Title

Tools That Teach Too: Exploring the Role of Interaction in the Development of Useful Cognitive Residuals

Author
Fleming, Morgan Magnus

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, MERCED

Tools That Teach Too: Exploring the Role of Interaction in the Development of Useful Cognitive Residuals

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

In

Cognitive and Information Sciences

by

Morgan Magnus Fleming

Committee in charge:
Professor Paul Maglio, Chair
Professor Teenie Matlock
Professor Carolyn Jennings

Dedicated to my committee, my family, my friends, and my acquaintances.

Without you, neither this document nor I would be what we are today.

The Dissertation of Morgan Magnus Fleming is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Carolyn Jennings

Teenie Matlock

Paul Maglio, Chair

University of California, Merced
2019

Table of Contents

ACKNOWLEDGEMENTS X

CURRICULUM VITAE XI

ABSTRACT XII

CHAPTER 1 - INTRODUCTION 1

THE COGNITIVE CONSEQUENCES OF TOOL USE 2
INTERACTIVE SPELL-CHECKER, PASSIVE AUTOCORRECT: FINDINGS 4
A NEW WAY TO INTERACT WITH SPELL-CHECKING: FINDINGS 5
CONCLUSION 6

CHAPTER 2 - TOOL USE AND HUMAN COGNITION 7

HUMAN COGNITION IS AFFECTED BY OUR TOOLS 8
COGNITIVE RESIDUALS AND TOOL-SHAPED INTERACTIONS 9
DETRIMENTAL COGNITIVE RESIDUALS 11
EXPLAINING THE DETRIMENTS 13
Technical Occlusion 13
User Responsibilities 15
CONCLUSION 16

CHAPTER 3 - STUDIES 17

STUDY 1 - A COMPARISON OF EDITING SKILL DEVELOPMENT USING MICROSOFT WORD'S AUTOCORRECT AND SPELL-CHECKING SUPPORT SYSTEMS 17
Methods 19
Results and Discussion 23
Conclusions for Study 1 26
STUDY 2 – INTERACTION STYLES AND WORD LEARNING: COMPARING AUTOCORRECT AND SPELL-CHECK WHEN SUPPORT SYSTEM PERFORMANCE IS CONTROLLED 27
Methods 27
Results and Discussion 31
Conclusions for Study 2 32
STUDY 3 - INTERACTION STYLES AND WORD LEARNING: COMPARING SPELL-CHECK TO AUTOCORRECT WITH DIFFICULT WORDS 33
Methods 33
Results and Discussion 37
Conclusions for Study 3 40
STUDY 4- INTERACTION STYLES AND WORD LEARNING: A COMPARISON OF SELECTING A WORD WITH RECEIVING A WORD 41
Methods 42
Results and Discussion 48
Conclusions for Study 4 51
STUDY 5 - BUILDING A SPELL-CHECK THAT TEACHES: INSTRUCTIVE DESIGN THROUGH PROCEDURE OR PERCEPTION 52
Methods 56
Results and Discussion 60
Conclusions for Study 5 62

