of Stylus choices, 20.32 ($SD = 1.81$), $r(29) = .48$, $p = .006$. This shows that participants in GGood have some valid memory of their decision-making during the task. For GBad however, the group that encountered the poorer performing version of Stylus, there was no statistically significant correlation between their estimated number of Stylus responses, $M = 12.31$ ($SD = 5.22$), and the real number, $M = 12.71$ ($SD = 1.95$), $r(29) = .26$, $p = .156$. The latter may be due to the low number of data points, but we want to reiterate that all results of this study have to be treated with caution, and that this also pertains for statistically significant ones.

3.4.5.1.2 Comparing average single trial confidence and post-hoc estimates of performance

The literature suggests that people are overconfident in single events, and better calibrated after a completed task (Gigerenzer 1991, 1994, Gigerenzer et al. 2008).

We compared participants' average confidence for each trial over the whole experiment (which is indeed higher than warranted by their performance) with their average estimated number of correct responses (converted into percentages), which we treat as a post-task frequency confidence measure. A paired samples *t*-test shows a statistically significant difference between GGood participants' average confidence from each trial over the whole

experiment ($M = 91.36$, $SD = 5.06$) and their average estimated percentage of correct responses ($M = 69.09$, $SD = 23.25$), $t(30) = 5.47$, $p < .001$, $d = 0.982$. Similarly, there is a statistically significant difference between GBad participants' average confidence from each trial over the whole experiment ($M = 91.20$, $SD = 5.44$) and their average estimated percentage of correct responses ($M = 65.32$, $SD = 23.85$), $t(30) = 6.41$, $p < .001$, $d = 1.151$. This means that estimation of percentage correct is less inflated than trial-by-trial confidence ratings, which suggests that this observation from the literature in other domains also applies to the domain of aided grammar and spelling checking.

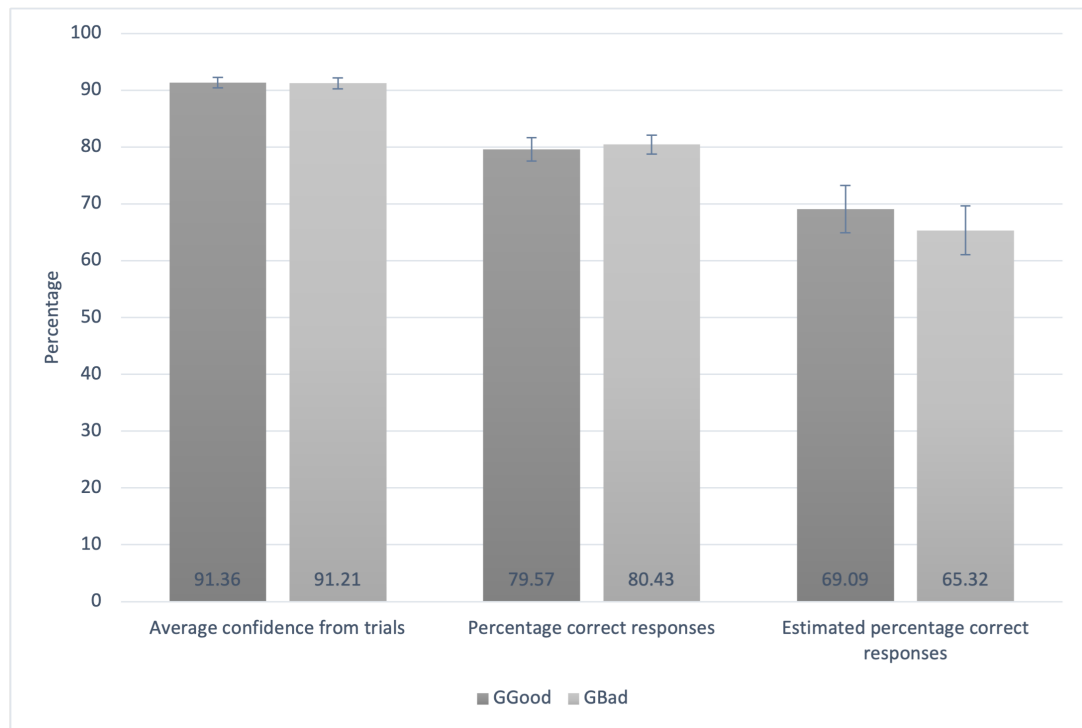


Figure 3.6 – S1 comparison of percentage trial confidence, percentage correct responses, and overall estimated percentage correct responses, mean and standard error per group

S1-H6 The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

To test whether participants' post-task frequency estimations are indeed more realistic than the average confidence reported during the task, we test the difference between percentage average trial confidence minus percentage correct, and percentage correct minus average post-task estimation of percentage correct. The means of trial confidence, percentage correct, and post-task estimation of percentage correct, as shown in Figure 3.6, have been reported earlier.

In GGood, the difference between trial confidence and percentage correct was $M = 11.79$ ($SD = 11.69$), and the difference between percentage correct and post-task estimation of percentage correct $M = 10.48$ ($SD = 22.19$). A paired

samples *t*-test showed no statistically significant difference between the two measures, $t(30) = 0.27$, $p = .792$, $d = 0.048$. In GBad, the difference between trial confidence and percentage correct was $M = 10.78$ ($SD = 10.17$), and the difference between percentage correct and post-task estimation of percentage correct $M = 15.11$ ($SD = 20.48$). A paired samples *t*-test showed no statistically significant difference between the two measures $t(30) = -1.04$, $p = .308$, $d = -0.186$.

The results show that a post-hoc estimation of correct responses is no more accurate than the trial confidence, and that confidence is as poorly calibrated, which contradicts our expectations based on findings from the literature (Gigerenzer, Hoffrage, and Kleinbölting 1991, Gigerenzer 1994, and Kahneman and Tversky 1996), and means that, surprisingly, S1-H6 must be rejected.

3.4.5.2 Post-task Trust in Stylus

S1-H7 Participants' trust in Stylus during the experiment (measured post-task) will be higher in GGood than in GBad.

As expected, GGood's trust in Stylus, $M = 68.77$ ($SD = 15.33$) is statistically significantly higher than that of GBad, $M = 49.79$ ($SD = 19.58$) as an independent samples *t*-test showed, $t(60) = 4.25$, $p < .001$, $d = 1.080$. Thus S1-H7 is confirmed, which suggests that while not given any feedback during the task, users are able to recognise the aid's reliability. This test also shows the effect of Stylus' reliability on participants' post task trust, with participants that encountered the more reliable version of Stylus trusting it more than the group that worked with the poorly performing aid.

3.4.5.3 Believability of Stylus as an automated system

The question “do you find it plausible that the Stylus suggestions are created by an automated system” served to confirm participants' engagement with the task. GGood judged the plausibility of Stylus' suggestions being created by an automated system at $M = 77.06$ ($SD = 18.52$), the average GBad judgement was 70.81 ($SD = 19.17$). A one-sample *t*-test above 50 showed a statistically significant result for GGood, $t(30) = 8.14$, $p < .001$, $d = 1.461$, as well as for GBad, $t(30) = 6.04$, $p < .001$, $d = 1.086$, which suggests that Stylus' performance was in line with what participants expected from this type of aid. We assumed beforehand that it would be reasonable to expect that participants in GGood found it more plausible that the “Stylus suggestions” came from an automated system than those in GBad, because performance of the version of Stylus GGood encountered is much closer to that of a real-world system than that of GBad, but an independent samples *t*-test showed no statistically significant difference between groups, $t(61) = 1.31$, $p = .196$, $d = 0.332$.

3.5 Summary of results

Summarising, we can conclude that our experimental design is promising because we observed several suggestions of trends, although we found not all of our hypotheses to be confirmed. Our main reservations are that the trials were perhaps too easy, which led to a ceiling effect in performance and confidence measures, and that the sample size was perhaps too small, which negatively affected the meaningfulness of our analyses.

We repeat the most important statistically significant results that we expect to hold up in an improved version of this experiment here, but not without repeating the proviso that results should be interpreted with caution due to the relatively low number of data points.

3.5.1 Sensitivity and bias

Analyses showed a statistically significant difference in bias between the groups, which indicated that participants in GGood, who encountered the better performing version of Stylus, were more willing to follow Stylus, independently of the number of correct Stylus suggestions. This effect was not matched by participants' sensitivity, which is their ability to select the correct answer, corrected for their bias.

3.5.2 Confidence

In both groups, there was a statistically significant effect of correctness, in that confidence in correct responses was higher than that in incorrect ones, although we did not find a difference between the groups in this case.

We observed a statistically significant difference between participants' mean reported trials confidence, and their estimated percentage of correct responses in both groups. In GBad there was a statistically significant correlation between participants' estimations of their number of correct responses and the real frequencies, but this was not the case in GGood. The inverse was true for the correlation between participants' estimations of their number of Stylus responses and the real frequencies, where GGood's results were statistically significant and not those of GBad.

3.5.3 Trust in Stylus

In GGood, participants' trust in Stylus was statistically significantly higher than in GBad, which is justifiable by the difference in Stylus' performance between the groups. Both groups reported it to be highly plausible that the Stylus suggestions were indeed created by an automated system, which we treat as a confirmation of their engagement with the task.

3.6 Conclusions and discussion

3.6.1 Stylus performance

Although discretely different between the two groups, the level of performance of the aid they encountered was deliberately weak in both to a) test the effect of assistance from imperfect automation even with an exaggerated level of imperfection, and b) to make sure we generated enough data points in all cells of the SDT matrix as described in Chapter 2. Interaction with better performing aids has been tested in S3, 4, and 5, and the effects on participants' performance and confidence, as well as implications for the experimental design, will be discussed in chapters 5, 6 and 7.

The distribution between the proportions of signal and noise trials in S1, as well as in S2 – 4, is asymmetric, which although not ideal (Macmillan and Creelman 2005), we believe is a good compromise with a relatively low number of trials, because it reinforces the experimental differences between the aids the two groups encounter.

3.6.2 Trial order and randomisation

To avoid question order bias, for example participants quitting after several trials due to a series of difficult items at the beginning of the experiment, the study was organised in five randomly ordered blocks of six internally randomised trials of different levels of complexity and difficulty. Although we had access to the randomisation order data per participant, we have not studied this in relation to their results because we have no reason to believe we would find any effect on performance or confidence.

3.6.3 Results and implications for the design of Stylus 2

After analysing the data, we observed several flaws in the design of our first experiment. As has already been noted, an important issue was that performance was very high, seemingly caused by the trials being too easy for our sample. We also observed a ceiling effect in the confidence scores, which could either be a reflection of easy trials, or an indication that the wrong type of measuring scale was used (Macmillan and Creelman 2005). Since this type of scale (50 – 100%) is commonly used to measure confidence (see e.g., Gigerenzer, Hoffrage, and Kleinbölting 1991), we assume that high confidence was indeed related to easy trials.

An improved version of the experiment was run as Stylus 2, which is described in Chapter 4. To combat the ceiling effects in the performance and confidence scores we observed in S1 and to warrant more reliable results, the trials were made more difficult, and the power of the experiment was increased.

Chapter 4 – *Stylus 2: Testing effects of trust and perceived self-efficacy with an improved experiment*

4.1 Introduction

This chapter details the methods and results of Stylus 2 (S2). The objective of S2 was to replicate S1, which is discussed in Chapter 3, but with an improved design that delivers more robust data. Although S2 is very similar to S1, we made several important improvements. Firstly, we made the trials more difficult because a lower level of performance increases the reliability of the sensitivity metric d' and the bias measure c (Macmillan and Creelman 2005), and might curb the ceiling effect in confidence scores. We also increased the number of participants to increase the number of data points, which benefits the power of the experiment and should benefit reliability of the results.

S2, like S1, focuses on the influence of the interplay between participants' prior trust and perceived self-efficacy on accepting/ rejecting suggestions from an automated writing style checker.

4.2 Method

4.2.1 *Task design, variables, and hypotheses*

4.2.1.1 Task design

S2 used a between-subjects design, for which participants were split into two almost equally sized groups (59 vs. 61 participants). The design of the task was identical to that of S1.

4.2.1.2 Variables

The variables measured in this experiment were identical to those in S1.

4.2.1.3 Hypotheses

For our second study we statistically tested the same seven predictions we made for S1, and we carried out the same additional statistical analyses about the relationship between variables as well. For descriptions of the purpose of the hypotheses, see Chapter 3, section 3.2.1.

S2-H1 Participants' prior perceived self-efficacy in the domain of writing will be greater than their estimation of the efficacy of the average British English speaker.

S2-H2 Participants' performance, in terms of percentage correct, will be better in GGood than in GBad.

S2-H3 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

S2-H4 Participants' perceived self-efficacy in the domain of writing will be negatively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

S2-H5 Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.

S2-H6 The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

S2-H7 Participants' trust in Stylus during the experiment (measured post-task) will be higher in GGood than in GBad.

4.2.2 Participants

120 participants were recruited on Prolific.ac in the same way as in S1, and the same participant screeners were used. Additionally, participants were also pre-screened with regards to participation in previous studies, and those who participated in S1 were excluded.

To guarantee an equal gender balance, the survey was run as two identical separate surveys with the screener "gender identity" selecting only male participants for one version, and only females for the other. In total 60 subjects were male and 60 were female. Ages ranged from 19 to 71 ($M = 8.95$, $SD = 13.03$) (Split out per group: GGood, M30 / F29, age $M = 40.46$, $SD = 12.86$; GBad, M30 / F31, age $M = 37.44$, $SD = 13.20$).

4.2.3 Materials

Items were created following the same methodology as for S1, but the items' difficulty level was increased. Correct and incorrect items were intuitively assigned to Original and Stylus sentences so they would look "believable" as human errors or automated system limitations. A pilot was run with one participant who was well-versed in British English to check this believability, and several minor changes were made on basis of their feedback.

4.2.3.1 Pre-task perceived self-efficacy, pre-task trust, confidence, post task estimations of frequency, post task trust in Stylus, and post-task believability of Stylus being an automated system.

All these variables were tested the same way as in S1.

4.2.4 Procedure

The procedure of S2 was the same as that of S1, with the following exceptions. Participants were paid an average of £1.80 (based on £6/hour), the estimated time for completing the survey was 18 minutes (automatically estimated by Qualtrics), and the actual average completion time was 18.19 minutes ($SD = 5.67$).

4.3 Analysis strategy

The analysis strategy for S2 was the same as for S1.

4.3.1 Participants returning the study, or timing out

Four participants in the male sample and two in the female sample "returned" the study, and a further three participants in each sample "timed out". Partial data of uncompleted tasks was not stored by Qualtrics, hence it has not been used in our analyses.

4.4 Results

The most important S2 data can be found in tabular form in Appendix A4, including breakdowns of aggregated variables. A table of all hypotheses from this thesis can be found in Appendix D.

4.4.1 Reliability testing

Reliability of prior perceived self-efficacy was $\alpha = .95$ for GGood and $\alpha = .95$ for GBad, that of prior perceived efficacy of average British English speakers $\alpha = .95$ and $\alpha = .88$, and that of prior trust $\alpha = .88$ and $\alpha = .93$ respectively. The Cronbach α scores for the post-task variable *post-trust* were respectively $\alpha = .83$ for GGood and $\alpha = .68$ for GBad.

4.4.2 Pre-task measures

Because allocation to groups was by order of participation, the two groups were expected to be more or less equal in terms of socio-demographics and pre-task efficacy.

4.4.2.1 Perceived self-efficacy, efficacy of others, and prior trust

Group Good (GGood) reported a perceived self-efficacy of $M = 73.18$ ($SD = 14.83$), an estimation of the efficacy of the average British English speakers' efficacy of $M = 60.07$ ($SD = 18.86$) and a rating of prior trust in writing suggestions of $M = 76.42$ ($SD = 17.07$).

Group Bad (GBad) reported a perceived self-efficacy of $M = 73.24$ ($SD = 14.25$), an estimation of the efficacy of the average British English speakers' efficacy of $M = 54.62$ ($SD = 17.65$) and a rating of prior trust in writing suggestions of $M = 75.60$ ($SD = 16.56$).

Independent samples *t*-tests showed that the differences between the groups (prior trust, $t(118) = 0.281$, $p = .779$, $d = 0.051$; prior perceived self-efficacy $t(118) = -0.02$, $p = .981$, $d = -0.004$), and between genders (prior trust, $t(118) = 0.03$, $p = .98$, $d = 0.006$; prior perceived self-efficacy, $t(118) = 0.71$, $p = .48$, $d = 0.129$) were not statistically significant.

4.4.2.2 The above average effect

S2-H1 Participants' prior perceived self-efficacy in the domain of writing will be greater than their estimation of the efficacy of the average British English speaker.

In both groups, perceived self-efficacy was statistically significantly higher than perceived efficacy of others, GGood, $t(60) = 5.70$, $p < 0.001$, $d = 0.742$; GBad, $t(60) = 9.18$, $p < 0.001$, $d = 1.176$. Thus, S2-H1, is confirmed, and it replicates the well-known phenomenon of the above average and the earlier S1-H1

result. This means that S2 participants think, on average, that they are better at spelling and grammar than the average British English speaker.

4.4.3 Performance

Any statistically significant differences between the groups in performance during the task may be explained by the manipulation, but we found only a limited number of the comparisons to be statistically significant. GGood took an average of 18.01 minutes ($SD = 5.13$) to complete the study and GBad an average of 18.37 minutes ($SD = 6.18$).

S2-H2 Participants' performance, in terms of percentage correct, will be better in GGood than in GBad.

The average percentage correct for GGood was $M = 64.97$ ($SD = 16.23$), and for GBad $M = 61.04$ ($SD = 9.95$). There was no statistically significant difference between the groups, $t(118) = 1.08$, $p = .283$, $d = 0.197$, i.e., there is no evidence for S2-H2. A comparison of the means with the S1 data (percentage correct GGood $M = 79.57$ ($SD = 11.38$), GBad $M = 80.43$ ($SD = 9.38$)) suggests we succeeded in creating a more difficult task.

Performance can be described according to 2x2 a version of the grid used in Signal Detection Theory (see Table 4.1), to allow bias towards accepting Stylus advice to be separated from ability to distinguish correct sentences. Following SDT, we name the four categories Hit (H), Miss (M), False Alarm (FA), and Correct Rejection (CR).

GGood	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 13.34$ ($SD = 4.12$)	(FA) $M = 3.85$ ($SD = 1.86$)
<i>Original sentence selected by participant</i>	(M) $M = 6.66$ ($SD = 4.12$)	(CR) $M = 6.15$ ($SD = 1.86$)

GBad	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 5.79$ ($SD = 1.46$)	(FA) $M = 7.48$ ($SD = 2.66$)
<i>Original sentence selected by participant</i>	(M) $M = 4.21$ ($SD = 1.46$)	(CR) $M = 12.52$ ($SD = 2.66$)

Table 4.1 – S2 number of responses per category per group

4.4.3.1 Testing sensitivity and bias, parametric vs. non-parametric approach

As discussed in Chapter 2, *Research approach and methodology*, there is a lot of disagreement in the SDT literature about the appropriateness of either parametric or non-parametric measures. As in Chapter 3, we decided to use both measures alongside each other again to check how they might affect

analysing our hypotheses. Mathematical formulae can be found in Chapter 2 as well. Figures 4.1a and 4.1b show comparisons of the parametric and non-parametric means and standard deviations of sensitivity, and 4.2a and 4.2b show comparisons of bias measures in GGood and GBad.

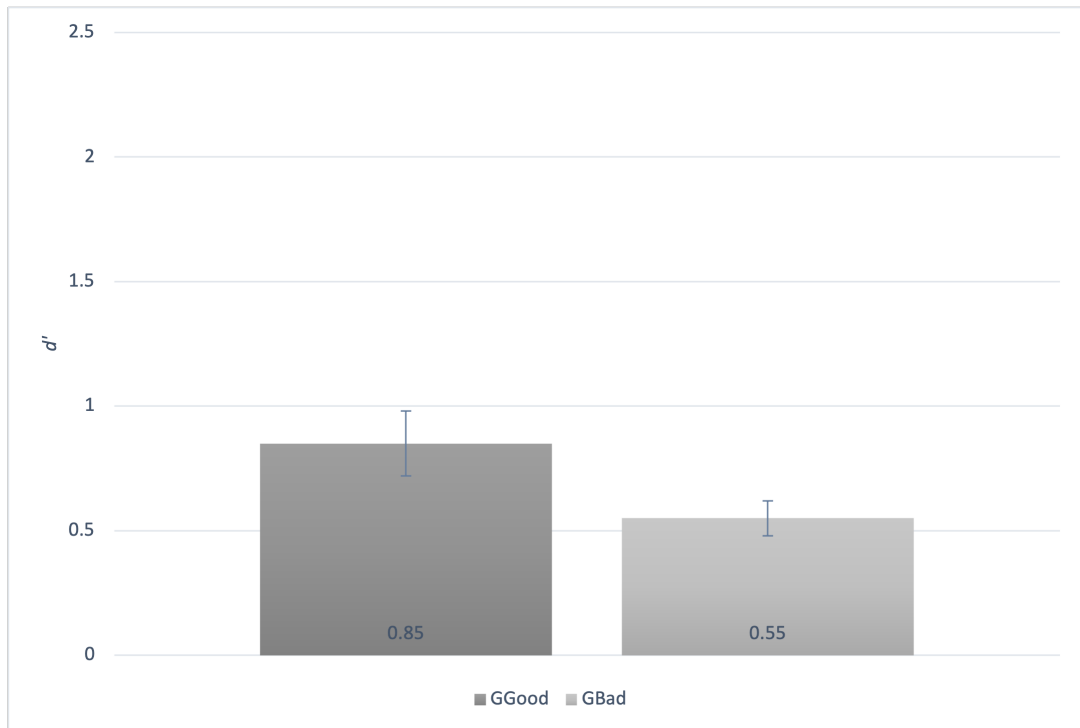


Figure 4.1a – S2 parametric sensitivity (d'), mean and standard error per group

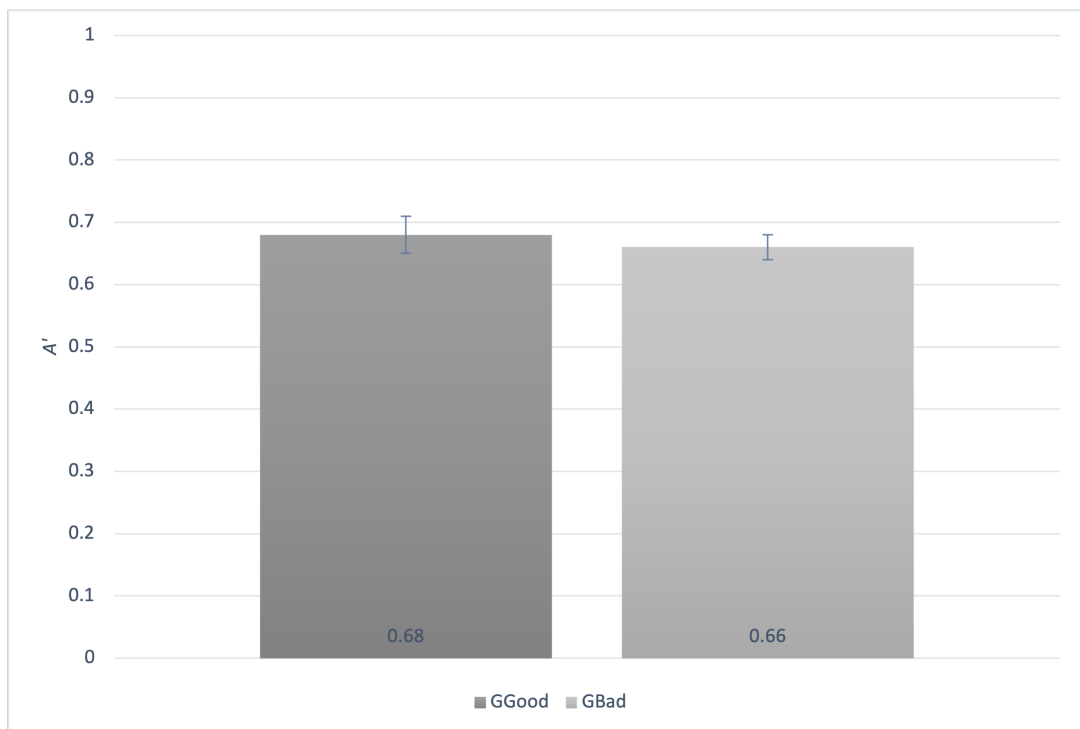


Figure 4.1b – S2 non-parametric sensitivity (A'), mean and standard error per group

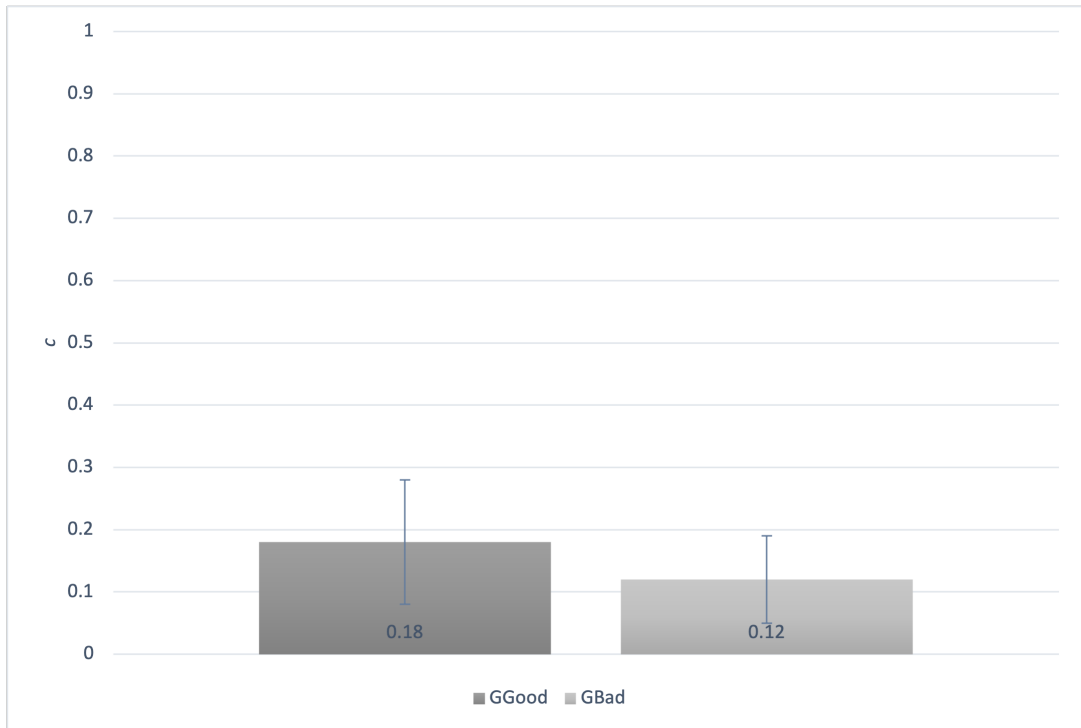


Figure 4.2a – S2 parametric bias (c), mean and standard error per group

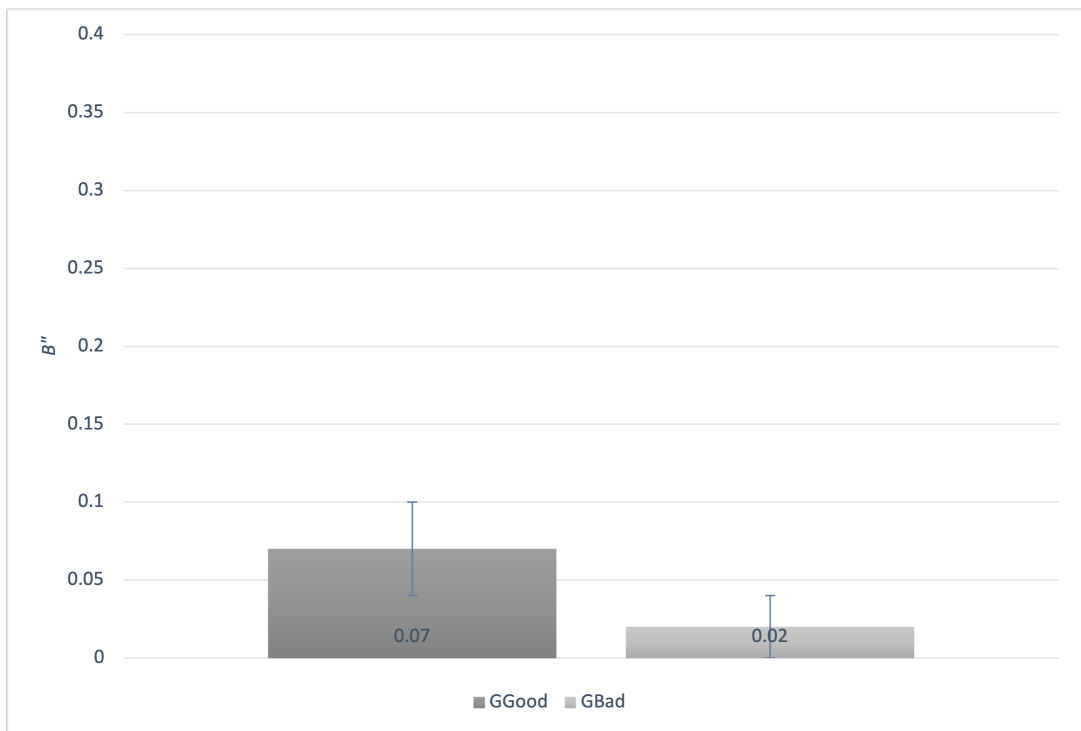


Figure 4.2b – S2 non-parametric bias (B''), mean and standard error per group

4.4.3.1.1 Testing sensitivity and bias with parametric measures

As discussed in Chapter 2, typical d' -scores are values up to 2, with positive scores meaning participants are sensitive to telling signal from noise, and $d' =$

0 meaning participants cannot discriminate between them. The higher the d' -score, the better calibrated participants are in choosing the correct answer. Because d' is not an intuitive metric, it is difficult to tell immediately how well the participants in GGood and GBad did, and the same is true for c . The higher the d' -score, the better calibrated participants are in choosing the correct answer (i.e., either the Original sentence or the Stylus alternative). The higher their c -score, the more they tend to choose the Stylus sentence, independent of its correctness. The average d' of 0.85 ($SD = 0.98$) for GGood and $M = 0.55$ ($SD = 0.53$) for GBad shows both groups have a reasonable ability to answer correctly. An independent samples t -test of the groups' d' -scores shows that GGood, the group that encountered the better performing version of Stylus, was statistically significantly better at responding correctly than GBad, $t(118) = 2.043$, $p = .043$, $d = 0.373$, which suggests that Stylus' performance affects participants' sensitivity. However, it must be noted that there was no equal variance between the groups and Levene's test was statistically significant ($p < .05$), hence we treat the resulting statistic with caution.

GGood's average c -score was 0.18 ($SD = 0.76$), and for GBad it was 0.12 ($SD = 0.56$). An independent samples t -test showed a statistically significant difference between the groups, which confirmed our prediction that participants in GBad notice Stylus' poor performance and therefore feel less inclined to accept its advice, $t(118) = 2.52$, $p = .013$, $d = 0.460$.

The bias score, c , has a zero point, at which a participant shows no overall preference for choosing Stylus suggestions (independently of their correctness). In both groups the mean of c was greater than zero. One-sample t -tests in both groups compared c with zero and showed no statistically significant effect, GGood, $t(58) = 1.83$, $p = .072$, $d = 0.238$; GBad, $t(60) = -1.75$, $p = .085$, $d = -0.224$.

4.4.3.1.2 Testing sensitivity and bias with non-parametric measures

The non-parametric equivalent sensitivity measure A' and bias measure B'' are easier to interpret, although they still need explanation. The closer A' is to 1, the better participants' performance (choosing the correct sentence, either Original or Stylus). Positive B'' indicates a bias towards Stylus, negative B'' a bias towards Original. In S2 a statistically significantly different B'' -score between the groups, with GBad scoring lower than GGood, would indicate that participants in GGood notice the fact that correct Stylus suggestions are more prevalent than Original ones, and vice versa in GBad.

GGood's average A' -score was 0.68 ($SD = 0.20$), that of GBad was $M = 0.66$ ($SD = 0.14$). An independent samples t -test showed no statistically significant difference between the groups, $t(118) = 0.84$, $p = .402$, $d = 0.154$. GGood's average B'' -score was 0.07 ($SD = 0.27$), and GBad scored $M = 0.02$ ($SD = 0.13$). An independent samples t -test showed a statistically significant difference between the groups, which confirmed our prediction that GBad participants notice Stylus' poor performance, $t(118) = 2.33$, $p = .021$, $d = 0.426$.

4.4.3.1.3 Testing the role pre-task trust and self-efficacy

The bias scores also allow us to test the main hypotheses concerning the role of prior trust and self-efficacy in determining the propensity to accept advice (S2-H3 and S2-H4).

S2-H3 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

The acceptance of Stylus recommendations is the degree of bias participants displayed. We tested S2-H3 with both the parametric bias variable c and the non-parametric variable B'' to understand how they would compare. In GGood there was a statistically significant correlation between prior trust in writing style checkers and c , $r(57) = -.27$, $p = .042$. This suggests that the higher GGood participants' level of trust in similar systems is, the more likely they are to accept the aid's advice, and vice versa, as observed earlier by others in other domains (Lee and Moray 1994, Moray et al. 1994, and Wiczorek and Meyer 2019). However, there was no statistically significant correlation between prior trust and B'' , $r(57) = .25$, $p = .061$ in GGood. Likewise in GBad, there was a statistically significant correlation between prior trust in writing style checkers and c in GBad, $r(59) = .26$, $p = .042$, but again there was no statistically significant correlation between prior trust and B'' , $r(59) = -.20$, $p = .124$.

Keeping in mind the issues around the application of SDT for a cognitive task and the associated low number of data points as discussed in Chapter 2, we believe these tests provide some confirmation of the predicted pattern in our data, although they do not irrefutably confirm S2-H3.

S2-H4 Participants' perceived self-efficacy in the domain of writing will be negatively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

In GGood there was no statistically significant correlation between perceived self-efficacy and bias toward accepting Stylus suggestions, c , $r(57) = -.04$, $p = .771$, or B'' , $r(57) = -.10$, $p = .458$. There was no statistically significant correlation either between perceived self-efficacy and c in GBad, $r(59) = -.04$, $p = .776$, and not between perceived self-efficacy and B'' in this group, $r(59) = -.09$, $p = .508$. Thus, S2-H4 is rejected.

Additionally, the sensitivity scores allow us to check whether participants' prior perceived self-efficacy, which although as already shown is likely overestimated, predict their level of performance. In GGood, there was no statistically significant correlation between participants' prior perceived self-efficacy and their sensitivity d' , $r(57) = .14$, $p = .309$, and between their prior perceived self-efficacy and A' , $r(57) = .11$, $p = .429$. In GBad there also was no statistically significant correlation between prior perceived self-efficacy and d' , $r(59) = -.01$, $p = .917$, nor between prior perceived self-efficacy and A' , $r(59) = -.05$, $p = .714$.

4.4.4 Confidence analysis

4.4.4.1 Confidence during the task

The average self-reported confidence across the task was 89.45 ($SD = 7.15$) for GGood, and $M = 86.70$ ($SD = 8.11$) for GBad. An independent samples t -test showed a statistically significant difference in average reported confidence between the groups, $t(118) = 5.90$, $p < .001$, $d = 1.08$, which suggests that using a more reliable aid increases participants' confidence.

S2-H5 Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.

Across participants, these average confidences can be compared with percentage correct to test the standard overconfidence finding for trial-by-trial confidence measures. A paired samples t -test showed a statistically significant difference between confidence and percentage correct in GGood, $t(58) = 10.54$, $p < .001$, $d = 1.372$, as well as in GBad, $t(60) = 16.18$, $p < .001$, $d = 2.072$, which clearly demonstrates that confidence is overall higher than warranted by performance, and confirms S2-H5 in line with earlier findings from the literature (e.g., Gigerenzer, Hoffrage, and Kleinbölting 1991) and the result of S1-H5.

4.4.4.1.1 Confidence ratings for H, M, FA and CR and applied corrections

In GGood, 2 participants had no Ms, hence no M confidence ratings, and 2 participants had no FAs, hence no FA confidence ratings. In GBad, 2 participants had no Ms, hence no M confidence ratings. To substitute a missing confidence value, the mean of a participant's own present values was used (see section 2.4.1).

Table 4.2 shows a breakdown the confidence scores for H, M, FA and CR for both groups, which is also shown in graphical form in Figure 4.3 for ease of comparison.

GGood	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 90.16$ ($SD = 7.70$)	(FA) $M = 89.58$ ($SD = 8.18$)
<i>Original sentence selected by participant</i>	(M) $M = 84.87$ ($SD = 10.25$)	(CR) $M = 90.58$ ($SD = 7.73$)
GBad	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 87.39$ ($SD = 8.38$)	(FA) $M = 85.01$ ($SD = 9.50$)
<i>Original sentence selected by participant</i>	(M) $M = 85.87$ ($SD = 10.01$)	(CR) $M = 87.50$ ($SD = 8.80$)

Table 4.2 – S2 mean confidence percentage per category per group

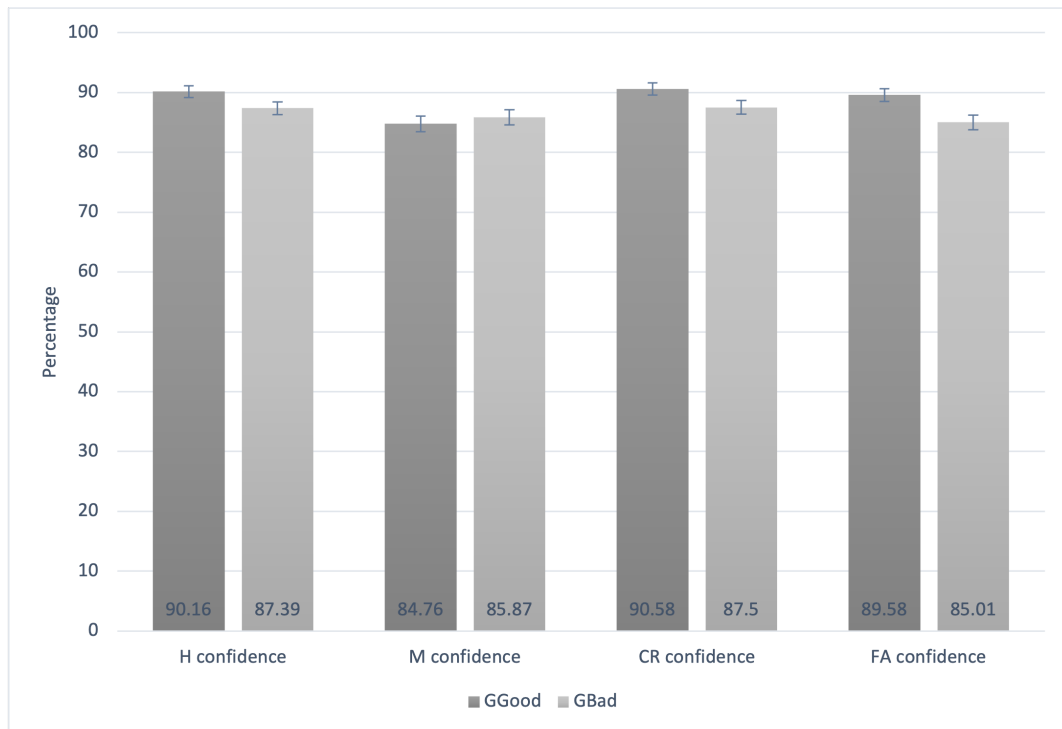


Figure 4.3 – S2 percentage H, M, FA and CR confidence, mean and standard error per group

The data in Table 4.2 were analysed in a 2x2x2 mixed ANOVA, with Group (GGood vs. GBad) as a between-subjects factor, and Correctness of response (Correct v. Incorrect) and Type of response (Stylus vs. Original) as within-subjects factors.

4.4.4.2 Confidence main effects

The ANOVA showed a statistically significant main effect of Correctness of response, in that correct responses ($M = 88.89$, $SD = 7.62$) were assigned more confidence than incorrect responses ($M = 86.29$, $SD = 8.67$), $F(1, 118) = 30.166$, $p < .001$, $partial \eta^2 = 0.204$. There was no statistically significant main effect for Type of response, i.e., Stylus ($M = 88.01$, $SD = 8.03$) vs. Original ($M = 87.17$, $SD = 8.47$), $F(1, 118) = 1.603$, $p = .208$, $partial \eta^2 = 0.026$. There was no statistically significant between-subjects main effect between the groups, $F(1, 118) = 2.711$, $p = .102$, $partial \eta^2 = 0.022$.

4.4.4.3 Confidence interaction effects

The repeated measures ANOVA also revealed three statistically significant two-way interaction effects. The first one is for Type of response x Group, $F(1, 118) = 7.60$, $p = .007$, $partial \eta^2 = 0.06$. If we examine the simple main effects of the Type of response (i.e., Stylus vs. Original) and of the two Groups (i.e., GGGood, the one that encountered the better performing version of Stylus vs. GBad, the group that encountered the version that performed poorly), we note that GGGood was marginally more confident than GBad in Stylus responses (GGGood: $M = 89.87$, $SD = 7.39$; GBad: $M = 86.20$, $SD = 8.28$), and in Original responses as well (GGGood: $M = 87.67$, $SD = 8.00$; GBad: $M = 86.68$, SD

=8.94), which suggests that participants' confidence is at least to some extent meaningfully associated with the aid's reliability. Looking at simple main effects: Type of response had a statistically significant effect on confidence for GGood, $F(1, 58) = 10.70$, $p = .002$, $partial \eta^2 = 0.156$, but not for GBad, $F(1, 60) = 0.47$, $p = .497$, $partial \eta^2 = 0.009$.

The second statistically significant interaction effect was Correctness of response x Type of response, $F(1, 118) = 7.436$, $p = .007$, $partial \eta^2 = 0.059$. We note that overall confidence in correct Stylus responses (H; $M = 88.76$, $SD = 8.14$) is statistically significantly higher than confidence in incorrect Stylus responses (FA; $M = 87.28$, $SD = 9.13$), $F(1, 118) = 6.50$, $p = .012$. Confidence in correct Original responses (CR; $M = 89.02$) is statistically significantly higher than confidence in incorrect Original responses (M; $M = 85.29$), $F(1, 118) = 30.76$, $p < .001$. The significant interaction effect suggests that participants' confidence is positively influenced by the correctness of their response, and more strongly so when rejecting the aid's advice. When accepting the aid's advice, participants are less influenced by the correctness of the aid's advice, perhaps because they are slightly more attuned to their own knowledge in this scenario.