CHAPTER 4 - CONCLUSION 64

REVIEW 64
METHODOLOGICAL ISSUES 66

Future Directions 67

SUMMARY **69**

REFERENCES **70**

APPENDIX A **76**

A.1 Materials 76

A.1.1 Stimuli *76*

A.2 Model Verification 78

A.2.1 Data Exploration *78*

A.2.2 Model Validation *90*

A.2.3 Model Fit *94*

APPENDIX B **97**

B.1 Materials 97

B.1.1 Website and Support Systems *97*

B.1.2 Stimuli *98*

B.2 Model Verification 99

B.2.1 Data Exploration *99*

B.2.2 Model Validation *106*

B.2.3 Model Fit *109*

APPENDIX C **111**

C.1 Materials 111

C.1.1 Website and Support Systems *111*

C.1.2 Stimuli *113*

C.2 Model Verification 114

C.2.1 Data Exploration *114*

C.2.2 Model Validation *122*

C.2.3 Model Fit *124*

APPENDIX D **126**

D.1 Materials 126

D.1.1 Website and Devices *126*

D.1.2 Stimuli *128*

D.2 Model Verification 129

D.2.1 Data Exploration *129*

D.2.2 Model Validation *139*

D.2.3 Model Fit *146*

APPENDIX F **149**

F.1 Materials 149

F.1.1 Website and Devices *149*

F.1.2 Stimuli *152*

F.2 Model Verification 154

F.2.1 Data Exploration *154*

F.2.2 Model Validation *160*

F.2.3 Model Fit *161*

Table of Figures

Figure 1 - Diagram of a Set 20
Figure 2 - Diagram of Experiment Flow 20
Figure 3 -Pre-Test and Post-Test Scores, by Condition 23
Figure 4 - Distribution of Scores (Number of Errors) on Manipulated Trials 25
Figure 5 - Screenshot of Website Used to Gather Data 29
Figure 6 -Pre-test and Post-test Scores, grouped by Condition 31
Figure 7- Screenshot of Website Used to Gather Data 34
Figure 8 - Scores on the Pre-Test and Post-Test, Grouped by Condition 37
Figure 9-Number of Correctly Spelled Words on Trials with Assistant, by Assistant Type 39
Figure 10 – Screenshot of Webpage Used for Data Collection 43
Figure 11 - Pre-Test and Post-Test Scores, by Condition (Hard Words) 48
Figure 12 - Pre-Test and Post-Test Scores, by Condition (Easy Words) 49
Figure 13- Violin Plot of Scores on Trials with Support Systems Enabled 51
Figure 14 - Screenshot of the Website Used to Gather Data 56
Figure 15 - Pre-Test Scores Grouped by Device 59
Figure 16 - Post-Test Scores Grouped by Device 59
Figure 17 - Predicted Post-Test Score Grouped by Type of Spelling Support System Used During Training Trials 62
Figure 18 - Distribution of Scores (Number of Errors) on Manipulated Trials 78
Figure 19 - Results from the Kruskal-Wallis Rank Sum Test- 79
Figure 20 - Population chart of Pre-Post Score Differences 80
Figure 21 - Violin Plot of Pre-Test 81
Figure 22 - Pre-Test vs Post-Test Score, by Condition 82
Figure 23 - Pre-Test vs Post-Test Score, by Condition (With Guiding Line) 83
Figure 24- Measure of Poissonality of Response Variable 84
Figure 25 - Cleaveland Dot Chart of Scores on Post-Test 85
Figure 26 - Cleaveland Dot Chart of Scores on Post-Test, Grouped by Condition 86
Figure 27 -Cleaveland Dot Chart of Pre-Test Scores 87
Figure 28 - Cleaveland Dot Chart of Pre-Test Scores, by Condition 88
Figure 29 - Influential Observation Diagnostic Plot 89
Figure 30-Output from Levene Test of Homogeneity of Variance on Pearson residuals, by Condition 90
Figure 31 - Pearson Residuals vs Fitted Post-Test Scores 91
Figure 32 - Pearson Residuals vs Pre-Test Scores 92
Figure 33 -Pearson Residuals vs Transfer Test Scores 93
Figure 34 - Results of GLMM 94
Figure 35 - R^2 of GLMM 95
Figure 36 - Result of GLM 95
Figure 37 - GLM for Transfer Test 96
Figure 38 – R^2 for Transfer Test Model 96
Figure 39 - Screenshot of Website Used to Gather Data 97
Figure 40- Violin Plot of Scores on Trials with Device, by Condition 99
Figure 41 - Kruskal-Wallis Rank Sum Test Between Conditions 100
Figure 42 - Score on Pre-Test vs Score on Post Test, Grouped by Condition 100
Figure 43 - Score on Pre-Test vs Score on Post-Test, Grouped by Condition 101
Figure 44 - Q-Q Plot of Estimated Poisson Quantiles and Observed Scores on Post-Test 102
Figure 45 - Cleaveland Plot of Pre-Test Scores 103

Figure 46 - Cleaveland Plot of Post-Test Scores ... 104
Figure 47 - Cleaveland Plot of Pre-Test Scores, Grouped by Condition ... 104
Figure 48 - Cleaveland Plot of Post-Test Scores, Grouped by Condition ... 105
Figure 49 - Outlier Diagnostic Charts for Poisson Model with OLRE ... 105
Figure 50- Levene's Test of Homogeneity of Variance for Poisson Regression with OLRE ... 106
Figure 51 - Pearson Residuals from Poisson Regression with OLRE vs Observed Score on Pre-Test ... 108
Figure 52 - Pearson Residuals from Poisson Regression with OLRE vs Predicted Score on Post-Test ... 108
Figure 53 - Results from GLMM (Poisson) Regression with OLRE ... 109
Figure 54 -R^2 of GLMM ... 109
Figure 55 - Results from Quasipoisson Regression ... 110
Figure 56- Screenshot of Website Used to Gather Data ... 111
Figure 57 - Kruskal-Wallis Rank Sum Test ... 114
Figure 58-Number of Correctly Spelled Words on Trials with Assistant, by Assistant Type ... 115
Figure 59 - Post-Test Scores vs Pre-Test Scores, Grouped by Condition ... 116
Figure 60 - Post-Test Scores vs Pre-Test Scores, Grouped by Condition (line) ... 117
Figure 61 - Q-Q Plot of Estimated Poisson Quantiles vs Observed Score on Post-Test ... 118
Figure 62-Cleaveland Plot of Post-Test Scores ... 119
Figure 63- Cleaveland Plot of Post-Test Scores, Grouped by Condition ... 120
Figure 64 - Cleaveland Plot of Pre-Test Scores ... 120
Figure 65 - Cleaveland Plot of Pre-Test Scores by Condition ... 121
Figure 66-Outlier Diagnostics for Poisson Model with OLRE ... 121
Figure 67 - Levene's Test for Homogeneity in Poisson Model with OLRE ... 122
Figure 68-Pearson Residuals from Poisson Regression vs Predicted Number of Target Words Spelled Correctly on Post-Test ... 123
Figure 69- Pearson Residuals for Model vs Score on Pre-Test, Grouped by Condition ... 123
Figure 70 - Results from Poisson Regression with OLRE ... 124
Figure 71- Results from Quasipoisson Regression ... 124
Figure 72 -R^2 for GLMM ... 125
Figure 73 - Multiple Comparisons Between Conditions,Output From 'glht' ... 125
Figure 74 - Screenshot of Website Used to Collect Data ... 126
Figure 75 - Rank Sum Test (Hard Words) ... 129
Figure 76- Rank Sum Test (Easy Words) ... 130
Figure 77 - Violin Plot of Score on Test with Support System (Easy Words) ... 130
Figure 78- Violin Plot of Score on Test with Support System (Hard Words) ... 131
Figure 79 - Plot of Post-Test Scores for Hard Words vs Pre-Test Scores for Hard Words, Grouped by Condition ... 132
Figure 80 - Plot of Post-Test Scores for Easy Words vs Pre-Test Scores for Easy Words, Grouped by Condition ... 132
Figure 81 - Q-Q Plots of Poisson Quantile Estimates of Post-Test Scores vs Observed Post ... 133
Figure 82 - Q-Q Plots of Poisson Quantile Estimates of Post-Test Scores vs Observed Post-Test Scores With Easy Words ... 134
Figure 83-Cleaveland Plot of Post-Test Scores With Hard Words ... 135
Figure 84-Cleaveland Plot of Post-Test Scores with Hard Words, Grouped by Condition ... 135
Figure 85 - Diagnostic Plot for Outliers, Hard Words ... 136