Correctness of response x Type of response x Group, $F(1, 118) = 13.251$, $p < .001$, $partial \eta^2 = 0.101$ was statistically significant too. Three-way interactions are notoriously hard to interpret; in this case the interaction seems to be due to the fact that only in GGood were incorrect responses associated with higher confidence when the response was Stylus than when it was Original. Lastly, the interaction of Correctness of response x Group was not statistically significant, $F(1, 118) = 1.603$, $p = .208$, $partial \eta^2 = 0.013$.

4.4.5 Post-task measures

4.4.5.1 Post-task confidence

4.4.5.1.1 Post-task estimation of number correct and number Stylus responses

GGood's subjective estimation of the number of times they selected the correct answer (either Original or Stylus), $M = 21.05$ ($SD = 5.85$), is higher than their real number of correct responses, $M = 19.49$ ($SD = 4.87$). Across participants, estimated performance did not correlate statistically significantly with objective performance, $r(57) = .04$, $p = .74$. GBad's average estimation was 21.43 ($SD = 5.32$), was again higher than actual performance, $M = 18.31$ ($SD = 2.99$). Again, there was no statistically significant correlation between estimated and actual performance across participants in the group, $r(59) = -.16$, $p = .219$.

We also compared the number of times participants thought they chose the Stylus suggestion over the Original sentence with the objective frequencies. In GGood, the average number of times participants thought they chose the Stylus suggestion was 17.53 ($SD = 5.59$), which is slightly higher than their real number of Stylus choices, 17.19 ($SD = 4.15$). However, across

participants the estimated number of Stylus choices did not correlate statistically significantly with the objective number, $r(57) = .03$, $p = .836$. In GBad, the group that encountered the poorer performing version of Stylus, the average estimated number of Stylus choices was 15.23 ($SD = 5.26$), which is higher than the real number, $M = 13.26$ ($SD = 3.08$). Again, there was no statistically significant correlation between the estimated number of Stylus responses and the real number in this group either, $r(59) = .10$, $p = .457$.

4.4.5.1.2 Comparing average single trial confidence and post-hoc estimates of performance

We compared participants' average confidence for each trial over the whole experiment (which is indeed higher than warranted by their performance) with their average estimated number of correct responses (converted into percentages), which we treat as a post-task frequency confidence measure. A paired samples t -test shows a statistically significant difference between GGood participants' average confidence from each trial over the whole experiment ($M = 89.45$, $SD = 7.15$) and their average estimated percentage of correct responses ($M = 70.17$, $SD = 19.51$), $t(58) = 8.27$, $p < .001$, $d = 1.077$. Similarly, there is a statistically significant difference between GBad participants' average confidence from each trial over the whole experiment ($M = 86.70$, $SD = 8.11$) and their average estimated percentage of correct responses ($M = 71.42$, $SD = 17.73$), $t(60) = 8.41$, $p < .001$, $d = 1.077$. These results are in line with observations from the literature in other domains.

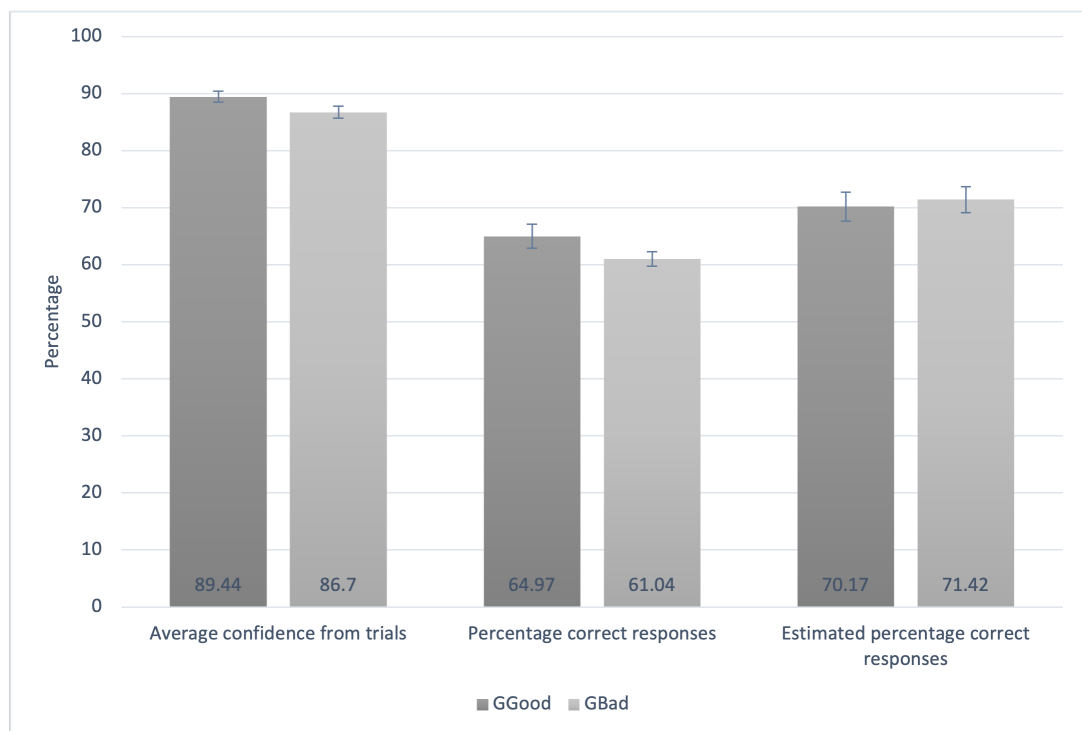


Figure 4.4 – S2 percentage average trial confidence, percentage correct responses, and overall estimated percentage correct responses, mean and standard error per group

S2-H6 The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

To test whether participants' post-task frequency estimations are indeed more realistic than the average confidence reported during the task, we test the difference between percentage average trial confidence minus percentage correct, and percentage correct minus average post-task estimation of percentage correct. The means of trial confidence, percentage correct, and post-task estimation of percentage correct, as shown in Figure 4.4, have been reported earlier.

In GGood, the difference between trial confidence and percentage correct was $M = 24.47$ ($SD = 17.84$), and the difference between percentage correct and post-task estimation of percentage correct $M = -5.20$ ($SD = 24.68$). A paired samples t -test showed a statistically significant difference between the two measures, $t(58) = 8.82$, $p < .001$, $d = 0.757$. In GBad, the difference between trial confidence and percentage correct was $M = 25.66$ ($SD = 12.39$), and the difference between percentage correct and post-task estimation of percentage correct $M = -10.38$ ($SD = 17.85$). A paired samples t -test showed a statistically significant difference between the two measures in this group too, $t(60) = 10.33$, $p < .001$, $d = 1.323$.

The results show that a post-hoc estimation of correct responses is indeed more accurate than the trial confidence, which confirms our expectations based on findings from the literature, and thus S2-H6 is confirmed.

4.4.5.2 Post-task Trust in Stylus

S2-H7 Participants' trust in Stylus during the experiment (measured post-task) will be higher in GGood than in GBad.

As expected, GGood's trust in Stylus, $M = 67.97$ ($SD = 17.07$) is statistically significantly higher than that of GBad, $M = 58.22$ ($SD = 17.20$) as an independent samples t -test showed, $t(118) = 3.10$, $p = .002$, $d = 0.57$. Thus S2-H7 is confirmed, which suggests that users recognise the aid's reliability, and shows the effect of Stylus' reliability on participants' post task trust. These results are in line with the earlier results of S1-H7.

4.4.5.3 Believability of Stylus as an automated system

The question "do you find it plausible that the Stylus suggestions are created by an automated system" served to confirm participants' engagement with the task. It was reasonable to expect that participants in GGood found it more plausible that the "Stylus suggestions" came from an automated system than those in GBad, because performance of the first is much closer to that of a real-world system than that of the latter. A one-sample t -test confirmed this with a statistically significant effect for Stylus' performance on how this plausibility was rated between groups, $t(119) = 39.38$, $p < .001$, $d = 3.595$, although the differences were smaller than expected (GGood: $M = 70.95$ ($SD = 15.66$), GBad: $M = 65.46$ ($SD = 21.47$)).

4.5 Conclusions and discussion

4.5.1 Summary and discussion of key findings

4.5.1.1 Sensitivity and bias

We statistically compared the sensitivity and bias of two socio-demographically equal groups who performed the same task, but were assisted by a purported automated system that performed at a different level for each group.

By analysing bias, we could compare participants' preference for Stylus suggestions independent of their correctness. Although there was no statistically significant preference towards Stylus above the 0-point in either of the groups, there was a statistically significant difference in bias between the groups. This confirms that participants in GGood recognised the fact that correct Stylus responses were more prevalent than Original ones, and vice versa in GBad. Scores were close to 0 in both groups, which suggests, unsurprisingly, that participants' own judgement of sentences was the primary step, with the Stylus suggestion serving as a secondary line of advice. It seems likely that the rather weak performance of Stylus in both conditions contributes to this finding.

The violation of equal variance in some of the data dictates that we treat above test results with caution, and interpret them as suggestions or patterns, rather than as statistically significant results.

4.5.1.2 Effects of trust and perceived self-efficacy

In both groups there was a statistically significant correlation between prior trust in writing style checkers and parametric bias, but not with the non-parametric bias measure. We believe the results provide some confirmation of the predicted pattern in our data. We have no evidence at all that perceived self-efficacy affected bias, this means that only one half of the proposed balance between trust and self-efficacy in the use of automated aids received any support.

4.5.1.3 Confidence

If confidence is well-calibrated, participants' confidence in their correct responses should be higher than their confidence in incorrect responses. We found a statistically significant difference between the groups in participants' average confidence from the trials, with GGood being overall more confident than GBad. We also found that participants assigned more confidence to their correct responses than to incorrect ones. There were three statistically significant confidence interaction effects as well, Type of response x Group (i.e., Stylus vs. Original in each group), Correctness of response x Type of response (i.e., Hits, Misses, Correct Rejections and False Alarms), and Correctness of response x Type of response x Group (i.e., Hits, Misses, Correct Rejections and False Alarms between the groups).

We also found that confidence measured as participants' average confidence from each trial over the whole experiment, was statistically significantly higher than their estimated total number of correct responses, which we treated as a measure of post-test confidence.

4.5.1.4 The above average effect, and the overconfidence effect during and after the task

In both groups, perceived self-efficacy was statistically significantly higher than perceived efficacy of others, which confirms the well-known phenomenon of the above average effect, one important version of overconfidence.

Another overconfidence measure is participants' average confidence in their responses in each trial, which can be compared with post-task estimation of the number of correct responses. In both groups, there was a statistically significant difference between average trial confidence and post-task confidence. The results from our tests show that a post-hoc estimation of correct responses is indeed more accurate than the trial confidence, which confirms our expectations based on findings from the literature

4.5.1.5 Trust and the plausibility of Stylus as an automated system

We assumed that participants who received suggestions from a better performing system would develop more trust in the system than participants who worked with a system that performed worse. We found that GGood's trust in Stylus was indeed statistically significantly higher than GBad's.

We also expected that the level of believability participants assigned to Stylus would indicate their level of engagement with the task. We expected GGood to find Stylus more believable as an automated system than GBad, because the first group encountered a system that performed on a level more in keeping with that of real-world systems than the latter. We found a statistically significant effect for Stylus' performance on participants' believability rating, with GGood finding it more believable that the Stylus suggestions were created by an automated system than GBad. Despite the statistically significant difference between the groups, the difference between the groups was smaller than we initially expected. This could perhaps be explained because participants were fully aware of the experimental setting and the fact that the task was fabricated for research purposes, as opposed to the suggestions coming from a real-world system.

4.5.2 Discussion of method

4.5.2.1 Trial order and randomisation

To avoid question order bias, for example participants quitting after several trials due to a series of difficult trials at the beginning of the experiment, the study was organised in five randomly ordered blocks of six internally randomised trials of different levels of complexity and difficulty. Although we had access to the randomisation order data per participant, we have only

casually studied this in relation to their results and we found no indication further analysis might reveal any effect on performance or confidence. Hence, in the following studies the randomisation is simplified, as explained in the relevant chapters.

4.5.2.2 Item difficulty potentially affecting performance and signal/noise bias

In S2, sentences were presumably at different levels of difficulty, and item difficulty might have been a confounding factor for some of the analyses of confidence. We believe our main between-groups comparisons, and our correlational analyses of trust and perceived self-efficacy cannot be affected by item-effects. However, in all following studies items are systematically rotated across trial-types, as explained in the relevant chapters.

4.5.3 Implications for the design of S3

4.5.3.1 Task design and interface

Following the suggestion that the *reliability* of Stylus suggestions (percentage correct Stylus suggestions, or signal frequency) made a statistically significant difference to participants' performance (most interestingly, their bias toward Stylus), our aim with S3 was to investigate how the *strength* of the advice given might improve the calibration of participants' subjective judgements. To test this, a statistically accurate representation of Stylus' own estimation of the likeliness of its suggestions being correct was added to the test interface. The cover story was updated to better reflect the fact that we were testing an imaginary system, rather than a real one. The same sentence pairs were used as in S2, but different versions of the survey were created to make sure item difficulty would not affect any aspect of participants' performance in terms of the effects we wished to test. Differences between the Original sentence and the Stylus suggestion were highlighted in the test interface so participants would not have to guess the locale or basis of Stylus' judgements.

4.5.3.2 Pre and post-task design

To reduce the completion time, which should make it more attractive for prospective participants to take part and improve the quality of the data because there is less distraction around the task, the sociodemographic questions we used in S1 and S2 at the start of the survey were removed in S3. Data from participants' Prolific profiles was used instead after we verified that there was a near-perfect overlap in S2. The pre and post-task social comparison questions were also removed because we concluded after S2 we would not use them in our analyses. The pre-task grammar and spelling efficacy and trust in systems questions were simplified, and the post-trust questions were restructured as well.

4.5.3.3 Trial randomisation

From S3 onward, trials are no longer grouped in internally randomised blocks, but simply randomised through the task. Items, i.e., sentence pairs, have been

systematically rotated through the trials, see Appendix B5 for S3 item randomisation, B6 for S4, and B7 for S5.

4.5.3.4 Parametric and non-parametric analysis of sensitivity and bias

The lack of equal distribution in our data limited the usefulness of some of our parametric analyses, which are supposed to guarantee more robust results than the non-parametric methods we used in parallel. We will carefully monitor potential violations of the equal variance assumption in our S3 data and base the selection of the most appropriate tests on our findings.

4.5.3.5 Gender balance

S2 was run with a 50/50 male/ female participant ratio. Because we observed no statistically significant effects for gender, this approach was abandoned for the following studies.

Chapter 5 – *Stylus 3: Introducing an aid that communicates its uncertainty*

5.1 Introduction

The results of our Stylus 2 study suggested that participants' prior trust in writing checkers may have a small effect on their bias towards choosing Stylus suggestions over Original sentences, but not necessarily on their sensitivity, nor on the confidence they have in their own judgements. We did not observe any effect of prior perceived self-efficacy on their performance or bias.

Even though the *reliability* of Stylus' suggestions (percentage correct Stylus suggestions, or signal frequency) was not directly communicated and Stylus was only ever partially reliable, Stylus' reliability made a difference to how participants performed in S2. The next step was to investigate how representations of a system's own "confidence" judgement, which would affect the *strength* of the advice given, might improve the calibration of participants' subjective judgements, so as to enhance their (metacognitive) performance in a series of decision-making tasks. As discussed in Chapter 1, *Subject area background and literature review*, it is plausible for an automated "learning" system like Stylus to have information about its own reliability, or its confidence in any particular guess. By communicating this information from the automation to human users, they might be able to make better decisions, and be more confident in their responses.

The general aim of Stylus 3 (S3), the experimental study we discuss in this chapter, is to test the suggestion that participants will be able to interpret and make use of Stylus suggestions and Stylus' own estimation of the likelihood of its suggestions being correct. We do this by testing our suggestion that an interactive system like Stylus making an uncertain recommendation, should effectively communicate its level of uncertainty. If this communicated uncertainty is valid, and can be understood, then participants should use it to adjust their own decisions and the confidence they have in their decisions: performance (number of correct responses) should improve, and confidence should become better calibrated (i.e., close to a level warranted by their performance).

5.1.1 Main differences in pre and post-task design between S1/ S2 and S3

This paragraph highlights the main changes we made in the design of the S3 pre and post-task survey, in comparison with S1/S2. Firstly, we removed the sociodemographic questions at the start of the survey to reduce the completion time. When comparing participants' submitted sociodemographic data in S1 and S2 with the sociodemographic information in their Prolific profiles, we found the overlap to be nearly perfect, so in this study we rely only on the latter. We also removed the pre-task social comparison questions, because the first two experiments so clearly support our hypothesis that the *above average effect* generalises to this domain, and because they are not central enough to our concerns to require further replication. The pre-task grammar and spelling efficacy questions were simplified, and so were the pre-task trust in systems questions. We streamlined the post-trust questions as well.

5.1.2 Experimental design comparison between S1/S2 and S3

Instead of the 'Wizard of Oz' cover story of S1 and S2, in S3 we introduced the Stylus sentences with the phrase 'imagine that [...]', with no hint at all that the Stylus suggestions had been created by a real automated system.

The sentence pairs we used in the S3 trials were the same as in S2 because overall performance level in that study was satisfactory to believe the sentences had a difficulty level that was in line with Stylus' performance (roughly 2/3 correct) and avoided potential ceiling effects in performance and confidence. To rule out item difficulty as having an influence on differential performance and confidence in the different cells of our various analyses, eight different versions of the experiment were created, to allow sentence pairs to be systematically rotated through the four conditions (see Appendix B5).

Three extra sentence pairs (one of which was treated as a dummy trial) were added because, as explained in the following section, S3 required 33 trials, instead of thirty in S1 and S2. Furthermore, four new sentence pairs were added as practise trials at the start of the survey to better ensure participants understood the task, before starting the experiment proper.

Like the response scales participants used to rate their confidence, Stylus indicates its own estimation of the likelihood of its suggestions being correct on a 50% – 100% scale. Less than 50% would mean the system estimates the Original sentence to be better than its own suggestion, 50% was not used in this experiment because it comprises a zero-information condition, and 100% was excluded from consideration because a 100% likelihood estimation is not realistic. If likelihood ratings between 51% and 99% were used on a linear scale in this study, varying trial-by-trial, arbitrary cut-offs (e.g. between 59% and 60%, or between 60% and 61%) between conditions would have been necessary to effectively analyse the results with a limited number of data points. To end up with enough data points to be able to meaningfully distinguish between conditions, S3 only works with likelihood ratings of 53% and 75%. The lower likelihood of 53% is so close to chance performance that

it renders Stylus advice almost useless. Thus, this condition serves as a proxy for unaided performance, with a task environment completely comparable with the other condition of 75% reliability. Although 75% is still somewhat low for a would-be useful automated advice system, it is above the reliability threshold (Wickens and Dixon 2007).

5.1.3 Statistical validity of conditions and need for one dummy trial

The experiment was designed so that Stylus' communicated likelihood estimation was completely accurate. The 32 trials each fall into one of the following four statistically accurate conditions: Original sentence is correct, Stylus' likelihood estimation is 53% (O53); the Stylus suggestion is correct, Stylus' likelihood estimation is 53% (S53); Original sentence is correct, Stylus' likelihood estimation is 75% (O75); the Stylus suggestion is correct, Stylus' likelihood estimation is 75% (S75) (See Table 5.1).

To avoid having a 50% (zero-information) condition, a 33rd trial was added, so S50 and O50 become S53 and O53; this trial was deliberately easy so (almost) all participants would respond correctly, and the responses to this single item were not used in further analyses, except for the analysis of post-task number of correct and number of Stylus response frequency estimations.

5.1.4 Test interface improvements

The S3 test interface, an example of which is shown in Figure 5.1, was improved by highlighting the differences between Original and Stylus sentences, so it was immediately clear for participants where the potential error in the two sentences was.

Which is better?

Original sentence: One of the best loved bands of the 1980s ~~were~~ The Smiths; would they have been able to carry over their success into the '90s had they not split up?

Stylus suggestion (53%) : One of the best loved bands of the 1980s ~~was~~ The Smiths; would they have been able to carry over their success into the '90s had they not split up?

How confident are you in your answer?

I guessed 50 55 60 65 70 75 80 85 90 95 100 I'm certain I got it right

Next

Figure 5.1 – S3 trial example screenshot

5.2 Method

5.2.1 Task design

S3 used a within-subjects design with 128 participants. Each participant was presented with 4 practise trials and 33 test trials, each made up of two alternative sentences presented one below the other. Of the first alternatives (labelled "Original sentence" in each trial) they were told to imagine they had just typed them into a word processor, of the second that they were to imagine they were 'suggested improvements from a writing aid called Stylus'. Each Stylus suggestion was accompanied by a statistically accurate percentage that represented Stylus' estimation of the likelihood of its suggestion being correct. This percentage was represented as a green rounded rectangle with a percentage. In each trial the potential errors participants should focus on, were highlighted in yellow in both the Original sentence and in the Stylus sentence. Participants were asked to indicate which sentence was better, the Original sentence or the Stylus suggestion, and how confident they were in their response.

Eight versions of the experiment were produced, with sentence pairs rotated through conditions so that across participants each item (sentence pair) appeared equally often in each of the four conditions, Stylus correct vs. incorrect X 53 vs. 75. A roughly equal number of participants was assigned to each one of these eight versions of the experiment according to the order in which they participated. The order of the trials was randomised for each participant.

In the overall experiment, across the two conditions of Stylus reliability, Stylus correctly suggested an alternative for 21 sentences with errors ("signal trials"), and incorrectly suggested alternatives for 12 correct original sentences ("noise trials"), which means that Stylus' overall reliability is close to the "reliability threshold" (Wickens and Dixon 2007), which is explained in Chapter 1.

5.2.2 Variables and hypotheses

5.2.2.1 Pre-task measures

The independent variables measured prior to the task were participants' prior trust in automated writing aids, and their perceived self-efficacy as checkers of grammar and spelling.

5.2.2.2 Performance

The independent variable manipulated during the experiment was the level of correctness of the Stylus recommendations, which is represented to participants by a Stylus likelihood estimation percentage (either 53% or 75%).

The dependent variables measured were overall percent correct and proportion of acceptance of Stylus recommendations, as well as sensitivity (ability to distinguish correct sentences independently of Stylus recommendations) and bias (tendency to accept Stylus recommendations independently of their correctness). We also recorded and analysed participants' confidence in their own responses.

5.2.2.3 Hypotheses

Our experimental hypotheses all derive from the overarching hypothesis that participants will be able to interpret and make use of Stylus suggestions and Stylus' own estimation of the likelihood of its suggestions being correct. Hypotheses S3-H1 and H2 test the main Human Factors claim that such systems will be useful. Hypotheses H3–H5 further test the theory that use of advice is moderated by prior trust in the automation and by perceived self-efficacy (a theory that has received only weak support so far). Other hypotheses test the role of confidence and overconfidence in decision making with advice, and additional statistical analyses will be undertaken to investigate the relationship between variables.

S3-H1 Participants' performance, in terms of percentage correct, will be better when Stylus' reliability is 75% than when it is 53%.

This hypothesis concerns whether participants' performance might be affected by the aid's reliability. If confirmed, this suggests that users might benefit from advice from imperfect automated aids (Wickens and Dixon 2007) and that higher aid reliability might positively affect users' performance.

S3-H2 Participants will be more inclined to accept Stylus suggestions when these have higher likelihood ratings, and this will be true independently of correctness.

This hypothesis concerns whether the strength of the aid's advice might affect participants' willingness to follow its suggestions. If confirmed, this suggests that improving the strength of the aid's advice might positively affect users' acceptance of its suggestions.

S3-H3 Participants' prior perceived self-efficacy in the domain of writing will be positively correlated with their sensitivity, independent of Stylus' reliability. This hypothesis concerns whether participants' prior perceived self-efficacy in the domain of writing might affect their ability to correctly follow the aid's advice. If confirmed, this suggests that the higher is participants' level of perceived self-efficacy, the less likely they are to correctly follow the aid's advice, and vice versa, in line with what was observed earlier by among others Lee and Moray 1994, Moray et al. 1994, and Wiczorek and Meyer 2019 in other domains.

S3-H4 Participants' prior perceived self-efficacy in the domain of writing will be negatively correlated with their bias, independent of Stylus' reliability.

This hypothesis concerns whether participants' prior perceived self-efficacy in the domain of writing might affect their tendency to follow the aid's

suggestions, regardless of the correctness of the advice. If confirmed, this suggests that the higher is participants' level of perceived self-efficacy, the less likely they are to follow the aid's advice independent of its correctness, and vice versa, in line with what was observed earlier by among others Lee and Moray 1994, Moray et al. 1994, and Wiczorek and Meyer 2019 in other domains.

S3-H5 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

This hypothesis replicates S1-H3 and S2-H3 with an aid that not only suggests an alternative, but also communicates the strength of its advice.

S3-H6 Participants will be more confident when using more reliable Stylus suggestions.

This hypothesis concerns whether the aid's reliability affects participants' confidence. If confirmed, this suggests that a more reliable aid positively affects participants' confidence.

S3-H7 Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.

This hypothesis replicates S1-H5 with an aid that not only suggests an alternative, but also communicates the strength of its advice.

S3-H8 Participants will be more confident when responding correctly.

This hypothesis concerns whether participants are aware of their own performance, even though they are not given any performance feedback during the task. If confirmed, this suggests that participants have an awareness of their own performance, despite the absence of feedback, and that their level of performance might positively affect their confidence.

S3-H9 The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

This hypothesis replicates S1-H6 with an aid that not only suggests an alternative, but also communicates the strength of its advice.

5.2.3 Participants

128 participants were recruited on Prolific.ac, the screeners used were current country of residence (registered as United Kingdom residents), country of birth (registered as born in the UK) first language (English) and participation in previous studies (S1 and S2 participants were excluded). Of the participants 47 were male, 81 female and ages ranged from 18 to 72 ($M = 37.11$; $SD = 12.07$).

5.2.4 Materials

The same sentences were used as in S2, with several additions as explained under *Experimental design comparison between S1/S2 and S3* in the introduction to this chapter.

5.2.4.1 Pre-task perceived self-efficacy

We measured participants' pre-task perceived self-efficacy with the questions 'When thinking of how good I am at English grammar, I would class myself as [0; Not very good at all] – [100; Very good]' and 'When thinking of how good I am at English spelling, I would class myself as [0; Not very good at all] – [100; Very good]'. After the internal reliability of the results was checked (Cronbach α), the average was used to define participants' level of prior perceived linguistic self-efficacy.

5.2.4.2 Pre-task trust, confidence, post-task estimations of frequency

All these variables were tested the same way as in S1/2.

5.2.4.3 Post-task trust in Stylus

Participants' retrospective trust in Stylus' suggestions during the task was measured with the four questions 'When thinking of the usefulness of Stylus' suggestions during this experiment, I would class them as [0; Not very useful at all] – [100; Very useful]', 'When thinking of the trustworthiness of Stylus' performance during this experiment, I would class it as [0; Not very trustworthy at all] – [100; Very trustworthy]', 'When thinking of the consistency of Stylus' performance during this experiment, I would class it as [0; Not very consistent at all] – [100; Very consistent]' and 'When thinking of Stylus' performance in general during this experiment, I would class it as [0; Not very good at all] – [100; Very good]'. The results were tested for internal reliability (Cronbach α) and averaged as a single measure of post-task trust in Stylus.

5.2.4.4 Perceived plausibility of Stylus suggestions being created by an automated system as evidence of engagement with the task

This variable was tested the same way as in S1/2

5.2.5 Procedure

The procedure of S3 was largely the same as that of S1/S2, with the following exceptions.

Participants were paid an average of £1.75 (based on £7/hour) on completion of the survey, the estimated time for completing the survey was 15 minutes (automatically estimated by Qualtrics), and the actual average completion time was 16.62 minutes ($SD = 7.41$).

5.2.5.1 Experimental task

The task consisted of four practise trials, followed by 33 experimental trials, presented one after another. The experimental trials were presented in random order after the practise trials. There were eight different versions of the experiment that were each assigned to an approximately equal number of participants. All versions used the same sentence pairs in the trials, but signal and noise trials were systematically formatted in different ways in each trial in each of the versions, as we explain below. The practise trials were identical for all versions. In S3, Stylus indicates its own estimation of the likelihood of its suggestions being correct with a 53% or a 75% label.

	Category	Number total	Number Original correct	Number Stylus correct
	53%	17 (<i>C53</i>)	8 (<i>O53</i>)	9 (<i>S53</i>)
	75%	16 (<i>C75</i>)	4 (<i>O75</i>)	12 (<i>S75</i>)
TOTAL		33	12	21
<i>Percentage</i>			36.36	64.64

Table 5.1 – S3 Stylus likelihood estimation distribution

5.2.5.1.1 Counterbalancing items across conditions

To ensure that individual item difficulty could not affect comparisons between experimental conditions, sentence pairs in S3 were systematically arranged in eight different versions of the experiment (see distribution in Appendix B5) so that the different sentence-pairs were used equally in the four conditions of the experiment (53 vs. 75 x Stylus correct vs. Original correct).

The rest of the procedure is identical to that of S1/S2.

5.3 Analysis strategy

The analysis strategy for S3 was largely the same as that for S1/S2, with the following exceptions.

5.3.1 Acceptance and rejection of data

The results of 1 participant were removed and a replacement was sought, because their percentage correct responses was 2 standard deviations below the mean, and they completed the survey unrealistically fast at 5.15 minutes.

Two participants “returned” the study; partial data of uncompleted tasks was not stored by Qualtrics, hence it has not been used in our analyses.

5.3.1.1 Dummy trial

The data from the dummy trial that was needed to arrive at an accurate 53% distribution, was removed from the results. Since this was a deliberately easy trial, all participants responded correctly, so this does not affect any of our analyses. For the comparisons between participants' estimated number of correct responses and their estimated number of Stylus responses with the actual frequencies, the data from the dummy trial was used.

5.3.1.2 Parametric and non-parametric sensitivity and bias measures

As discussed in Chapter 2, and in Chapter 4 discussing S2, there is a lot of disagreement in the SDT literature about the appropriateness of either parametric or non-parametric methods. In S1 and S2 both parametric measures d' and c , and their non-parametric equivalents A' and B'' were reported and used in analyses. Because the latter yielded no different insights, only parametric measures, the least controversial of the two (Macmillan and Creelman 2005), are reported from now on. For completeness, A' and B'' analyses can be found in Appendix C5.

5.3.4 Analysing confidence

Participants' confidence ratings were analysed using a 2x2x2 repeated measures ANOVA. The within-subjects variables that were tested were Correctness of the response (i.e., "Correct" vs. "Incorrect"), Type of response (i.e., "Stylus" vs. "Original"), and Strength of recommendation (i.e., 53% vs. 75%). Participants' average confidence in each cell of the design was entered into this ANOVA.

5.4 Results

The most important S3 data can be found in tabular form in Appendix A5, including breakdowns of aggregated variables. A table of all hypotheses from this thesis can be found in Appendix D.

5.4.1 Pre-task measures

5.4.1.1 Reliability testing

Reliability of prior perceived self-efficacy was $\alpha = .74$, that of prior trust $\alpha = .81$, and of the post-task variable *post-trust* it was $\alpha = .93$.

5.4.1.2 Prior perceived self-efficacy and Prior trust

Participants reported a prior perceived self-efficacy of $M = 75.00$ ($SD = 14.64$) and a rating of prior trust in automated suggestions of $M = 75.27$ ($SD = 15.64$).

5.4.2 Performance

Table 5.2 reports performance in absolute numbers in all four categories (H, M, FA, CR) in both conditions (C53 and C75), as this serves as the basis for our further analyses.

C53	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 6.03$ ($SD = 4.12$)	(FA) $M = 2.78$ ($SD = 1.86$)
<i>Original sentence selected by participant</i>	(M) $M = 1.97$ ($SD = 4.12$)	(CR) $M = 5.22$ ($SD = 1.86$)

C75	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 9.23$ ($SD = 1.82$)	(FA) $M = 1.48$ ($SD = 1.10$)
<i>Original sentence selected by participant</i>	(M) $M = 2.77$ ($SD = 1.82$)	(CR) $M = 2.52$ ($SD = 1.10$)

Table 5.2 – S3 mean number of responses per category per condition

5.4.2.1 Proportion of correct responses between conditions

Percentage correct ($(N_H + N_{CR}) / N_{Total} * 100$) overall was $M = 71.90$ ($SD = 12.56$).

S3-H1 Participants' performance, in terms of percentage correct, will be better when Stylus' reliability is 75% than when it is 53%.

Average percentage correct in C53 was 70.31 ($SD = 14.82$), in C75 it was $M = 73.49$ ($SD = 14.39$). A paired samples t -test showed participants' performance was statistically significantly better in the condition in which they received more reliable Stylus suggestions, $t(127) = -2.41$, $p = .017$, $d = 0.213$. Thus, S3-H1 was confirmed, which suggests that users benefited from Stylus' advice in line with similar findings in other domains in the literature (Wickens and Dixon 2007), and that higher aid reliability positively affected users' performance.

5.4.2.2 Testing sensitivity and bias

5.4.2.2.1 Testing sensitivity

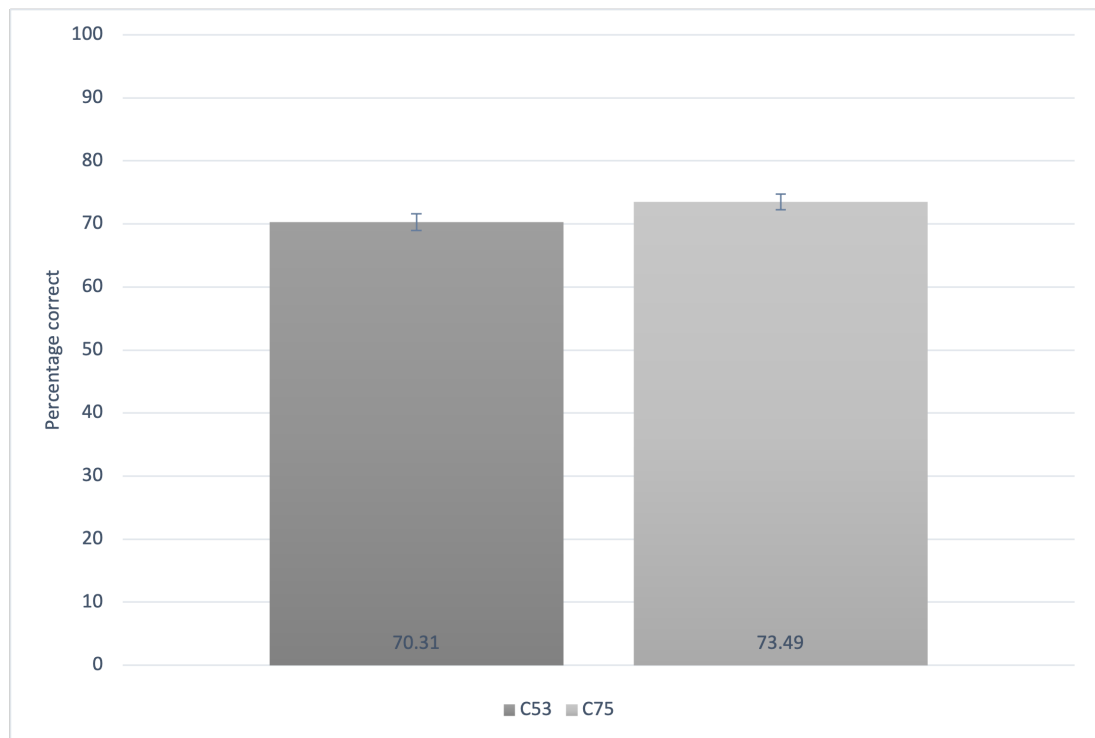


Figure 5.2 – S3 parametric sensitivity (d'), mean and standard error per condition

Figure 5.2 shows the of the sensitivity metric d' in C53 and C75. As discussed in Chapter 2, typical d' -scores are values up to 2, with positive scores meaning participants are sensitive to telling signal from noise, and $d' = 0$ meaning participants cannot discriminate between them. The higher the d' -score, the better calibrated participants are at choosing the correct answer (i.e., either the Original sentence or the Stylus alternative), independently of their propensity to accept Stylus advice. There is not expected to be any effect of condition on this measure. The average d' of 1.19 ($SD = 0.90$) in C53 and $M = 1.15$ ($SD = 0.91$) in C75 show participants have a reasonable ability to respond correctly in either condition. A paired samples t -test did not show a statistically significant difference in d' between the conditions, $t(127) = -0.58$, $p = .566$, $d = -0.05$. The seeming discrepancy, between the statistically significant difference in performance measured as "percentage correct" between C53 and C75, and the lack of a statistically significant difference in d' between the conditions, can be explained by the fact that d' corrects for the effect of bias on correctness. Guessing that Stylus is correct will yield a high percentage correct in C75 because participants will be biased towards Stylus, but this will not be the case in C53. Correcting for bias makes the difference disappear.

5.4.2.2.2 Testing bias

Participants' level of bias shows to what extent they lean towards choosing Stylus or choosing Original, independent of correctness of the response. For our bias analyses we followed the same steps as above for sensitivity. The higher their *c*-score, the more participants tend to choose the Stylus sentence, independent of its correctness. A negative score indicates they lean towards choosing Original.

In terms of our measures, there should be a statistically significant difference between the two conditions in both proportion of Stylus responses, and in bias. The average proportion of Stylus responses in C53 was 55.08 ($SD = 11.33$), and it was 66.94 ($SD = 12.05$) in C75. A paired samples *t*-test showed a statistically significant difference in the average percentage of Stylus responses between the conditions, $t(127) = -8.30$, $p < .001$, $d = -0.734$. However, this difference will be influenced by the fact that there were more correct Stylus responses in C75. To test whether participants' bias toward Stylus, independent of correctness, is different in the two conditions, we turn to the SDT measure.

S3-H2 Participants will be more inclined to accept Stylus suggestions when these have higher likelihood ratings, and this will be true independently of correctness.

The average *c* of 0.31 ($SD = 0.67$) in C53 (one-sample *t*-test showed this to be statistically significantly above 0, $t(127) = 5.21$, $p < 0.001$, $d' = 0.460$), and $M = 0.50$ ($SD = 0.77$) in C75 (statistically significant above 0 as well, $t(127) = 7.38$, $p < 0.001$, $d' = 0.652$), show that participants lean, on average, towards choosing Stylus in both conditions. A paired samples *t*-test revealed a statistically significant difference in *c* between the conditions, $t(127) = -2.16$, $p = .033$, $d = -0.191$, which demonstrated that the stronger Stylus' advice is, the more likely participants are to follow it. Thus, S3-H2 was confirmed, which suggests that the strength of the Stylus' advice positively affected participants' acceptance of its suggestions.

5.4.2.2.3 Testing the role of pre-task trust and self-efficacy

S3-H3 Participants' prior perceived self-efficacy in the domain of writing will be positively correlated with their sensitivity, independent of Stylus' reliability. In C53, there was a statistically significant correlation between participants' prior perceived self-efficacy and their sensitivity d' , $r(126) = .29$, $p = .001$. In C75 however, there was no statistically significant correlation between prior perceived self-efficacy and d' , $r(126) = .09$, $p = .318$. As d' is corrected for any bias toward accepting Stylus' advice, there is no obvious reason why perceived self-efficacy should be a more valid predictor in one condition than in the other. Thus, we must reject S3-H3.

S3-H4 Participants' prior perceived self-efficacy in the domain of writing will be negatively correlated with their bias, independent of Stylus' reliability.

In C53, there was no statistically significant correlation between participants' prior perceived self-efficacy and their bias *c*, $r(126) = -.06$, $p = .532$. Likewise, in C75 we did not find a statistically significant correlation either between prior

perceived self-efficacy and c , $r(126) = -.05$, $p = .588$. This suggests that contrary to what we hypothesised, participants' bias is not statistically significantly affected by their prior perceived self-efficacy, but only, or primarily, by their perception of the system's performance and/ or by the strength of its advice. Thus, we must reject S3-H4.

S3-H5 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

The acceptance of Stylus recommendations is the degree of bias participants displayed. We did not find a statistically significant correlation between prior trust in writing style checkers and bias c in either condition; C53, $r(126) = -.04$, $p = .628$; C75, $r(126) = -.01$, $p = .900$. Thus, S3-H5 is not supported.

5.4.3 Confidence

5.4.3.1 Confidence during the task

S3-H6 Participants will be more confident when using more reliable Stylus suggestions.

The average self-reported confidence across the task was 89.15 ($SD = 6.99$), and it was $M = 88.63$ ($SD = 7.37$) in C53, and $M = 89.67$ ($SD = 7.06$) in C75. A paired samples t -test showed a statistically significant difference in average reported confidence between the conditions, $t(127) = -3.27$, $p = .001$, $d = -2.089$, with participants being more confident in their responses in trials with a higher Stylus likelihood rating. Thus, S3-H6 was confirmed, which suggests that more reliable Stylus performance positively affected participants' confidence.

S3-H7 Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.

It is clear that in both conditions, confidence overestimates the percentage of correct trials. A 2x2 mixed repeated measures ANOVA (C53 vs. C75 x Average trial confidence vs. Percentage correct) revealed a statistically significant effect of Condition, $F(1,127) = 9.61$, $p = .002$, $partial \eta^2 = 0.070$, as well as of Confidence vs. Performance, $F(1,127) = 203.94$, $p < .001$, $partial \eta^2 = 0.616$, and thus S3-H7, and with that the *overconfidence effect*, was confirmed, in line with earlier findings from the literature (e.g., Gigerenzer, Hoffrage, and Kleinbölting 1991) and the results of S1-H5 and S2-H5. There was no statistically significant effect of Condition x Average trial confidence vs. Percentage correct, $F(1,127) = 2.49$, $p = .117$, $partial \eta^2 = 0.019$.

5.4.3.2 Confidence ratings for H, M, FA and CR and applied corrections

In C53, 18 participants had no Ms, hence no M confidence ratings, and seven participants had no FAs, hence no FA confidence ratings. In C75, seven participants had no Ms, hence no M confidence ratings, 26 had no FAs and FA confidence ratings, and seven had no CRs and CR confidence ratings. To substitute a missing confidence value, the mean of a participant's own present values was used (see section 2.4.1).

When we break down the confidence scores for H, M, FA and CR, they are as shown in Table 5.3, and in graphical form in Figure 5.3 for ease of comparison.