FIGURE 86- CLEAVELAND PLOT OF POST-TEST SCORES WITH EASY WORDS 137
FIGURE 87- CLEAVELAND PLOT OF POST-TEST SCORES WITH EASY WORDS, GROUPED BY CONDITION 137
FIGURE 88 - DIAGNOSTIC OUTLIER PLOT, EASY WORDS 138
FIGURE 89- LEVENE'S TEST, EASY WORDS 139
FIGURE 90- LEVENE'S TEST, HARD WORDS 139
FIGURE 91- PEARSON RESIDUALS FROM EASY MODEL VS PREDICTOR VARIABLE: PRE-TEST..... 140
FIGURE 92- PEARSON RESIDUALS FROM EASY MODEL VS PREDICTED POST-TEST SCORES 140
FIGURE 93 - PEARSON RESIDUALS FROM HARD MODEL VS PREDICTED POST-TEST SCORES...... 141
FIGURE 94 - PEARSON RESIDUALS FROM HARD MODEL VS PREDICTOR: PRE-TEST 143
FIGURE 95 - PEARSON RESIDUALS FROM EASY MODEL VS NON-MODEL CO-VARIATE 144
FIGURE 96 - PEARSON RESIDUALS FROM HARD MODEL VS NON-MODEL CO-VARIATE............... 145
FIGURE 98 - QUASIPOISSON FIT FOR EASY WORDS 147
FIGURE 99 - R^2 FOR MODEL OF EASY WORDS 147
FIGURE 100- POISSON REGRESSION WITH OLRE FOR HARD WORDS 148
FIGURE 101 - R^2 FOR HARD WORD MODEL 148
FIGURE 101 - R^2 FOR HARD WORD MODEL 148
FIGURE 102 - SCREENSHOT OF THE WEBSITE USED TO GATHER DATA 149
FIGURE 103 – RANK SUM TEST (INCLUDES CORRECTED CONTROL) 154
FIGURE 104 - PRE-TEST SCORES GROUPED BY DEVICE 155
FIGURE 105 - POST-TEST SCORES GROUPED BY DEVICE 155
FIGURE 106 – Q-Q PLOT OF ESTIMATED POISSON DISTRIBUTION AND OBSERVED VALUES, BY CONDITION 156
FIGURE 107- CLEAVELAND DOT CHART OF PRE-TEST SCORES... 157
FIGURE 108 - CLEAVELAND DOT CHART OF PRE-TEST SCORES BY CONDITION 158
FIGURE 109- CLEAVELAND DOT CHART OF POST-TEST SCORES BY CONDITION 159
FIGURE 110 - LEVENE'S TEST FOR HOMOGENEITY OF VARIANCE FOR GLMM 160
FIGURE 111 - LEVENE'S TEST FOR HOMOGENEITY OF VARIANCE FOR QUASIPOISSON GLM 160
FIGURE 112- GENERALIZED POISSON MIXED MODEL FOR POST-TEST SCORES 161
FIGURE 113 -R^2 FOR GENERALIZED POISSON MIXED MODEL (POST-TEST SCORES) 161
FIGURE 114 – GENERALIZED POISSON MIXED MODEL FOR TRANSFER TEST SCORES 162
FIGURE 115 – R^2 FOR GENERALIZED POISSON MIXED MODEL (TRANSFER TEST)...................... 162
FIGURE 116 -QUASI-POISSON MODEL FOR TRANSFER TEST SCORES... 163
FIGURE 117 - R^2 FOR QUASI-POISSON MODEL (TRANSFER TEST SCORES)................................. 163
FIGURE 118 - TUKEY POST-HOC TEST FOR POST-TEST SCORES 164
FIGURE 119 -TUKEY POST-HOC TEST FOR TRANSFER TEST SCORES (QUASI-POISSON) 165

Table of Tables

TABLE 1 - PHASES OF THE SPELLING ASSISTANCE PROCESS.. 19
TABLE 2 - TABLE OF REGRESSION COEFFICIENTS FOR STUDY 1 24
TABLE 4 - TABLE OF REGRESSION COEFFICIENTS FOR STUDY 3 38
TABLE 5 - SOFTWARE ASSISTANCE TABLE 42
TABLE 6 - TARGET WORDS 45
TABLE 7 - TABLE OF COEFFICIENTS, EASY WORDS IN STUDY 4 50
TABLE 8 - TABLE OF COEFFICIENTS, HARD WORDS IN STUDY 4.. 50
TABLE 9 - TABLE OF REGRESSION COEFFICIENTS... 60
TABLE 10 - TARGET WORDS 76
TABLE 11 - PARAMETER ESTIMATIONS WITH INFLUENTIAL OBSERVATIONS REMOVED 90
TABLE 12 - TABLE OF GENERALIZED POISSON MIXED MODE REGRESSION COEFFICIENTS 94
TABLE 13 - TARGET WORDS 98
TABLE 14- TABLE OF REGRESSION COEFFICIENTS AFTER OUTLIER REMOVAL........................... 106
TABLE 15- TABLE OF GLMM COEFFICIENTS 109
TABLE 16 - TABLE OF USE STEPS FOR SPELLING SUPPORT SYSTEMS 112
TABLE 17 - TARGET WORDS 113
TABLE 18- TABLE OF REGRESSION COEFFICIENTS AFTER REMOVAL OF OUTLIERS.................... 122
TABLE 19- PHASES OF THE SPELLING ASSISTANCE PROCESS 127
TABLE 20 - TABLE OF TARGET WORDS FOR STUDY 4.. 128
TABLE 21 - TABLE OF REGRESSION COEFFICIENTS AFTER REMOVAL OF OUTLIERS, HARD WORDS 136
TABLE 22 - TABLE OF REGRESSION COEFFICIENTS AFTER REMOVAL OF OUTLIERS, EASY WORDS 138
TABLE 23 - PHASES OF THE SPELLING ASSISTANCE PROCESS.. 150
TABLE 24 - TARGET WORDS 152
TABLE 25 - PARAMETER ESTIMATES WITH OUTLIERS REMOVED 159

Acknowledgements

I would first like to thank my committee. Dr. Paul Maglio, Dr. Teenie Matlock, and Dr. Carolyn Jennings have shown me great patience, wisdom, knowledge, and kindness during my time here at UC Merced. Without their guidance and support, I would not have been able to complete this project. Thank you all so much, it has been wonderful working with you.

I would also like to thank the many research assistants that helped me gather data. These are Lamar Williams, Cesar Gamez, Milagros Lopez Obeso, Tyler Jones, Reed Urman, Kaitlyn Silvey, and Chris Reding. Aside from being wonderfully diligent and capable, each also contributed crucial observations and indispensable input. I do not know how I managed to get so lucky as to have had the opportunity to work with all of them.

I want to thank Mitch Ylarregui, who assisted me through several bureaucratic hang-ups during my time at UC Merced. I will also thank Dr. Ramesh Balasubramaniam and Dr. Michael Spivey for similar assistance, though their help also included some stellar academic guidance as well. I will also thank Benjamin St Claire, Melissa Eisner, Molly Caroline, and Devon Batey, who have been thrilling conversationalists, supportive friends, and able competitors in my spare time. A thank you to Umesh Krishnamurthy as well. It was a blast palling around the department with you. I would also like to thank the UC Merced Cognitive and Information Sciences department, professors and students alike. This community has been immensely enjoyable, genuinely kind, and stunningly supportive. Thank you all!

Curriculum Vitae

Morgan Magnus Fleming

University of California, Merced

2013	Bachelor of Arts	Cognitive Science

University of California, Merced

2014-2015	Graduate Researcher	Cognitive and Information Sciences
2013 -2019	Teaching Assistant	Cognitive and Information Sciences
2017	Instructor	Cognitive and Information Sciences
2019	Doctorate of Philosophy	Cognitive and Information Sciences

Publications

Fleming, M. M., & Maglio, P. (2014, November 6). *Building Words through Physical Action.* Poster session presented at the meeting of CSDL, Santa Barbara, CA.

Fleming, M. M. & Maglio, P. (2015, July). *How Interaction helps performance in a Scrabble-like task*. Poster session presented at the 37th Annual Cognitive Science Meeting, Pasadena, CA.

Fleming, M. M., Maglio, P. (2016, August). *On Measuring the Difficulty of Scrabble-like Problems*. Poster session presented at the meeting of the Cognitive Science Society, Philadelphia, PA.