C53	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 89.80$ ($SD = 7.43$)	(FA) $M = 86.94$ ($SD = 10.53$)
<i>Original sentence selected by participant</i>	(M) $M = 85.34$ ($SD = 11.90$)	(CR) $M = 89.43$ ($SD = 8.64$)

C75	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 90.90$ ($SD = 6.79$)	(FA) $M = 86.86$ ($SD = 12.12$)
<i>Original sentence selected by participant</i>	(M) $M = 86.44$ ($SD = 10.71$)	(CR) $M = 89.52$ ($SD = 9.78$)

Table 5.3 – S3 mean confidence percentage per category per condition

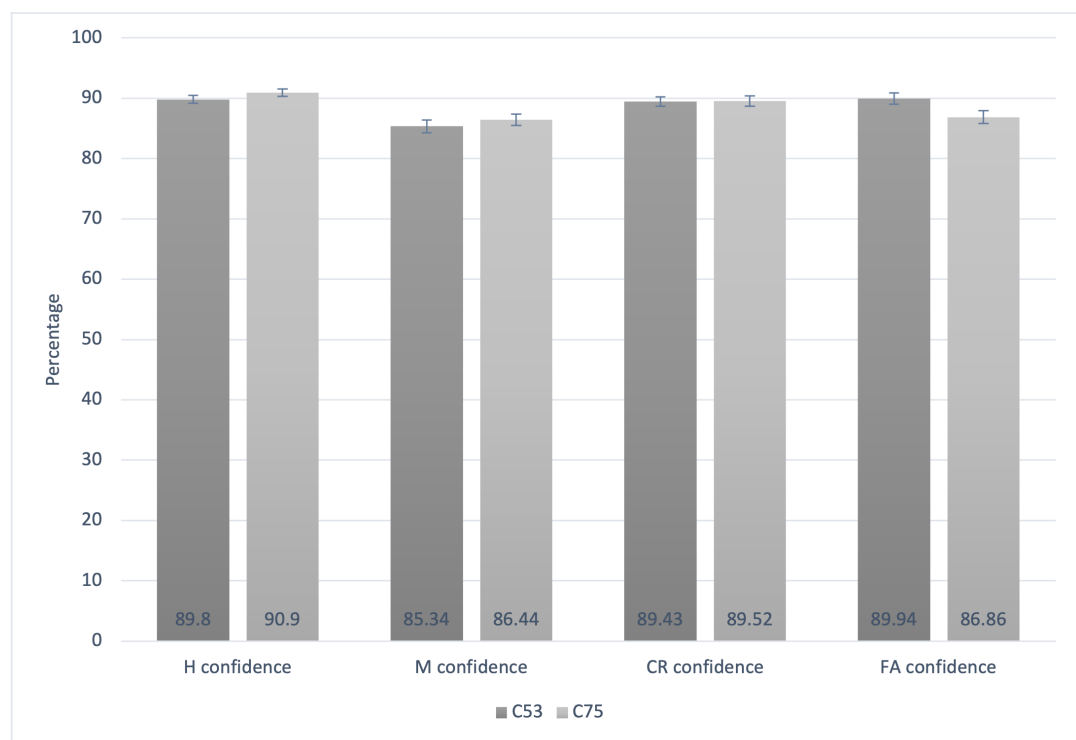


Figure 5.3 – S3 percentage H, M, FA and CR confidence, mean and standard error per condition

Participants' confidence ratings were analysed using a 2x2x2 repeated measures ANOVA. The within-subjects variables that were tested were Correctness of the response (i.e., "Correct" vs. "Incorrect"), Type of response (i.e., "Stylus" vs. "Original"), and Strength of recommendation (i.e., 53% vs.

75%). Participants' average confidence in each cell of the design was entered into this ANOVA.

S3-H8 Participants will be more confident when responding correctly. The ANOVA showed a statistically significant main effect of Correctness of response, in that correct responses (C53: $M = 89.64$ ($SD = 7.36$); C75: $M = 90.63$ ($SD = 6.94$)) were assigned more confidence than incorrect responses (C53: 85.86 ($SD = 10.15$); C75: $M = 86.07$ ($SD = 9.99$)), $F(1, 127) = 48.96$, $p < .001$, $partial \eta^2 = 0.278$. Thus, S3-H8 was supported, which suggests that participants had an awareness of their own performance, even if they were not given any feedback during the task, and that their level of performance might have positively affected their confidence.

There was no statistically significant main effect for Strength of recommendation on participants' confidence, C53 ($M = 88.63$, $SD = 7.37$) vs. C75 ($M = 89.67$, $SD = 7.06$), $F(1, 127) = 1.96$, $p = .165$, $partial \eta^2 = 0.015$. This seems at odds with our earlier finding that participants' average reported confidence in responses in trials with a higher Stylus likelihood rating was statistically significantly higher than that in trials with weaker Stylus advice. However, the earlier result relied on participant-averages that were not separated in terms of correct vs. incorrect trials. It is important to note that this does not mean that the earlier finding of an overall difference in average confidence between conditions is in any sense an artefact of correctness. The participants did not know which items they got correct, and the items in both conditions (across participants) were the same. Rather, it is that the ANOVA obscures the real difference between conditions because conditions are correlated with correctness.

There was no statistically significant main effect for Type of response ((Stylus C53: 89.04 ($SD = 7.42$); C75: $M = 90.52$ ($SD = 6.67$)) vs. Original (C53: 88.31 ($SD = 8.53$); C75: $M = 88.15$ ($SD = 9.10$)) responses, $F(1, 127) = 3.44$, $p = .066$, $partial \eta^2 = 0.026$).

5.4.3.3 Confidence interaction effects

The ANOVA revealed no statistically significant interaction effects for Strength of recommendation x Type of response, $F(1, 127) = 0.008$, $p = .927$, $partial \eta^2 = 0.000$. This seems at odds with S3-H6, but can be explained by the asymmetry in the number of signal and noise trials between the conditions.

The ANOVA revealed no statistically significant interaction effects for Strength of recommendation x Correctness of response either, $F(1, 127) = 0.011$, $p = .918$, $partial \eta^2 = 0.000$, nor for Correctness of response x Type of response, $F(1, 127) = 0.028$, $p = .866$, $partial \eta^2 = 0.000$, or Strength of recommendation x Correctness of response x Type of response, $F(1, 127) = 2.217$, $p = .139$, $partial \eta^2 = 0.017$.

5.4.4 Post-task measures

5.4.4.1 Confidence

5.4.4.1.1 Post-task estimation of number correct and number Stylus responses

Participants' subjective estimation of the number of times they selected the correct answer (either Original or Stylus) was $M = 23.27$ ($SD = 7.46$). There was a statistically significant correlation between the average number of times participants thought they selected the correct answer, and the objective frequency (including dummy trial), $M = 24.01$ ($SD = 4.02$), $r(126) = .23$, $p = .011$. For ease of comparison, the frequencies were converted into percentages. The average difference between the actual percentage correct, $M = 72.75$, $SD = 12.18$, and the estimated percentage correct, $M = 70.53$, $SD = 22.59$, was 2.23% ($SD = 23.13$). A paired samples t -test on the means confirmed there was no statistically significant difference between participants' estimated percentage correct and the actual percentage correct responses $t(127) = -1.09$, $p = .278$, $d = -0.096$. The most optimistic estimate was 36.36% too high (estimated 12 more correct trials than actual number correct), the most pessimistic estimate was 72.73% too low (estimated 24 fewer correct trials than actual number).

We also compared the number of times participants thought they chose the Stylus suggestion over the Original sentence with the objective frequencies. The average estimated number of Stylus responses across participants was 19.80 ($SD = 6.59$), and the actual number (including dummy trial) was $M = 20.52$ ($SD = 2.71$). There was a statistically significant correlation between the average number of times participants thought they chose the Stylus suggestion, and the objective frequency, $r(126) = .82$, $p = .041$. For ease of comparison, the frequencies were converted into percentages. The average difference between the actual percentage Stylus responses, $M = 62.19$, $SD = 8.20$, and the estimated percentage, $M = 59.99$, $SD = 19.96$, was 2.20% ($SD = 0.16$). A paired samples t -test showed no statistically significant difference between participants' estimated percentage Stylus responses and the actual percentage, $t(127) = -1.24$, $p = .219$, $d = -0.109$. The most optimistic estimate was 36.36% (estimated 12 more Stylus responses than actual number Stylus), the most pessimistic estimate was 72.73% (estimated 24 fewer Stylus responses than actual number).

We should note that 24 participants had percentage correct or percentage Stylus estimations that were over or below $1.5 \times SD$. Six of these participants estimated they had (an absolute number of) 1 correct response and 1 Stylus response, which suggested they had no idea and just indicated they did very badly, or they had not taken the task seriously in general. If we remove all results over or below $1.5 \times SD$, the mean difference between the percentage of actual correct responses and the estimated percentage is -3.22 ($SD = 15.69$), and the mean difference between the percentage of actual Stylus responses and the estimated percentage is -2.08 ($SD = 12.16$).

5.4.4.1.2 Comparing average single trial confidence and overall confidence

A paired samples *t*-test shows a statistically significant difference between participants' average confidence from each trial over the whole experiment ($M = 89.15$, $SD = 6.99$; this is indeed higher than warranted by their performance) and their average estimated percentage of correct responses (reported above), which we treat as a post-task frequency confidence measure, $t(127) = 9.92$, $p < .001$, $d = 0.877$. This confirms that there is a statistically significant difference between participants' overconfidence on item-to-item basis, and their overall confidence, which is better calibrated.

5.4.4.1.3 Comparing average single trial confidence, overall confidence, and percentage correct responses

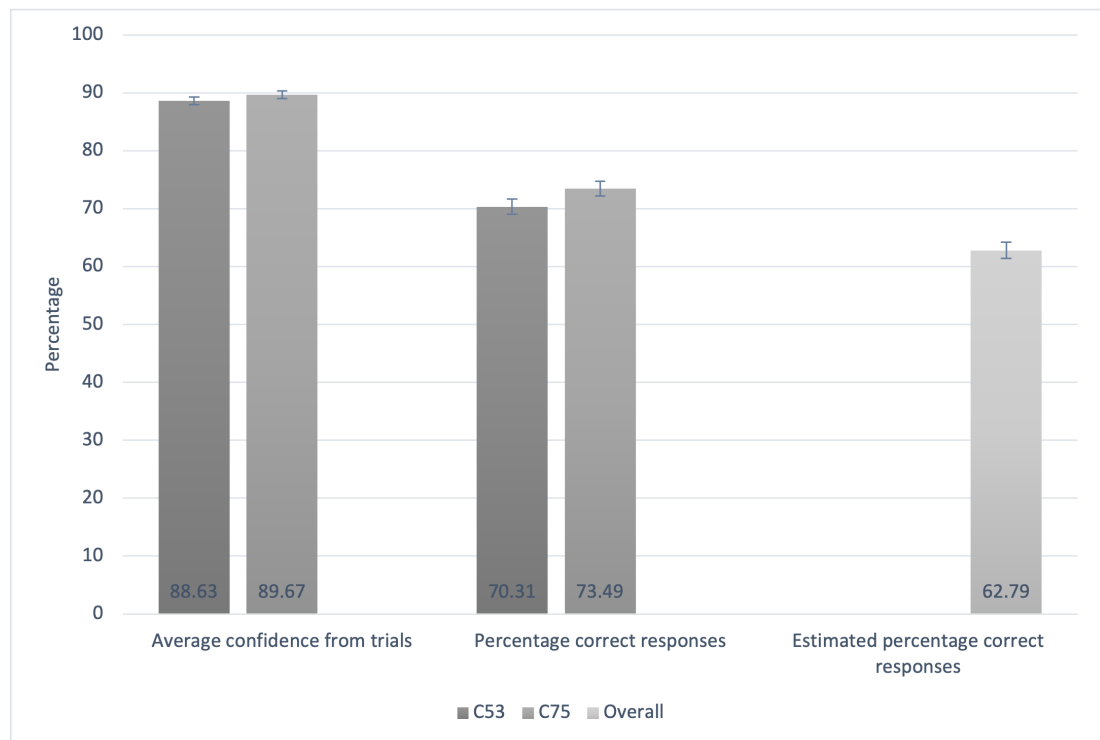


Figure 5.4 – S3 percentage average trial confidence, percentage correct responses, and overall estimated percentage correct responses, mean and standard error per condition

S3-H9 The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

To test whether participants' post-task frequency estimations are indeed more realistic than the average confidence reported during the task, we test the difference between percentage average trial confidence minus percentage correct, and percentage correct minus average post-task estimation of percentage correct. The means of trial confidence, percentage correct, and post-task estimation of percentage correct, as shown in Figure 5.4, have been reported earlier.

The difference between participants' trial confidence and their percentage correct response was $M = 17.25$ ($SD = 13.67$), and the difference between percentage correct and post-task estimation of percentage correct $M = 1.37$ ($SD = 23.25$). A paired samples t -test showed a statistically significant difference between the two measures, $t(127) = 5.67$, $p < .001$, $d = 0.501$.

The results show that a post-hoc estimation of correct responses is statistically significantly more accurate than the poorly calibrated trial confidence, which confirms our expectations based on findings from the literature and the results of S2-H6, and means that S3-H9 was confirmed.

5.4.4.2 Trust

5.4.4.2.1 Performance and post-task trust in Stylus

As noted before, we did not find support for the idea that prior trust will be associated with willingness to accept Stylus advice. Perhaps this is because trust in checkers *in general* transfers only weakly, if at all, to the experimental set up. Post-task trust in Stylus cannot be said to cause bias toward Stylus, of course, but it may nevertheless expose the plausible relation between trust and bias.

In C53, there was a statistically significant correlation between c , which is participants' propensity to accept Stylus suggestions independently of their correctness, and their post-task trust in Stylus, $r(126) = .23$, $p = .008$. In C75 we also observed a statistically significant correlation between c and post-task trust in Stylus, $r(126) = .22$, $p = .012$. These results suggest that the post-task trust rating is indeed tapping participants' trust in the aid while making their decisions.

5.4.4.2.2 Post-task trust in Stylus reliability

As reported in section 5.4.1, the post-task aggregated variable Post trust was reliable at $\alpha = .93$. We assumed that perceived trustworthiness of Stylus' suggestions might be one of the preconditions for participants' perception of the usefulness of Stylus' performance, which also includes an opinion on Stylus performing better in C75 than in C53. Therefore, we expected participants' rating of Stylus' trustworthiness to be statistically significantly higher than that of its usefulness. However, contrary to what we expected, the average trustworthiness rating was 64.45 ($SD = 16.10$), and usefulness was rated at $M = 70.09$ ($SD = 15.25$), and a paired samples t -test rejected our assumption, $t(127) = 5.86$, $p < .001$, $d = 0.518$. We have no explanation for this counter-intuitive result, but we believe the phrasing of the questions was potentially too subtle for participants to properly recognise what was asked.

5.4.4.3 Believability of Stylus as an automated system

The question “do you find it plausible that the Stylus suggestions are created by an automated system” served to confirm participants' engagement with the task; the average response was 71.59 ($SD = 19.02$), demonstrating that participants indeed found it plausible that Stylus' suggestions were made by an automated system.

5.5 Conclusion and discussion

5.5.1 Summary and discussion of key findings

5.5.1.1 Performance: sensitivity and bias

We observed that participants had a statistically significantly higher percentage of correct responses in the condition in which Stylus performed better and gave more reliable advice than in the condition where it performed poorly and gave low-information advice. Participants' bias towards Stylus was statistically significantly higher in C75 than in C53, which suggests that the stronger Stylus' advice is, the more likely participants are to accept it.

5.5.1.2 Confidence

Participants' confidence was statistically significantly higher in their correct responses than in incorrect ones, and it was statistically significantly higher in the condition where Stylus' advice was better (C53 vs. C75) as well. As expected on basis of the literature, there were statistically significant correlations between participants' estimations of the number of times they chose the correct response and the Stylus response, and their real frequencies. We also found statistically significant differences between participants' average confidence from all trials and their post-task estimation of the number of correct responses, which is better calibrated. This confirms that the overconfidence effect can indeed reduce if approached as frequencies rather than probabilities (Gigerenzer 1991).

5.5.1.3 Believability of Stylus as an automated system

Participants generally found it believable that the Stylus suggestions were created by an automated system, which confirms their engagement with the task.

5.5.1.4 Effects of trust

There was a statistically significant correlation between participants' prior trust and their confidence in responses where Stylus' advice was correctly followed. Like in S2, we also found a statistically significant correlation between participant's bias towards Stylus and their post-task trust in Stylus in both conditions in S3.

5.5.2 Implications for the design of S4

Our S4 study was largely the same as S3, and was intended to confirm the S3 findings with Stylus performing at a more realistic level.

5.5.2.1 Introducing a Stylus condition closer to real-world performance

In S3, Stylus performed at 53% and 75% (average 64%) statistically accurate advice. Even the higher of the two is a level of performance well below what can perhaps be expected from many real-world systems, and the average is

also below the reliability threshold of 70%, which is suggested to be required for useful automated aids (Wickens and Dixon 2007). Performance in C53 suggests that, on average, participants “know” more than 50% of the answers. If they accepted Stylus advice for all the rest, they would perform at circa 90%. If they simply accepted all Stylus suggestions, they would perform at 75%, which is slightly better than they actually performed. These data clearly suggest an underuse of Stylus' advice, even while at the same time showing that the acceptance of advice is positively encouraged if it is more accurate. Perhaps the underuse of the more reliable advice is in part explained by the low average reliability of Stylus' advice, alongside the reliable overconfidence effects that have been observed in all three studies so far. To test this possibility, and to gain insight into performance with more realistically reliable advice, we re-ran the study with a 53% and a 94% (average 73.5%, which is above the reliability threshold of 70%) condition to test whether an overall more reliable version of the aid would promote better use in participants; this study, S4, will be discussed in the following chapter.

Chapter 6 – *Stylus 4: Increasing the reliability of the aid and the strength its advice*

6.1 Introduction

The results of S3 suggested that participants used, but underutilised, and perhaps mistrusted, Stylus' advice, even when its reliability was above the estimated threshold for useful alarms (Wickens and Dixon 2007). The general aim of Stylus 4 (S4), the experimental study we discuss in this chapter, was to test whether an even greater strength of Stylus advice than in S3 might affect the calibration of participants' judgements and the confidence in their judgements.

6.1.1 Main differences in survey between S3 and S4

The design of S4 was largely the same as S3 and the same sentence pairs were used, but C75 has been replaced with a statistically accurate 94% Stylus likelihood estimation condition (C94). For S3 we needed eight different versions of the survey to guarantee an equal distribution of sentences through the four conditions (low likelihood vs. high likelihood x Stylus correct vs. Original correct), for S4 we needed 32 versions to assure this, due to the number of trials in each condition. The test interface remained unchanged.

6.2 Method

Because S4 is very similar to S3, instead of describing the method in full, we will only highlight the relevant differences between the two experiments in this Method section.

6.2.1 Task design, variables and hypotheses

S4 used a within-subjects design with 140 participants. The design was nearly identical to S3, with the only difference being the strength of the Stylus advice. The same variables were used (although the high Stylus likelihood estimation

was 94% instead of 75%), and the same hypotheses were tested as far as possible.

Because there was only one trial in the O94 condition (i.e., when Stylus advice was given with 94% strength, but was in fact incorrect), analysing sensitivity and bias measures in this condition is meaningless due to participants only having either a single CR, or a single FA. Therefore, several of the hypotheses from the previous chapter could not be re-tested. What follows is a list of all hypotheses tested in this chapter, named by experiment number and in sequence (S4-H1, etc.) as in previous chapters. For ease of comparison of hypotheses between Chapter 5 and this chapter, the S4 hypotheses have been numbered to mirror the S3 ones; ergo, not all numbers are present, and H9 is tested after H10 and H11. Only new hypotheses are explained below, for descriptions of the purpose of the other hypotheses, see Chapter 5, section 5.2.2. As in the previous chapters, as well as testing these main hypotheses, further statistical analyses will be undertaken to investigate the relationship between variables.

S4-H1 Participants' performance, in terms of percentage correct, will be better when Stylus' reliability is 94% than when it is 53%.

S4-H6 Participants will be more confident when using more reliable Stylus suggestions.

S4-H7 Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.

S4-H8 Participants will be more confident when responding correctly.

S4-H9 The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

S4-H10 In the single *Original is correct – strength of Stylus advice is high* trial, participants who respond incorrectly (Stylus is correct) will have rated their prior perceived self-efficacy in the domain of writing lower than participants who respond correctly.

This hypothesis serves to gain further insight in whether participants' perceived self-efficacy might affect their acceptance of the aid's advice. If confirmed, this suggests that if participants are less certain of their own efficacy and they are not sure of the correct response, they are more likely to rely on the aid's advice when the system shares a high rating of the likelihood of its advice being correct.

S4-H11 In the single *Original is correct – strength of Stylus advice is high* trial, participants who respond incorrectly (Stylus is correct) will have rated their prior trust in automated writing style checkers higher than participants who respond correctly.

This hypothesis serves to gain further insight in whether participants' trust in similar systems might affect their acceptance of the aid's advice. If confirmed, this suggests that if participants have expressed a higher trust in similar

systems in general and they are not sure of the correct response, they are more likely to rely on the aid's advice when the system shares a high rating of the likelihood of its advice being correct.

6.2.2 Participants

140 participants were recruited on Prolific.ac; the recruitment and screening process was the same as for S3, but this time those who participated in S3 were also excluded. Of the participants 58 were male, 82 female and ages ranged from 18 to 75 ($M = 36.19$; $SD = 12.77$).

6.2.3 Materials

The design of the decision tasks was identical to that of S3 and the same sentence pairs were used, the only difference was the strength of the Stylus advice in the higher condition.

6.2.4 Procedure

The procedure of S4 was largely the same as that of S1–3, with the following exceptions. Participants were paid an average of £2.50 (based on £7.50/hour) on completion of the survey, the estimated time for completing the survey was 15 minutes (automatically estimated by Qualtrics), and the actual average completion time was 16.14 minutes ($SD = 6.71$).

6.2.4.1 Pre-task and word processing software use

These sections of the survey were identical to S3/S4.

6.2.4.2 Experimental task

As in S3, the task consisted of four practise trials, followed by 33 experimental trials, presented one after another in random order. There were 32 different versions of the experiment that were each randomly assigned to an approximately equal number of participants. As in S3, all versions used the same sentence pairs in the trials, but signal and noise trials were systematically formatted in each of the versions, as we explain below. The practise trials were identical to S3 for all versions in S4 as well, and so was the dummy trial.

6.2.4.2.1 Stylus likelihood estimations

In S4, Stylus indicates its own estimation of the likelihood of its suggestions being correct with a 53% or a 94% label, the distribution of the number of trials per condition is shown in Table 6.1.

	Condition	Number total	Number Original correct	Number Stylus correct
	53%	17 (C53)	8 (O53)	9 (S53)
	94%	16 (C94)	1 (O94)	15 (S94)
TOTAL		33	9	24
<i>Percentage</i>			<i>27.27</i>	<i>72.72</i>

Table 6.1 – S4 item distribution over Stylus reliability conditions

6.2.4.2.2 Counterbalancing items across conditions

To validate that individual item difficulty cannot affect comparisons between experimental conditions, sentence pairs in S4 were systematically arranged in 32 different versions of the experiment (see distribution in Appendix B6) so that sentence-pairs were used equally in the four conditions of the experiment (53 vs. 94 x Stylus correct vs. Original correct).

6.2.4.2.3 Participants' confidence judgements

Participants rated their level of confidence in each response and post-task in the same way as in S3.

6.3 Analysis strategy

The analysis strategy for S4 was the same as for S3, with the following exceptions.

6.3.1 Acceptance and rejection of data

The results of all participants were used, and no data were rejected.

Three participants “returned” the study and one participant “timed out”; partial data of uncompleted tasks was not stored by Qualtrics, hence it has not been used in our analyses.

6.3.2 Signal Detection Theory to analyse task data

Because there are 15 correct Stylus suggestions and only one correct Original sentence in C94, meaningful sensitivity and bias analyses could not be performed in this chapter.

6.3.3 Analysing confidence

No changes were made in comparison with S3, but a between-subjects comparison of O94 data was added.

6.3.4 Comparing S3 and S4

To test the effect of a better performing aid on participants' performance and confidence, an S3 and S4 between-experiment comparison of percentage correct and confidence was performed and reported after the S4 analyses.

6.4 Results

The most important S4 data can be found in tabular form in Appendix A6, including breakdowns of aggregated variables. A table of all hypotheses from this thesis can be found in Appendix D.

6.4.1 Pre-task measures

6.4.1.1 Reliability testing

Reliability of prior perceived self-efficacy was $\alpha = .80$, that of prior trust $\alpha = .78$, and of the post-task variable *post-trust* it was $\alpha = .95$.

6.4.1.2 Prior perceived self-efficacy and Prior trust

Participants reported a prior perceived self-efficacy of $M = 74.85$ ($SD = 14.10$) and a rating of prior trust in automated suggestions of $M = 75.64$ ($SD = 15.91$).

6.4.2 Performance

Table 6.2 reports performance in absolute numbers in all four categories (H, M, FA, CR) in both conditions (C53 and C94), as this serves as the basis for our further analyses.

C53	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 5.89$ ($SD = 1.56$)	(FA) $M = 2.84$ ($SD = 1.63$)
<i>Original sentence selected by participant</i>	(M) $M = 2.11$ ($SD = 1.56$)	(CR) $M = 5.16$ ($SD = 1.63$)
C94	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 11.85$ ($SD = 2.10$)	(FA) $M = 0.61$ ($SD = 0.49$)
<i>Original sentence selected by participant</i>	(M) $M = 3.15$ ($SD = 2.10$)	(CR) $M = 0.39$ ($SD = 0.49$)

Table 6.2 – S4 mean number of responses per category per condition

6.4.2.1 Proportion of correct responses between conditions

Percentage correct $((N_H + N_{CR}) / N_{Total} * 100)$ overall was $M = 73.48$ ($SD = 11.77$).

S4-H1 Participants' performance, in terms of percentage correct, will be better when Stylus' reliability is 94% than when it is 53%.

Average percentage correct was 69.06 ($SD = 14.77$), in C53, in C94 it was $M = 77.90$ ($SD = 13.95$). A paired samples t -test showed participants' performance was statistically significantly better in the condition in which they received more reliable Stylus suggestions, $t(139) = -6.34$, $p < .001$, $d = -0.536$. Thus, S4-H1 was supported, which suggests that users benefited from Stylus' advice in line with similar findings in other domains in the literature (Wickens and Dixon 2007), and that higher aid reliability positively affected users' performance. This finding replicates the result of S3-H1.

6.4.3 Confidence

6.4.3.1 Confidence during the task

S4-H6 Participants will be more confident when using more reliable Stylus suggestions.

The average self-reported confidence across the task was 89.20 ($SD = 6.81$), and it was $M = 88.34$ ($SD = 7.32$) in C53, and $M = 90.06$ ($SD = 6.95$) in C94. A paired samples t -test showed a statistically significant difference in average reported confidence between the conditions, $t(139) = -4.75$, $p < .001$, $d = -0.402$, with participants being more confident in their responses in trials with a higher Stylus likelihood rating. Thus, S4-H6 was confirmed, which suggests that more reliable Stylus performance positively affected participants' confidence. This is in line with the earlier S3-H6 findings.

S4-H7 Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.

It is clear that in both conditions, confidence overestimates the percentage of correct trials. A 2x2 mixed ANOVA (C53 vs. C94 x Average trial confidence vs. Percentage correct) revealed a statistically significant effect of Condition, $F(1,139) = 45.84$, $p < .001$, $partial \eta^2 = 0.248$, as well as of Confidence vs. Performance, $F(1,139) = 196.18$, $p < .001$, $partial \eta^2 = 0.585$, and of Condition x Average trial confidence vs. Percentage correct, $F(1,139) = 29.54$, $p < .001$, $partial \eta^2 = 0.175$. Thus, S4-H7 was confirmed, in earlier results of S1-H5, S2-H5, S3-H7 and the findings in the literature we mentioned in relation to those results.

6.4.3.2 Confidence ratings for H, M, FA and CR and applied corrections

In C53, 21 participants had no Ms, hence no M confidence ratings and 11 participants had no FAs, hence no FA confidence ratings. In C94, 14 participants had no Ms, hence no M confidence ratings. To substitute a missing confidence value, the mean of a participant's own present values was used (see section 2.4.1).

Because there was only a single *Original is correct* trial in C94, CR and FR scores are mutually exclusive. 54 Participants had no CR, and 86 had no FA and associated confidence ratings in this condition.

When we break down the confidence scores for H, M, FA and CR, they are as shown in table 6.3, as well as graphically in Figure 6.1 for ease of comparison.

C53	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 90.31$ ($SD = 7.34$)	(FA) $M = 86.31$ ($SD = 9.78$)
<i>Original sentence selected by participant</i>	(M) $M = 84.13$ ($SD = 12.74$)	(CR) $M = 89.05$ ($SD = 8.40$)

C94	<i>Stylus suggestion correct</i>	<i>Original sentence correct</i>
<i>Stylus suggestion selected by participant</i>	(H) $M = 91.30$ ($SD = 6.72$)	(FA) $M = 88.60$ ($SD = 9.84$)
<i>Original sentence selected by participant</i>	(M) $M = 85.60$ ($SD = 11.46$)	(CR) $M = 88.79$ ($SD = 11.80$)

Table 6.3 – S4 mean confidence percentage per category per condition

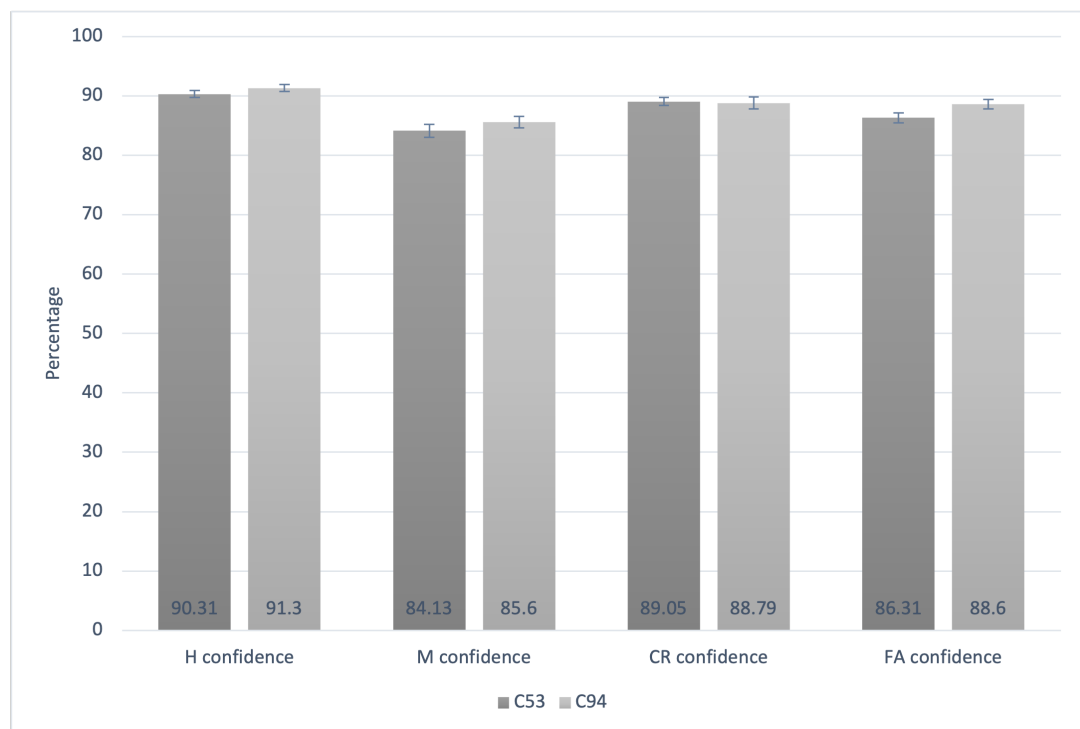


Figure 6.1 – S4 percentage H, M, FA and CR confidence, mean and standard error per condition

Participants' confidence ratings were analysed using a 2x2x2 repeated measures ANOVA. The within-subjects variables that were tested were

Correctness of the response (i.e., "Correct" vs. "Incorrect"), Type of response (i.e., "Stylus" vs. "Original"), and Strength of recommendation (i.e., 53% vs. 94%). Participants' average confidence in each cell of the design was entered into this ANOVA.

S4-H8 Participants will be more confident when responding correctly. The ANOVA did not show a statistically significant main effect for Correctness of response (correct (C53: 89.73 ($SD = 7.04$); C94: $M = 91.19$ ($SD = 6.67$) vs. incorrect (C53: 85.29 ($SD = 9.70$); C94: $M = 91.23$ ($SD = 6.74$)), $F(1, 139) = 0.68$, $p = .41$, *partial* $\eta^2 = 0.005$. Thus, S4-H8 was not supported, which is surprising given the reliability of this effect in other experiments of this thesis, see the results of e.g., S3-H8 and S5-H4.

There was a statistically significant main effect for Strength of recommendation on participants' confidence, C53 ($M = 88.34$, $SD = 7.31$) vs. C94 ($M = 90.06$, $SD = 6.95$), $F(1, 139) = 5.35$, $p = .0225$, *partial* $\eta^2 = 0.037$. This shows that participants are overall more confident when receiving stronger advice. The ANOVA also revealed a statistically significant main effect for Type of response ((Stylus C53: 89.06 ($SD = 7.12$); C94: $M = 91.23$ ($SD = 6.74$)) vs. Original (C53: 87.47 ($SD = 8.76$); C94: $M = 85.93$ ($SD = 11.27$))) responses, $F(1, 139) = 92.32$, $p < .001$, *partial* $\eta^2 = 0.399$. This shows that participants are overall more confident when accepting than when rejecting the aid's advice.

6.4.3.3 Confidence interaction effects

The ANOVA revealed a statistically significant interaction effect for Strength of recommendation x Type of response $F(1, 139) = 3.97$, $p = .048$, *partial* $\eta^2 = 0.028$, which means that the effect of Stylus' likelihood estimations was greater if participants responded "Stylus" than if they responded "Original". It also revealed a statistically significant interaction effect for Correctness of response x Type of response $F(1, 139) = 27.50$, $p < .001$, *partial* $\eta^2 = 0.165$, this means that the effect of correct responses on confidence was greater if participants responded "Stylus" than if then responded "Original". There was no effect for Strength of recommendation x Correctness of response $F(1, 139) = 0.07$, $p = .794$, *partial* $\eta^2 = 0.000$, nor for Strength of recommendation x Correctness of response x Type of response $F(1, 139) = 2.50$, $p = .116$, *partial* $\eta^2 = 0.018$.

6.4.3.4 O94 confidence between-subjects analysis

Although the results of the O94 trial, where *Original is correct – strength of Stylus advice is high*, could not meaningfully be compared with those of S94, or with C53, the results of this single trial warranted independent between-subjects exploration. 86 (61%) of the 140 participants correctly responded 'Original is correct' (CR), and 54 (39%) thought Stylus' advice was correct (FA).

There was no statistically significant difference in participants' confidence ratings between those who thought the Original sentence was correct, $M = 88.44$, $SD = 13.75$, $min = 50$, $max = 100$, and those who believed Stylus was

correct, $M = 89.43$, $SD = 11.20$, $\min = 59$, $\max = 100$, $t(138) = -0.44$, $p = .659$, $d = -0.077$. This result is in line with our earlier finding that over the whole of the experiment there was no statistically significant main effect for Correctness of response on confidence, although it should be treated with caution because it is based on just a single trial.

S4-H10 In the single *Original is correct – strength of Stylus advice is high* trial, participants who respond incorrectly (Stylus is correct) will have rated their prior perceived self-efficacy in the domain of writing lower than participants who respond correctly.

There was no statistically significant difference in prior perceived self-efficacy between participants who thought the Original sentence was correct ($M = 75.68$, $SD = 14.28$) and those who believed Stylus was correct ($M = 73.54$, $SD = 13.85$), $t(138) = 0.874$, $p = .383$, $d = 0.152$. Thus, S4-H10 was rejected.

S4-H11 In the single *Original is correct – strength of Stylus advice is high* trial, participants who respond incorrectly (Stylus is correct) will have rated their prior trust in automated writing style checkers higher than participants who respond correctly.

There was no statistically significant difference in confidence between participants who thought the Original sentence was correct ($M = 74.94$, $SD = 16.03$) and those who believed Stylus was correct ($M = 76.76$, $SD = 15.81$), $t(138) = -0.66$, $p = .513$, $d = -0.114$. Thus, S4-H11 was rejected.

6.4.4 Post-task measures

6.4.4.1 Confidence

6.4.4.1.1 Post-task estimation of number correct and number Stylus responses

Participants' subjective estimation of the number of times they selected the correct answer (either Original or Stylus) was $M = 23.42$ ($SD = 7.24$). There was a statistically significant correlation between the average number of times participants thought they selected the correct answer, and the objective frequency (including dummy trial), $M = 24.51$ ($SD = 3.77$), $r(138) = .43$, $p < .001$, which, once more, shows that participants have a level of awareness of their performance. For ease of comparison, the frequencies were converted into percentages. The average difference between the actual percentage correct, $M = 74.26$, $SD = 11.43$, and the estimated percentage correct, $M = 70.97$, $SD = 21.93$, was 3.29% ($SD = 19.96$). The most optimistic estimate was 39.39% too high (estimated 13 more correct trials than actual number correct), the most pessimistic estimate was 72.73% too low (estimated 24 fewer correct trials than actual number).

We also compared the number of times participants thought they chose the Stylus suggestion over the Original sentence with the objective frequencies. The estimated average number of Stylus choices across participants was 19.39 ($SD = 6.04$), and the real number (including dummy trial) was $M = 21.96$ ($SD = 2.87$). There was no statistically significant correlation between the

average number of times participants thought they chose the Stylus suggestion, and the objective frequency, $r(138) = .15$, $p = .084$. For ease of comparison, the frequencies were converted into percentages. The average difference between the actual percentage Stylus responses, $M = 66.56$, $SD = 8.71$, and the estimated percentage, $M = 58.77$, $SD = 18.29$, was 7.79% ($SD = 19.07$). The most optimistic estimate was 39.39% (estimated 13 more Stylus responses than actual number Stylus), the most pessimistic estimate was 75.76% (estimated 25 fewer Stylus responses than actual number).

We should note that 28 participants had percentage correct or percentage Stylus estimations that were over or below $1.5*SD$. 3 of these participants estimated they had (an absolute number of) 1 correct response and 1 Stylus response, which suggested they had no idea and just indicated they did very badly, or they had not taken the task seriously in general. If we remove all results over or below $1.5*SD$, the mean difference between the percentage of actual correct responses and the estimated percentage is 0.32 ($SD = 13.53$), and the mean difference between the percentage of actual Stylus responses and the estimated percentage is 3.35 ($SD = 12.33$).

6.4.4.1.2 Comparing average single trial confidence and overall confidence

A paired samples t -test shows a statistically significant difference between participants' average confidence from each trial over the whole experiment ($M = 89.20$, $SD = 6.81$; this is indeed higher than warranted by their performance) and their average estimated percentage of correct responses (reported above), which we treat as a post-task frequency confidence measure, $t(139) = 10.18$, $p < .001$, $d = 0.861$. This confirms that there is a statistically significant difference between participants' overconfidence on item-to-item basis, and their overall confidence, which is better calibrated.

6.4.4.1.3 Comparing average single trial confidence, overall confidence, and percentage correct responses

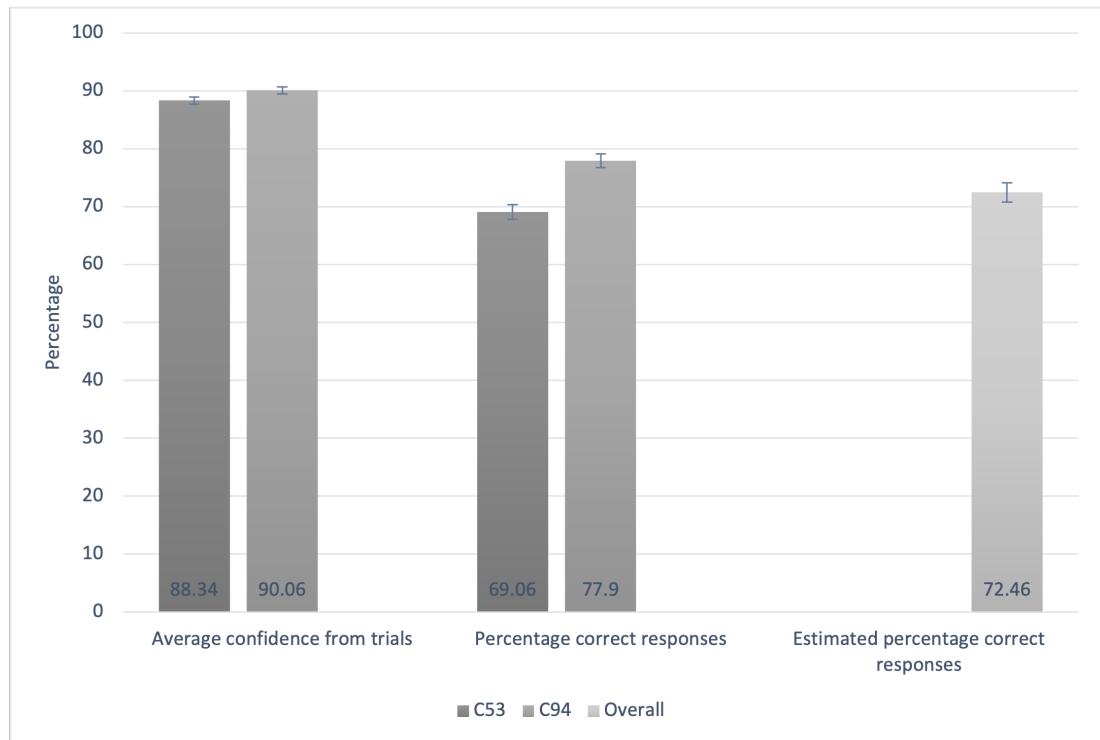


Figure 6.2 – S4 percentage average trial confidence, percentage correct responses, and overall estimated percentage correct responses, mean and standard error per condition

S4-H9 The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

To test whether participants' post-task frequency estimations are indeed more realistic than the average confidence reported during the task, we test the difference between percentage average trial confidence minus percentage correct, and percentage correct minus average post-task estimation of percentage correct. The means of trial confidence, percentage correct, and post-task estimation of percentage correct, as shown in Figure 6.2, have been reported earlier.

The difference between participants' percentage trial confidence and their percentage correct responses was $M = 15.71$ ($SD = 13.27$), and the difference between percentage correct responses and percentage post-task estimation of correct responses was $M = 2.51$ ($SD = 20.04$). A paired samples t -test showed no statistically significant difference between the two measures, $t(139) = 5.87$, $p < .001$, $d = 0.496$.

The results show that S4 participants' post-hoc estimation of correct responses is statistically significantly more accurate than their poorly calibrated trial confidence, which confirms S4-H9. This is in line with the earlier results of S2-H6 and S3-H9, and the findings from the literature we mentioned in relation to those results.

6.4.4.2 Trust

6.4.4.2.1 Performance and post-task trust in Stylus

We failed, again, to find support for the idea that prior trust will be associated with willingness to accept Stylus advice. Perhaps this is because trust in checkers *in general* transfers only weakly, if at all, to the experimental set up. Post-task trust in Stylus cannot be said to cause bias toward Stylus, of course, but it may nevertheless expose the plausible relation between trust and bias.

In C53, there was no statistically significant correlation between c , which is participants' propensity to accept Stylus suggestions independently of their correctness, and their post-task trust in Stylus, $r(138) = .04$, $p = .653$. In C94 there also was no statistically significant correlation between c and trust in Stylus, $r(138) = .02$, $p = .814$.

6.4.4.2.2 Post-task trust in Stylus reliability

As reported in section 6.4.1, the post-task aggregated variable Post trust was reliable at $\alpha = .95$. In S3 we expected participants' average rating of Stylus' trustworthiness to be statistically significantly higher than that of its usefulness, but found the opposite was the case. In S4 the average trustworthiness rating was 62.68 ($SD = 18.40$), and usefulness was rated at $M = 69.64$ ($SD = 16.07$), and a paired samples t -test showed a result comparable to that in S3, $t(139) = 7.54$, $p < .001$, $d = 0.637$. If we compare participants' Stylus usefulness ratings between S3 and S4, an independent samples t -test shows no statistically significant difference, $t(266) = 0.24$, $p = .814$, $d = 0.029$. A comparison between trustworthiness ratings in both studies does not show a statistically significant difference either, $t(266) = 0.84$, $p = .404$, $d = 102$. This means that we assume that the strength of Stylus' advice does not statistically significantly affect participants' ratings of Stylus' usefulness and trustworthiness.

6.4.4.2.4 Believability of Stylus as an automated system

The question “do you find it plausible that the Stylus suggestions are created by an automated system” served to confirm participants' engagement with the task; the average response was 71.13 ($SD = 16.32$), demonstrating that participants indeed found it plausible that Stylus' suggestions were made by an automated system.

6.5 Comparing S3 and S4

The similar designs of S3 and S4 allowed us to make a comparison between the results of both experiments and examine the effects of a better performing aid on participants' performance and confidence.

6.5.1 Performance

Percentage correct in S53 in S3 was $M = 70.31$ ($SD = 14.82$), and it was $M = 69.06$ ($SD = 14.77$) in S4. In the condition where Stylus performed well, S75 in S3 and S94 in S4, percentage correct was respectively $M = 73.49$ ($SD = 14.39$) in S3, and $M = 77.90$ ($SD = 13.95$) in S4.

The percentage correct data were analysed in a 2x2 mixed ANOVA, with Experiment as a between-subjects factor, and Stylus reliability as a within-subjects factor. There was a statistically significant effect of Stylus reliability (S53 vs. S75/94), $F(1, 266) = 38.88$, $p < .001$, $partial \eta^2 = 0.128$, and there also was a statistically significant effect for Reliability x Experiment, $F(1, 266) = 8.65$, $p = .004$, $partial \eta^2 = 0.031$. There was no between-subject effect for Experiment, $F(1, 266) = 1.13$, $p = .288$, $partial \eta^2 = 0.004$. Taken together, these results show that participants perform better when using a more reliable aid.

6.5.2 Confidence

The average confidence from the trials in S53 in S3 was $M = 88.63$ ($SD = 7.37$), and it was $M = 88.37$ ($SD = 7.32$) in S4. In the condition where Stylus performed well, S75 in S3 and S94 in S4, confidence was $M = 89.67$ ($SD = 7.06$) in S3, and $M = 90.06$ ($SD = 6.95$) in S4.

A 2x2 Experiment x Reliability ANOVA on confidence revealed a statistically significant effect of Reliability (S53 vs. S75/94), $F(1, 266) = 32.34$, $p < .001$, $partial \eta^2 = 0.108$, which shows that participants were more confident when using a more reliable aid. However, there was no statistically significant effect for Reliability x Experiment, $F(1, 266) = 1.96$, $p = .163$, $partial \eta^2 = 0.007$, nor was there a between-subjects effect for Experiment, $F(1, 266) = 10.003$, $p = .959$, $partial \eta^2 = 0.000$.

6.6 Conclusions and discussion

The general aim of S4, was to test whether an even greater strength of Stylus advice than in S3 might affect the calibration of participants' subjective judgements and their confidence in their judgements.

6.6.1 Summary and discussion of key findings, and comparison with S3

6.6.1.1 Performance

Just as in S3, we observed that in S4 participants had a statistically significantly higher percentage of correct responses in the condition in which

Stylus performed better and where it gave higher likelihood estimations, than in the condition where it performed poorly and gave low-information advice. We could not compare sensitivity and bias between the conditions, and hence between the studies, because there was not enough data to draw meaningful conclusions from in one of the S4 conditions.

6.6.1.2 Confidence

Participants' confidence was statistically significantly higher in correct responses than in incorrect ones in S3, but we did not observe this effect in S4, a discrepancy we cannot explain. We found in S4 that participants had statistically significantly more confidence in their Stylus responses than when they chose "Original is better", while this was not the case in S3. This finding suggests that a better performing version of Stylus boosts participants' confidence.

In S3 we found no statistically significant interaction effects for Strength of recommendation x Correctness of response, Strength of recommendation x Type of response, Correctness of response x Type of response, and Strength of recommendation x Correctness of response x Type of response. Interestingly, the first two effects were statistically significant in S4.

As expected on basis of the literature, also in S4 there were statistically significant correlations between participants' estimations of the number of times they chose the correct response and the Stylus response, and their real frequencies. This means that although they are overconfident in each individual trial, overall, the overconfidence effect disappears and they turn out to be well-calibrated, with their estimates just below the real frequencies. The relationship between this overconfidence at trial level and the overall calibration is consistent, as demonstrated by a statistically significant correlation between the two.

6.6.1.3 The plausibility of Stylus as an automated system

Participants found it believable that the Stylus suggestions were created by an automated system, which confirms their engagement with the task, and a test showed a statistically significant effect for Stylus' performance.

6.6.1.4 Effects of perceived self-efficacy, trust, and system likelihood estimations

Our finding in S3 that there is a statistically significant effect of the level of Stylus' likelihood estimations on participants' confidence, with them being more confident in responses in trials where Stylus displayed a high likelihood estimation and gave reliable advice, was confirmed in S4. It is not clear what causes this effect, the strength of the advice, or the reliability of Stylus' performance.

6.6.1.5 Effects of a better performing aid

The comparison of S3 and S4 data shows that the aid is still radically underused if its performance is in fact better than that of participants. In other

words, blindly accepting the aid's judgements rather than judging them on a case-by-case basis would have benefited their performance. Tests showed that the level of Stylus' performance statistically significantly affects participants' performance, yet it has no statistically significant effect on their confidence.

6.6.2 Discussion of method

There were several hypotheses from Chapter 5 that could not be tested in this chapter due to the low number of data points in the O94 condition, which makes analysing d' and c for C94 meaningless. While it was one of our objectives to dramatically increase Stylus' performance in this experiment in comparison with S3 in the condition where it performed well, the issue of missing data arguably affected the usefulness of the current design. The only solution to this problem would be to vastly increase the number of trials, which is what we do in our next experiment, Stylus 5.

6.6.3 Implications for the design of S5

With S4, we believe we have run into a limitation of the usefulness of SDT with our current within-subjects design, so we redefined our approach for S5. Because meaningful analysis of some of our S4 data was impossible, we have designed S5 as a between-subjects study with three groups, and 100 trials for each group. In S3 and S4 participants encountered a 2AFC-like design and a system that indicated the likelihood of its advice being correct in each trial. In S5, two groups each encounter a yes-no design with 50 signal and 50 noise trials, and a system that performs at a single level, either 53% or 90%. The system's performance level is explained to participants in the brief. The third group is a control group, who complete the experiment without help from Stylus. Although this approach, which is much closer to classic perceptual SDT research (see e.g., Macmillan and Creelman 2005), also has its inherent limitations, which are discussed in chapter 7, it vastly increases the number of data points (from $128 * 32 = 4,096$ in S3, $140 * 32 = 4,480$ in S4, to $114 * 100 = 11,400$ in S5), and eradicates many of the methodological issues we discussed in relation to S1 – S4.

Chapter 7 – *Stylus 5: Testing models of aided interaction*

7.1 Introduction

In S3, and more especially in S4, we noticed we were running into the limits of what can be reliably computed about participant use of advice from SDT-metrics such as bias. In particular there are limits on the 2AFC-design employed so far, due to the limited number of trials we could reasonably present to participants. S5 therefore has a slightly different approach.

In S5 we use a design that is much closer to a classic perceptual SDT task, like Rice and McCarley's baggage screening task (2011) described in Chapter 1, section 1.2.3.3. Instead of a 2AFC-like design, we use a Yes/No (Y/N) design. Furthermore, with respect to the reliability of the aid, we use an overall indication of the system's performance rather than an item-by-item system likelihood estimation. This seems important, as the underuse of reliable aids that has been documented so far might be influenced by the co-presence of less reliable versions. By moving to a Y/N design we are able more directly to compare participant behaviour against the predictions of the idealised models presented by Bartlett and McCarley (2017, 2019).

The type of errors used in the sentences in S5 was limited to one kind that is common in everyday language. The main benefits of this approach were that it was easier to generate a larger number of sentences, and that each trial was much quicker to process by participants, which enabled a larger number of trials, hence an increased number of data points. The latter should allow us to improve the reliability of the SDT metrics. When selecting example words and sentences to create items, we observed two different types of errors that people make when writing, namely slips (e.g., typos) and mistakes (e.g., confused words). Our new design uses "homophones", words that sound similar to other words but have a different meaning, and may fall in either or both of these categories. Homophones can either be confused words, or misspellings that are meaningful themselves (but a meaning different to the one the writer intended).

As discussed in the literature review in Chapter 1, misspellings are normally easily picked up by simple spell checkers that parse from dictionaries, but sentence context can make them difficult to spot. The same is true for confused words, because their meaning is context dependent by definition.

Some of our examples might fall in both categories. An error like "dose" instead of "does" ("Dose he live here?") is very likely to be a misprint, but we cannot be certain some people would just use the wrong word. While it is not directly relevant in the generation of the items, it probably will be in how participants process them. If someone would never mistake one word for the other, it is reasonable to expect they will pick up this error when checking a sentence, no matter what a checker suggests. When on the other hand someone genuinely believes "does" could be spelled as "dose", they might rely more on the system's suggestion. In S5 we used a between-subjects design with two groups that each encounter a system with a discrete difference in performance level between them. We also used a control group of participants who were asked to process the same sentences without Stylus suggestions and rate their confidence in their responses.

Another methodological difference with S3/S4, is that participants in S5 are told about the reliability of Stylus' judgement (either 70% or 90%) in the pre-task briefing, rather than in each individual trial. A danger with this design is that the difference in Stylus reliability may be less salient. We attempted to ameliorate this danger by employing post-hoc questions which acted, in a sense, as manipulation checks. As before, the information about Stylus' reliability was entirely accurate, so that participants in the different groups certainly experienced advice of different reliability.

The test interface was changed to reflect the new Y/N-design, an example of a trial where Stylus does not indicate an error is shown in Figure 7.1a, Figure 7.1b shows an example of a trial where Stylus has detected a purported error.

Younger voters may form an important voting bloc in the next elections.

Do you think this is a correct sentence?

Yes

No

How confident are you in your answer? Please manipulate the slider to select a percentage.

I guessed 50 55 60 65 70 75 80 85 90 95 100 I'm certain I got it right

Next

Figure 7.1a – S5 trial example screenshot; Stylus indicates no error

This **balmy** policy seems to be a waste of energy.

Do you think this is a correct sentence?

Yes

No

How confident are you in your answer? Please manipulate the slider to select a percentage.

I guessed 50 55 60 65 70 75 80 85 90 95 100 I'm certain I got it right

Next

Figure 7.1b – S5 trial example screenshot; Stylus indicates supposed error

7.2 Method

Since there were several major differences in the design of S5 compared to S3 and S4, we will run through the method in greater detail than in the previous chapter.

7.2.1 Task design

S5 used a between-subjects design with 114 participants, divided into three equally sized groups ($N = 38$ per group). Each participant was presented with four practise trials and one hundred test trials, each made up of a sentence followed by the question 'Do you think this is a correct sentence? [Yes / No]'. Participants were told that 50 sentences were correct, and 50 contained a homophone error; homophones are words that sound similar but have a different meaning. Participants in two groups (G90 and G70) were briefed '*In some sentences you will see a word highlighted yellow, these are suggestions of errors from Stylus, an imaginary text editing aid based on artificial intelligence technology. This means that if Stylus were a real system, it would be self-learning and its judgements would be informed by an algorithm. Stylus is not perfect, it occasionally suggests there is an error whilst there isn't one, and sometimes Stylus misses an error.*' Participants in one of the groups (G90) were also told that '*Stylus's judgements are known to be 90% accurate in the sample of sentences you see. This means that out of each 10 sentences you see, Stylus's judgement (it either highlights an error or it doesn't) is correct in 9*', and in the other (G70) that Stylus was 70% accurate. The third group (GC) was a control group that encountered 50 correct and 50 incorrect sentences without suggestions from Stylus. All participants were also asked to indicate how confident they were in their response.

Ten versions of the experiment were produced for G90 and four versions for G70, with sentences rotated through the trials so that each sentence appeared equally often as a correct or incorrect sentence and with or without a Stylus recommendation. A roughly equal number of participants within each group was assigned to each one of these versions of the experiment according to the order in which they participated. For the control group there was just one version of the experiment. The order of the trials was randomised for each participant.

7.2.2 Variables and hypotheses

7.2.2.1 Pre-task measures

The independent variables measured prior to the task were participants' prior trust in automated writing checkers, and their perceived self-efficacy as checkers of grammar and spelling.

7.2.2.2 Performance

The between-subjects independent variable we manipulated during the experiment was the level of correctness of the Stylus suggestions, which is represented to participants by a Stylus performance percentage (either 70% or 90%) in the introductory briefing.

The dependent variables we measured were overall percent correct and proportion of acceptance of Stylus recommendations, as well as sensitivity (ability to distinguish correct sentences, independently of Stylus recommendations) and bias (favour of Yes or No responses, and tendency to accept Stylus recommendations independently of their correctness). We also recorded and analysed participants' confidence in their own responses, as in earlier experiments.

7.2.2.3 Hypotheses

Our experimental hypotheses all derive from the overarching hypothesis that participants will be able to interpret and make use of Stylus suggestions and Stylus' own estimation of the likelihood of its suggestions being correct. In this experiment, because there is no longer a 2AFC design, the main hypothesis of improved performance will be tested primarily with a measure of sensitivity rather than pure percentage correct (although percentage correct is in itself not an uninteresting measure). Some of our hypotheses will be very familiar by now. We re-test several hypotheses concerning the relation between trust, perceived self-efficacy, and tendency to accept advice, that have received no support or very patchy support in the experiments reported thus far.

S5-H1 Participants' sensitivity will be higher when Stylus' reliability is 90% than when it is 70%.

This hypothesis concerns whether the aid's reliability might affect participants' sensitivity. If confirmed, this suggests that a higher reliability of the aid might positively affect participants' ability to correctly follow its advice.

S5-H2 Participants' acceptance of Stylus suggestions will be higher when Stylus' reliability is 90% than when it is 70%, independently of correctness of the advice.

This hypothesis replicates S3-H2 with a Y/N design instead of the earlier 2AFC-like design.

S5-H3 Participants' will be more confident when Stylus' reliability is 90% than when it is 70%.

This hypothesis is a variant of S3-H6 that replicates this earlier hypothesis with a Y/N design instead of the earlier 2AFC-like design; the wording has been adjusted to fit the design of the current experiment. If confirmed, this suggests that a more reliable aid might positively affect participants' confidence.

S5-H4 Participants will be more confident when responding correctly.

This hypothesis replicates S3/S4-H8 with a Y/N design instead of the earlier 2AFC-like design.

S5-H5 Participants' prior perceived self-efficacy in the domain of writing will be positively correlated with their sensitivity, independent of Stylus' reliability. This hypothesis replicates S3-H3 with a Y/N design instead of the earlier 2AFC-like design.

S5-H6 Participants' prior perceived self-efficacy in the domain of writing will be negatively correlated with their bias, their tendency to accept Stylus advice.

This hypothesis replicates S3-H6 with a Y/N design instead of the earlier 2AFC-like design.

S5-H7 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations.

This hypothesis concerns whether there might be a potential relationship between trust in writing style checkers in general and acceptance of correct suggestions. If confirmed, this suggests that participants' acceptance of the aid's suggestions might be positively affected by their trust in writing style checkers in general.

S5-H8 Participants' trust in Stylus during the experiment (measured post-task) will be positively correlated with acceptance of correct Stylus suggestions.

This hypothesis concerns whether there might be a potential relationship between trust in the aid and acceptance of correct suggestions. If confirmed, this suggests that users recognise the aid's reliability, and that this in turn might influence to what extent they accept its suggestions.

7.2.3 Participants

114 participants were recruited on Prolific.ac. Participants were screened on current country of residence (registered as United Kingdom residents), country of birth (registered as born in the UK), nationality (UK), first language (English), and participation in previous studies, and those who participated in S1–4 were excluded. Of the participants 32 were male and 82 female, and ages ranged from 18 to 74 ($M = 35.27$; $SD = 13.45$).

7.2.4 Materials

A new set of items was created for this experiment. Sources used were the researchers' own knowledge, Lexico (lexico.com) and Scribendi (scribendi.com). 120 items were created and tested in a pilot with 14 participants. Four participants' data were rejected because they displayed chance performance, combined with unrealistically long or short completion times. From the data of the ten participants that were accepted, the 20 items that on average received the highest combined percentage correct and confidence scores, were rejected, which left 100 items that were the used in the experiment proper.

7.2.4.1 Pre-task perceived self-efficacy, pre-task trust, confidence, and post task frequency estimations

These factors were measured the same way as in S1–S4.

7.2.4.2 Post-task trust in Stylus

Participants' retrospective trust in Stylus' suggestions during the task was measured with the four questions 'When thinking of the usefulness of Stylus' suggestions during this experiment, I would class them as [0; Not very useful at all] – [100; Very useful]', 'When thinking of the trustworthiness of Stylus' performance during this experiment, I would class it as [0; Not very trustworthy at all] – [100; Very trustworthy]', 'When thinking of the consistency of Stylus' performance during this experiment, I would class it as [0; Not very consistent at all] – [100; Very consistent]' and 'When thinking of Stylus' performance in general during this experiment, I would class it as [0; Not very good at all] – [100; Very good]'. The results were tested for internal reliability (Cronbach α) and averaged as a single measure of post-task trust in Stylus.

7.2.4.3 Round-up questions

Before debriefing the participants, we asked them 'When assessing the sentences, did you remember the Stylus accuracy rate? [0; I did NOT remember the Stylus accuracy rate] – [100; I did remember the Stylus accuracy rate]', 'In your answers, did you consider the Stylus accuracy rate? [0; I did NOT consider the Stylus accuracy rate] – [100; I did consider the Stylus accuracy rate]', 'After assessing 100 sentences, in which Stylus indicated potential errors in 50, do you believe that Stylus was 90% [70%] accurate? [0; I do NOT believe that Stylus was 90% [70%] accurate] – [100; I do believe that Stylus was 90% [70%] accurate]' and lastly, 'How plausible do you find it that the Stylus suggestions were created by an automated system? [0; Not very plausible at all] – [100; Very plausible]'

7.2.5 Procedure

The procedure of S5 was largely the same as that of S1–4, with the following exceptions. Participants were paid an average of £3.15 (based on £7.80/hour) on completion of the survey, the estimated time for completing the survey was 30 minutes (automatically estimated by Qualtrics), and the actual average completion time was 26.17 minutes ($SD = 10.38$).

7.2.5.1 Pre-task and word processing software use

These sections of the survey were identical to S3/S4.

7.2.5.2 Experimental task

The task consisted of four practise trials, followed by one hundred experimental trials, presented one after another. The experimental trials were presented in random order after the practise trials. There were ten different versions of the experiment for G90, four for G70, and one for GC, which were

each assigned to an approximately equal number of participants. All versions used the same sentences in the trials, but signal and noise trials and correct and incorrect Stylus suggestions were systematically formatted in different ways in each trial in each of the versions, as we explain below. The practise trials were identical for all versions.

In each trial, participants were shown a sentence, of which they were asked if it was "correct", and to which they could respond by ticking a "Yes" or a "No" box. Each of the test groups (G70 and G90) encountered a version of the experiment where the Stylus suggestions were statistically accurate as briefed in the introduction. If a sentence was correct but Stylus indicated a perceived homophone error, the word in question was highlighted in yellow. If a sentence was correct and Stylus detected no homophone error, no words were highlighted. If a sentence was incorrect and Stylus indicated a perceived homophone error, the word in question was highlighted in yellow. If a sentence was incorrect and Stylus detected no homophone error, no words were highlighted.

As in the previous experiments, no performance feedback was provided to participants during the experiment.

7.2.5.2.1 Stylus performance

Other than the four practise trials, there are 100 trials in total in this study, of which 50 are "signal" trials, and 50 "noise" trials. In S5 participants are briefed about Stylus' level in the introduction to the task; G90 encountered a system that was correct in nine out of ten trials, G70's Stylus gave 70% correct suggestions, and GC did not get any suggestions from Stylus. Table 7.1 shows the distribution of trials over the four conditions for both groups.

G90		Number total	<i>Stylus advice is GOOD</i>	<i>Stylus advice is BAD</i>
	<i>Sentence is correct</i>	50	45	5
	<i>Sentence is incorrect</i>	50	45	5
TOTAL/ PERCENTAGE		100	90	10

G70		Number total	<i>Stylus advice is GOOD</i>	<i>Stylus advice is BAD</i>
	<i>Sentence is correct</i>	50	35	15
	<i>Sentence is incorrect</i>	50	35	15
TOTAL/ PERCENTAGE		100	70	30

Table 7.1 – S5 number of trials per condition, per group

7.2.5.2.2 Counterbalancing items across trials

To ensure that individual item difficulty could not affect comparisons between experimental conditions, sentences in S5 were systematically arranged in ten different versions of the experiment for G70 (A–J), four for G90 (A–D), and one for GC (see distribution in Appendix B7) so that sentences were presented equally in '*sentence is correct – Stylus advice is good*', '*sentence is correct – Stylus advice is bad*', '*sentence is incorrect – Stylus advice is good*', and '*sentence is incorrect – Stylus advice is bad*' trials.

7.2.5.2.3 Participants' confidence judgements

After responding "Yes" or "No", participants were asked to rate their level of confidence in their response following the same procedure as in S1–4.

7.2.5.3 Post-task

After the sequence of decision tasks, participants' estimations of their own and Stylus' performance were measured. Participants were also asked if and how they considered Stylus' reliability level, and how likely they thought it was the Stylus suggestions were created by an automated system.

7.3 Analysis strategy

7.3.1 Acceptance and rejection of data

Seven participants “returned” the study, and two participants “timed out”; partial data from uncompleted tasks was not stored by Qualtrics, hence it has not been used in our analyses.

7.3.2 Aggregated variables

Aggregated variables were processed following the same procedure as in S1–4 (see section 3.4.1).

7.3.3 Signal Detection Theory to analyse task data

7.3.3.1 Signal detection labels and conditions

We used Signal Detection Theory to analyse the S5 task data. Because the labels *signal* and *noise* are arbitrary in our Y/N design (*signal* could for example be defined as "good Stylus advice" or as "correct sentence"), the H, M, FA and CR labels need to be explained. We initially used the following labels, which also allowed us to compare the groups that received Stylus advice with the control group where required.

- H – The sentence is correct, the participant responds "Yes"
- M – The sentence is correct, the participant responds "No"
- FA – The sentence is incorrect, the participant responds "Yes"
- CR – The sentence is incorrect, the participant responds "No"

These labels are independent of Stylus' advice, which means the groups who encounter a version of the experiment with Stylus advice have H, M, FA and CR scores in two different conditions (Stylus advice is good vs. Stylus advice is bad), see Table 7.2.

G90	<i>Stylus advice GOOD</i>		<i>Stylus advice BAD</i>	
	<i>Sentence correct</i>	<i>Sentence incorrect</i>	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged YES</i>	H	FA	H	FA
<i>Judged NO</i>	M	CR	M	CR

G70	<i>Stylus advice GOOD</i>		<i>Stylus advice BAD</i>	
	<i>Sentence correct</i>	<i>Sentence incorrect</i>	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged YES</i>	H	FA	H	FA
<i>Judged NO</i>	M	CR	M	CR

GC	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged YES</i>	H	FA
<i>Judged NO</i>	M	CR

Table 7.2 – S5 H, M, FA, CR distribution over conditions per group

7.3.3.2 Two different measures of sensitivity and bias

In the S3 and S4 2AFC-like within-subjects design, the bias measure *c* indicated bias towards or against following Stylus' judgement, independent of Stylus' advice being good or bad (we computed these measures for both Stylus performance conditions). S5 uses a Y/N between-subjects design,

which means that if c is computed the same way, it indicates bias towards or against responding Yes or No. Although marginally relevant as an attention check (participants should have no bias either way), a more relevant measure is that of bias towards or away from Stylus. Thus, we computed two different sets of sensitivity measures, d'_{YN} , ability to identify correct sentences, and d'_{Stylus} , ability to identify good Stylus suggestions, and two bias measures as well, c_{YN} , tendency to favour Yes or No responses, and c_{Stylus} , tendency to follow Stylus advice.

d'_{YN} is important as a measure of performance, and will be used as the basis for testing S5-H1, and for testing performance against predictions from statistical models of how participants might combine their own judgments with those of Stylus. c_{YN} is of marginal relevance. In contrast, as in previous analyses, d'_{Stylus} is of marginal relevance, but c_{Stylus} is important for testing participants' propensity to accept stylus advice and for testing S5-H3.

d'_{YN} and c_{YN} were computed from the matrices in Table 7.2. The d'_{Stylus} and c_{Stylus} measures were computed by rearranging the SDT matrix as shown in Table 7.3 and in simplified form in Table 7.4, but only for the purpose of these measures. d' and c were then calculated via the usual procedure, as explained in Chapter 2, *Research approach and methodology*.

- H – Stylus advice that correct sentence is correct is GOOD, participant responds "YES" + Stylus advice that incorrect sentence is incorrect is GOOD, participant responds "NO"
- M – Stylus advice that correct sentence is correct is GOOD, participant responds "NO" + Stylus advice that incorrect sentence is incorrect is GOOD, participant responds "YES"
- FA – Stylus advice that correct sentence is incorrect is BAD, participant responds "NO" + Stylus advice that incorrect sentence is correct is BAD, participant responds "YES"
- CR – Stylus advice that correct sentence is incorrect is BAD, participant responds "YES" + Stylus advice that incorrect sentence is correct is BAD, participant responds "NO"

This rearrangement allowed us to compute d'_{Stylus} and c_{Stylus} for G90 and G70; GC did not receive any help from Stylus, so the grid cannot be used in that case.

G90	<i>Sentence correct</i>		<i>Sentence incorrect</i>	
	<i>Stylus advice GOOD</i>	<i>Stylus advice BAD</i>	<i>Stylus advice GOOD</i>	<i>Stylus advice BAD</i>
<i>Judged YES</i>	H	CR	M	FA
<i>Judged NO</i>	M	FA	H	CR

G70	<i>Sentence correct</i>		<i>Sentence incorrect</i>	
	<i>Stylus advice GOOD</i>	<i>Stylus advice BAD</i>	<i>Stylus advice GOOD</i>	<i>Stylus advice BAD</i>
<i>Judged YES</i>	H	CR	M	FA
<i>Judged NO</i>	M	FA	H	CR

Table 7.3 – S5 H, M, FA, CR distribution over sentence correct and incorrect conditions for the purpose of d'_{Stylus} and c_{Stylus} per group

This manipulation reduces the grid to the following 2x2 grid:

	<i>Stylus advice GOOD</i>	<i>Stylus advice BAD</i>
<i>Participant agrees with Stylus</i>	H	FA
<i>Participant disagrees with Stylus</i>	M	CR

Table 7.4 – S5 H, M, FA, CR distribution for the purpose of d'_{Stylus} and c_{Stylus} per group

7.3.3.3 Testing interaction models

Team sensitivity for the interaction models presented in Chapter 1 can be computed with the formulae given in Chapter 2 and the results compared with G90 and G70 participants' $d'_{Y/N}$ -scores. In order to compute d'_{team} for both groups for each of the models, the results from the CG participants were used to represent $d'_{Y/N}$. d'_{team} was computed for each GC participant for all four models, and then the mean score from all participants was used to compare each of the model predictions with the G90 and G70 Stylus-aided participants' $d'_{Y/N}$.

Because the equations used to test the models are unusual, we reprint them here for reference, now with the specific wording we use in this thesis. $d'_{Y/N} = d'_{participant}$.

Coin Flip (CF) model

$$\begin{aligned} \rho_{H-CF} &= 0.5 * (\rho_{H-participant} + \rho_{H-Stylus}) \\ \rho_{FA-CF} &= (\rho_{FA-participant}) * (\rho_{FA-Stylus} + (0.5 * \rho_{FA-participant})) * (1 - \rho_{FA-Stylus}) + (0.5 * (1 - \rho_{FA-participant})) * \rho_{FA-Stylus} \\ d'_{CF} &= Z_{\rho(H)-CF} - Z_{\rho(FA)-CF} \\ C_{CF} &= -0.5 (Z_{\rho(H)-CF} - Z_{\rho(FA)-CF}) \end{aligned}$$

Probability Matching (PM) model

R_{Stylus} is the Stylus' average reliability rate

$$\begin{aligned} \rho_{H-PM} &= R_{Stylus} * \rho_{H-Stylus} + (1 - R_{Stylus}) * \rho_{H-participant} \\ \rho_{FA-PM} &= R_{Stylus} * \rho_{FA-Stylus} + (1 - R_{Stylus}) * \rho_{FA-participant} \\ d'_{PM} &= Z_{\rho(H)-PM} - Z_{\rho(FA)-PM} \\ C_{PM} &= -0.5 * (Z_{\rho(H)-PM} - Z_{\rho(FA)-PM}) \end{aligned}$$

Optimal Weighting (OW) model

$$d'_{OW} = \text{sqrt}(d'_{participant}^2 + d'_{Stylus}^2)$$

Uniform Weighting (UW) model

$$d'_{UW} = (d'_{participant} + d'_{Stylus}) / \text{sqrt}2$$

7.3.3.4 Parametric and non-parametric measures

In this chapter we no longer report analyses with non-parametric sensitivity and bias measures, as previous studies showed no benefit in doing so.

7.3.4 Analysing confidence

Participants' confidence ratings were analysed using a 4x2 mixed repeated measures ANOVA, all factors of which are listed in Table 7.5. The between-subjects factor was Group (G90 v G70), and the repeated measures factors were Stylus advice (Good vs. Bad), Sentence correctness (correct vs. incorrect) and Response correctness (Correct vs. Incorrect). Participants' average confidence in each cell of the design was entered into the ANOVAs. Table 7.8 shows the distribution of average confidence ratings.

Confidence ANOVA	<i>Factors</i>	<i>Levels</i>
<i>Between-subjects factor</i>	Group	G90
		G70
<i>Within-subjects factors</i>	Stylus advice	GOOD
		BAD
	Sentence correctness	Sentence correct
		Sentence incorrect
	Response correctness	Responded correctly
		Responded incorrectly

Table 7.5 – S5 confidence ANOVA factors and levels

7.4 Results

The most important S5 data can be found in tabular form in Appendix A7, including breakdowns of aggregated variables. A table of all hypotheses from this thesis can be found in Appendix D.

7.4.1 Pre-task measures

7.4.1.1 Reliability testing

Reliability of prior perceived self-efficacy was $\alpha = .83$, that of prior trust $\alpha = .81$ and of the post-task variable *post-trust* lastly, it was $\alpha = .95$.

7.4.1.2 Prior perceived self-efficacy and Prior trust

Participants reported a prior perceived self-efficacy of $M = 76.96$ ($SD = 14.54$; G90, $M = 76.38$ ($SD = 11.77$); G70, $M = 78.37$ ($SD = 15.74$); GC, $M = 76.13$ ($SD = 16.11$)), and a rating of prior trust in automated suggestions of $M = 77.46$ ($SD = 14.03$; G90, $M = 77.24$ ($SD = 14.99$); G70, $M = 8.38$ ($SD = 14.86$); GC, $M = 76.75$ ($SD = 12.23$)).

7.4.2 Performance

Table 7.6 shows participants' performance in mean absolute numbers in all four categories (H, M, FA, CR) for all three groups, split into conditions of good and bad Stylus advice for the G70 and G90, as this serves as the basis for our further analyses. Table 7.7 shows the same data converted into rates, as this makes it easier to compare the raw data at a glance.

G90	<i>Stylus advice GOOD</i>		<i>Stylus advice BAD</i>	
	<i>Sentence correct</i>	<i>Sentence incorrect</i>	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(H) $M = 38.45$ ($SD = 4.83$)	(FA) $M = 9.63$ ($SD = 5.72$)	(H) $M = 3.82$ ($SD = 1.18$)	(FA) $M = 1.71$ ($SD = 1.41$)
<i>Judged incorrect</i>	(M) $M = 6.55$ ($SD = 4.83$)	(CR) $M = 35.37$ ($SD = 5.72$)	(M) $M = 1.18$ ($SD = 1.18$)	(CR) $M = 3.29$ ($SD = 1.41$)

G70	<i>Stylus advice GOOD</i>		<i>Stylus advice BAD</i>	
	<i>Sentence correct</i>	<i>Sentence incorrect</i>	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(H) $M = 29.11$ ($SD = 4.11$)	(FA) $M = 7.18$ ($SD = 4.35$)	(H) $M = 11.45$ ($SD = 2.63$)	(FA) $M = 4.39$ ($SD = 3.11$)
<i>Judged incorrect</i>	(M) $M = 5.89$ ($SD = 4.11$)	(CR) $M = 27.82$ ($SD = 4.35$)	(M) $M = 3.55$ ($SD = 2.63$)	(CR) $M = 10.61$ ($SD = 3.11$)

GC	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(H) $M = 41.63$ ($SD = 3.68$)	(FA) $M = 13.55$ ($SD = 6.16$)
<i>Judged incorrect</i>	(M) $M = 8.37$ ($SD = 3.68$)	(CR) $M = 36.45$ ($SD = 6.16$)

Table 7.6 – S5 mean absolute performance per category per group

G90	<i>Stylus advice GOOD</i>		<i>Stylus advice BAD</i>	
	<i>Sentence correct</i>	<i>Sentence incorrect</i>	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(p_H) $M = 0.85$ ($SD = 0.10$)	(p_{FA}) $M = 0.21$ ($SD = 0.13$)	(p_H) $M = 0.76$ ($SD = 0.24$)	(p_{FA}) $M = 0.34$ ($SD = 0.28$)
<i>Judged incorrect</i>	(p_M) $M = 0.15$ ($SD = 0.10$)	(p_{CR}) $M = 0.79$ ($SD = 0.13$)	(p_M) $M = 0.24$ ($SD = 0.24$)	(p_{CR}) $M = 0.66$ ($SD = 0.28$)

G70	<i>Stylus advice GOOD</i>		<i>Stylus advice BAD</i>	
	<i>Sentence correct</i>	<i>Sentence incorrect</i>	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(p_H) $M = 0.83$ ($SD = 0.12$)	(p_{FA}) $M = 0.21$ ($SD = 0.12$)	(p_H) $M = 0.76$ ($SD = 0.18$)	(p_{FA}) $M = 0.29$ ($SD = 0.21$)
<i>Judged incorrect</i>	(p_M) $M = 0.17$ ($SD = 0.12$)	(p_{CR}) $M = 0.79$ ($SD = 0.12$)	(p_M) $M = 0.24$ ($SD = 0.18$)	(p_{CR}) $M = 0.71$ ($SD = 0.21$)

GC	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(p_H) $M = 0.83$ ($SD = 0.07$)	(p_{FA}) $M = 0.27$ ($SD = 0.12$)
<i>Judged incorrect</i>	(p_M) $M = 0.17$ ($SD = 0.07$)	(p_{CR}) $M = 0.73$ ($SD = 0.12$)

Table 7.7 – S5 mean performance rates per category per group

7.4.2.1 Percentage of correct responses per group

Over the three groups, the average percentage of correct responses ($(N_H + N_{CR}) / N_{Total} * 100$) was $M = 79.32$ ($SD = 8.92$).

G90 participants' average percentage correct was 80.92 ($SD = 9.20$), for G70 it was $M = 78.97$ ($SD = 9.61$) and for GC $M = 78.08$ ($SD = 7.85$). An ANOVA showed no statistically significant difference in performance between the three groups, $F(2, 111) = 1.01$, $p = .368$, $partial \eta^2 = 0.018$.

7.4.2.2 Sensitivity, d'_{YN}

S5-H1 Participants' sensitivity will be higher when Stylus' reliability is 90% than when it is 70%.

As discussed in Chapter 2, typical d' -scores are values up to 2, with positive scores meaning participants are sensitive to telling signal from noise, and $d' = 0$ meaning participants cannot discriminate between them. The higher the d' -score, the better participants are in choosing the correct answer (i.e., either

"Yes" or "No"). The average d'_{YN} of 1.95 ($SD = 0.80$) for G90, $M = 1.80$ ($SD = 0.74$) for G70, and $M = 1.67$ ($SD = 0.57$) for GC show participants have a high ability to answer correctly in all three groups. An ANOVA did not show a statistically significant difference in d' between the groups, $F(2, 111) = 1.44$, $p = .242$, $partial \eta^2 = .025$. This means participants' sensitivity in the groups does not statistically significantly differ and S5-H1 is not supported.

7.4.2.3 Interaction models

The models of aided sensitivity listed by Bartlett and McCarley (2017, 2019). compute predicted d' for an aided participant by combining performance of the unaided participant with performance of the aid. In our case, we must use a between-groups prediction of the unaided participant, by using the data from the GC participants, as described in more detail in section 7.3.3.3 of this chapter.

In G90, Stylus' sensitivity, $d' = 2.56$, was higher than that of the participants, $d'_{YN} = 1.95$, yet in G70 participants had a better sensitivity, $d'_{YN} = 1.80$, than Stylus, $d' = 1.05$. Figure 7.2 shows, for the two groups that received assistance from Stylus, participants' and Stylus' sensitivity, and d'_{team} , the predicted combined sensitivity of the user and the aid, of the four models that were tested.

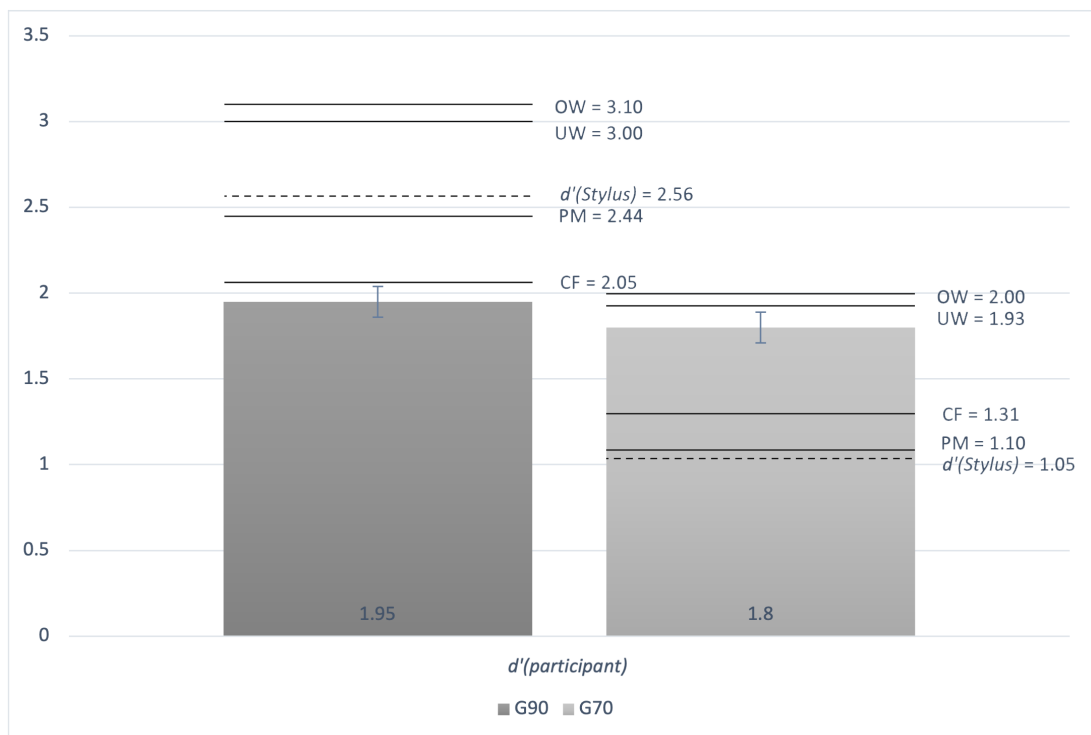


Figure 7.2 – S5 $d'_{participant}$ mean and standard error, d'_{Stylus} , and interaction models d'_{team} predictions per group. CF = Coin Flip; PM = Probability Matching; OW = Optimal Weighting; UW = Uniform Weighting.

In G90, Stylus was actually more sensitive than aided participants, whereas in G70 aided participants had the higher sensitivity. For G90, all models

predicted better performance than was observed. The CF model made the best prediction about d'_{team} with this highly reliable aid. For G70, the PM and CF models predicted lower sensitivity than was observed. The UW model predicted sensitivity best for this group, i.e., when the aid was reliable just above the reliability threshold (Wickens and Dixon 2007).

In summary, none of the models do a good job of describing or predicting aided sensitivity. OW and UW consistently predict too-good performance. CF and PM over or under-predict according to group. UW is the most accurate prediction for G70, and almost the worst for G90. CF is the most accurate for G90 and close to worst for G70. There is no clear sign in these data that any of the models is the best to pursue or develop.

Perhaps the clearest conclusion from all the comparisons implicit in Figure 7.2, is that the aid is radically underused in the G90 condition. Participants' aided sensitivity is substantially lower than the aid's sensitivity, as is anyway evident from participants' percent correct of < 90 .

7.4.2.4 Sensitivity, d'_{Stylus}

d'_{Stylus} shows the ability to correctly follow Stylus' advice, corrected for bias. The average d'_{Stylus} of 1.62 ($SD = 0.96$) for G90 and $M = 1.67$ ($SD = 0.83$) for G70 show participants have a high ability to follow Stylus correctly in all three groups. An independent samples t -test did not show a statistically significant difference in d'_{Stylus} between the groups, $t(74) = -0.22$, $p = .831$, $d = -0.049$. This means participants' sensitivity to correct Stylus judgements after a correction for bias does not statistically significantly differ between the two groups.

7.4.2.5 Bias $c_{Y/N}$: bias towards responding Yes or No

Participants' level of bias $c_{Y/N}$ shows to what extent they lean towards choosing "Yes" or choosing "No", independent of correctness of the response. The higher their $c_{Y/N}$ score, the more participants tend to choose "Yes", independent of its correctness. A low score indicates they lean towards choosing "No".

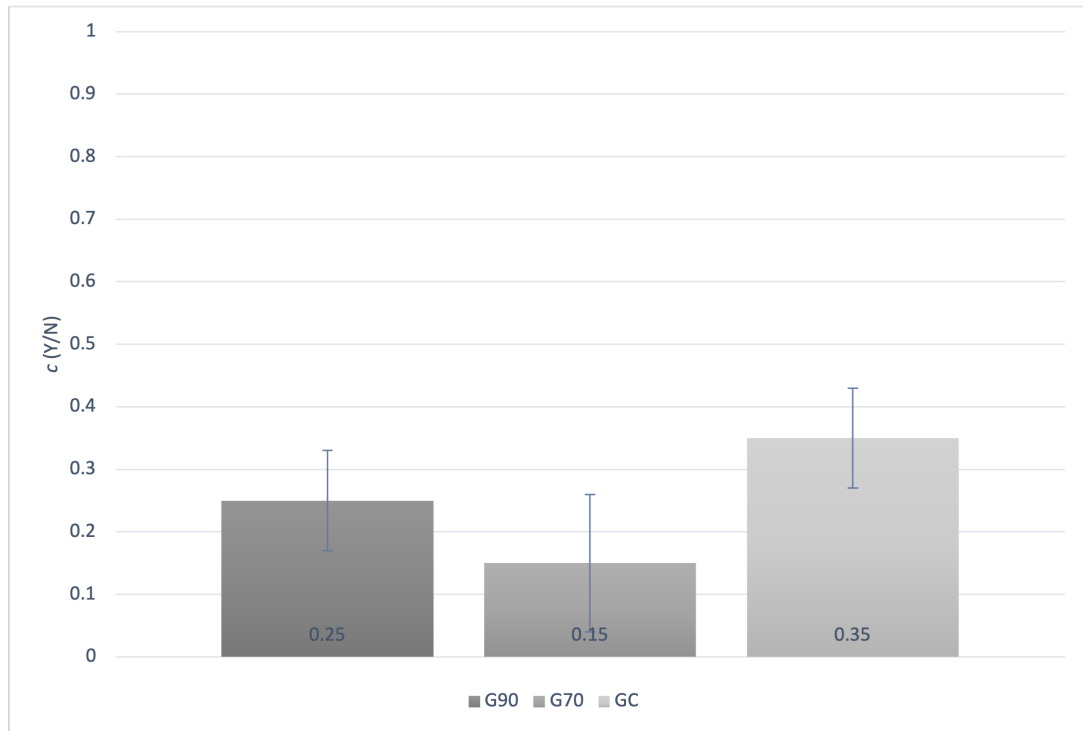


Figure 7.3 – S5 Yes/ No bias ($c_{Y/N}$), mean and standard error per group

The average $c_{Y/N}$ of 0.25 ($SD = 0.47$) in G90 shows that participants lean slightly towards responding "Yes", a one-sample t -test above 0 shows this bias to be statistically significant, $t(37) = 3.16$, $p = .003$, $d = 0.513$. In G70 however, there was no statistically significant bias towards responding "Yes", $M = 0.15$ ($SD = 0.66$), $t(37) = 1.36$, $p = .182$, $d = 0.221$. In GC lastly, there was again a statistically significant bias towards responding "Yes", $M = 0.35$ ($SD = 0.47$), $t(37) = 4.65$, $p < .001$, $d = 0.754$. An ANOVA did not show a statistically significant difference in $c_{Y/N}$ between the groups, $F(2, 111) = 1.35$, $p = .264$, $partial \eta^2 = 0.024$. A visual comparison of the $c_{Y/N}$ -scores of all three groups is shown in Figure 7.3.

The suggestion in these bias scores is, perhaps unsurprisingly, that participants will assume a sentence is correct unless they have some reason to think otherwise.

7.4.2.6 d'_{Stylus} in correct vs. incorrect sentences

Both groups that received advice from Stylus encountered fifty correct sentences and fifty incorrect ones during the experiment. G90 had an average d'_{Stylus} of 1.84 ($SD = 0.81$) for trials with a correct sentence, and $M = 1.35$ ($SD = 1.08$) in the trials with an incorrect sentence. A paired samples t -test showed a statistically significant difference between d'_{Stylus} for correct and incorrect sentences, $t(36) = -3.13$, $p = .003$, $d = 0.515$, which means that participants' sensitivity was higher when they faced a correct sentence. The situation was different for G70, with an average d'_{Stylus} of 1.80 ($SD = 0.90$) for trials with a correct sentence, and $M = 1.59$ ($SD = 1.11$) for incorrect sentences. In this group, a paired samples t -test showed no statistically significant difference

between d'_{Stylus} for correct and incorrect sentences, $t(35) = 0.927$, $p = .360$, $d = 0.154$.

7.4.2.7 Bias, c_{Stylus} : bias towards following Stylus' judgements

Participants' level of bias c_{Stylus} shows to what extent they lean towards choosing to follow Stylus, independent of the correctness of its judgements. The higher their c_{Stylus} score, the more participants tend to agree with Stylus. A low score indicates that they lean towards not following Stylus. The average c_{Stylus} of 0.34 ($SD = 0.61$) in G90 shows that participants lean slightly towards following Stylus, a one-sample t -test above 0 shows this bias to be statistically significant, $t(37) = 3.48$, $p = .001$, $d = 0.564$. In G70 there also was a statistically significant bias towards following Stylus, $M = 0.22$ ($SD = 0.41$), $t(37) = 3.37$, $p = .002$, $d = 0.547$. An independent samples t -test did not show a statistically significant difference in c_{Stylus} between the groups, $t(74) = 0.1$, $p = .321$, $d = 0.229$.

S5-H2 Participants' acceptance of Stylus suggestions will be higher when Stylus' reliability is 90% than when it is 70%, independently of correctness of the advice.

In terms of our measures, there should be a statistically significant difference between G90 and G70 in both proportion of Stylus responses, and in bias towards Stylus because of the difference in Stylus' performance level. In 90 the proportion of Stylus (H + FA) responses was $M = 0.54$ ($SD = 0.07$), and in G70 it was $M = 0.52$ ($SD = 0.09$). An independent samples t -test did not reveal a statistically significant difference between the groups, $t(74) = 0.81$, $p = .421$, $d = 0.186$, thus S5-H2 was not supported.

7.4.3 Confidence

Table 7.8 shows the average confidence in responses across the three groups, according to Stylus advice, Sentence correctness and Participant response. A graphical representation of the same data of the groups that received help from Stylus is shown in Figure 7.4 for ease of comparison. Initially the overall confidence across the three groups of the experiment was compared, with a single-factor between-groups ANOVA.

S5-H3 Participants' will be more confident when Stylus' reliability is 90% than when it is 70%.

The average self-reported confidence across the task was 91.03 ($SD = 6.98$), and it was $M = 90.93$ ($SD = 7.46$) in G90, $M = 90.68$ ($SD = 7.43$) in G70, and in GC it was $M = 91.49$ ($SD = 6.15$). The ANOVA showed no statistically significant difference in average reported confidence between the groups, $F(2, 111) = 6.42$, $p = .878$, $partial \eta^2 = 0.002$, with participants in the group that encountered the better performing version of Stylus not being statistically significantly more confident in their responses, and no noticeable difference between the groups who received Stylus suggestions and the control group, who did not. Thus S5-H3 was not supported.

Next, self-reported confidence across the task was analysed with a 2x2x2x2 mixed ANOVA, with factors of Group (G90 vs. G70), Stylus advice (Good vs. Bad), Sentence correctness (correct vs. incorrect) and Response correctness (Correct vs. Incorrect). A full ANOVA table for this analysis can be found in Appendix C7.

G90	<i>Stylus advice GOOD</i>		<i>Stylus advice BAD</i>	
	<i>Sentence correct</i>	<i>Sentence incorrect</i>	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(H) <i>M</i> = 92.28 (<i>SD</i> = 7.42)	(FA) <i>M</i> = 88.25 (<i>SD</i> = 16.46)	(H) <i>M</i> = 90.87 (<i>SD</i> = 9.54)	(FA) <i>M</i> = 91.00 (<i>SD</i> = 8.97)
<i>Judged incorrect</i>	(M) <i>M</i> = 82.49 (<i>SD</i> = 12.78)	(CR) <i>M</i> = 91.73 (<i>SD</i> = 6.63)	(M) <i>M</i> = 88.50 (<i>SD</i> = 10.95)	(CR) <i>M</i> = 91.67 (<i>SD</i> = 9.10)

G70	<i>Stylus advice GOOD</i>		<i>Stylus advice BAD</i>	
	<i>Sentence correct</i>	<i>Sentence incorrect</i>	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(H) <i>M</i> = 92.43 (<i>SD</i> = 6.97)	(FA) <i>M</i> = 89.81 (<i>SD</i> = 9.77)	(H) <i>M</i> = 91.54 (<i>SD</i> = 7.88)	(FA) <i>M</i> = 89.95 (<i>SD</i> = 9.24)
<i>Judged incorrect</i>	(M) <i>M</i> = 81.08 (<i>SD</i> = 12.34)	(CR) <i>M</i> = 92.06 (<i>SD</i> = 7.21)	(M) <i>M</i> = 82.69 (<i>SD</i> = 13.51)	(CR) <i>M</i> = 90.86 (<i>SD</i> = 7.98)

GC	<i>Sentence correct</i>	<i>Sentence incorrect</i>
<i>Judged correct</i>	(H) <i>M</i> = 92.15 (<i>SD</i> = 6.05)	(FA) <i>M</i> = 90.02 (<i>SD</i> = 6.31)
<i>Judged incorrect</i>	(M) <i>M</i> = 82.98 (<i>SD</i> = 10.67)	(CR) <i>M</i> = 92.79 (<i>SD</i> = 6.52)

Table 7.8 – S5 mean confidence per category per group

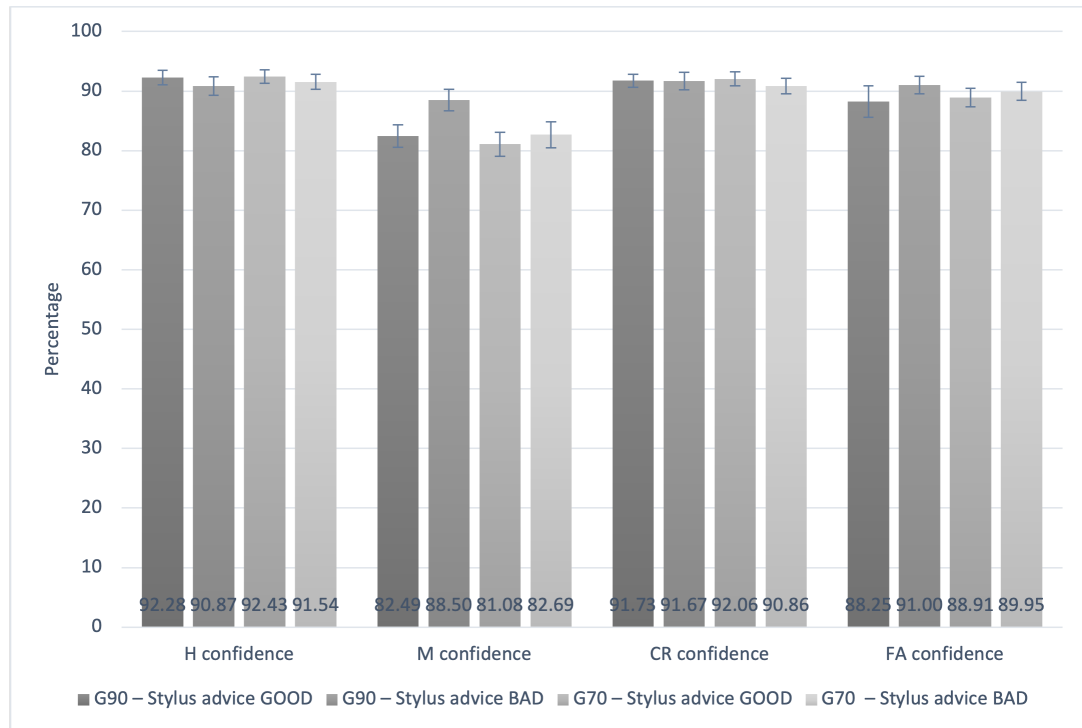


Figure 7.4 – S5 effect of Stylus advice on percentage H, M, FA and CR confidence, mean and standard error per group

The pattern of confidence judgments according to these factors can be seen in Figure 7.4, but it is somewhat complex. First, it seems clear that there is an effect of Response Correctness on confidence: Ms have lower confidence than Hs, and, but less markedly, FAs have lower confidence than CRs. What is also striking, is that Ms have the lowest confidence, in particular lower than FAs, which is an effect of Sentence Correctness. The ANOVA showed no main effect of Group, $F(1, 74) = 0.25, p = .622, \text{partial } \eta^2 = 0.002$, but all other factors had statistically significant main effects, and there were statistically significant interactions which alter the interpretation of these effects.

S5-H4 Participants will be more confident when responding correctly. The ANOVA showed a statistically significant main effect of Correctness of response, in that correct responses (G90: $M = 92.01 (SD = 6.91)$; G70: $M = 92.01 (SD = 6.83)$) were assigned more confidence than incorrect responses (G90: $86.45 (SD = 10.91)$; G70: $M = 85.52 (SD = 9.37)$), $F(1, 74) = 68.60, p < .001, \text{partial } \eta^2 = 0.110$. Thus S5-H4 was supported, which suggests that S5 participants had an awareness of their own performance, even if they were not given any feedback during the task, and that their level of performance might have positively affected their confidence. This finding is in line with the earlier S3-S8 result. There was no statistically significant interaction effect for Correctness of response x Group, $F(1, 74) = 0.40, p = .528, \text{partial } \eta^2 = 0.001$.

There was also a statistically significant main effect for Sentence correctness, with confidence being statistically significantly higher when the sentence is correct, $F(1, 74) = 25.99, p < .001, \text{partial } \eta^2 = 0.260$, but again no statistically

significant effect for Sentence correctness x Group, $F(1, 74) = 1.52, p = .222, \text{partial } \eta^2 = 0.020$.

There was a statistically significant interaction between Sentence correctness and Response correctness, which is explained by the particularly depressed confidence associated with Ms. None of the other interactions with Sentence correctness approached statistical significance.

Finally, and perhaps most interestingly, the ANOVA showed a statistically significant effect for Stylus advice on confidence, $F(1, 74) = 6.02, p = .017, \text{partial } \eta^2 = 0.075$. There also was a statistically significant interaction effect for Stylus advice x Group, $F(1, 74) = 4.21, p = .044, \text{partial } \eta^2 = 0.054$, and for Stylus advice x Response correctness, $F(1,74) = 17.96, d = < .001, \text{partial } \eta^2 = 0.009$.

Eyeballing Figure 7.4 allows some understanding of these effects. When the response is correct, confidence is slightly higher when Stylus advice is good, see Hs and CRs. But when the response is incorrect, confidence is higher when Stylus advice is bad, See Ms and FAs. Both these tendencies can be reduced to the observation that *confidence is higher when participants are accepting Stylus advice*, so giving correct answers when Stylus advice is good, and wrong answers when Stylus advice is bad.

That there is a reliable interaction between Stylus advice and Group, suggests that the boost to confidence of agreeing with Stylus is, as seems plausible, greater in G90.

The most striking finding in S5 is that confidence was higher when Stylus advice was good. No other interaction effects approached statistical significance. A full ANOVA table can be found in Appendix C7.

7.4.4 Post-task measures

7.4.4.1 Confidence

7.4.4.1.1 Post-task estimation of number correct and number Stylus responses

Participants' subjective estimation of the number of times they selected the correct answer (either "Yes" or "No") was $M = 75.27 (SD = 6.02)$. There was no statistically significant correlation between the average number of times participants thought they selected the correct answer, and the objective frequency $M = 70.79 (SD = 2.56), r(112) = .02, p = .854$. The most optimistic estimate was 33 more correct trials than actual number correct, and the most pessimistic estimate was 87 fewer correct trials than actual number.

We also split out the results per group. For G90 the scores were estimated number correct $M = 73.00 (SD = 25.76), MAX = 28, MIN = -87$. There was no statistically significant correlation between the average number of times G90

participants thought they selected the correct answer, and the objective frequency $M = 80.92$ ($SD = 9.20$), $r(36) = .14$, $p = .405$. G70 reported an estimated number correct of correct response of $M = 77.17$ ($SD = 14.49$), $MAX = 18$, $MIN = -65$. There was a statistically significant correlation between the average number of times G70 participants thought they selected the correct answer, and the objective frequency $M = 78.97$ ($SD = 9.61$), $r(36) = .38$, $p = .019$. And lastly, GC thought they responded correctly $M = 75.63$ ($SD = 16.46$), $MAX = 33$, $MIN = -54$. Also for this group there was no statistically significant correlation between the average number of times participants thought they selected the correct answer, and the objective frequency $M = 78.08$ ($SD = 7.85$), $r(36) = .18$, $p = .289$. The non-significant results are surprising given the significant findings in earlier studies, and the null-effects may simply indicate a lack of statistical power.

We also compared the number of times participants thought they chose to follow Stylus' judgement with the objective frequencies. The average estimated number of "Stylus responses" (agreeing with either Stylus flagging up an error, or it not doing so) across G90 participants was 52.45 ($SD = 33.09$), and the actual number was $M = 76.71$ ($SD = 7.41$). There was no statistically significant correlation between the average number of times participants thought they followed Stylus' judgement, and the objective frequency, $r(36) = -.002$, $p = .989$. The most optimistic estimate was 31 more Stylus trials than actual number of times they agreed with Stylus, and the most pessimistic estimate was 81 fewer Stylus trials than actual number. For G70 the average estimated number was 50.66 ($SD = 22.40$), the actual number $M = 64.87$ ($SD = 3.79$), $MAX = 32$, $MIN = -68$, $r(36) = .05$, $p = .760$.

We should note that in the total sample, nine participants had a number correct estimations that was over or below $1.5*SD$, and 18 participants in G90 and G70 had a plus or minus $1.5*SD$ number of Stylus estimations. For example, three of these participants estimated they had just a single correct response and a single Stylus response, and one estimated respectively 0 and 98, which suggested they had no idea and just indicated they did very badly, or they had not taken the task seriously in general. If we remove all results over or below $1.5*SD$, the average difference between the number of actual correct responses and the estimated number over the three groups is -0.14 ($SD = 11.01$), $MIN = -28$, $MAX = 25$, with a statistically significant correlation between the two, $r(112) = .55$, $p < .001$.

In all three groups overall single-event confidence was statistically significantly higher than warranted by participants' performance (percentage correct G90 $M = 80.92$ ($SD = 9.20$), $t(37) = -6.25$, $p < .001$, $d = -1.013$; G70 $M = 78.97$ ($SD = 9.61$), $t(37) = -8.68$, $p < .001$, $d = -1.408$; GC $M = 78.08$ ($SD = 7.85$), $t(37) = -9.74$, $p < .001$, $d = -1.579$). Paired samples t -tests show that this discrepancy disappears when we compare performance and estimated number correct (corrected for $\pm 1.5*SD$ as reported earlier), which results in statistically non-significant results for G90, $t(33) = 0.056$, $p = .956$, $d = 0.010$; and GC, $t(34) =$

0.80, $p = .430$, $d = 0.135$). In G70 however, the result was statistically significant, $t(37) = -8.68$, $p < .001$, $d = -1.408$.

The mean difference between the number of actual Stylus responses and the estimated number in G90 and G70 is 5.48 ($SD = 17.29$), $MIN = -32$, $MAX = 43$, there was no statistically significant correlation between actual and estimated number of Stylus responses at $r(74) = .20$, $p = .116$.

7.4.4.1.2 Comparing average single trial confidence, overall confidence, and performance

A 3x3 repeated measures ANOVA was performed with factors of Group (G90 vs. G70 vs. GC) and Confidence/ Performance (Confidence from trials vs. Estimated percentage correct vs. Percentage correct). The average Estimated percentage correct was corrected for $\pm 1.5*SD$ as explained in section 7.4.4.1.1, and the average Confidence from trials and Percentage correct values from participants that were rejected were removed as well, but only for this analysis.

The ANOVA shows a statistically significant effect of Confidence/ Performance, $F(2, 204) = 149.58$, $p < .001$, $partial \eta^2 = 0.595$. This confirms that there is a statistically significant difference between participants' average confidence trial-by-trial basis, their overall post-hoc confidence, and their performance. There was no statistically significant interaction effect for Confidence/ Performance x Group, $F(4, 204) = 0.95$, $p = .438$, $partial \eta^2 = 0.018$.

Figure 7.5 offers a graphical comparison between participants' average confidence from the trials, their performance measured as the percentage of correct responses, and their post-task estimated percentage of correct responses (not corrected for $\pm 1.5*SD$).

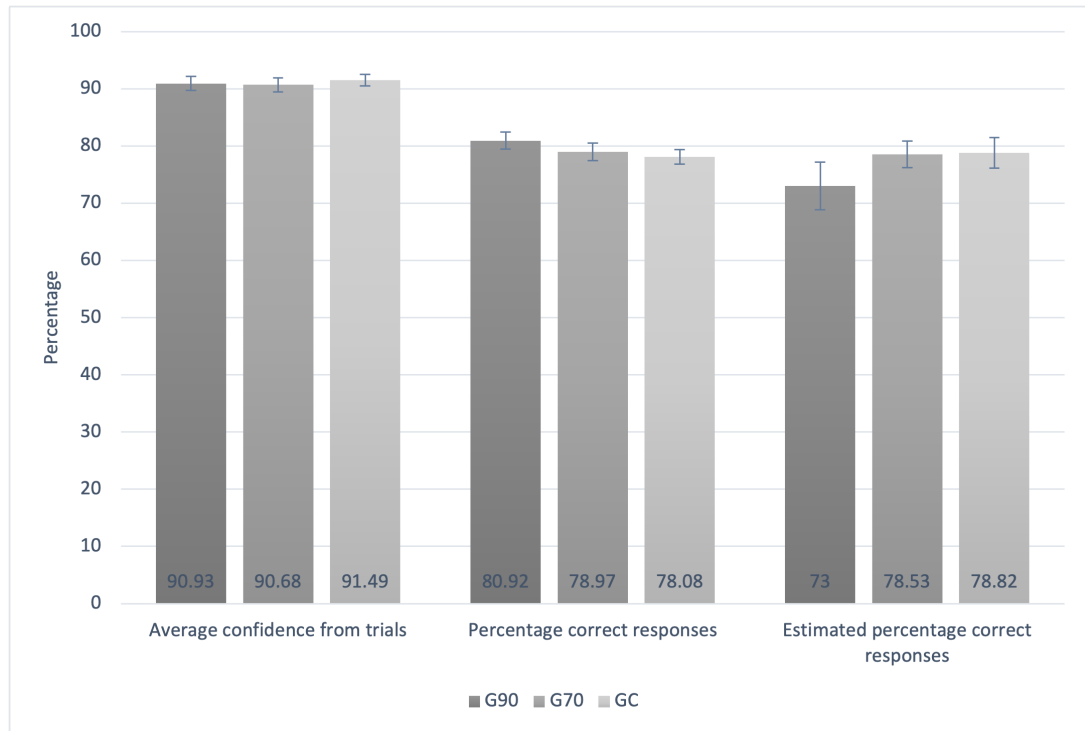


Figure 7.5 – S5 percentage average trial confidence, percentage correct responses, and overall estimated percentage correct responses (= post-task frequency confidence measure), mean and standard error per group (not corrected for $\pm SD \cdot 1.5$)

7.4.4.2 Post-task trust in Stylus

Participants' trust in Stylus after the task was $M = 60.46$ ($SD = 22.38$) in G90 and it was $M = 56.96$ ($SD = 20.64$) in G70. An independent samples t -test shows the level of post-task trust in Stylus is not statistically significantly different between G90 and G70, which suggests trust built during the task is not statistically significantly affected by the aid's performance, $t(74) = 0.71$, $p = .481$, $d = 0.163$.

7.4.4.3 Engagement with Stylus

The questions [1] "When assessing the sentences, did you remember the Stylus accuracy rate?" and [2] "In your answers, did you consider the Stylus accuracy rate?" were asked to better understand participants' engagement with the task. In G90, the average response to the first question was 75.55 ($SD = 25.32$), in G70 it was $M = 67.97$ ($SD = 31.01$). Responses to the second question were $M = 59.03$ ($SD = 31.29$) in G90, and $M = 52.58$ ($SD = 28.46$) in G70.

7.4.4.4 Believability of Stylus' accuracy (= reliability)

The average G90 response to the question "After assessing 100 sentences, in which Stylus indicated potential errors in 50, do you believe that Stylus was 90% accurate?" was 52.47 ($SD = 30.33$). G70's response to the question "After assessing 100 sentences, in which Stylus indicated potential errors in 50, do you believe that Stylus was 70% accurate?" was $M = 54.16$ ($SD = 27.40$).

7.4.4.4.1 *Believability of Stylus as an automated system*

The question “do you find it plausible that the Stylus suggestions are created by an automated system” served to confirm participants' engagement with the task; the average response was 69.05 ($SD = 17.74$) in G90, and $M = 66.16$ ($SD = 18.74$) in G70.

7.4.5 *Effects of trust, perceived self-efficacy, and system likelihood estimations*

After examining pre-task measures, performance, confidence, and post-task results individually, this section discusses how we tested a series of hypotheses across the phases of the experiment, and the conclusions we drew from these analyses.

7.4.5.1 Prior perceived self-efficacy and performance

S5-H5 Participants' prior perceived self-efficacy in the domain of writing will be positively correlated with their sensitivity, independent of Stylus' reliability. In G90, there was a statistically significant correlation between participants' prior perceived self-efficacy and their sensitivity $d'_{Y/N}$, $r(36) = .42$, $p = .009$, in G70 there also was a statistically significant correlation between prior perceived self-efficacy and $d'_{Y/N}$, $r(36) = .45$, $p = .005$, as well as in the control group, GC, $r(36) = .42$, $p = .009$. This means that S5-H5 was supported, which suggests that the higher participants' level of perceived self-efficacy was, the less likely they were to correctly follow the aid's advice, and vice versa. This is in line with what was observed earlier by among others Lee and Moray 1994, Moray et al. 1994, and Wiczorek and Meyer 2019 in other domains, but a surprising result nonetheless given that our earlier hypothesis S3-H3 and other hypotheses about effects of perceived self-efficacy (S1-H4 and S2-H4) were rejected.

S5-H6 Participants' prior perceived self-efficacy in the domain of writing will be negatively correlated with their bias, their tendency to accept Stylus advice.

In G90, there was no statistically significant correlation between participants' prior perceived self-efficacy and their bias c_{Stylus} , $r(36) = -.03$, $p = .873$, nor in G70, $r(36) = -.21$, $p = .198$. This suggests that contrary to what we hypothesised, participants' bias towards following Stylus is not statistically significantly affected by their prior perceived self-efficacy, and thus S5-H6 was not supported.

7.4.5.2 Prior trust and performance

S5-H7 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations.

The acceptance of Stylus recommendations is the degree of bias towards Stylus participants displayed. There was no statistically significant correlation between participants' prior trust and c_{Stylus} in G90, $r(36) = -.09$, $p = .593$, nor in G70, $r(36) = .12$, $p = .463$, thus S5-H7 was not supported.

7.4.5.3 Performance and post trust in Stylus

S5-H8 Participants' trust in Stylus during the experiment (measured post-task) will be positively correlated with acceptance of correct Stylus suggestions.

In G90, there was no statistically significant correlation between participants' post-task trust in Stylus and their acceptance of correct Stylus suggestions (Stylus advice is correct H + CR), $r(36) = .21$, $p = .197$. In G70 there also was no statistically significant correlation between post-task trust in Stylus and acceptance of correct Stylus suggestions, $r(36) = -.11$, $p = .525$, thus S5-Hb was rejected.

7.5 Conclusions and discussion

7.5.1 Summary of key findings

7.5.1.1 Performance

There was no statistically significant difference in percentage correct responses between groups G90, G70 and GC, nor in the sensitivity measure d'_{YN} .

In both G90 and G70 there was a statistically significant bias towards following Stylus' judgements, c_{Stylus} , but no statistically significant difference between the groups, nor between correct and incorrect sentences.

7.5.1.2 Interaction models

None of the four interaction models we tested were very accurate in their predictions, neither for G90 nor G70, and it seems participants made their judgements largely independent of Stylus' judgements. Although they took Stylus seriously as a believable automated aid, which is also supported by the responses to the post-task questions, the results suggest they somewhat ignored it. In Chapter 8 we discuss a proposal for a model that better describes interaction with Stylus.

7.5.1.3 Confidence

Participants were, on average, slightly more confident when their response was correct and Stylus advice was good than when it was bad. When their response was incorrect though, they were more confident when Stylus advice was bad. This means that, overall, confidence is higher when participants are accepting Stylus advice, and in G90 statistically significantly more so than in G70. Confidence was also statistically significantly higher in both groups for correct sentences, and when responding "Yes" rather than "No". An ANOVA also revealed statistically significant effects for Sentence correctness x Type of response, and Sentence correctness x Type of response x Correctness of

response. There was no statistically significant difference in average reported confidence between the groups.

There was no statistically significant correlation between participants' average single event confidence (probability) and their overall confidence (frequency estimation) in G90, G70 as well as in GC. We treat participants' estimation of their total number of correct responses as the overall confidence measure, and note a statistically significant discrepancy between probabilities and frequencies in all groups. We also note that there is no statistically significant difference in G90 and GC between actual and estimated performance, which means that as expected on basis of the literature (e.g., Gigerenzer 1994), the overconfidence effect disappears. In G70 there was a statistically significant difference, which we cannot currently explain.

7.5.1.4 Trust

There were no statistically significant differences between the groups in prior trust in checkers in general, nor between their post-task trust in Stylus. The fact that G70 performed better than Stylus by casting its advice aside, whereas G90 would have performed better than they did if they had just followed Stylus' advice, underlines the issue of lack of trust from different angles. In the case of G70 the low level of trust was warranted by participants' level of performance being higher than Stylus', in case of G90 this was certainly not the case.

7.5.1.5 Effects of perceived self-efficacy, trust, and sensitivity

In all three groups, there was a statistically significant correlation between participants' prior perceived self-efficacy and their sensitivity d'_{NY} ; this was the first time we observed this effect in any of our studies. Tests did reveal a statistically significant correlation between prior perceived self-efficacy and bias c_{NY} in G90, but not in G70 and GC. No statistically significant correlation between prior trust and c_{Stylus} was observed, nor between prior trust and bias.

7.5.1.6 Effects of prior trust and confidence

There was a statistically significant correlation between participants' prior trust in automated writing checkers and their confidence in correct Stylus responses in G90, but not in G70.

7.5.1.7 Effects of trust in Stylus and confidence

In both G90 and G70 there was no statistically significant correlation between participants' post-task trust in Stylus and their acceptance of correct Stylus suggestions.

7.5.1.8 Engagement with Stylus

Participants in G90 and G70 did remember, but not necessarily consider Stylus' reliability rate during the task. On average, both groups found it believable that Stylus' performance was respectively 90% and 70%, and that the Stylus suggestions were indeed created by an automated system.

Although most participants reported post-task that they did remember Stylus' reliability rate, at the same time they said they did not all necessarily consider its advice during the task. This, in conjunction with the fact that they also found it believable that the Stylus suggestions were created by an automated system, suggests that although they have taken Stylus seriously, they largely ignored or rejected it because they did not trust the likelihood of its judgements being correct (strength of the advice), and/or its performance level (reliability of the advice), which were interlinked. This is confirmed by the level of trust in Stylus they reported post-task, which is low in comparison with Stylus' likelihood estimation and performance level.

7.5.2 Discussion of method

The results of S5 suggest that participants interacted radically differently with this Y/N design than they did with the 2AFC-like design of the previous two experiments, and that Stylus in this experiment had a credibility problem because its reliability was not believed. The question is, if this was caused by the experimental design, or by Stylus' performance. To start with the latter, Stylus' performance in S3 was 53% and 75%, and it was 53% and 94% in S4; this means the average Stylus performance in G70 was similar to the average Stylus performance in S4, while in S90 Stylus performance was much higher than its overall performance in any of the previous experiments. This suggests that performance is not the reason participants did not trust Stylus enough per se.

We should note that in this experiment participants' performance was on average higher than in the previous experiments. The literature suggests that Y/N experiments usually yield higher percentage correct rates than 2AFC(-like) experiments (see e.g., Macmillan and Creelman 2005), which was one of the reasons we initially opted for a 2AFC-like design. Ceiling effects in performance can be problematic because they lead to (close to) infinite sensitivity, and lack of FAs, as observed in the S4 94 condition. Equally problematic is the opposite, inability to discriminate, and a low proportion of Hs because the task is too difficult.

7.5.3 Implications for the design of future studies

This study underlined the complexity of using SDT for the analysis of aided judgements in text editing tasks. On the one hand the previous studies showed that it is difficult to produce enough trials with a satisfactory difficulty level in a 2AFC-like design and that the number of tasks that can be presented to participants is low due to the high workload, but it served well to demonstrate differences between different Stylus performance levels in a within-subjects design. Although the Y/N design in this study was chosen because it allowed us to present a vastly greater number of trials to participants, it proved not very suitable to discriminate between good and bad Stylus advice conditions, and

between different Stylus performance levels, even though the post-task reported engagement tests suggest participants considered Stylus' advice and found its behaviour during the task believable.

In an ideal scenario, an aided text editing judgement experiment would be a within-subjects 2AFC-like design, because of its high discriminability between conditions, with a very high number of trials to warrant reliability, Stylus suggestions with likelihood estimations at any level between 50% and 100%, and items systematically rotated through noise and alarm trials and Stylus likelihood levels. Unfortunately, this ideal scenario is hardly feasible in the real world, because it requires hundreds of trials and potentially thousands of versions. Our S5 Y/N experiment confirms that, although perhaps less than ideal and with caveats, the 2AFC-like design used in S3 and S4 is a workable compromise for this type of research. However, at a minimum, the number of participants should be increased to offset the, in comparison with most classic SDT perceptual experiments, relatively low number of trials to compensate for the inherent noisiness of the experiment (see Macmillan and Creelman 2005), although this will most likely only marginally improve reliability of the findings. Another justification for the (2 or more) AFC(-like) design, is that it resembles real world conditions better than the Y/N design. In text editors the system usually does not only sound the alarm (like in a Y/N experimental design), but it often presents alternatives as well (like in an AFC design).

Chapter 8 – *Conclusions and discussion*

8.1 Introduction

It has been argued in the Human Factors literature that one plausible model for the use of automated assistance in decision making is that it is influenced by the interplay of users' trust in comparable automated aids, their perceived self-efficacy, and the confidence they have in their own decisions (Lee and Moray 1994, Moray et al. 1994, Wiczorek and Meyer 2019). This very simple model of trust and confidence operating in balance has some support in the area of perceptual decision making and simple control tasks, but it has hardly been explored at all in the domain of knowledge-dependent cognitive skills of the kind for which automated assistance is becoming more prevalent. Thus, the first and over-riding contribution of this thesis is to begin an exploration of personal beliefs in relation to performance under uncertainty (Edwards 1954), and with support from an imperfect automated aid (Wickens and Dixon 2007) in the domain of text writing and editing, and in particular spelling and grammar checking. Not only are writing and editing aids a common current example of automated assistance that perhaps represent aspects of the use of similar aids in other domains, but this application also allows for a great level of experimental flexibility and efficiency at reasonable cost.

In our experimental research, we presented participants with different versions of a notional text writing or editing aid that shares characteristics with commonly used systems such as found in among others Microsoft Word and similar word processing packages. In our first two studies (Stylus 1 (S1) and Stylus (S2)) the aid, Stylus, offers an alternative to an existing sentence in which it has detected a purported error, and participants are asked to indicate which of the two alternatives they think is better. They are also asked to rate the confidence they have in their own response. In the following two studies (S3 and S4) participants perform a similar task, and this time Stylus also gives users information about its own estimation of the likelihood of its suggestions being correct. In the final study (S5), participants' task is to indicate whether a given sentence is correct. In half of the trials Stylus indicates a homophone error in the sentence, in the other half Stylus' judgement is that there is no error. In this last experiment, Stylus' average reliability throughout the task is shared with participants prior to the task. A general characteristic of Stylus in all the studies is that its judgements are not always correct, although its level of performance varies.

S1–4 used a Two Alternative Forced Choice (2AFC)-like design, based on Signal Detection Theory (SDT, Macmillan and Creelman 2005), and S5 used an SDT Yes/ No (Y/N) SDT design. S1, 2 and 5 were between-subjects studies, whereas S3 and 4 were within-subjects.

This concluding chapter is organised as follows. In section 8.2, we summarise and discuss the most prominent innovations and findings in our five experimental studies and show how they relate to the literature in the field. In 8.3, we outline a sketch of a new model of interaction with an imperfect automated aid under uncertainty in a text editing task. In 8.4 we discuss opportunities and limitations of our methodology, some inherent to the method and some the result of our experimental design, and how it might be built upon in future work. In 8.5 we discuss potential for further research and potential for implementation of our findings, and in 8.6 lastly, we round off this thesis with a few concluding remarks.

8.2 Overview and discussion of findings

Although it is hardly contested that *trust in automation*, a factor with characteristics similar to interpersonal trust (Moray et al. 1994, Muir 1987), and the confidence factors *perceived self-efficacy* (Bandura 1997), and *confidence in one's own decisions* (Gigerenzer, Hoffrage, and Kleinbölting 1991, Gigerenzer 1991, 1994) play important roles in aided decision making tasks, it is not yet clear quite *how* these factors affect decision making, and how they interrelate. Wiczorek and Meyer's research (2019) suggests that the often-observed miscalibrations of confidence and trust in fact always mean *overconfidence* and *undertrust*, and that if two (human or automated) decision makers of different sensitivity levels collaborate, sensitivity of the joint human-machine system remains below that of the better one. A reasonably high level of trust tends to encourage the use of advice, and high levels of confidence or perceived self-efficacy tend to discourage it by obviating the perceived need for it. Our novel methodology enabled us to analyse different aspects of performance, trust, and confidence of users interacting with imperfect automated writing aids.

8.2.1 Methodological innovations

The methodological contributions of our work can be described in terms of a series of innovations. In S1–2, we introduced an SDT 2AFC-like task design to investigate user behaviour in aided text editing tasks. This allowed us to explore the effects of individual differences, as well as the effects of the *reliability*, or *accuracy*, of automated advice, by comparing performance and confidence measures of two groups of participants that used versions of an imperfect aid that performed at discretely different levels of reliability. We also

introduced a framework to measure and compare three different forms of confidence: *prior perceived self-efficacy* (pre-task), *confidence in each individual response*, and *post-task confidence*. This method not only confirmed the well-known *above average* and *overconfidence* effects, but it also provided evidence of users' awareness of the aid's and of their own performance, which are important signifiers of users' meta-cognition.

In S3–4, we introduced an aid that communicates the likelihood of its judgements being correct, which allowed us to explore the effects of individual differences as in S1–2, but also the effects of the *strength* of the aid's advice. In S5 lastly, the introduction of homophone pairs, instead of different types of error categories, made it much easier to generate large numbers of items that were reasonably believable, both in "correct" sentences and in "error" sentences. Homophone errors are necessarily context dependent, which makes this category of errors extra interesting in this type of research, because this is exactly where automated aids often underperform, usually with large proportions of false negatives as a result. With a design that focuses on this type of *real-word error* (Kukich 1992) and where users are reliant on their own knowledge, we also made an attempt to better control experimental noise.

Throughout these studies we performed some analyses of performance by partitioning the data into an SDT-style grid in which a H was a correct agreement with the automated advice, an M was an incorrect failure to accept advice, an FA was an incorrect acceptance of faulty advice and a CR was a correct rejection of faulty advice. This allowed a computation of sensitivity independent of the advice (for which we had no major hypotheses, except that it should be related to participants' perceived self-efficacy; as it transpired this relation was very weak or non-existent). It further allowed a computation of bias toward accepting the advice, separated from sensitivity. This bias index allowed us directly to test the hypothesised effects of trust, and the effects of the aid's overall accuracy on participant decision making.

8.2.2 Overview of main findings

With the first two studies we demonstrated that the influence of trust in similar systems on participants' performance during the task is smaller than we anticipated on basis of the literature (Muir 1987, Lee and Moray 1992, Moray et al. 1994, Moray and Inagaki 1999, Bisantz and Seong 2001, Dzindolet et al. 2003, Chavaillaz, Wastell, and Sauer 2016). Similarly, the influence of participants' perceived self-efficacy on their performance appears to be small, however, our results suggest it may play a more important role if the advice from the system is weak and users must be more reliant on their own knowledge. In the groups that received help from the most reliable version of Stylus, participants' level of confidence in their own responses was the highest. On average, confidence was higher in correct responses than in incorrect ones.

With our third and fourth studies, where we introduced the novel feature of the system sharing an estimate of its suggestions being correct with users, we

demonstrated in a comparative analysis that improving the reliability of the system's advice positively affects users' performance. We also demonstrated that an automated system that is very accurate still gets underused. In other words, if the system's advice is very accurate, blindly following it would benefit users more than judging the advice on a case-by-case basis, provided they recognise the systems' high level of performance.

In the last study, we presented a redesigned task with a design that is closer to how classic perceptual Signal Detection experiments are usually conducted. This design proved to confirm that participants had an awareness of Stylus, even if they largely ignored its advice. For example, participants' confidence was reliably affected by Stylus advice, and post-task they reported sufficiently high levels of awareness and consideration of Stylus' reliability during the task. This awareness suggests that ignoring Stylus is an effect of users' lack of trust in the aid, and/or of overconfidence in their own efficacy, rather than any artifact of a flaw in the experimental design.

Throughout our experiments, we confirmed two different versions of overconfidence: the first one is that people assume that their own ability is on average higher than that of others (*above average effect*, Dunning, Meyerowitz, and Holzberg 1989, Hoorens 1993), and the second is that they overestimate their performance if measured as a probability measure of confidence during a task, but that it is better calibrated if measured as a frequency post-task (*overconfidence effect*, Gigerenzer, Hoffrage, and Kleinbölting 1991, Gigerenzer 1994, Kahneman and Tversky 1996, Ayton and McClelland 1997, Gigerenzer et al. 2008).

Another characteristic that this series of experiments demonstrates, is that users can recognise how well a system is doing, even if they do not receive any feedback on the system's performance, as in S1–2. Users of a better performing system showed to be more willing to accept the aid's advice, which suggests an effect of the *reliability* of the automation's advice. In S3–4, where the system indicates the *strength* of its advice, represented by a likelihood estimation that is shared with users, this observation is confirmed. Without receiving feedback about their own performance, users also show they have an awareness of how well they themselves are doing, which is demonstrated by a higher level of confidence in correct responses than in incorrect ones.

8.2.3 Reliability of findings

Some of the findings were inconsistent between experiments. E.g., in S3 and S5 participants were statistically significantly more confident when responding correctly than when giving an incorrect response, but this was not the case in S4. Further, we found a reliable correlation between prior perceived self-efficacy and correct responses in S5 that we did not observe in S3, where this was also tested. We acknowledge that inconsistent and unusual findings, especially those that seem to clash with findings from the literature, should not be taken at face value, but researched in more depth in the future.

8.3 Models of aided performance

In the spirit of Box' adage that '*all models are wrong but some are useful*' (Box 1976, Box and Draper 1987) that was mentioned in Chapter 1, we tried to understand the potential value of the four interaction models that were described and tested by Bartlett and McCarley (2017, 2019), and tested again by us with our own data in Chapter 7. We concluded that none of the models predicted the combined sensitivity of the user and the aid very well independently of the aid's reliability. In the group that received 90% reliable advice from Stylus, the Coin Flip model's prediction was the closest to participants' aided sensitivity. The Probability Matching, Optimal Weighting, and Uniform Weighting models all predicted a d'_{team} -value that was considerably higher than that of the participants. For the group that received 70% reliable help from Stylus, all four models predicted a team sensitivity that was considerably different from participants' aided sensitivity. The CF and PM model's predictions were lower than participants' sensitivity, and the UW and OW models were overly optimistic in their forecast. As there was no statistically significant difference in participants' sensitivity between the groups, the level of success in the prediction of any of the models primarily depends on the sensitivity of the aid, assuming the models can successfully predict the interaction at all at any level of sensitivity. This limited usefulness in our cognitive task comes as no real surprise, as Bartlett and McCarley (2017, 2019) already found none of the seven models they tested to reliably predict team sensitivity in a perceptual task.

We are not aware of any other single overarching model that can describe the interactions in our experiments, and we therefore believe there is potential for a new type of model that will at be least useful to understand and predict the interaction between user and aid in a text editing task to some degree.

The aforementioned models' predictions are made up from just two factors: the user's sensitivity, and the sensitivity of the aid. The CF and PM models work by considering participant and aid agreement, they assume the participant makes independent judgments and then checks the aid's advice. In the case of agreement, the judgment is confirmed, in the case of disagreement that must be resolved. The CF model resolves the disagreement by choosing at random, the PM model by choosing according to the aid's reliability. The OW and UW models instead predict team sensitivity by combining the participant's and aid's overall sensitivity.

One limitation, therefore, of all the models is that they work at the level of aggregate data, i.e., performance aggregated across the whole set of trials. They do not really function as psychological models which predict responses to individual trials, that may then be aggregated to produce quantitative predictions. We propose that the development of such cognitive models might be a more promising avenue to better understanding and predicting aided

interaction (compare e.g., Wiegman, Rich and Zhang 2001, see Chapter 1, Section 1.2.3.4).

A full cognitive model of the decisions made by participants in the studies would of course have to include a model of the linguistic skills involved; such a model would not transfer to other aided decision-making contexts. Instead, what we have in mind is a model that, theorises *how* a participant's confidence and trust impact their use of decision aids and how they are adjusted according to experience with the aid. This is unlike the models Bartlett and McCarley described, yet in keeping with, but beyond the initial orientation of the thesis.

Rather than a statistical model such as the ones that Bartlett and McCarley (2017, 2019) described and tested and that we used in Chapter 7, the type of model we propose might, for example, look like a decision tree, or a more compact representation in the form of an influence diagram or relevance diagram (Howard 2007) that describes predictors of successful team (i.e., user + aid) performance. The model could perhaps be expressed as a combination of objective factors such as the user's and the aid's abilities, the user's personal beliefs, and chance factors that represent uncertainties such as environment, potential distractions, level of fatigue, etc. Howard mentions four pillars of decision analysis that all might contribute to drafting a new model of aided interaction in the domain of text writing and editing: Systems Analysis, Decision Theory, Epistemic Probability, and Cognitive Psychology (2007). Whether a new model of the interactions in this domain should draw on e.g., users' mental models (Payne 2003), heuristics (Kahneman et al. 1982), or formal inference rules (e.g., if > then) will require further research, and careful consideration of usefulness and relevance.

More specifically, we suggest that among the inputs to such models might be the factors of personal beliefs that we researched in our experiments. For trust in the automated aid, this may begin at a level determined by prior trust in similar systems, which we know from the literature can affect the propensity to accept advice (Li, Hess, and Valacich 2008, Manchon, Bueno, and Navarro 2021), although we did not find strong support for such an effect in our studies. Trust will also need to be adjusted according to current experience with the aid at hand, as we know from among others Moray, Lootsteen, and Pajak (1986) and Hutton and Klein (1999). Yet exactly when and how this adjustment is made remains to be investigated, as shown by our finding that reliability of the advice affects trust in Stylus (S1 and S2), yet the seemingly incongruent observation in S5 that trust in Stylus is not statistically significantly correlated with acceptance of correct Stylus suggestions. We found little effect of the confidence measure of prior perceived self-efficacy on decision-making, but weighting of advice surely must also be affected by the user's confidence in a particular judgment. It seems very likely that overconfidence, as observed in every Stylus experiment, is one of the root causes of underuse of advice, as observed in S3, 4, and 5 (also see Wiczorek and Meyer 2019). Like trust, confidence is presumably not only an input to judgments (measured as perceived self-efficacy prior to the task), but also subject to adjustment after-

the-fact, as shown by the finding that confidence can be higher when the response is consistent with the aid's advice, as shown in S4 and S5.

8.4 Discussion of method

8.4.1 The use of online surveys and crowd working platforms, and potential differences with supervisory control tasks

The online crowd working platform Prolific was successfully used for all five of our experimental studies. An advantage of using this type of platform, is having access to a demographically varied sample. The combination of a large participant pool, the relatively low cost, and incredible speed at which studies can be conducted, perhaps at the cost of uncertainties about participants' identity and lack of control over the experimental environment, justifies the use of Prolific for our specific research.

As mentioned in Chapter 1, section 1.3.4, interaction with spelling and grammar checkers can fundamentally differ in nature from assisted supervisory control tasks, in the sense that users are provided with direct feedback on their own writing to assist them with text editing, rather than them receiving feedback on an external process they are supervising. If and how potential differences between inspecting someone else's work and one's own writing might affect the interaction with the automated aid, is not something we have tested because it is outside the scope of this thesis. We would welcome research that can shine a light on potential effects though.

On balance we believe the reliability of our experimental findings will match or exceed that of research carried out with a limited number of (undergraduate) participants in a physical lab.

8.4.2 The use of quantitative methods

In this thesis we have focused on testing and developing the theory of fundamental relationships between users making uncertain language judgements, and advice they receive from an aid that is itself somewhat uncertain. We studied, among others, participants' sensitivity, bias, and confidence in an online experimental setting. Human behaviour is complex by its very nature, and research that attempts to explain isolated behavioural processes is sometimes, usually not very favourably, being called "reductionist" by some who argue that complexity is intrinsically irreducible. They do this either with more or less scientific arguments, e.g., based on the Duhem–Quine thesis and philosophers such as Leibnitz, Kant, and even Aristoteles (Orman Quine 1976), or with more ideological motives that are perhaps not very relevant in this context. Simon (1962) powerfully rebukes the allegation of "reductionism" by explaining that every system is itself both part of a larger system, as well as it having its own subsystems. I.e., every system

is part of a hierarchical structure, and therefore, it is only a matter of pragmatism to study parts of the whole in isolation. In Simon's own words: *'In the face of complexity, an in-principle reductionist may be at the same time a pragmatic holist. [...] How complex or simple a structure is depends critically upon the way in which we describe it. Most of the complex structures found in the world are enormously redundant, and we can use this redundancy to simplify their description.'* This doesn't mean it's a free-for-all though: *'But to use it, to achieve the simplification, we must find the right representation.'*

In Chapter 2, section 2.1.3, we explained that we opted to use quantitative methods in experiments with a controlled context, because they allowed us to study fundamental principles of interaction with automated aids across domains and without the context obscuring our findings. To satisfy the need for 'the right representation', we have also laid out how the methods have been used, and under what conditions. We do not mean to take a principled stance or choose a side in the debate about reductionism vs. holism, instead we believe it is much more productive to combine the strengths of lab and real-world research and of quantitative and qualitative methods, to attempt to better understand both theory and practice of interaction with automated aids. We therefore welcome future research that tests and complements our findings with qualitative methods or from different perspectives.

8.4.3 The use of Signal Detection Theory

With our experiments we have demonstrated that, overall, our analysis based on Signal Detection Theory is a viable method to research participants' sensitivity and bias, as well as different factors related to trust and confidence in user interaction with imperfect automated writing aids under uncertainty.

The main experimental challenge our work has brought to light, is the need for an experimental design with an aid that performs reasonably well (like in S4 and S5), and at the same time generates enough data points in each of the four cells of the SDT-matrix to be able to compute reliable sensitivity and bias measures. In our studies with a 2AFC-like design, the relatively small number of trials, compared with most perceptual SDT-studies, resulted in low numbers of false-positives and false-negatives, leading to missing data in some of the cells if the aid was highly reliable. With the number of trials in our studies, the need for enough data points in all four cells of the SDT-matrix assumes a design with a poorly performing aid (such as the one in S1 and S2). With S3, we demonstrated that a compromise is possible: in this experiment, where participants receive assistance from an effectively no-information aid in one condition, and one that performs just above the reliability threshold (Wickens and Dixon 2007) in the other, we created a design in which participants' interaction with a reasonable and a very poor aid could be compared.

8.4.4 Missing data and ceiling effects

The ability to generate enough data points in all cells of the SDT-matrix is a general concern in SDT-based experiments, although it is perhaps easier resolved in perceptual experiments with a large enough participant sample, a

sufficiently large number of trials, or a combination of both (Macmillan and Creelman 2005). Although our cognitive task design suffers from the same conundrum, the potential for solutions is intrinsically more limited in cognitive task research because the number of trials cannot be increased indefinitely. If the number of trials is increased, so will completion time, and we believe an increase in completion time may negatively affect attrition rates and/ or the reliability of the findings. Although we are not aware of any literature that has conclusively investigated this specific subject, we assume that experimental task load research from other domains can be generalised to our type of experiments (see e.g., Schatz, Egger, and Masuch 2012).

In general, SDT measures are most reliable with low participant number – high trial number ratios (Macmillan and Creelman 2005), but if the number of trials cannot be increased dramatically, that leaves the alternative of increasing the participant sample. However, a vast increase in the number of participants will still not be sufficient if in one hundred trials, just ten percent of the aid's suggestions are false-negatives and false-positives, such as in S5. Because in this scenario proportions of FAs and CRs will be very low by default, the reliability of sensitivity and bias measures will still be compromised. The only thing an increased number of participants will probably do, is promote regression toward the mean.

Although combatting ceiling performance with a much more difficult task to improve equal data distribution over the cells of the SDT-matrix would perhaps slightly reduce the amount of missing FA data points, it introduces two new problems in return: the difficulty to generate enough items that are difficult enough and have a level of difficulty that is known and can be controlled, and the fact that tasks that are too difficult exaggerate bias because they encourage guessing (Macmillan and Creelman 2005). It is also important to note that making the task more difficult can never be a real solution anyway, because although the number of missing FA data points may be reduced, conversely, the number of missing CR data points will increase because the two cells are effectively communicating vessels.

8.4.5 Effect of proportion of signal and noise trials on bias, and on trust in a system

The distribution between the proportions of signal and noise trials, or *likelihood ratio*, in S1–4 is asymmetric as dictated by the reliability levels of the aid. This is not ideal because it inherently introduces noise in the results (Macmillan and Creelman 2005), but we believe it is an acceptable compromise with a relatively low number of trials because it reinforces the experimental differences between the aids the two groups encounter.

8.5 Some implications of findings, and recommendations for further research and future systems development

In the introduction to this thesis, we mentioned that we initially planned to research user interface design variations of automated decision aids but decided to focus on the theory of the interaction between humans and uncertain systems first. We believe that having answered some of these fundamental questions qualifies us to make several recommendations for future research as well as for potential implications for systems design.

8.5.1 Future research

8.5.1.1 Refining interaction models

In section 8.3 we laid the foundation for an improved interaction model. It is clear that although imperfect like any such model, there is a lot of potential for further development.

8.5.1.2 Using machine learning to generate controlled trials

A difficulty we observed while designing our experiments, is the generation of sentences that are a) difficult enough, and b) have a known and controlled level of difficulty so as to generate a body of trials with enough variation to be representative of real-world text. We noticed that some of our items in S3–4 were believable as errors, but not when reversed, i.e., they worked if they were presented as a sentence where Stylus indicated an error but were not believable as sentences that Stylus suggested were OK, because they contained an error that real systems would always flag up. The use of only homophone errors in S5 solved this problem because homophone pairs are always made up of words that are intrinsically valid as words, just potentially not in a particular context (*real-word errors*, Kukich 1992). Another advantage of the use of homophones is that they are context dependent, which is exactly the area where automated writing aids usually struggle. We believe that Machine Learning (ML) could potentially be useful to identify existing sentences that contain words with homophone counterparts in heterogenous corpora, such as the British National Corpus. If a system is fed with a lexicon of homophone words (Dautriche, Fibla, Fievet and Christophe (2018) identified 10,652 unique English homophone pairs), this can be used to find and isolate sentences that contain matching words from a corpus. These sentences can then be used to generate a vast body of trials, in which correct or incorrect use of the homophone counterparts can automatically be randomly distributed and rotated. The advantage of this method would not only be that a vast number of trials can be generated with relatively little effort, but that the level of difficulty of each trial can be individually tested and classified as well (see e.g., Balyan, McCarthy, and McNamara 2020). Another advantage would be that "real world" sentences can be used in the trials, instead of sentences made up by the researchers. It seems possible that this method would make an experiment easier to set up, more realistic in terms of sentences used, and give the

researchers more control over the level of difficulty as well as the opportunity to systematically adjust the aid's sensitivity.

8.5.1.3 Individual or adjustable alarm thresholds in writing aids

Although the effects moderated by the manipulation of the aid's sensitivity in our experiments are noticeable, they are relatively modest. Where we have explored sharing likelihood information with participants on basis of randomly assigned and systematically distributed thresholds, Gadala, Strigini, and Ayton (2021) have found evidence for the use of specific alerting thresholds in specific situations. Because individual differences on the other hand are comparatively large, we believe it would be valuable to explore our methodology in conjunction with Gadala, Strigini, and Ayton's research (2021), of which we only became aware after finishing our experiments, and Meyer and Sheridan's exploration of auto-adjusting thresholds (2017). This research suggests there is merit in automated aids with individual or adjustable alerting thresholds that help moderate trust in a system, which in turn might discourage disuse (Parasuraman and Riley 1997) and positively affect performance (de Visser and Parasuraman 2011, Parasuraman and Manzey 2010). What we envision, is an aid that communicates its estimate of the likelihood of its performance being correct as in our S3 and S4 studies, but with a dynamic alerting threshold, i.e., an M/ FA-ratio that can be adjusted based on users' performance.

8.5.1.4 Better understanding reliability and strength of advice

In S1 and S2 there were clear effects of the reliability of the aid. In S3 and S4, where Stylus also shared with users an estimation of the likelihood of its suggestions being correct, we cannot be sure if the findings should be attributed to the *reliability* of the aid (i.e., its performance level), or the *strength* of its advice (i.e., its own estimation of its level of performance as communicated to users), because the two were statistically matched, as explained in Chapter 5, section 5.1.3. We would welcome experimental research that seeks to differentiate effects of the aid's reliability from effects of the strength of its advice. A good start, we imagine, would be to run two versions of the same experiment in parallel, where the aids have identical reliability, but one comes with, and the other without likelihood estimation labels.

8.5.2 Future systems

We believe that some of our findings may be generalised to knowledge-dependent automation-aided human decision-making under uncertainty in other domains, very much like Gadala, Strigini, and Ayton (2021) suggest with a comparison between behaviour patterns in spell-checking and mammography reading tasks. This may potentially result in interesting opportunities for developers of interactive sociotechnical systems to test in their applications. Other domains in which some of our findings may be applied, are for example financial decision making (e.g., comparing insurance

quotes), wayfinding (e.g., comparing travel distance and travel time), judging CVs, judging and editing automated translations, and checking automated coursework marking. In all these domains AI systems are currently being used to aid human decision makers, and the issues addressed in this thesis are therefore relevant in principle.

However, it is presently unclear how much the confidence and trust effects we observed will be generalised across the whole class of technology, so that prior experience with similar aids might colour judgments of any new one. It is also unknown to what extent trust will be undermined by any error on behalf of the automated aid; perhaps imperfect performance will discourage use disproportionately as suggested in some of the literature (e.g., Dzindolet et al. 2003, Wickens et al. 2021), in which case aids might be much less useful in practice than in theory. It seems plausible a priori that overconfidence effects, if they exist, will lead to the underuse of aids, compared with what would be optimal for aided performance. Lastly, potential distortions, and factors not present in spelling and grammar checking tasks, such as concurrent task-load and the requirement to observe operational safety protocols, could change the pattern of aided behaviour.

Although some aspects of our experimental design were different from users interacting with real aids, we believe our research shows, at a minimum, potential for the application of system likelihood information communication in spelling and grammar checkers. Although we currently lack the opportunity to attempt to implement such a feature in a real-world system, we suggest that there may be potential to explore the scope for integration with open-source word processors like Apache Open Office, even if only for more in-depth testing purposes.

8.6 Concluding remarks

In this thesis we set out to contribute to better understanding complex cognitive tasks of interaction with imperfect automated aids in the domain of spelling and grammar checking, by developing a novel experimental paradigm based on Signal Detection Theory. In a series of five closely related experimental studies where participants had to judge sentences with assistance from a purported automated aid, we tested hypotheses around effects of performance, trust, and confidence. The use of the SDT-construct of bias, as a measure of users' propensity to accept suggestions from an automated aid, is one of the major novel methodological contributions of the thesis.

Our innovative research has successfully demonstrated opportunities and limitations of using Signal Detection Theory in the context of a judgement task to research effects of performance, trust, and confidence in aided cognitive

tasks, that will be valuable for future research, especially in combination with other research that explores individual or adjustable alerting thresholds.

Although we made a few important methodological and empirical contributions, more research is of course required to fully understand all the interactions that play a role in cognitive processes in aided interaction. We therefore welcome future research that builds upon our work, by testing it with quantitative as well as with qualitative methods, and by exploring how the opportunities and limitations of the use of the Signal Detection Theory methods we identified for research of cognitive tasks translate to other domains. We also hope our research inspires developers and designers of automated aids in interactive systems to implement and test design interventions based on our findings.

Bibliography

- 2018a. "The Guardian." <https://www.theguardian.com/uk>.
- 2018b. "Retro Rides." <http://retrorides.proboards.com/>.
- Abdi, Hervé. 2007. "Signal detection theory (SDT)." *Encyclopedia of measurement and statistics*:886-889.
- Agarwal, Harish, John E. Renaud, Evan L. Preston, and Dhanesh Padmanabhan. 2004. "Uncertainty quantification using evidence theory in multidisciplinary design optimization." *Reliability Engineering & System Safety* 85 (1):281-294. doi: <https://doi.org/10.1016/j.ress.2004.03.017>.
- Alberdi, EPA, L Strigini, and P Ayton. 2010. "CAD: risks and benefits for radiologists' decision." *The handbook of medical image perception and techniques*:326-330.
- Allwood, Carl Martin, and Henry Montgomery. 1987. "Response selection strategies and realism of confidence judgments." *Organizational Behavior and Human Decision Processes* 39 (3):365-383.
- Ayton, Peter, and Alastair G. R. McClelland. 1997. "How Real is Overconfidence?" *Journal of Behavioral Decision Making* 10 (3):279-285. doi: [https://doi.org/10.1002/\(SICI\)1099-0771\(199709\)10:3<279::AID-BDM280>3.0.CO;2-N](https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<279::AID-BDM280>3.0.CO;2-N).
- Bahrami, Bahador, Karsten Olsen, Peter E Latham, Andreas Roepstorff, Geraint Rees, and Chris D Frith. 2010. "Optimally interacting minds." *Science* 329 (5995):1081-1085.
- Balfe, Nora, John R. Wilson, Sarah Sharples, and Theresa Clarke. 2012. "Development of design principles for automated systems in transport control." *Ergonomics* 55 (1):37-54. doi: [10.1080/00140139.2011.636456](https://doi.org/10.1080/00140139.2011.636456).
- Balyan, Renu, Kathryn S McCarthy, and Danielle S McNamara. 2020. "Applying natural language processing and hierarchical machine learning approaches to text difficulty classification." *International Journal of Artificial Intelligence in Education* 30 (3):337-370.
- Bandura, Albert. 1984. "Recycling misconceptions of perceived self-efficacy." *Cognitive therapy and research* 8 (3):231-255.
- Bandura, Albert. 1997. *Self-efficacy : the exercise of control*. New York: New York : W. H. Freeman.
- Bandura, Albert. 2006. "Guide for constructing self-efficacy scales." *Self-efficacy beliefs of adolescents* 5 (1):307-337.
- Barber, Bernard. 1983. *The logic and limits of trust*. New Brunswick, N.J.: New Brunswick, N.J. : Rutgers University Press.
- Bartlett, Megan L, and Jason S McCarley. 2017. "Benchmarking Aided Decision Making in a Signal Detection Task." *Human Factors: The*

- Journal of Human Factors and Ergonomics Society* 59 (6):881-900. doi: 10.1177/0018720817700258.
- Bartlett, Megan L, and Jason S McCarley. 2019. "No effect of cue format on automation dependence in an aided signal detection task." *Human Factors* 61 (2):169-190.
- Bartlett, Megan L, and Jason S McCarley. 2021. "Ironic efficiency in automation-aided signal detection." *Ergonomics* 64 (1):103-112.
- Begoli, Edmon, Tanmoy Bhattacharya, and Dimitri Kusnezov. 2019. "The need for uncertainty quantification in machine-assisted medical decision making." *Nature Machine Intelligence* 1 (1):20-23. doi: 10.1038/s42256-018-0004-1.
- Birn, Juhani. 2000. "Detecting grammar errors with Lingsoft's Swedish grammar checker." Proceedings of the 12th Nordic Conference of Computational Linguistics (NODALIDA 1999).
- Bisantz, Ann M, and Younho Seong. 2001. "Assessment of operator trust in and utilization of automated decision-aids under different framing conditions." *International Journal of Industrial Ergonomics* 28 (2):85-97.
- Bliss, James P, Richard D Gilson, and John E Deaton. 1995. "Human probability matching behaviour in response to alarms of varying reliability." *Ergonomics* 38 (11):2300-2312.
- Box, George EP. 1976. "Science and statistics." *Journal of the American Statistical Association* 71 (356):791-799.
- Box, George EP, and Norman R Draper. 1987. *Empirical model-building and response surfaces*: John Wiley & Sons.
- Brann, David B, David A Thurman, and Christine M Mitchell. 1996. "Human interaction with lights-out automation: A field study." Proceedings Third Annual Symposium on Human Interaction with Complex Systems. HICS'96.
- Brehmer, Berndt, and Dietrich Dörner. 1993. "Experiments with computer-simulated microworlds: escaping both the narrow straits of the laboratory and the deep blue sea of the field study." *Computers in Human Behavior* 9 (2-3):171-184. doi: [http://dx.doi.org/10.1016/0747-5632\(93\)90005-D](http://dx.doi.org/10.1016/0747-5632(93)90005-D).
- Breznitz, Shlomo. 2013. *Cry wolf: The psychology of false alarms*: Psychology Press.
- Carroll, J. M., and D. D. Reese. 2003. Community collective efficacy: structure and consequences of perceived capacities in the Blacksburg Electronic Village. USA.
- Cattell, Raymond B. 1957. "Personality and motivation structure and measurement."
- Cattell, Raymond B, and Paul Ed Kline. 1977. *The scientific analysis of personality and motivation*: Academic Press.
- Chavaillaz, Alain, David Wastell, and Jürgen Sauer. 2016. "System reliability, performance and trust in adaptable automation." *Applied Ergonomics* 52:333-342. doi: 10.1016/j.apergo.2015.07.012.
- Cox, Michael T. 2005. "Metacognition in computation: A selected research review." *Artificial intelligence* 169 (2):104-141.

- Dautriche, Isabelle, Laia Fibla, Anne-Caroline Fievet, and Anne Christophe. 2018. "Learning homophones in context: Easy cases are favored in the lexicon of natural languages." *Cognitive Psychology* 104:83-105. doi: <https://doi.org/10.1016/j.cogpsych.2018.04.001>.
- Davis, James H. 1992. "Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples 1950–1990." *Organizational Behavior and Human Decision Processes* 52 (1):3-38.
- Davis, Joshua M, and Brad M Tuttle. 2013. "A heuristic–systematic model of end-user information processing when encountering IS exceptions." *Information & Management* 50 (2-3):125-133.
- De Visser, Ewart, and Raja Parasuraman. 2011. "Adaptive Aiding of Human-Robot Teaming: Effects of Imperfect Automation on Performance, Trust, and Workload." *Journal of Cognitive Engineering and Decision Making* 5 (2):209-231. doi: 10.1177/1555343411410160.
- De Vries, Peter, Cees Midden, and Don Bouwhuis. 2003. "The effects of errors on system trust, self-confidence, and the allocation of control in route planning." *International Journal of Human-Computer Studies* 58 (6):719-735.
- DeJoy, David M. 1989. "The optimism bias and traffic accident risk perception." *Accident Analysis & Prevention* 21 (4):333-340.
- Dixon, Stephen R, and Christopher D Wickens. 2006. "Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload." *Human factors* 48 (3):474-486.
- Donders, A Rogier T, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. 2006. "A gentle introduction to imputation of missing values." *Journal of clinical epidemiology* 59 (10):1087-1091.
- Doran, Derek, Sarah Schulz, and Tarek R Besold. 2017. "What does explainable AI really mean? A new conceptualization of perspectives." *arXiv preprint arXiv:1710.00794*.
- Douer, Nir, and Joachim Meyer. 2019. Theoretical, Measured and Subjective Responsibility in Aided Decision Making. *arXiv e-prints*. Accessed April 01, 2019.
- Dourish, Paul. 2001. *Where the action is: the foundations of embodied interaction*. Cambridge, Mass.
- Duncan-Reid, Jackson, and Jason S. McCarley. 2021. "Strategy Use in Automation-Aided Decision Making." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65 (1):96-100. doi: 10.1177/1071181321651259.
- Dunning, David. 2011. "The Dunning–Kruger effect: On being ignorant of one's own ignorance." In *Advances in experimental social psychology*, 247-296. Elsevier.
- Dunning, David, Judith A Meyerowitz, and Amy D Holzberg. 1989. "Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability." *Journal of personality and social psychology* 57 (6):1082.

- Dzindolet, Mary T, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 1999. "Misuse and disuse of automated aids." Proceedings of the Human Factors and Ergonomics Society Annual Meeting.
- Dzindolet, Mary T, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. 2002. "The perceived utility of human and automated aids in a visual detection task." *Human Factors* 44 (1):79-94.
- Dzindolet, Mary T., Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. "The role of trust in automation reliance." *International journal of human-computer studies* 58 (6):697-718. doi: 10.1016/S1071-5819(03)00038-7.
- Edwards, Ward. 1954. "The theory of decision making." *Psychological bulletin* 51 (4):380.
- Elvers, Greg C, and Paul Elrif. 1997. "The effects of correlation and response bias in alerted monitor displays." *Human factors* 39 (4):570-580.
- Evans, Joel R, and Anil Mathur. 2005. "The value of online surveys." *Internet research*.
- Fantino, Edmund, and Ali Esfandiari. 2002. "Probability matching: Encouraging optimal responding in humans." *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 56 (1):58-63. doi: 10.1037/h0087385.
- Figueredo, Lauren, and Connie K Varnhagen. 2004. "Detecting a problem is half the battle: The relation between error type and spelling performance." *Scientific Studies of Reading* 8 (4):337-356.
- Figueredo, Lauren, and Connie K. Varnhagen. 2005. "Didn't You Run the Spell Checker? Effects of Type of Spelling Error and Use of a Spell Checker on Perceptions of the Author." *Reading Psychology* 26 (4-5):441-458. doi: 10.1080/02702710500400495.
- Fischer, F William, Donald Shankweiler, and Isabelle Y Liberman. 1985. "Spelling proficiency and sensitivity to word structure." *Journal of memory and language* 24 (4):423-441.
- Fleming, Stephen M., and Hakwan C. Lau. 2014. "How to measure metacognition." *Frontiers in Human Neuroscience* 8 (443). doi: 10.3389/fnhum.2014.00443.
- Gadala, Marwa, Lorenzo Strigini, and Peter Ayton. 2021. "Improving Human Decisions by Adjusting the Alerting Thresholds for Computer Alerting Tools According to User and Task Characteristics." *arXiv preprint arXiv:2106.13544*.
- Galletta, Dennis, Alexandra Durcikova, Andrea Everard, and Brian Jones. 2005. "Does spell- checking software need a warning label?" *Communications of the ACM* 48 (7):82-86. doi: 10.1145/1070838.1070841.
- Garcia-Retamero, Rocio, Edward T. Cokely, and Ulrich Hoffrage. 2015. "Visual aids improve diagnostic inferences and metacognitive judgment calibration." *Frontiers in Psychology* 6 (932). doi: 10.3389/fpsyg.2015.00932.
- Geels-Blair, Kasha, Stephen Rice, and Jeremy Schwark. 2013. "Using system-wide trust theory to reveal the contagion effects of automation false alarms and misses on compliance and reliance in a simulated

- aviation task." *The International Journal of Aviation Psychology* 23 (3):245-266.
- Ghanem, Roger, David Higdon, and Houman Owhadi. 2017. *Handbook of uncertainty quantification*. Vol. 6: Springer.
- Gigerenzer, Gerd. 1991. "How to make cognitive illusions disappear: Beyond "heuristics and biases"." *European review of social psychology* 2 (1):83-115.
- Gigerenzer, Gerd. 1994. "Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa)." In *Subjective probability.*, 129-161. Oxford, England: John Wiley & Sons.
- Gigerenzer, Gerd, Ralph Hertwig, Ulrich Hoffrage, and Peter Sedlmeier. 2008. "Cognitive illusions reconsidered." *Handbook of experimental economics results* 1:1018-1034.
- Gigerenzer, Gerd, Ulrich Hoffrage, and Heinz Kleinbölting. 1991. "Probabilistic mental models: a Brunswikian theory of confidence." *Psychological Review* 98:506-529. doi: 10.1037/0033-295X.98.4.506.
- Gignac, Gilles E, and Marcin Zajenkowski. 2020. "The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data." *Intelligence* 80:101449.
- Glikson, Ella, and Anita Williams Woolley. 2020. "Human trust in artificial intelligence: Review of empirical research." *Academy of Management Annals* 14 (2):627-660.
- Graber, Mark L, and Eta S Berner. 2008. *Diagnostic error: is overconfidence the problem?*: Excerpta Medica.
- Green, David Marvin, and John A Swets. 1966. *Signal detection theory and psychophysics*. Vol. 1: Wiley New York.
- Greenbaum, Joan, and Morten Kyng. 1991. *Design at work: cooperative design of computer systems*. Hillsdage, N.J.: Lawrence Erlbaum Associates.
- Guznov, Svyatoslav, Joseph Lyons, Alexander Nelson, and Montana Woolley. 2016. "The Effects of Automation Error Types on Operators' Trust and Reliance." Cham.
- Hall, Crystal C., Lynn Ariss, and Alexander Todorov. 2007. "The illusion of knowledge: When more information reduces accuracy and increases confidence." *Organizational Behavior and Human Decision Processes* 103 (2):277-290. doi: <https://doi.org/10.1016/j.obhdp.2007.01.003>.
- Hautus, MJ, D Van Hout, and H-S Lee. 2009. "Variants of A Not-A and 2AFC tests: Signal detection theory models." *Food Quality and Preference* 20 (3):222-229.
- Heath, Christian, and Paul Luff. 2000. *Technology in action, Learning in doing*. Cambridge: Cambridge University Press.
- Heath, Chip, and Amos Tversky. 1991. "Preference and belief: Ambiguity and competence in choice under uncertainty." *Journal of risk and uncertainty* 4 (1):5-28.
- Hoorens, Vera. 1993. "Self-enhancement and Superiority Biases in Social Comparison." *European Review of Social Psychology* 4 (1):113-139. doi: 10.1080/14792779343000040.

- Howard, Ronald Arthur. 2007. *The foundations of decision analysis revisited*: Citeseer.
- Humphreys, Lloyd G. 1939. "Acquisition and extinction of verbal expectations in a situation analogous to conditioning." *Journal of Experimental Psychology* 25 (3):294.
- Hutchins, Edwin. 1995. "How a cockpit remembers its speeds." *Cognitive Science* 19 (3):265-288. doi: 10.1207/s15516709cog1903_1.
- Hutchinson, Jack, Luke Strickland, Simon Farrell, and Shayne Loft. 2022. "The Perception of Automation Reliability and Acceptance of Automated Advice." *Human Factors* 0 (0):00187208211062985. doi: 10.1177/00187208211062985.
- Hutton, R. J. B., and G. Klein. 1999. "Expert decision making." *Systems Engineering* 2 (1):32-45.
- Johnson, Edgar M, Raymond C Cavanagh, Ronald L Spooner, and Michael G Samet. 1973. "Utilization of reliability measurements in Bayesian inference: Models and human performance." *IEEE Transactions on Reliability* 22 (3):176-183.
- Kahneman, Daniel, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*: Cambridge university press.
- Kahneman, Daniel, and Amos Tversky. 1973. "On the psychology of prediction." *Psychological review* 80 (4):237.
- Kahneman, Daniel, and Amos Tversky. 1996. "On the reality of cognitive illusions." *Psychological Review* 103 (3):582-591. doi: 10.1037/0033-295X.103.3.582.
- Keynes, John Maynard. 1921. *A treatise on probability*: Macmillan and Company, limited.
- Kidwell, Brian, Gloria L Calhoun, Heath A Ruff, and Raja Parasuraman. 2012. "Adaptable and adaptive automation for supervisory control of multiple autonomous vehicles." Proceedings of the human factors and Ergonomics society annual meeting.
- King, G. 2009a. *Collins Improve Your Grammar*: HarperCollins Publishers Limited.
- King, G. 2009b. *Collins Improve Your Punctuation*: Collins.
- King, G. 2009c. *Collins Improve Your Writing Skills*: Collins.
- Kline, Paul. 2000. *A psychometrics primer*: free Assn books.
- Knight, Frank Hyneman. 1921. *Risk, uncertainty and profit*. Vol. 31: Houghton Mifflin.
- Koehler, Derek J, and Greta James. 2009. "Probability matching in choice under uncertainty: Intuition versus deliberation." *Cognition* 113 (1):123-127.
- Kubovy, Michael, Amnon Rapoport, and Amos Tversky. 1971. "Deterministic vs probabilistic strategies in detection." *Perception & Psychophysics* 9 (5):427-429.
- Kukich, Karen. 1992. "Techniques for automatically correcting words in text." *Acm Computing Surveys (CSUR)* 24 (4):377-439.

- Lee, John D. 2008. "Review of a Pivotal Human Factors Article: "Humans and Automation: Use, Misuse, Disuse, Abuse"." *Human factors* 50 (3):404-410. doi: 10.1518/001872008X288547.
- Lee, John D., and Neville Moray. 1994. "Trust, self-confidence, and operators' adaptation to automation." *International Journal of Human - Computer Studies* 40 (1):153-184. doi: 10.1006/ijhc.1994.1007.
- Lee, John D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors: The Journal of Human Factors and Ergonomics Society* 46 (1):50-80. doi: 10.1518/hfes.46.1.50_30392.
- Lee, John, and Neville Moray. 1992. "Trust, control strategies and allocation of function in human-machine systems." *Ergonomics* 35 (10):1243-1270. doi: 10.1080/00140139208967392.
- Lefever, Samuel, Michael Dal, and Ásrún Matthíasdóttir. 2007. "Online data collection in academic research: advantages and limitations." *British Journal of Educational Technology* 38 (4):574-582.
- lexico.com. <https://www.lexico.com/grammar/commonly-confused-words>.
- Li, Xin, Traci J. Hess, and Joseph S. Valacich. 2008. "Why do we trust new technology? A study of initial trust formation with organizational information systems." *The journal of strategic information systems* 17 (1):39-71. doi: 10.1016/j.jsis.2008.01.001.
- Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D Phillips. 1977. "Calibration of probabilities: The state of the art." *Decision making and change in human affairs*:275-324.
- Lin, Po-Han. 2017. "Effects of spell checkers on English as a second language students' incidental spelling learning: a cognitive load perspective." *Reading & Writing* 30 (7):1501-1526. doi: 10.1007/s11145-017-9734-4.
- Lumsden, Jim. 2018. "What are the advantages and limitations of an online sample?", accessed 2 November 2018. <http://help.prolific.ac/prolific-s-best-practice-guide/statistical-and-methodological-concepts/what-are-the-advantages-and-limitations-of-an-online-sample>.
- MacArthur, Charles A, Steve Graham, Jacqueline B Haynes, and Susan DeLaPaz. 1996. "Spelling checkers and students with learning disabilities: Performance comparisons and impact on spelling." *The Journal of Special Education* 30 (1):35-57.
- Macmillan, N.A., and C.D. Creelman. 2005. *Detection Theory: A User's Guide*: Lawrence Erlbaum Associates.
- Madhavan, Poornima, Douglas A Wiegmann, and Frank C Lacson. 2006. "Automation failures on tasks easily performed by operators undermine trust in automated aids." *Human factors* 48 (2):241-256.
- Major, David. 2010. "How Computer Editing Responds to Types of Writing Errors." *Issues in Writing* 18 (2):146-167.
- Manchon, JB, Mercedes Bueno, and Jordan Navarro. 2021. "Automation, expert systems calibration of trust in automated driving: a matter of initial level of trust and automated driving style?" *Human factors*:00187208211052804.
- Manzey, Dietrich, Nina Gérard, and Rebecca Wiczorek. 2014. "Decision-making and response strategies in interaction with alarms: the impact

- of alarm reliability, availability of alarm validity information and workload." *Ergonomics* 57 (12):1833-1855.
- Mayer, Roger C. 1995. "An integrative model of organizational trust." *Academy of Management Review* 20 (3):709-735. doi: 10.5465/AMR.1995.9508080335.
- McDonnell, John D. 1969. "An application of measurement methods to improve the quantitative nature of pilot rating scales." *IEEE Transactions on man-machine systems* 10 (3):81-92.
- McGuirl, John M., and Nadine B. Sarter. 2006. "Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information." *Human Factors: The Journal of Human Factors and Ergonomics Society* 48 (4):656-665. doi: 10.1518/001872006779166334.
- Meyer, J., and J. K. Kuchar. 2021. "Maximal benefits and possible detrimental effects of binary decision aids." 2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS), 8-10 Sept. 2021.
- Meyer, Joachim, and Yuval Bitan. 2002. "Why better operators receive worse warnings." *Human Factors* 44 (3):343-353.
- Meyer, Joachim, and Thomas B Sheridan. 2017. "The intricacies of user adjustments of alerting thresholds." *Human factors* 59 (6):901-910.
- Moore, Don A., and Paul J. Healy. 2008. "The trouble with overconfidence." *Psychological Review* 115 (2):502-517. doi: 10.1037/0033-295X.115.2.502.
- Moray, Neville, Douglas Hiskes, John Lee, and Bonnie M. Muir. 1994. *Trust and human intervention in automated systems*.
- Moray, Neville, and Toshiyuki Inagaki. 1999. "Laboratory studies of trust between humans and machines in automated systems." *Transactions of the Institute of Measurement and Control* 21 (4):203-211.
- Moray, Neville, Pam Lootsteen, and Jan Pajak. 1986. "Acquisition of Process Control Skills." *Systems, Man and Cybernetics, IEEE Transactions on* 16 (4):497-504. doi: 10.1109/TSMC.1986.289252.
- Moray, Neville, Penelope M Sanderson, and Kim J Vicente. 1992. "Cognitive task analysis of a complex work domain: A case study." *Reliability Engineering & System Safety* 36 (3):207-216.
- Mousavi, Shabnam, and Gerd Gigerenzer. 2014. "Risk, uncertainty, and heuristics." *Journal of Business Research* 67 (8):1671-1678.
- Mousavi, Shabnam, and Gerd Gigerenzer. 2017. "Heuristics are tools for uncertainty." *Homo Oeconomicus* 34 (4):361-379.
- Muir, Bonnie M. 1987. "Trust between humans and machines, and the design of decision aids." *International Journal of Man-Machine Studies* 27 (5):527-539. doi: 10.1016/S0020-7373(87)80013-5.
- Muir, BM. 1989. "Operators' trust in and percentage of time spent using the automatic controllers in supervisory process control task (Doctoral dissertation, University of Toronto, 1989)."
- Nayak, M Siva Durga Prasad, and KA Narayan. 2019. "Strengths and weaknesses of online surveys." *technology* 6:7.

- Newstead, S. K. 1986. *Human assessment : cognition and motivation*. Edited by S. H. Irvine and P. L. Dann. 1st ed. 1986. ed. Dordrecht: Dordrecht : Springer.
- Norman, Donald A. 2002. *The design of everyday things*. New York: Basic Books.
- Norman, Donald A. 2011. *Living with complexity*. Cambridge, Mass.: MIT Press.
- Nowotny, Helga. 2021. *IN AI WE TRUST: power, illusion and control of predictive algorithms*: John Wiley & Sons.
- Orman Quine, Willard van. 1976. "Two dogmas of empiricism." In *Can theories be refuted?*, 41-64. Springer.
- Palan, Stefan, and Christian Schitter. 2018. "Prolific.ac—A subject pool for online experiments." *Journal of Behavioral and Experimental Finance* 17:22-27. doi: <https://doi.org/10.1016/j.jbef.2017.12.004>.
- Parasuraman, Raja. 2000. "Designing automation for human use: empirical studies and quantitative models." *Ergonomics* 43 (7):931-951.
- Parasuraman, Raja, and Dietrich H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52 (3):381-410. doi: 10.1177/0018720810376055.
- Parasuraman, Raja, and Victor Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39 (2):230-253. doi: 10.1518/001872097778543886.
- Parasuraman, Raja, Thomas B Sheridan, and Christopher D Wickens. 2000. "A model for types and levels of human interaction with automation." *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans* 30 (3):286-297.
- Parasuraman, Raja, and Christopher D Wickens. 2008. "Humans: still vital after all these years of automation." *Human factors* 50 (3):511-520.
- Parent, Kevin. 2012. "The most frequent English homonyms." *RELC Journal* 43 (1):69-81.
- Pastore, Richard E., Edward J. Crawley, Melody S. Berens, and Michael A. Skelly. 2003. "'Nonparametric' A' and other modern misconceptions about signal detection theory." *Psychonomic bulletin & review* 10 (3):556-569. doi: 10.3758/BF03196517.
- Payne, Stephen J. 2003. "Users' mental models: The very ideas." *HCI models, theories, and frameworks: Toward a multidisciplinary science*:135-156.
- Payne, Stephen J, and Andrew Howes. 2013. "Adaptive interaction: A utility maximization approach to understanding human interaction with technology." *Synthesis Lectures on Human-Centered Informatics* 6 (1):1-111.
- Pedler, Jennifer. 2001. "Computer spellcheckers and dyslexics—A performance survey." *British Journal of Educational Technology* 32 (1):23-37.
- Peterson, WWTG, T Birdsall, and We Fox. 1954. "The theory of signal detectability." *Transactions of the IRE professional group on information theory* 4 (4):171-212.
- Pollack, Irwin, and Allan B Madans. 1964. "On the performance of a combination of detectors." *Human Factors* 6 (5):523-531.

- Pollack, Irwin, and Donald Norman. 1964. "Non-parametric analysis of recognition experiments." *Psychonomic Science* 1. doi: 10.3758/BF03342823.
- Prinzel, Lawrence J., III. 2002. The Relationship of Self- Efficacy and Complacency in Pilot-Automation Interaction - NASA/TM-2002-211925. Sponsoring Organization: NASA Langley Research Center.
- Quattrini Li, Alberto, Licia Sbattella, and Roberto Tedesco. 2013. "Polispell: an adaptive spellchecker and predictor for people with dyslexia." International Conference on User Modeling, Adaptation, and Personalization.
- Rello, Luz, Miguel Ballesteros, and Jeffrey P Bigham. 2015. "A spellchecker for dyslexia." Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility.
- Riano, Adriana, and Sara Margolin. 2018. "But spell checker always corrects witch words eye misspelled: Spell checker use among good and poor spellers." *Written Language & Literacy* 20:129-146. doi: 10.1075/wll.00001.ria.
- Rice, Stephen, and Jason S McCarley. 2011. "Effects of response bias and judgment framing on operator use of an automated aid in a target detection task." *Journal of Experimental Psychology: Applied* 17 (4):320.
- Rigas, Georgios, Eva Carling, and Berndt Brehmer. 2002. "Reliability and validity of performance measures in microworlds." *Intelligence* 30 (5):463-480.
- Robinson, DE, and RD Sorkin. 1985. "A contingent criterion model of computer assisted detection." *Trends in ergonomics/human factors II*:75-82.
- Rosenfeld, Avi. 2021. "Better metrics for evaluating explainable artificial intelligence." Proceedings of the 20th international conference on autonomous agents and multiagent systems.
- Rotter, Julian B. 1980. "Interpersonal trust, trustworthiness, and gullibility." *American psychologist* 35 (1):1.
- Rovira, Ericka, Kathleen McGarry, and Raja Parasuraman. 2007. "Effects of imperfect automation on decision making in a simulated command and control task." *Human factors* 49 (1):76-87.
- Roy, Christopher J, and William L Oberkamp. 2011. "A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing." *Computer methods in applied mechanics and engineering* 200 (25-28):2131-2144.
- Rumsfeld, Donald H.; Myers, Richard. 2002. "DoD News Briefing - Secretary Rumsfeld and Gen. Myers." U.S. Department of Defense. <https://archive.ph/20180320091111/http://archive.defense.gov/Transcripts/Transcript.aspx?TranscriptID=2636>.
- Sanchez, Julian. 2006. *Factors that affect trust and reliance on an automated aid*: Georgia Institute of Technology.
- Schatz, Raimund, Sebastian Egger, and Kathrin Masuch. 2012. "The impact of test duration on user fatigue and reliability of subjective quality ratings." *Journal of the Audio Engineering Society* 60 (1/2):63-73.

- Schoenherr, Jordan. 2022. *Ethical Artificial Intelligence from Popular to Cognitive Science: Trust in the Age of Entanglement*. Routledge.
- Schraw, Gregory, and David Moshman. 1995. "Metacognitive theories." *Educational psychology review* 7 (4):351-371.
- Scott-Sharoni, Sidney T., Yusuke Yamani, Cara M. Kneeland, Shelby K. Long, Jing Chen, and Joseph W. Houpt. 2021. "Exploring the Effects of Perceptual Separability on Human-Automation Team Efficiency." *Computational Brain & Behavior* 4 (4):486-496. doi: 10.1007/s42113-021-00108-z.
- scribendi.com.
https://www.scribendi.com/academy/articles/guide_to_commonly_confused_words.en.html.
- Sheridan, Thomas B. 1992. *Telerobotics, automation, and human supervisory control*. MIT press.
- Sheridan, Thomas B, and William L Verplank. 1978. Human and computer control of undersea teleoperators. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.
- Simon, Herbert A. 1962. "The Architecture of Complexity." *Proceedings of the American Philosophical Society* 106 (6):467-482.
- Simon, Herbert A. 1996. *The sciences of the artificial*. 3rd ed. ed. Cambridge, Mass. ; London: MIT Press.
- Skitka, Linda J, Kathleen L Mosier, and Mark Burdick. 1999. "Does automation bias decision-making?" *International Journal of Human-Computer Studies* 51 (5):991-1006.
- Smith, Ralph C. 2013. *Uncertainty quantification: theory, implementation, and applications*. Vol. 12: Siam.
- Solomon, Jacob. 2014. "Customization bias in decision support systems." Proceedings of the SIGCHI conference on human factors in computing systems.
- Sorkin, Robert D, and Huanping Dai. 1994. "Signal detection analysis of the ideal group." *Organizational Behavior and Human Decision Processes* 60 (1):1-13.
- Sorkin, Robert D, Christopher J Hays, and Ryan West. 2001. "Signal-detection analysis of group decision making." *Psychological review* 108 (1):183.
- Sorkin, Robert D., Barry H. Kantowitz, and Susan C. Kantowitz. 1988. "Likelihood Alarm Displays." *Human Factors* 30 (4):445-459. doi: 10.1177/001872088803000406.
- Sorkin, Robert D., and David D. Woods. 1985. "Systems with Human Monitors: A Signal Detection Analysis." *Human-Computer Interaction* 1 (1):49-75. doi: 10.1207/s15327051hci0101_2.
- Spiegelhalter, David J, and Hauke Riesch. 2011. "Don't know, can't know: embracing deeper uncertainties when analysing risks." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 369 (1956):4730-4750.
- Stanislaw, Harold, and Natasha Todorov. 1999. "Calculation of signal detection theory measures." *Behavior Research Methods, Instruments, & Computers* 31 (1):137-149. doi: 10.3758/BF03207704.

- Stankov, Lazar, Sabina Kleitman, and Simon A Jackson. 2015. "Measures of the trait of confidence." *Measures of personality and social psychological constructs*:158-189.
- Suchman, Lucy A. 1987. *Plans and situated actions: the problem of human-machine communication*. Cambridge: Cambridge : Cambridge University Press.
- Sullivan, Timothy John. 2015. *Introduction to uncertainty quantification*. Vol. 63: Springer.
- Svenson, Ola. 1981. "Are we all less risky and more skillful than our fellow drivers?" *Acta psychologica* 47 (2):143-148.
- Swets, John A. 1992. "The science of choosing the right decision threshold in high-stakes diagnostics." *American Psychologist* 47 (4):522.
- Tomsett, Richard, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. 2020. "Rapid trust calibration through interpretable and uncertainty-aware AI." *Patterns* 1 (4):100049.
- Trentesaux, Damien, Neville Moray, and Christian Tahon. 1998. "Integration of the human operator into responsive discrete production management systems." *European Journal of Operational Research* 109 (2):342-361.
- Tversky, Amos, and Daniel Kahneman. 1989. "Rational choice and the framing of decisions." In *Multiple criteria decision making and risk analysis using microcomputers*, 81-126. Springer.
- Wang, Xinru, and Ming Yin. 2021. "Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making." International Conference on Intelligent User Interfaces, 2021.
- Weisberg, Herbert I. 2014. *Willful ignorance: The mismeasure of uncertainty*: John Wiley & Sons.
- Wickens, Christopher, and Angela Colcombe. 2007. "Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information." *Human factors* 49 (5):839-850.
- Wickens, Christopher D. 2020. *Processing resources and attention*: CRC Press.
- Wickens, Christopher D, William S Helton, Justin G Hollands, and Simon Banbury. 2021. *Engineering psychology and human performance*: Routledge.
- Wickens, Christopher D, Stephen Rice, David Keller, Shaun Hutchins, Jamie Hughes, and Krisstal Clayton. 2009. "False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect?" *Human factors* 51 (4):446-462.
- Wickens, Christopher D., and Stephen R. Dixon. 2007. "The benefits of imperfect diagnostic automation: a synthesis of the literature." *Theoretical Issues in Ergonomics Science* 8 (3):201-212. doi: 10.1080/14639220500370105.
- Wiczorek, Rebecca. 2017. "Investigating users' mental representation of likelihood alarm systems with different thresholds." *Theoretical Issues in Ergonomics Science* 18 (3):221-240.

- Wiczorek, Rebecca, and Joachim Meyer. 2019. "Effects of Trust, Self-Confidence, and Feedback on the Use of Decision Automation." *Frontiers in psychology* 10:519-519. doi: 10.3389/fpsyg.2019.00519.
- Wiegmann, Douglas A. 2002. "Agreeing with automated diagnostic aids: A study of users' concurrence strategies." *Human Factors* 44 (1):44-50.
- Wiegmann, Douglas A, Aaron Rich, and Hui Zhang. 2001. "Automated diagnostic aids: The effects of aid reliability on users' trust and reliance." *Theoretical Issues in Ergonomics Science* 2 (4):352-367.
- Wohleber, Ryan W, and Gerald Matthews. 2014. "Individual differences in driver over-confidence: implications for stress, error and managing impairments." Proceedings of the human factors and ergonomics society annual meeting.
- Woods, David D, Leila Johannesen, and Scott S Potter. 1991. "Human interaction with intelligent systems: An overview and bibliography." *ACM SIGART Bulletin* 2 (5):39-50.
- Woods, David D. . 1985. "Cognitive Technologies: The Design of Joint Human-Machine Cognitive Systems." *AI Magazine* 6 (4). doi: 10.1609/aimag.v6i4.511.
- Zhang, Jun, and Shane T. Mueller. 2005. "A note on ROC analysis and non-parametric estimate of sensitivity." *Psychometrika* 70 (1):203-212. doi: 10.1007/s11336-003-1119-8.

Appendices

- Appendix A – Stylus studies data overview
- Appendix B – Stylus studies item distribution
- Appendix C – Stylus studies supplementary information
- Appendix D – Hypotheses overview
- Appendix E – Ethics, consent, and debrief

Appendix A3 – Stylus 1 data overview

STYLUS 1, GroupGood						
<i>N</i> = 31						
		<i>Units</i>	<i>Scale</i>	<i>Mean average</i>	<i>Standard deviation</i>	
1	Male/ female ratio					13/18
2	Age			33.48	10.42	
3	Duration	Minutes		17.36	6.72	
4	When thinking of my own knowledge of British English grammar, I would class myself as	Percentage	0 – 100	73.06	14.79	
5	When thinking of my own capability to spot grammar errors in British English texts, I would class myself as	Percentage	0 – 100	75.39	14.52	
6	When thinking of my own capability to correct grammar errors in British English texts, I would class myself as	Percentage	0 – 100	75.06	13.94	
7	Prior perceived self-efficacy	Percentage	0 – 100	74.51	14.42	Aggregated variable from 4, 5, 6 (Cronbach α = .96)
8	When thinking of the average native British English speaker's knowledge of British English grammar, I would class them as	Percentage	0 – 100	60.94	20.05	
9	When thinking of the average native British English speaker's capability to spot grammar errors in British English texts, I would class them as	Percentage	0 – 100	63.94	21.24	
10	When thinking of the average native British English	Percentage	0 – 100	60.42	22.00	

	speaker's capability to correct grammar errors in British English texts, I would class them as					
11	Prior perceived efficacy of others	Percentage	0 – 100	61.76	20.05	Aggregated variable from 8, 9, 10 (Cronbach α = .95)
12	When thinking of the trustworthiness of spelling suggestions in word processing software packages or internet browsers in general, I would class them as	Percentage	0 – 100	75.42	18.95	
13	When thinking of the trustworthiness of grammar suggestions in word processing software packages or internet browsers in general, I would class them as	Percentage	0 – 100	73.16	19.64	
14	Average prior trust	Percentage	0 – 100	74.29	18.51	Aggregated variable from 12, 13 (Cronbach α = .91)
15	Average post confidence (from 30 trials)	Percentage	50 – 100	91.36	5.06	
16	I would estimate the number of times I chose the Stylus suggestion over the original sentence at	Trials	0 – 30	19.29	4.65	
17	Actual number of Stylus	Trials	0 – 30	20.32	1.81	
18	I would estimate the number of times I chose the correct answer (either the original sentence or the Stylus suggestion) at	Trials	0 – 30	20.73	6.98	
19	Actual number correct	Trials	0 – 30	23.87	3.41	

20	Percentage correct	Percentage	0 – 100	79.57	11.38	
21	So far, 183 people have rated the same Stylus suggestions. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	67.68	15.65	
22	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	68.45	16.28	
23	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	69.10	16.31	
24	Post trust	Percentage	0 – 100	68.77	15.33	Aggregated variable from 22, 23 (Cronbach $\alpha = .87$)
25	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	77.07	18.52	
26	Sensitivity	d'		1.67	0.65	
27	Bias	c		0.65	0.51	
28	Sensitivity	A'		0.83	0.13	
29	Bias	B''		0.28	0.29	
30	Number H	Trials	0 – 20	17.10 / 20	2.12	
31	Number M	Trials	0 – 20	2.90 / 20	2.12	
32	Number CR	Trials	0 – 10	6.77 / 10	1.73	
33	Number FA	Trials	0 – 10	3.23 / 10	1.73	
34	Percentage H	Percentage	0 – 100	56.99	7.06	
35	Percentage M	Percentage	0 – 100	9.68	7.06	
36	Percentage CR	Percentage	0 – 100	22.58	5.75	
37	Percentage FA	Percentage	0 – 100	10.75	5.75	
38	Percentage confidence H	Percentage	50 – 100	92.97	5.37	
39	Percentage confidence M	Percentage	50 – 100	86.31	11.13	

40	Percentage confidence CR	Percentage	50 – 100	90.51	5.71	
41	Percentage confidence FA	Percentage	50 – 100	88.74	8.55	
42	Confidence correct	Percentage	50 – 100	91.47	5.54	
43	Confidence incorrect	Percentage	50 – 100	87.53	9.84	
44	Confidence Stylus	Percentage	50 – 100	90.86	6.96	
45	Confidence Original	Percentage	50 – 100	88.41	8.42	

STYLUS 1, GroupBad						
<i>N = 31</i>						
		<i>Units</i>	<i>Scale</i>	<i>Mean average</i>	<i>Standard deviation</i>	
1	Male/ female ratio					6/25
2	Age	Years		39.97	14.06	
3	Duration	Minutes		16.94	5.82	
4	When thinking of my own knowledge of British English grammar, I would class myself as	Percentage	0 – 100	71.48	15.87	
5	When thinking of my own capability to spot grammar errors in British English texts, I would class myself as	Percentage	0 – 100	69.42	12.36	
6	When thinking of my own capability to correct grammar errors in British English texts, I would class myself as	Percentage	0 – 100	68.58	13.18	
7	Prior perceived self-efficacy	Percentage	0 – 100	69.83	13.80	Aggregated variable from 4, 5, 6 (Cronbach $\alpha = 0.96$)
8	When thinking of the average native British English speaker's knowledge of British English grammar, I would class them as	Percentage	55.84	60.94	18.07	
9	When thinking of the average native British English speaker's capability	Percentage	0 – 100	54.00	18.10	

	to spot grammar errors in British English texts, I would class them as					
10	When thinking of the average native British English speaker's capability to correct grammar errors in British English texts, I would class them as	Percentage	0 – 100	51.32	16.67	
11	Prior perceived efficacy of others	Percentage	0 – 100	53.72	16.99	Aggregated variable from 8, 9, 10 (Cronbach α = .96)
12	When thinking of the trustworthiness of spelling suggestions in word processing software packages or internet browsers in general, I would class them as	Percentage	0 – 100	69.71	17.34	
13	When thinking of the trustworthiness of grammar suggestions in word processing software packages or internet browsers in general, I would class them as	Percentage	0 – 100	64.84	18.13	
14	Average prior trust	Percentage	0 – 100	67.27	17.17	Aggregated variable from 12, 13 (Cronbach α = .93)
15	Average post confidence (from 30 trials)	Percentage	50 – 100	91.21	5.44	
16	I would estimate the number of times I chose the Stylus suggestion over the original sentence at	Trials	0 – 30	12.31	5.22	
17	Actual number of Stylus	Trials	0 – 30	12.71	1.95	
18	I would estimate the number of times I chose the correct answer (either the original sentence or the Stylus suggestion) at	Trials	0 – 30	19.60	7.16	

19	Actual number correct	Trials	0 – 30	24.13	2.81	
20	Percentage correct	Percentage	0 – 100	80.43	9.38	
21	So far, 183 people have rated the same Stylus suggestions. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	63.39	16.26	
22	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	50.65	20.24	
23	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	48.94	20.69	
24	Post trust	Percentage	0 – 100	49.79	19.58	Aggregated variable from 22, 23 (Cronbach $\alpha = .91$)
25	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	70.81	19.17	
26	Sensitivity	d'		1.89	0.63	
27	Bias	c		0.21	0.41	
28	Sensitivity	A'		0.88	0.08	
29	Bias	B''		0.12	0.23	
30	Number H	Trials	0 – 10	8.42 / 10	1.06	
31	Number M	Trials	0 – 10	1.58 / 10	1.06	
32	Number CR	Trials	0 – 20	15.71 / 20	2.18	
33	Number FA	Trials	0 – 20	4.29 / 20	2.18	
34	Percentage H	Percentage	0 – 100	28.06	3.52	
35	Percentage M	Percentage	0 – 100	5.27	3.52	
36	Percentage CR	Percentage	0 – 100	52.37	7.26	
37	Percentage FA	Percentage	0 – 100	14.30	7.26	
38	Percentage confidence H	Percentage	50 – 100	92.20	4.99	
39	Percentage confidence M	Percentage	50 – 100	89.32	12.69	

40	Percentage confidence CR	Percentage	50 – 100	91.90	5.47	
41	Percentage confidence FA	Percentage	50 – 100	86.28	8.20	
42	Confidence correct	Percentage	50 – 100	92.05	5.23	
43	Confidence incorrect	Percentage	50 – 100	87.80	10.44	
44	Confidence Stylus	Percentage	50 – 100	89.24	6.59	
45	Confidence Original	Percentage	50 – 100	90.61	9.08	

Appendix A4 – Stylus 2 data overview

STYLUS 2, GroupGood						
<i>N</i> = 59						
		<i>Units</i>	<i>Scale</i>	<i>Mean average</i>	<i>Standard deviation</i>	
1	Male/ female ratio					30/29
2	Age			40.46	12.86	
3	Duration	Minutes		18.01	5.13	
4	When thinking of my own knowledge of British English grammar, I would class myself as	Percentage	0 – 100	73.92	14.41	
5	When thinking of my own capability to spot grammar errors in British English texts, I would class myself as	Percentage	0 – 100	73.00	14.76	
6	When thinking of my own capability to correct grammar errors in British English texts, I would class myself as	Percentage	0 – 100	72.61	15.33	
7	Prior perceived self-efficacy	Percentage	0 – 100	73.18	14.83	Aggregated variable from 4, 5, 6 (Cronbach α = .95)
8	When thinking of the average native British English speaker's knowledge of British English grammar, I would class them as	Percentage	0 – 100	63.10	17.09	
9	When thinking of the average native British English speaker's capability to spot grammar errors in British English texts, I would class them as	Percentage	0 – 100	59.14	20.23	
10	When thinking of the average native British English	Percentage	0 – 100	57.98	19.25	

	speaker's capability to correct grammar errors in British English texts, I would class them as					
11	Prior perceived efficacy of others	Percentage	0 – 100	60.07	18.86	Aggregated variable from 8, 9, 10 (Cronbach α = .95)
12	When thinking of the trustworthiness of spelling suggestions in word processing software packages or internet browsers in general, I would class them as	Percentage	0 – 100	77.59	16.71	
13	When thinking of the trustworthiness of grammar suggestions in word processing software packages or internet browsers in general, I would class them as	Percentage	0 – 100	73.61	16.40	
14	Average prior trust	Percentage	0 – 100	76.42	17.07	Aggregated variable from 12, 13 (Cronbach α = .88)
15	Average post confidence (from 30 trials)	Percentage	50 – 100	89.44	7.15	
16	I would estimate the number of times I chose the Stylus suggestion over the original sentence at	Trials	0 – 30	17.53	5.59	
17	Actual number of Stylus	Trials	0 – 30	17.19	4.15	
18	I would estimate the number of times I chose the correct answer (either the original sentence or the Stylus suggestion) at	Trials	0 – 30	21.05	5.85	
19	Actual number correct	Trials	0 – 30	19.49	4.87	

20	Percentage correct	Percentage	0 – 100	64.97	16.23	
21	So far, 183 people have rated the same Stylus suggestions. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	67.34	14.66	
22	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	69.19	16.97	
23	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	66.76	19.23	
24	Post trust	Percentage	0 – 100	67.97	17.07	Aggregated variable from 22, 23 (Cronbach $\alpha = .83$)
25	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	70.95	15.66	
26	Sensitivity	d'		0.85	0.98	
27	Bias	c		0.18	0.76	
28	Sensitivity	A'		0.68	0.20	
29	Bias	B''		0.07	0.27	
30	Number H	Trials	0 – 20	13.34 / 20	4.12	
31	Number M	Trials	0 – 20	6.66 / 20	4.12	
32	Number CR	Trials	0 – 10	6.15 / 10	1.86	
33	Number FA	Trials	0 – 10	3.85 / 10	1.86	
34	Percentage H	Percentage	0 – 100	44.46	13.74	
35	Percentage M	Percentage	0 – 100	22.20	13.74	
36	Percentage CR	Percentage	0 – 100	20.51	6.21	
37	Percentage FA	Percentage	0 – 100	12.82	6.21	
38	Percentage confidence H	Percentage	50 – 100	90.16	7.70	

39	Percentage confidence M	Percentage	50 – 100	84.87	10.25	
40	Percentage confidence CR	Percentage	50 – 100	90.58	7.73	
41	Percentage confidence FA	Percentage	50 – 100	89.58	8.18	
42	Confidence correct	Percentage	50 – 100	90.37	7.71	
43	Confidence incorrect	Percentage	50 – 100	87.17	9.21	
44	Confidence Stylus	Percentage	50 – 100	89.87	7.94	
45	Confidence Original	Percentage	50 – 100	87.67	8.99	

STYLUS 2, GroupBad						
<i>N = 61</i>						
		<i>Units</i>	<i>Scale</i>	<i>Mean average</i>	<i>Standard deviation</i>	
1	Male/ female ratio					30/31
2	Age	Years		37.44	13.20	
3	Duration	Minutes		18.37	6.18	
4	When thinking of my own knowledge of British English grammar, I would class myself as	Percentage	0 – 100	76.00	13.62	
5	When thinking of my own capability to spot grammar errors in British English texts, I would class myself as	Percentage	0 – 100	73.05	13.84	
6	When thinking of my own capability to correct grammar errors in British English texts, I would class myself as	Percentage	0 – 100	70.66	15.28	
7	Prior perceived self-efficacy	Percentage	0 – 100	73.23	14.25	Aggregated variable from 4, 5, 6 (Cronbach $\alpha = .95$)
8	When thinking of the average native British English speaker's knowledge of British English grammar, I would class them as	Percentage	0 – 100	57.38	19.98	

9	When thinking of the average native British English speaker's capability to spot grammar errors in British English texts, I would class them as	Percentage	0 – 100	53.97	16.15	
10	When thinking of the average native British English speaker's capability to correct grammar errors in British English texts, I would class them as	Percentage	0 – 100	52.51	16.82	
11	Prior perceived efficacy of others	Percentage	0 – 100	54.62	17.65	Aggregated variable from 8, 9, 10 (Cronbach α = .88)
12	When thinking of the trustworthiness of spelling suggestions in word processing software packages or internet browsers in general, I would class them as	Percentage	0 – 100	77.59	16.71	
13	When thinking of the trustworthiness of grammar suggestions in word processing software packages or internet browsers in general, I would class them as	Percentage	0 – 100	73.61	16.40	
14	Average prior trust	Percentage	0 – 100	75.60	16.56	Aggregated variable from 12, 13 (Cronbach α = .93)
15	Average post confidence (from 30 trials)	Percentage	50 – 100	86.70	8.11	
16	I would estimate the number of times I chose the Stylus suggestion over the original sentence at	Trials	0 – 30	15.23	5.26	
17	Actual number of Stylus	Trials	0 – 30	13.26	3.08	

18	I would estimate the number of times I chose the correct answer (either the original sentence or the Stylus suggestion) at	Trials	0 – 30	21.43	5.32	
19	Actual number correct	Trials	0 – 30	18.31	2.99	
20	Percentage correct	Percentage	0 – 100	61.04	9.95	
21	So far, 183 people have rated the same Stylus suggestions. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	64.46	17.15	
22	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	57.95	20.99	
23	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	58.49	19.07	
24	Post trust	Percentage	0 – 100	58.22	17.2	Aggregated variable from 22, 23 (Cronbach $\alpha = .68$)
25	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	65.46	21.47	
26	Sensitivity	d'		0.55	0.53	
27	Bias	c		-0.12	0.56	
28	Sensitivity	A'		0.66	0.14	
29	Bias	B''		-0.02	0.13	
30	Number H	Trials	0 – 10	5.79 / 10	1.46	
31	Trials	Trials	0 – 10	4.21 / 10	1.46	
32	Number CR	Trials	0 – 20	12.52 / 20	2.66	
33	Number FA	Trials	0 – 20	7.48 / 20	2.66	
34	Percentage H	Percentage	0 – 100	19.29	4.87	

35	Percentage M	Percentage	0 – 100	14.04	4.87	
36	Percentage CR	Percentage	0 – 100	41.75	8.85	
37	Percentage FA	Percentage	0 – 100	24.92	8.85	
38	Percentage confidence H	Percentage	50 – 100	87.39	8.38	
39	Percentage confidence M	Percentage	50 – 100	85.87	10.01	
40	Percentage confidence CR	Percentage	50 – 100	87.50	8.80	
41	Percentage confidence FA	Percentage	50 – 100	85.01	9.50	
42	Confidence correct	Percentage	50 – 100	87.44	8.59	
43	Confidence incorrect	Percentage	50 – 100	85.44	9.75	
44	Confidence Stylus	Percentage	50 – 100	86.20	8.94	
45	Confidence Original	Percentage	50 – 100	86.68	9.40	

Appendix A5 – Stylus 3 data overview

STYLUS 3, C53 + C75						
<i>N = 128</i>						
<i>Results for test trials only (practise and dummy trials excluded), unless indicated otherwise</i>						
		Units	Scale	Mean average	Standard deviation	
1	Male / female ratio					47 / 81
2	Age	Years		37.11	12.07	
3	Duration	Minutes		16.62	7.41	
4	When thinking of how good I am at English grammar , I would class myself as	Percentage	0 – 100	73.50	16.74	
5	When thinking of how good I am at English spelling , I would class myself as	Percentage	0 – 100	76.50	16.12	
6	Prior perceived self-efficacy	Percentage	0 – 100	75.00	14.64	Aggregated variable from 4 and 5 (Cronbach $\alpha = .74$)
7	When thinking of how good English language spell checkers are , I would class them as	Percentage	0 – 100	80.36	15.32	
8	When thinking of how good English language grammar checkers are , I would class them as	Percentage	0 – 100	70.19	18.70	
9	Average prior trust	Percentage	0 – 100	75.27	15.64	Aggregated variable from 7 and 8 (Cronbach $\alpha = .81$)
10	Average post confidence (from 33 trials)	Percentage	50 – 100	89.15	6.99	
11	I would estimate the number of times I chose the Stylus suggestion over the original sentence at	Trials	0 – 33	19.80	6.59	
12	Actual number of Stylus	Trials	0 – 33	20.52	2.71	
13	I would estimate the number of times I chose the correct answer (either the original sentence or the Stylus suggestion) at	Trials	0 – 33	23.27	7.46	

14	Actual number correct (incl. dummy trial)	Trials	0 – 33	24.01	4.02	
15	Percentage correct	Percentage	0 – 100	71.90	12.56	
16	So far, 183 people have rated the same Stylus suggestions. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	66.02	15.92	
17	When thinking of the usefulness of Stylus' suggestions during this experiment, I would class them as	Percentage	0 – 100	70.09	15.25	
18	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	64.45	16.10	
19	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	65.15	18.78	
20	When thinking of Stylus' performance in general during this experiment, I would class it as	Percentage	0 – 100	67.43	16.08	
21	Post trust	Percentage	0 – 100	66.78	15.07	Aggregated variable from 17, 18, 19, 20 (Cronbach $\alpha = .93$)
22	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	71.59	19.02	
23	Sensitivity	d'		1.22	0.85	
24	Bias	c		0.37	0.56	
25	Sensitivity	A'		0.77	0.14	
26	Bias	B''		0.12	0.24	
27	Number H	Trials	0 – 20	15.27/ 20	2.73	
28	Number M	Trials	0 – 20	4.73/ 20	2.73	
29	Number CR	Trials	0 – 12	7.74/12	2.07	
30	Number FA	Trials	0 – 12	4.26/12	2.07	
31	Confidence H	Percentage	50 – 100	90.51	6.57	
32	Confidence M	Percentage	50 – 100	84.96	10.57	
33	Confidence CR	Percentage	50 – 100	89.27	8.49	

34	Confidence FA	Percentage	50 – 100	86.84	9.97	
35	Confidence correct	Percentage	50 – 100	90.13	6.82	
36	Confidence incorrect	Percentage	50 – 100	85.68	9.15	
37	Confidence Stylus	Percentage	50 – 100	88.09	8.30	
38	Confidence Original	Percentage	50 – 100	89.88	6.67	

STYLUS 3, C53

N = 128

Results for test trials only (practise and dummy trials excluded), unless indicated otherwise

		Units	Scale	Mean average	Standard deviation	
1	Percentage correct	Percentage	0 – 100	70.31	14.82	
2	Sensitivity	d'		1.19	0.90	
3	Bias	c		0.31	0.67	
4	Sensitivity	A'		0.60	0.10	
5	Bias	B''		0.11	0.24	
6	Trials	Trials	0 – 20	6.03/ 8	1.46	
7	Number M	Trials	0 – 20	1.97/ 8	1.46	
8	Number CR	Trials	0 – 12	5.22/ 8	1.52	
9	Number FA	Trials	0 – 12	2.78/ 8	1.52	
10	Confidence average	Percentage	50 – 100	88.63	7.37	
11	Confidence H	Percentage	50 – 100	89.80	7.43	
12	Confidence M	Percentage	50 – 100	85.34	11.90	
13	Confidence CR	Percentage	50 – 100	89.43	8.64	
14	Confidence FA	Percentage	50 – 100	86.94	10.53	
15	Confidence correct	Percentage	50 – 100	89.64	7.36	
16	Confidence incorrect	Percentage	50 – 100	85.86	10.15	
17	Confidence Stylus	Percentage	50 – 100	89.04	7.42	
18	Confidence Original	Percentage	50 – 100	88.31	8.53	

STYLUS 3, C75

N = 128

Results for test trials only (practise and dummy trials excluded), unless indicated otherwise

		Units	Scale	Mean average	Standard deviation	
1	Percentage correct	Percentage	0 – 100	73.49	14.39	
2	Sensitivity	d'		1.15	0.91	
3	Bias	c		0.50	0.77	
4	Sensitivity	A'		0.74	0.20	
5	Bias	B''		0.20	0.36	

6	Number H	Trials	0 – 20	9.23 / 12	1.82	
7	Number M	Trials	0 – 20	2.77 / 12	1.82	
8	Number CR	Trials	0 – 10	2.52 / 4	1.10	
9	Number FA	Trials	0 – 10	1.48 / 4	1.10	
10	Confidence average	Percentage	50 – 100	89.67	7.06	
11	Confidence H	Percentage	50 – 100	90.90	6.79	
12	Confidence M	Percentage	50 – 100	86.44	10.71	
13	Confidence CR	Percentage	50 – 100	89.52	9.78	
14	Confidence FA	Percentage	50 – 100	86.86	12.12	
15	Confidence correct	Percentage	50 – 100	90.63	6.94	
16	Confidence incorrect	Percentage	50 – 100	86.07	9.99	
17	Confidence Stylus	Percentage	50 – 100	90.52	6.67	
18	Confidence Original	Percentage	50 – 100	88.15	9.10	

Appendix A6 – Stylus 5 data overview

STYLUS 4, C53 + C94						
<i>N = 140</i>						
<i>Results for test trials only (practise and dummy trials excluded), unless indicated otherwise</i>						
		Units	Scale	Mean average	Standard deviation	
1	Male / female ratio					58 / 82
2	Age	Years		36.19	12.77	
3	Duration	Minutes		16.62	7.41	
4	When thinking of how good I am at English grammar , I would class myself as	Percentage	0 – 100	73.86	14.69	
5	When thinking of how good I am at English spelling , I would class myself as	Percentage	0 – 100	75.84	16.14	
6	Prior perceived self-efficacy	Percentage	0 – 100	74.85	14.10	Aggregated variable from 4 and 5 (Cronbach $\alpha = .80$)
7	When thinking of how good English language spell checkers are , I would class them as	Percentage	0 – 100	80.50	15.56	
8	When thinking of how good English language grammar checkers are , I would class them as	Percentage	0 – 100	70.79	19.39	
9	Average prior trust	Percentage	0 – 100	75.64	15.91	Aggregated variable from 7 and 8 (Cronbach $\alpha = .78$)
10	Average post confidence (from 33 trials)	Percentage	50 – 100	89.15	6.99	
11	I would estimate the number of times I chose the Stylus suggestion over the original sentence at	Trials	0 – 33	18.39	6.04	
12	Actual number of Stylus (incl dummy trial)	Trials	0 – 33	21.96	2.87	
13	I would estimate the number of times I chose the correct answer (either the original	Trials	0 – 33	22.42	7.24	

	sentence or the Stylus suggestion) at					
14	Actual number correct (incl dummy trial)	Trials	0 – 33	24.51	3.77	
15	Percentage correct	Percentage	0 – 100	73.48	11.77	
16	So far, 183 people have rated the same Stylus suggestions. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	65.96	15.97	
17	When thinking of the usefulness of Stylus' suggestions during this experiment, I would class them as	Percentage	0 – 100	69.64	16.07	
18	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	62.68	18.40	
19	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	63.84	19.41	
20	When thinking of Stylus' performance in general during this experiment, I would class it as	Percentage	0 – 100	68.09	17.11	
21	Post trust	Percentage	0 – 100	66.06	16.54	Aggregated variable from 17, 18, 19, 20 (Cronbach $\alpha = .95$)
22	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	71.13	16.32	
23	Sensitivity	d'		1.16	0.79	
24	Bias	c		0.36	0.59	
25	Sensitivity	A'		0.77	0.14	
26	Bias	B''		0.11	0.29	
27	Number H	Trials	0 – 20	17.74/ 20	2.83	
28	Number M	Trials	0 – 20	5.26/ 20	2.83	
29	Number CR	Trials	0 – 12	5.77/12	1.78	
30	Number FA	Trials	0 – 12	3.23/12	1.78	
31	Confidence H	Percentage	50 – 100	91.00	6.38	
32	Confidence M	Percentage	50 – 100	84.08	11.33	

33	Confidence CR	Percentage	50 – 100	89.08	8.40	
34	Confidence FA	Percentage	50 – 100	86.49	9.44	
35	Confidence correct	Percentage	50 – 100	90.51	6.39	
36	Confidence incorrect	Percentage	50 – 100	84.93	9.50	
37	Confidence Stylus	Percentage	50 – 100	90.36	6.42	
38	Confidence Original	Percentage	50 – 100	86.91	8.76	

STYLUS 4 – C53

N = 140

Results for test trials only (practise and dummy trials excluded), unless indicated otherwise

		Units	Scale	Mean average	Standard deviation	
1	Percentage correct	Percentage	0 – 100	69.06	14.77	
2	Sensitivity	d'		1.12	0.89	
3	Bias	c		0.28	0.79	
4	Sensitivity	A'		0.74	0.16	
5	Bias	B''		0.09	0.25	
6	Number H	Trials	0 – 20	5.89/ 8	1.56	
7	Number M	Trials	0 – 20	2.11/ 8	1.56	
8	Number CR	Trials	0 – 12	5.16/ 8	1.63	
9	Number FA	Trials	0 – 12	2.84/ 8	1.63	
10	Confidence average	Percentage	50 – 100	88.34	7.31	
11	Confidence H	Percentage	50 – 100	90.31	7.34	
12	Confidence M	Percentage	50 – 100	84.14	12.76	
13	Confidence CR	Percentage	50 – 100	89.05	8.40	
14	Confidence FA	Percentage	50 – 100	86.31	9.78	
15	Confidence correct	Percentage	50 – 100	89.73	7.04	
16	Confidence incorrect	Percentage	50 – 100	85.29	9.70	
17	Confidence Stylus	Percentage	50 – 100	89.06	7.12	
18	Confidence Original	Percentage	50 – 100	87.47	8.76	

STYLUS 4 – C94

N = 140

Results for test trials only (practise and dummy trials excluded), unless indicated otherwise

		Units	Scale	Mean average	Standard deviation	
1	Percentage correct	Percentage	0 – 100	77.90	13.95	
2	Sensitivity	d'		n/a	n/a	
3	Bias	c		n/a	n/a	

4	Sensitivity	A'		n/a	n/a	
5	Bias	B''		n/a	n/a	
6	Number H	Trials	0 – 20	11.85 / 12	2.10	
7	Number M	Trials	0 – 20	3.15 / 12	2.10	
8	Number CR	Trials	0 – 10	0.61 / 4	0.49	
9	Number FA	Trials	0 – 10	0.39 / 4	0.49	
10	Confidence average	Percentage	50 – 100	90.06	6.95	
11	Confidence H	Percentage	50 – 100	91.30	6.72	
12	Confidence M	Percentage	50 – 100	85.60	11.46	
13	Confidence CR	Percentage	50 – 100	88.79	11.80	
14	Confidence FA	Percentage	50 – 100	88.60	9.84	
15	Confidence correct	Percentage	50 – 100	91.19	6.67	
16	Confidence incorrect	Percentage	50 – 100	91.23	6.74	
17	Confidence Stylus	Percentage	50 – 100	91.23	6.74	
18	Confidence Original	Percentage	50 – 100	85.93	11.27	

Appendix A7 – Stylus 5 data overview

STYLUS 5, G90						
<i>N</i> = 38						
		Units	Scale	Mean average	Standard deviation	
1	Male / female ratio					12 / 26
2	Age	Years		34.13	13.45	
3	Duration	Minutes		27.50	12.76	
4	When thinking of how good I am at English grammar , I would class myself as	Percentage	0 – 100	75.42	15.34	
5	When thinking of how good I am at English spelling , I would class myself as	Percentage	0 – 100	77.34	14.89	
6	Prior perceived self-efficacy	Percentage	0 – 100	76.38	11.77	Aggregated variable from 4 and 5 (Cronbach $\alpha = 0.84$)
7	When thinking of how good English language spell checkers are , I would class them as	Percentage	0 – 100	81.05	15.44	
8	When thinking of how good English language grammar checkers are , I would class them as	Percentage	0 – 100	73.42	18.80	
9	Average prior trust	Percentage	0 – 100	77.24	14.99	Aggregated variable from 7 and 8 (Cronbach $\alpha = 0.68$)
10	Average post confidence	Percentage	50 – 100	90.93	7.45	
11	I would estimate the number of times I followed Stylus' judgement	Trials	0 – 100	52.45	33.09	
12	Actual number Stylus	Trials	0 – 100	76.71	7.41	
13	I would estimate the number of times I chose the correct answer (either "Yes" or "No") at	Trials	0 – 100	73.00	25.76	
14	Actual number correct	Trials	0 – 100	80.92	9.20	

15	So far, 183 people have rated the same Stylus suggestions. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	64.87	16.82	
16	When thinking of the usefulness of Stylus' suggestions during this experiment, I would class them as	Percentage	0 – 100	62.34	5.72	
17	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	57.21	6.51	
18	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	61.26	22.11	
19	When thinking of Stylus' performance in general during this experiment, I would class it as	Percentage	0 – 100	63.03	22.87	
20	Post trust	Percentage	0 – 100	60.46	22.38	Aggregated variable from 17, 18, 19, 20 (Cronbach $\alpha = 0.94$)
21	When assessing the sentences, did you remember the Stylus accuracy rate?	Percentage	0 – 100	75.55	25.32	
22	In your answers, did you consider the Stylus accuracy rate?	Percentage	0 – 100	59.03	31.29	
23	After assessing 100 sentences, in which Stylus indicated potential errors in 50, do you believe that Stylus was 90% accurate?	Percentage	0 – 100	52.47	30.33	
24	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	69.05	17.74	
25	Sensitivity	d'_{NY}		1.95	0.80	
26	Bias	$c_{Y/N}$		0.25	0.49	
27	Sensitivity	d'_{Stylus}		1.62	0.62	
28	Bias	c_{Stylus}		0.34	0.61	
29	H rate	Trials		0.85	0.09	
30	M rate	Trials		0.15	0.09	

31	CR rate	Trials		0.77	0.13	
32	FA rate	Trials		0.23	0.13	
33	Confidence H	Percentage	50 – 100	92.15	7.52	
34	Confidence M	Percentage	50 – 100	83.17	11.41	
35	Confidence CR	Percentage	50 – 100	91.72	6.69	
36	Confidence FA	Percentage	50 – 100	88.96	15.82	
37	Confidence correct	Percentage	50 – 100	92.01	6.91	
38	Confidence incorrect	Percentage	50 – 100	86.45	10.91	
39	Confidence agreement with Stylus	Percentage	50 – 100	91.97	6.86	
40	Confidence disagreement with Stylus	Percentage	50 – 100	87.71	9.52	

STYLUS 5, G70						
<i>N = 38</i>						
		Units	Scale	Mean average	Standard deviation	
1	Male / female ratio					10 / 28
2	Age	Years		35.11	12.90	
3	Duration	Minutes		24.90	9.88	
4	When thinking of how good I am at English grammar , I would class myself as	Percentage	0 – 100	76.53	16.36	
5	When thinking of how good I am at English spelling , I would class myself as	Percentage	0 – 100	80.21	17.67	
6	Prior perceived self-efficacy	Percentage	0 – 100	78.37	15.74	Aggregated variable from 4 and 5 (Cronbach $\alpha = 0.83$)
7	When thinking of how good English language spell checkers are , I would class them as	Percentage	0 – 100	82.97	16.64	
8	When thinking of how good English language grammar checkers are , I would class them as	Percentage	0 – 100	73.79	17.04	
9	Average prior trust	Percentage	0 – 100	78.38	14.86	Aggregated variable

						from 7 and 8 (Cronbach $\alpha = 0.72$)
10	Average post confidence	Percentage	50 – 100	90.68	7.43	
11	I would estimate the number of times I followed Stylus' judgement	Trials	0 – 100	50.66	22.40	
12	Actual number Stylus	Trials	0 – 100	64.87	3.79	
13	I would estimate the number of times I chose the correct answer (either "Yes" or "No") at	Trials	0 – 100	77.17	14.49	
14	Actual number correct	Trials	0 – 100	78.97	9.61	
15	So far, 183 people have rated the same Stylus suggestions. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	67.95	13.80	
16	When thinking of the usefulness of Stylus' suggestions during this experiment, I would class them as	Percentage	0 – 100	54.16	24.15	
17	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	56.50	20.36	
18	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	57.11	20.96	
19	When thinking of Stylus' performance in general during this experiment, I would class it as	Percentage	0 – 100	60.08	21.69	
20	Post trust	Percentage	0 – 100	56.96	20.64	Aggregated variable from 17, 18, 19, 20 (Cronbach $\alpha = 0.96$)
21	When assessing the sentences, did you remember the Stylus accuracy rate?	Percentage	0 – 100	67.97	31.01	
22	In your answers, did you consider the Stylus accuracy rate?	Percentage	0 – 100	52.58	28.46	

23	After assessing 100 sentences, in which Stylus indicated potential errors in 50, do you believe that Stylus was 70% accurate?	Percentage	0 – 100	54.16	27.40	
24	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	66.16	18.74	
25	Sensitivity	d'_{NY}		1.80	0.73	
26	Bias	$C_{Y/N}$		0.15	0.66	
27	Sensitivity	d'_{Stylus}		1.67	0.82	
28	Bias	C_{Stylus}		0.22	0.41	
29	H rate	Trials		0.81	0.12	
30	M rate	Trials		0.19	0.12	
31	CR rate	Trials		0.77	0.14	
32	FA rate	Trials		0.23	0.14	
33	Confidence H	Percentage	50 – 100	92.18	6.99	
34	Confidence M	Percentage	50 – 100	81.65	11.64	
35	Confidence CR	Percentage	50 – 100	91.77	7.13	
36	Confidence FA	Percentage	50 – 100	88.98	8.87	
37	Confidence correct	Percentage	50 – 100	92.01	6.91	
38	Confidence incorrect	Percentage	50 – 100	86.45	10.91	
39	Confidence agreement with Stylus	Percentage	50 – 100	91.59	7.18	
40	Confidence disagreement with Stylus	Percentage	50 – 100	89.13	8.07	

STYLUS 5, GC (control group)						
<i>N</i> = 38						
		Units	Scale	Mean average	Standard deviation	
1	Male / female ratio					10 / 28
2	Age	Years		36.59	14.23	
3	Duration	Minutes		26.12	8.07	
4	When thinking of how good I am at English grammar , I would class myself as	Percentage	0 – 100	74.24	16.89	
5	When thinking of how good I am at English spelling , I would class myself as	Percentage	0 – 100	78.03	17.46	
6	Prior perceived self-efficacy	Percentage	0 – 100	76.13	16.11	Aggregated variable

						from 4 and 5 (Cronbach $\alpha = 0.86$)
7	When thinking of how good English language spell checkers are , I would class them as	Percentage	0 – 100	82.95	10.59	
8	When thinking of how good English language grammar checkers are , I would class them as	Percentage	0 – 100	70.55	16.69	
9	Average prior trust	Percentage	0 – 100	76.75	12.23	Aggregated variable from 7 and 8 (Cronbach $\alpha = 0.69$)
10	Average post confidence	Percentage	50 – 100	90.68	7.43	
11	I would estimate the number of times I followed Stylus' judgement	Trials	0 – 100	n/a	n/a	
12	Actual number Stylus	Trials	0 – 100	n/a	n/a	
13	I would estimate the number of times I chose the correct answer (either "Yes" or "No") at	Trials	0 – 100	75.63	16.46	
14	Actual number correct	Trials	0 – 100	78.08	7.85	
15	So far, 183 people have rated the same sentences. – When thinking of my own performance in relation to others during the experiment, I would estimate it as	Percentage	0 – 100	67.21	13.80	
16	When thinking of the usefulness of Stylus' suggestions during this experiment, I would class them as	Percentage	0 – 100	n/a	n/a	
17	When thinking of the trustworthiness of Stylus during this experiment, I would class it as	Percentage	0 – 100	n/a	n/a	
18	When thinking of the consistency of Stylus' performance during this experiment, I would class it as	Percentage	0 – 100	n/a	n/a	

19	When thinking of Stylus' performance in general during this experiment, I would class it as	Percentage	0 – 100	n/a	n/a	
20	Post trust	Percentage	0 – 100	n/a	n/a	Aggregated variable from 17, 18, 19, 20 (Cronbach $\alpha = n/a$)
21	When assessing the sentences, did you remember the Stylus accuracy rate?	Percentage	0 – 100	n/a	n/a	
22	In your answers, did you consider the Stylus accuracy rate?	Percentage	0 – 100	n/a	n/a	
23	After assessing 100 sentences, in which Stylus indicated potential errors in 50, do you believe that Stylus was XX% accurate?	Percentage	0 – 100	n/a	n/a	
24	When thinking of the plausibility of the Stylus suggestions being created by an automated system, I would class it as	Percentage	0 – 100	n/a	n/a	
25	Sensitivity	d'_{NY}		1.67	0.35	
26	Bias	$c_{Y/N}$		0.35	0.47	
27	Sensitivity	d'_{Stylus}		n/a	n/a	
28	Bias	c_{Stylus}		n/a	n/a	
29	H rate	Trials		0.83	0.07	
30	M rate	Trials		0.17	0.07	
31	CR rate	Trials		0.73	0.12	
32	FA rate	Trials		0.27	0.12	
33	Confidence H	Percentage	50 – 100	92.15	6.05	
34	Confidence M	Percentage	50 – 100	82.98	10.67	
35	Confidence CR	Percentage	50 – 100	92.79	6.52	
36	Confidence FA	Percentage	50 – 100	90.02	6.31	
37	Confidence correct	Percentage	50 – 100	92.50	5.96	
38	Confidence incorrect	Percentage	50 – 100	87.57	7.13	
39	Confidence agreement with Stylus	Percentage	50 – 100	n/a	n/a	
40	Confidence disagreement with Stylus	Percentage	50 – 100	n/a	n/a	

Appendix B5 – Stylus 3 item distribution

Correct Item	Likelihood percentage	Version A	Version B	Version C	Version D	Version E	Version F	Version G	Version H
Stylus	53%	1	29	25	21	17	13	9	5
Stylus	53%	2	30	26	22	18	14	10	6
Stylus	53%	3	31	27	23	19	15	11	7
Stylus	53%	4	32	28	24	20	16	12	8
Stylus	53%	5	1	29	25	21	17	13	9
Stylus	53%	6	2	30	26	22	18	14	10
Stylus	53%	7	3	31	27	23	19	15	11
Stylus	53%	8	4	32	28	24	20	16	12
Stylus	75%	9	5	1	29	25	21	17	13
Stylus	75%	10	6	2	30	26	22	18	14
Stylus	75%	11	7	3	31	27	23	19	15
Stylus	75%	12	8	4	32	28	24	20	16
Stylus	75%	13	9	5	1	29	25	21	17
Stylus	75%	14	10	6	2	30	26	22	18
Stylus	75%	15	11	7	3	31	27	23	19
Stylus	75%	16	12	8	4	32	28	24	20
Stylus	75%	17	13	9	5	1	29	25	21
Stylus	75%	18	14	10	6	2	30	26	22
Stylus	75%	19	15	11	7	3	31	27	23
Stylus	75%	20	16	12	8	4	32	28	24
Original	53%	21	17	13	9	5	1	29	25
Original	53%	22	18	14	10	6	2	30	26
Original	53%	23	19	15	11	7	3	31	27
Original	53%	24	20	16	12	8	4	32	28
Original	53%	25	21	17	13	9	5	1	29
Original	53%	26	22	18	14	10	6	2	30
Original	53%	27	23	19	15	11	7	3	31
Original	53%	28	24	20	16	12	8	4	32
Original	75%	29	25	21	17	13	9	5	1
Original	75%	30	26	22	18	14	10	6	2
Original	75%	31	27	23	19	15	11	7	3
Original	75%	32	28	24	20	16	12	8	4
Stylus	53%	33	33	33	33	33	33	33	33

Appendix B6 – Stylus 4 item distribution

Correct Item	Likelihood	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
Stylus	53%	1	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	53%	2	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	53%	3	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	53%	4	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	53%	5	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	53%	6	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	53%	7	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	53%	8	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	9	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	10	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	11	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	12	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	13	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	14	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	15	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	16	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	17	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	18	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	19	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	20	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	21	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	22	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	94%	23	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	53%	24	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	53%	25	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	53%	26	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	53%	27	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	53%	28	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	53%	29	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	53%	30	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	53%	31	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Original	94%	32	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
Stylus	53%	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2

Appendix B7 – Stylus 5 item distribution

G90												
Sentence correct	Stylus highlights a perceived error	Stylus judgement correct	Version A	Version B	Version C	Version D	Version E	Version F	Version G	Version H	Version I	Version J
Yes	Yes	No	Q1	Q46	Q41	Q36	Q31	Q26	Q21	Q16	Q11	Q6
Yes	Yes	No	Q2	Q47	Q42	Q37	Q32	Q27	Q22	Q17	Q12	Q7
Yes	Yes	No	Q3	Q48	Q43	Q38	Q33	Q28	Q23	Q18	Q13	Q8
Yes	Yes	No	Q4	Q49	Q44	Q39	Q34	Q29	Q24	Q19	Q14	Q9
Yes	Yes	No	Q5	Q50	Q45	Q40	Q35	Q30	Q25	Q20	Q15	Q10
Yes	No	Yes	Q6	Q1	Q46	Q41	Q36	Q31	Q26	Q21	Q16	Q11
Yes	No	Yes	Q7	Q2	Q47	Q42	Q37	Q32	Q27	Q22	Q17	Q12
Yes	No	Yes	Q8	Q3	Q48	Q43	Q38	Q33	Q28	Q23	Q18	Q13
Yes	No	Yes	Q9	Q4	Q49	Q44	Q39	Q34	Q29	Q24	Q19	Q14
Yes	No	Yes	Q10	Q5	Q50	Q45	Q40	Q35	Q30	Q25	Q20	Q15
Yes	No	Yes	Q11	Q6	Q1	Q46	Q41	Q36	Q31	Q26	Q21	Q16
Yes	No	Yes	Q12	Q7	Q2	Q47	Q42	Q37	Q32	Q27	Q22	Q17
Yes	No	Yes	Q13	Q8	Q3	Q48	Q43	Q38	Q33	Q28	Q23	Q18
Yes	No	Yes	Q14	Q9	Q4	Q49	Q44	Q39	Q34	Q29	Q24	Q19
Yes	No	Yes	Q15	Q10	Q5	Q50	Q45	Q40	Q35	Q30	Q25	Q20
Yes	No	Yes	Q16	Q11	Q6	Q1	Q46	Q41	Q36	Q31	Q26	Q21
Yes	No	Yes	Q17	Q12	Q7	Q2	Q47	Q42	Q37	Q32	Q27	Q22
Yes	No	Yes	Q18	Q13	Q8	Q3	Q48	Q43	Q38	Q33	Q28	Q23
Yes	No	Yes	Q19	Q14	Q9	Q4	Q49	Q44	Q39	Q34	Q29	Q24
Yes	No	Yes	Q20	Q15	Q10	Q5	Q50	Q45	Q40	Q35	Q30	Q25
Yes	No	Yes	Q21	Q16	Q11	Q6	Q1	Q46	Q41	Q36	Q31	Q26
Yes	No	Yes	Q22	Q17	Q12	Q7	Q2	Q47	Q42	Q37	Q32	Q27
Yes	No	Yes	Q23	Q18	Q13	Q8	Q3	Q48	Q43	Q38	Q33	Q28
Yes	No	Yes	Q24	Q19	Q14	Q9	Q4	Q49	Q44	Q39	Q34	Q29
Yes	No	Yes	Q25	Q20	Q15	Q10	Q5	Q50	Q45	Q40	Q35	Q30
Yes	No	Yes	Q26	Q21	Q16	Q11	Q6	Q1	Q46	Q41	Q36	Q31
Yes	No	Yes	Q27	Q22	Q17	Q12	Q7	Q2	Q47	Q42	Q37	Q32
Yes	No	Yes	Q28	Q23	Q18	Q13	Q8	Q3	Q48	Q43	Q38	Q33
Yes	No	Yes	Q29	Q24	Q19	Q14	Q9	Q4	Q49	Q44	Q39	Q34
Yes	No	Yes	Q30	Q25	Q20	Q15	Q10	Q5	Q50	Q45	Q40	Q35
Yes	No	Yes	Q31	Q26	Q21	Q16	Q11	Q6	Q1	Q46	Q41	Q36
Yes	No	Yes	Q32	Q27	Q22	Q17	Q12	Q7	Q2	Q47	Q42	Q37
Yes	No	Yes	Q33	Q28	Q23	Q18	Q13	Q8	Q3	Q48	Q43	Q38
Yes	No	Yes	Q34	Q29	Q24	Q19	Q14	Q9	Q4	Q49	Q44	Q39
Yes	No	Yes	Q35	Q30	Q25	Q20	Q15	Q10	Q5	Q50	Q45	Q40
Yes	No	Yes	Q36	Q31	Q26	Q21	Q16	Q11	Q6	Q1	Q46	Q41
Yes	No	Yes	Q37	Q32	Q27	Q22	Q17	Q12	Q7	Q2	Q47	Q42
Yes	No	Yes	Q38	Q33	Q28	Q23	Q18	Q13	Q8	Q3	Q48	Q43
Yes	No	Yes	Q39	Q34	Q29	Q24	Q19	Q14	Q9	Q4	Q49	Q44
Yes	No	Yes	Q40	Q35	Q30	Q25	Q20	Q15	Q10	Q5	Q50	Q45
Yes	No	Yes	Q41	Q36	Q31	Q26	Q21	Q16	Q11	Q6	Q1	Q46
Yes	No	Yes	Q42	Q37	Q32	Q27	Q22	Q17	Q12	Q7	Q2	Q47
Yes	No	Yes	Q43	Q38	Q33	Q28	Q23	Q18	Q13	Q8	Q3	Q48
Yes	No	Yes	Q44	Q39	Q34	Q29	Q24	Q19	Q14	Q9	Q4	Q49
Yes	No	Yes	Q45	Q40	Q35	Q30	Q25	Q20	Q15	Q10	Q5	Q50
Yes	No	Yes	Q46	Q41	Q36	Q31	Q26	Q21	Q16	Q11	Q6	Q1
Yes	No	Yes	Q47	Q42	Q37	Q32	Q27	Q22	Q17	Q12	Q7	Q2
Yes	No	Yes	Q48	Q43	Q38	Q33	Q28	Q23	Q18	Q13	Q8	Q3
Yes	No	Yes	Q49	Q44	Q39	Q34	Q29	Q24	Q19	Q14	Q9	Q4
Yes	No	Yes	Q50	Q45	Q40	Q35	Q30	Q25	Q20	Q15	Q10	Q5
No	No	No	Q51	Q96	Q91	Q86	Q81	Q76	Q71	Q66	Q61	Q56
No	No	No	Q52	Q97	Q92	Q87	Q82	Q77	Q72	Q67	Q62	Q57
No	No	No	Q53	Q98	Q93	Q88	Q83	Q78	Q73	Q68	Q63	Q58
No	No	No	Q54	Q99	Q94	Q89	Q84	Q79	Q74	Q69	Q64	Q59

No	No	No	Q55	Q100	Q95	Q90	Q85	Q80	Q75	Q70	Q65	Q60
No	Yes	Yes	Q56	Q51	Q96	Q91	Q86	Q81	Q76	Q71	Q66	Q61
No	Yes	Yes	Q57	Q52	Q97	Q92	Q87	Q82	Q77	Q72	Q67	Q62
No	Yes	Yes	Q58	Q53	Q98	Q93	Q88	Q83	Q78	Q73	Q68	Q63
No	Yes	Yes	Q59	Q54	Q99	Q94	Q89	Q84	Q79	Q74	Q69	Q64
No	Yes	Yes	Q60	Q55	Q100	Q95	Q90	Q85	Q80	Q75	Q70	Q65
No	Yes	Yes	Q61	Q56	Q51	Q96	Q91	Q86	Q81	Q76	Q71	Q66
No	Yes	Yes	Q62	Q57	Q52	Q97	Q92	Q87	Q82	Q77	Q72	Q67
No	Yes	Yes	Q63	Q58	Q53	Q98	Q93	Q88	Q83	Q78	Q73	Q68
No	Yes	Yes	Q64	Q59	Q54	Q99	Q94	Q89	Q84	Q79	Q74	Q69
No	Yes	Yes	Q65	Q60	Q55	Q100	Q95	Q90	Q85	Q80	Q75	Q70
No	Yes	Yes	Q66	Q61	Q56	Q51	Q96	Q91	Q86	Q81	Q76	Q71
No	Yes	Yes	Q67	Q62	Q57	Q52	Q97	Q92	Q87	Q82	Q77	Q72
No	Yes	Yes	Q68	Q63	Q58	Q53	Q98	Q93	Q88	Q83	Q78	Q73
No	Yes	Yes	Q69	Q64	Q59	Q54	Q99	Q94	Q89	Q84	Q79	Q74
No	Yes	Yes	Q70	Q65	Q60	Q55	Q100	Q95	Q90	Q85	Q80	Q75
No	Yes	Yes	Q71	Q66	Q61	Q56	Q51	Q96	Q91	Q86	Q81	Q76
No	Yes	Yes	Q72	Q67	Q62	Q57	Q52	Q97	Q92	Q87	Q82	Q77
No	Yes	Yes	Q73	Q68	Q63	Q58	Q53	Q98	Q93	Q88	Q83	Q78
No	Yes	Yes	Q74	Q69	Q64	Q59	Q54	Q99	Q94	Q89	Q84	Q79
No	Yes	Yes	Q75	Q70	Q65	Q60	Q55	Q100	Q95	Q90	Q85	Q80
No	Yes	Yes	Q76	Q71	Q66	Q61	Q56	Q51	Q96	Q91	Q86	Q81
No	Yes	Yes	Q77	Q72	Q67	Q62	Q57	Q52	Q97	Q92	Q87	Q82
No	Yes	Yes	Q78	Q73	Q68	Q63	Q58	Q53	Q98	Q93	Q88	Q83
No	Yes	Yes	Q79	Q74	Q69	Q64	Q59	Q54	Q99	Q94	Q89	Q84
No	Yes	Yes	Q80	Q75	Q70	Q65	Q60	Q55	Q100	Q95	Q90	Q85
No	Yes	Yes	Q81	Q76	Q71	Q66	Q61	Q56	Q51	Q96	Q91	Q86
No	Yes	Yes	Q82	Q77	Q72	Q67	Q62	Q57	Q52	Q97	Q92	Q87
No	Yes	Yes	Q83	Q78	Q73	Q68	Q63	Q58	Q53	Q98	Q93	Q88
No	Yes	Yes	Q84	Q79	Q74	Q69	Q64	Q59	Q54	Q99	Q94	Q89
No	Yes	Yes	Q85	Q80	Q75	Q70	Q65	Q60	Q55	Q100	Q95	Q90
No	Yes	Yes	Q86	Q81	Q76	Q71	Q66	Q61	Q56	Q51	Q96	Q91
No	Yes	Yes	Q87	Q82	Q77	Q72	Q67	Q62	Q57	Q52	Q97	Q92
No	Yes	Yes	Q88	Q83	Q78	Q73	Q68	Q63	Q58	Q53	Q98	Q93
No	Yes	Yes	Q89	Q84	Q79	Q74	Q69	Q64	Q59	Q54	Q99	Q94
No	Yes	Yes	Q90	Q85	Q80	Q75	Q70	Q65	Q60	Q55	Q100	Q95
No	Yes	Yes	Q91	Q86	Q81	Q76	Q71	Q66	Q61	Q56	Q51	Q96
No	Yes	Yes	Q92	Q87	Q82	Q77	Q72	Q67	Q62	Q57	Q52	Q97
No	Yes	Yes	Q93	Q88	Q83	Q78	Q73	Q68	Q63	Q58	Q53	Q98
No	Yes	Yes	Q94	Q89	Q84	Q79	Q74	Q69	Q64	Q59	Q54	Q99
No	Yes	Yes	Q95	Q90	Q85	Q80	Q75	Q70	Q65	Q60	Q55	Q100
No	Yes	Yes	Q96	Q91	Q86	Q81	Q76	Q71	Q66	Q61	Q56	Q51
No	Yes	Yes	Q97	Q92	Q87	Q82	Q77	Q72	Q67	Q62	Q57	Q52
No	Yes	Yes	Q98	Q93	Q88	Q83	Q78	Q73	Q68	Q63	Q58	Q53
No	Yes	Yes	Q99	Q94	Q89	Q84	Q79	Q74	Q69	Q64	Q59	Q54
No	Yes	Yes	Q100	Q95	Q90	Q85	Q80	Q75	Q70	Q65	Q60	Q55

G70						
Sentence correct	Stylus highlights a perceived error	Stylus judgement correct	Version A	Version B	Version C	Version D
Yes	Yes	No	Q1	Q46	Q41	Q36
Yes	Yes	No	Q2	Q47	Q42	Q37
Yes	Yes	No	Q3	Q48	Q43	Q38
Yes	Yes	No	Q4	Q49	Q44	Q39
Yes	Yes	No	Q5	Q50	Q45	Q40

Yes	Yes	No	Q6	Q1	Q46	Q41
Yes	Yes	No	Q7	Q2	Q47	Q42
Yes	Yes	No	Q8	Q3	Q48	Q43
Yes	Yes	No	Q9	Q4	Q49	Q44
Yes	Yes	No	Q10	Q5	Q50	Q45
Yes	Yes	No	Q11	Q6	Q1	Q46
Yes	Yes	No	Q12	Q7	Q2	Q47
Yes	Yes	No	Q13	Q8	Q3	Q48
Yes	Yes	No	Q14	Q9	Q4	Q49
Yes	Yes	No	Q15	Q10	Q5	Q50
Yes	No	Yes	Q16	Q11	Q6	Q1
Yes	No	Yes	Q17	Q12	Q7	Q2
Yes	No	Yes	Q18	Q13	Q8	Q3
Yes	No	Yes	Q19	Q14	Q9	Q4
Yes	No	Yes	Q20	Q15	Q10	Q5
Yes	No	Yes	Q21	Q16	Q11	Q6
Yes	No	Yes	Q22	Q17	Q12	Q7
Yes	No	Yes	Q23	Q18	Q13	Q8
Yes	No	Yes	Q24	Q19	Q14	Q9
Yes	No	Yes	Q25	Q20	Q15	Q10
Yes	No	Yes	Q26	Q21	Q16	Q11
Yes	No	Yes	Q27	Q22	Q17	Q12
Yes	No	Yes	Q28	Q23	Q18	Q13
Yes	No	Yes	Q29	Q24	Q19	Q14
Yes	No	Yes	Q30	Q25	Q20	Q15
Yes	No	Yes	Q31	Q26	Q21	Q16
Yes	No	Yes	Q32	Q27	Q22	Q17
Yes	No	Yes	Q33	Q28	Q23	Q18
Yes	No	Yes	Q34	Q29	Q24	Q19
Yes	No	Yes	Q35	Q30	Q25	Q20
Yes	No	Yes	Q36	Q31	Q26	Q21
Yes	No	Yes	Q37	Q32	Q27	Q22
Yes	No	Yes	Q38	Q33	Q28	Q23
Yes	No	Yes	Q39	Q34	Q29	Q24
Yes	No	Yes	Q40	Q35	Q30	Q25
Yes	No	Yes	Q41	Q36	Q31	Q26
Yes	No	Yes	Q42	Q37	Q32	Q27
Yes	No	Yes	Q43	Q38	Q33	Q28
Yes	No	Yes	Q44	Q39	Q34	Q29
Yes	No	Yes	Q45	Q40	Q35	Q30
Yes	No	Yes	Q46	Q41	Q36	Q31
Yes	No	Yes	Q47	Q42	Q37	Q32
Yes	No	Yes	Q48	Q43	Q38	Q33
Yes	No	Yes	Q49	Q44	Q39	Q34
Yes	No	Yes	Q50	Q45	Q40	Q35
No	No	No	Q51	Q96	Q91	Q86
No	No	No	Q52	Q97	Q92	Q87
No	No	No	Q53	Q98	Q93	Q88
No	No	No	Q54	Q99	Q94	Q89
No	No	No	Q55	Q100	Q95	Q90
No	No	No	Q56	Q51	Q96	Q91
No	No	No	Q57	Q52	Q97	Q92
No	No	No	Q58	Q53	Q98	Q93
No	No	No	Q59	Q54	Q99	Q94
No	No	No	Q60	Q55	Q100	Q95
No	No	No	Q61	Q56	Q51	Q96
No	No	No	Q62	Q57	Q52	Q97
No	No	No	Q63	Q58	Q53	Q98
No	No	No	Q64	Q59	Q54	Q99
No	No	No	Q65	Q60	Q55	Q100
No	Yes	Yes	Q66	Q61	Q56	Q51
No	Yes	Yes	Q67	Q62	Q57	Q52
No	Yes	Yes	Q68	Q63	Q58	Q53
No	Yes	Yes	Q69	Q64	Q59	Q54
No	Yes	Yes	Q70	Q65	Q60	Q55
No	Yes	Yes	Q71	Q66	Q61	Q56
No	Yes	Yes	Q72	Q67	Q62	Q57
No	Yes	Yes	Q73	Q68	Q63	Q58
No	Yes	Yes	Q74	Q69	Q64	Q59

No	Yes	Yes	Q75	Q70	Q65	Q60
No	Yes	Yes	Q76	Q71	Q66	Q61
No	Yes	Yes	Q77	Q72	Q67	Q62
No	Yes	Yes	Q78	Q73	Q68	Q63
No	Yes	Yes	Q79	Q74	Q69	Q64
No	Yes	Yes	Q80	Q75	Q70	Q65
No	Yes	Yes	Q81	Q76	Q71	Q66
No	Yes	Yes	Q82	Q77	Q72	Q67
No	Yes	Yes	Q83	Q78	Q73	Q68
No	Yes	Yes	Q84	Q79	Q74	Q69
No	Yes	Yes	Q85	Q80	Q75	Q70
No	Yes	Yes	Q86	Q81	Q76	Q71
No	Yes	Yes	Q87	Q82	Q77	Q72
No	Yes	Yes	Q88	Q83	Q78	Q73
No	Yes	Yes	Q89	Q84	Q79	Q74
No	Yes	Yes	Q90	Q85	Q80	Q75
No	Yes	Yes	Q91	Q86	Q81	Q76
No	Yes	Yes	Q92	Q87	Q82	Q77
No	Yes	Yes	Q93	Q88	Q83	Q78
No	Yes	Yes	Q94	Q89	Q84	Q79
No	Yes	Yes	Q95	Q90	Q85	Q80
No	Yes	Yes	Q96	Q91	Q86	Q81
No	Yes	Yes	Q97	Q92	Q87	Q82
No	Yes	Yes	Q98	Q93	Q88	Q83
No	Yes	Yes	Q99	Q94	Q89	Q84
No	Yes	Yes	Q100	Q95	Q90	Q85

GC	
Sentence correct	Version A
Yes	Q1
Yes	Q2
Yes	Q3
Yes	Q4
Yes	Q5
Yes	Q6
Yes	Q7
Yes	Q8
Yes	Q9
Yes	Q10
Yes	Q11
Yes	Q12
Yes	Q13
Yes	Q14
Yes	Q15
Yes	Q16
Yes	Q17
Yes	Q18
Yes	Q19
Yes	Q20
Yes	Q21
Yes	Q22
Yes	Q23
Yes	Q24
Yes	Q25
Yes	Q26
Yes	Q27
Yes	Q28
Yes	Q29
Yes	Q30
Yes	Q31
Yes	Q32
Yes	Q33
Yes	Q34
Yes	Q35
Yes	Q36
Yes	Q37
Yes	Q38

Yes	Q39
Yes	Q40
Yes	Q41
Yes	Q42
Yes	Q43
Yes	Q44
Yes	Q45
Yes	Q46
Yes	Q47
Yes	Q48
Yes	Q49
Yes	Q50
No	Q51
No	Q52
No	Q53
No	Q54
No	Q55
No	Q56
No	Q57
No	Q58
No	Q59
No	Q60
No	Q61
No	Q62
No	Q63
No	Q64
No	Q65
No	Q66
No	Q67
No	Q68
No	Q69
No	Q70
No	Q71
No	Q72
No	Q73
No	Q74
No	Q75
No	Q76
No	Q77
No	Q78
No	Q79
No	Q80
No	Q81
No	Q82
No	Q83
No	Q84
No	Q85
No	Q86
No	Q87
No	Q88
No	Q89
No	Q90
No	Q91
No	Q92
No	Q93
No	Q94
No	Q95
No	Q96
No	Q97
No	Q98
No	Q99
No	Q100

Appendix C5 – Stylus 3 analyses with non-parametric measures

Sensitivity

Participants' average A' was $M = 0.76$ ($SD = 0.16$) in C53, and only marginally lower in C75 ($M = 0.74$, $SD = 0.20$), and there was no statistically significant difference between the groups $t(127) = 0.44$, $p = .663$, $d = 0.039$.

Bias

S3-H2 Participants will be more inclined to accept Stylus suggestions when these have higher likelihood ratings, and this will be true independently of correctness.

Participants in C53 had a positive B'' -score in C53 ($M = 0.11$, $SD = 0.24$), as well as in C75 ($M = 0.20$, $SD = 0.36$), which indicates a bias towards Stylus. A paired samples t -test revealed a statistically significant effect of Stylus' likelihood estimations on participants' bias towards Stylus, $t(127) = -2.62$, $p = .010$, $d = -0.231$.

Interaction effects

S3-H3 Participants' prior perceived self-efficacy in the domain of writing will be positively correlated with their sensitivity, independent of Stylus' reliability. In C53, there was a statistically significant correlation between participants' prior perceived self-efficacy and their sensitivity A' , $r(126) = .25$, $p = .004$. In C75 however, there was no statistically significant correlation between prior perceived self-efficacy and A' , $r(126) = .04$, $p = .640$.

S3-H4 Participants' prior perceived self-efficacy in the domain of writing will be negatively correlated with their bias, independent of Stylus' reliability. In C53, there was no statistically significant correlation between participants' prior perceived self-efficacy and their bias B'' , $r(126) = -.06$, $p = .532$. In C75 we did not find a statistically significant correlation either between prior perceived self-efficacy and B'' , $r(126) = .04$, $p = .633$.

S3-H5 Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.

There was no statistically significant correlation between prior trust and B'' in C53, $r(126) = .00$, $p = .998$, nor in C75, $r(126) = .03$, $p = .700$.

Appendix C7 – Stylus 5 ANOVA table

Repeated Measures ANOVA confidence Within Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G
Stylus advice	149.699	1	149.699	6.018	0.017	0.002
Stylus advice * Group	104.821	1	104.821	4.214	0.044	0.002
Residual	1840.908	74	24.877			
Sentence correctness	1197.090	1	1197.090	25.985	< .001	0.020
Sentence correctness * Group	69.952	1	69.952	1.518	0.222	0.001
Residual	3409.033	74	46.068			
Response correctness	3925.469	1	3925.469	71.428	< .001	0.061
Response correctness * Group	154.319	1	154.319	2.808	0.098	0.003
Residual	4066.819	74	54.957			
Sentence correctness * Stylus advice	17.182	1	17.182	0.413	0.522	0.000
Sentence correctness * Stylus advice * Group	3.031	1	3.031	0.073	0.788	0.000
Residual	3075.691	74	41.563			
Sentence correctness * Response correctness	1373.676	1	1373.676	25.806	< .001	0.022
Sentence correctness * Response correctness * Group	152.470	1	152.470	2.864	0.095	0.003
Residual	3939.085	74	53.231			
Response correctness * Stylus advice	538.376	1	538.376	17.962	< .001	0.009
Response correctness * Stylus advice * Group	69.181	1	69.181	2.308	0.133	0.001
Residual	2218.001	74	29.973			
Sentence correctness * Response correctness * Stylus advice	54.295	1	54.295	1.445	0.233	0.001
Sentence correctness * Response correctness * Stylus advice * Group	46.537	1	46.537	1.239	0.269	0.001
Residual	2779.899	74	37.566			

Note. Type III Sum of Squares

Between Subjects Effects

	Sum of Squares	df	Mean Square	F	p	η^2_G
Group	128.129	1	128.129	0.245	0.622	0.002
Residual	38696.505	74	522.926			

Note. Type III Sum of Squares

Descriptives

Stylus advice	Sentence correctness	Response correctness	Group	Mean	SD	N
GOOD	Correct	Correct	G90	92.276	7.423	38
			G70	92.432	6.971	38
		Incorrect	G90	82.489	11.674	38
			G70	81.083	12.337	38
	Incorrect	Correct	G90	91.726	6.630	38
			G70	92.060	7.206	38
		Incorrect	G90	88.251	16.455	38
			G70	88.815	9.772	38
BAD	Correct	Correct	G90	90.869	9.542	38
			G70	91.538	7.877	38
		Incorrect	G90	88.497	10.953	38
			G70	82.692	13.510	38
	Incorrect	Correct	G90	91.666	9.099	38
			G70	90.864	7.982	38
		Incorrect	G90	91.002	8.970	38
			G70	89.945	9.243	38

Appendix D – Hypotheses overview

<i>Study</i>	S1	S1	S1	S1	S1	S1	S1
<i>Number</i>	H1	H2	H3	H4	H5	H6	H7
<i>Result</i>	Confirmed	Rejected	Rejected	Rejected	Confirmed	Rejected, surprising result	Confirmed
<i>Hypothesis</i>	Participants' prior perceived self-efficacy in the domain of writing will be greater than their estimation of the efficacy of the average British English speaker.	Participants' performance, in terms of percentage correct, will be better in GGood than in GBad.	Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.	Participants' perceived self-efficacy in the domain of writing will be negatively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.	Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses (the <i>overconfidence effect</i>).	The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.	Participants' trust in Stylus during the experiment (measured post-task) will be higher in GGood than in GBad.

<i>Study</i>	S2	S2	S2	S2	S2	S2	S2	S2	S2
<i>Number</i>	H1	H2	H3	H4	H5	H6	H7		
<i>Result</i>	Confirmed	Rejected	Confirmed with parametric measure only	Rejected	Confirmed	Confirmed	Confirmed	Confirmed	Confirmed
<i>Hypothesis</i>	Participants' prior perceived self-efficacy in the domain of writing will be greater than their estimation of the efficacy of the average British English speaker.	Participants' performance, in terms of percentage correct, will be better in GGood than in GBad.	Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.	Participants' perceived self-efficacy in the domain of writing will be negatively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.	Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses (the <i>overconfidence effect</i>).	The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.	Participants' trust in Stylus during the experiment (measured post-task) will be higher in GGood than in GBad.		

<i>Study</i>	S3	S3	S3	S3	S3	S3	S3	S3	S3
<i>Number</i>	H1	H2	H3	H4	H5	H6	H7	H8	H9
<i>Result</i>	Confirmed	Confirmed	Rejected	Rejected	Rejected	Confirmed	Confirmed	Confirmed	Confirmed
<i>Hypothesis</i>	Participants' performance, in terms of percentage correct, will be better when Stylus' reliability is 75% than when it is 53%.	Participants will be more inclined to accept Stylus suggestions when these have higher likelihood ratings, and this will be true independent of correctness.	Participants' prior perceived self-efficacy in the domain of writing will be positively correlated with their sensitivity, independent of Stylus' reliability.	Participants' prior perceived self-efficacy in the domain of writing will be negatively correlated with their bias, independent of Stylus' reliability.	Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations, independent of correctness of the latter.	Participants will be more confident when using more reliable Stylus suggestions.	Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses.	Participants will be more confident when responding correctly.	The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.

<i>Study</i>	S4	S4	S4	S4	S4	S4	S4	S4	S4
<i>Number</i>	H1	H6	H7	H8	H9	H10	H11		
<i>Result</i>	Confirmed	Confirmed	Confirmed	Rejected, surprising result	Confirmed	Rejected	Rejected	Rejected	Rejected
<i>Hypothesis</i>	Participants' performance, in terms of percentage correct, will be better when Stylus' reliability is 94% than when it is 53%.	Participants will be more confident when using more reliable Stylus suggestions.	Participants' mean percentage of trial confidence judgements will be higher than their percentage of correct responses (the overconfidence effect).	Participants will be more confident when responding correctly.	The overconfidence effect will be less marked in a comparison between participants' performance, in terms of percentage correct, and their post-hoc estimation of their own performance.	In the single Original is correct – strength of Stylus advice is high trial, participants who respond incorrectly (Stylus is correct) will have rated their prior perceived self-efficacy in the domain of writing lower than participants who respond correctly.	In the single Original is correct – strength of Stylus advice is high trial, participants who respond incorrectly (Stylus is correct) will have rated their prior trust in automated writing style checkers higher than participants who respond correctly.		

<i>Study</i>	S5	S5	S5	S5	S5	S5	S5	S5	S5
<i>Number</i>	H1	H2	H3	H4	H5	H6	H7	H8	
<i>Result</i>	Rejected	Rejected	Rejected	Confirmed	Confirmed, surprising result	Rejected	Rejected	Rejected	Rejected
<i>Hypothesis</i>	Participants' sensitivity will be higher when Stylus' reliability is 90% than when it is 70%.	Participants' acceptance of Stylus suggestions will be higher when Stylus' reliability is 90% than when it is 70%, independently of correctness of the latter.	Participants' will be more confident when Stylus' reliability is 90% than when it is 70%.	Participants will be more confident when responding correctly.	Participants' prior perceived self-efficacy in the domain of writing will be positively correlated with their sensitivity, independent of Stylus' reliability.	Participants' prior perceived self-efficacy in the domain of writing will be negatively correlated with their bias, their tendency to accept Stylus advice.	Participants' prior trust in automated writing style checkers will be positively correlated with their acceptance of Stylus recommendations.	Participants' trust in Stylus during the experiment (measured post-task) will be positively correlated with acceptance of correct Stylus suggestions.	

Appendix E1 – University of Bath Department of Computer Science ethics check list

UNIVERSITY OF BATH

Department of Computer Science

13-POINT ETHICS CHECK LIST

This document describes the 13 issues that need to be considered carefully before students or staff involve other people (“participants”) for the collection of information as part of their project or research.

1. Have you prepared a briefing script for volunteers?

The experiment starts with an introduction, explaining participants what they will be required to do during the experiment, what kind of data is recorded, how it is stored and how it will be used.

You must explain to people what they will be required to do, the kind of data you will be collecting from them and how it will be used.

2. Will the participants be using any non-standard hardware?

No.

Participants should not be exposed to any risks associated with the use of non-standard equipment: anything other than pen and paper or typical interaction with PCs on desks is considered non-standard.

3. Is there any intentional deception of the participants?

Participants are told this experiment is about text editing efficiency, whilst in fact trust and perceived self-efficacy will be measured. This neutral explanation is necessary to avoid biasing participants; I deem it unlikely this will cause any participant to object or show unease when debriefed.

Withholding information or misleading participants is unacceptable if participants are likely to object or show unease when debriefed.

4. How will participants voluntarily give consent?

Participants will have to consent to taking part in the experiment online, their responses being recorded and stored anonymously and securely in a database and their responses possibly being used anonymously for publications and future research before being able to start the experiment.

If the results of the evaluation are likely to be used beyond the term of the project (for example, the software is to be deployed, or the data is to be

published), then signed consent is necessary. A separate consent form should be signed by each participant.

5. Will the participants be exposed to any risks greater than those encountered in their normal work life?

No.

Investigators have a responsibility to protect participants from physical and mental harm during the investigation. The risk of harm must be no greater than in ordinary life.

6. Are you offering any incentive to the participants?

Not sure yet.

The payment of participants must not be used to induce them to risk harm beyond that which they risk without payment in their normal lifestyle.

7. Are any of your participants under the age of 16?

No.

Parental consent is required for participants under the age of 16.

8. Do any of your participants have an impairment that will limit their understanding or communication?

No.

Additional consent is required for participants with impairments.

9. Are you in a position of authority or influence over any of your participants?

No.

A position of authority or influence over any participant must not be allowed to pressurise participants to take part in, or remain in, any experiment.

10. Will the participants be informed that they could withdraw at any time?

Yes.

All participants have the right to withdraw at any time during the investigation. They should be told this in the introductory script.

11. Will the participants be informed of your contact details?

Yes.

All participants must be able to contact the investigator after the investigation. They should be given the details of the Unit Lecturer or Supervisor as part of the debriefing.

12. Will participants be de-briefed?

After they have completed the study, I will issue all participants with a brief description of the study and its hypotheses and purpose.

The student must provide the participants with sufficient information in the debriefing to enable them to understand the nature of the investigation.

13. Will the data collected from the participants be stored in an anonymous form?

Yes.

All participant data (hard copy and soft copy) should be stored securely, and in anonymous form.

NAME: Melle Zijlstra

SUPERVISOR (IF APPLICABLE): Prof Stephen Payne

SECOND READER (IF APPLICABLE):

PROJECT TITLE: Stylus

DATE: 23 March 2018

Appendix E2 – Stylus Qualtrics consent screen



Thank you for agreeing to take part in this experiment! Completing the survey should take approximately 30 minutes; you will have to complete it in one go.

This experiment has been designed following the University of Bath Department of Computer Science's local ethics guidelines and it is approved by the Psychology Research Ethics Committee (reference number 19-285). No harm will come to you from taking part in this experiment and you have the right to stop at any time.

Before you continue, we need your consent to the following:

1. I consent to taking part in this experiment online
2. I understand and consent to my responses being recorded and stored anonymously and securely in a database
3. I understand and consent to my responses may be used anonymously for publications and future research

I consent to the above

Please enter your Prolific ID:

Appendix E3 – Stylus Qualtrics debrief screen

Thank you for taking part in this experiment. In our research, we're interested in finding out about people's subjective thinking and their decision making when they're given suggestions by automated systems. We hope our research will lead to a better understanding of how humans make decisions when working with automated systems in general.

Melle Zijlstra – m.zijlstra@bath.ac.uk

Prof Stephen Payne – s.j.payne@bath.ac.uk

University of Bath, Department of Computer Science

Next