Obeso, M. F. L. & Fleming, M. M. (2015, July). *Modern Symbolic Communication Through Non-Word Text and Images.* Poster session presented at the 37th Annual Cognitive Science Meeting, Pasadena, CA

Matlock, T., Castro, S. C., Fleming, M., Gann, T. M., & Maglio, P. P. (2014). Spatial Metaphors of Web Use. *Spatial Cognition & Computation*, *14*(4), 306-320

Bunce, J., Gordon, C., Abney, D., Fleming, M. M., Greenwood, M., Chiu, E., Spivey, M., & Scott, R. (2015, November 14). *Mouse tracking reveals knowledge of multiple competing referents during cross-situational word learning*. Poster session presented at the meeting of Boston University Conference on Language Development, Boston, CA.

Lobato, E. J., Tabatabaeian, S., Fleming, M., Sulzmann, S., & Holbrook, C. (2019). Religiosity Predicts Evidentiary Standards. *Social Psychological and Personality Science*, 1948550619869613.

Abstract

Tools That Teach Too: Exploring the Role of Interaction in the Development of Useful Cognitive Residuals

Morgan Magnus Fleming

Cognitive and Information Sciences

University of California, Merced 2019

Professor Paul Maglio, Chair

Tools help us perform tasks. But they also change the way we think. In five studies, I explored the lasting effects — the cognitive residuals — of using two common forms of automated spelling assistance, spell-checking and autocorrect. These different spelling assistance systems aim to provide similar spelling outcomes, but users interact quite differently with each. By comparing differences in spelling skills after using these different types of spelling assistance, I assessed differences in the cognitive residuals that arise from direct interaction (spell-checking) and passive support (autocorrect). Results suggest that the passive support provided by autocorrect is just as good as for the development of spelling skill as the direct interaction offered by traditional forms of spell-checking. However, I also demonstrate a new type of direct-interaction spell-checking that outperforms traditional approaches in creating cognitive residuals for users that help improve their spelling skills.

Chapter 1 - Introduction

The studies presented here support the conclusion that both autocorrect and spell-check offer incidental learning opportunities that distinguish them from other modern tools. They also support a secondary claim, that there exist other ways of interacting with spelling support systems that can further enhance these incidental learning opportunities. Modern devices that utilize button-style interfaces lack many of the affordances for learning that were inherent to old “physical” tools (e.g., hammers and knives) (Osiurak, Navarro, and Reynaud 2018; Osiurak & Heinke, 2018). The process of use for these modern devices is divorced from the process of actually solving the task they aim to assist, and when these devices assist with cognitive tasks, the result can be a failure to develop task-relevant skills (Osiurak & Heinke, 2018). Consequently, these devices prove not only inept with regards to user skill development but can even prove to be detrimental to user skill development.

Researchers have found that after using modern button-style interfaces, people end up less capable in the tasks that the devices assisted them with (Ishikawa, Fujiwara, Imai, & Okabe, 2008; Fenech, Drews, & Bakdash, 2010; Sparrow et al., 2011; Henkel, 2014; Dong & Potenza 2015). However, in line with prior research on spelling support systems (Arif, Sylla, & Mazalek, 2016; Lin, Liu, & Paas, 2017), I found that interacting with spell-checking and autocorrect systems does not incur an associated loss of editing skill (Study 3 and 5). On the contrary, interacting with these systems produced users that were better able to spell after using these devices.

I had initially anticipated that differences in the method of interaction between autocorrect and spell-checking would produce different levels of user-skill development. This seemed like a sensible anticipation, given the importance of self-driven manual interaction and task-relevant feedback on the development of cognitive residuals in tool use (van Andel, Cole, & Pepping, 2017; D’Angelo, di Pellegrino, Seriani, Gallina, & Frassinetti, 2018; Cardinali, Jacobs, Brozzoli, Frassinetti, Roy, & Farnè, 2012). Autocorrect does not naturally afford involvement in the spelling correction process and does not guide user attention to any corrections it makes. I found that, within the Microsoft family of spelling support systems, participants that were provided access to a spell-checking support system performed better on an unsupported post-test than users that were provided an autocorrect support system (Study 1).

However, this difference does not appear to be driven by the method of interaction, but the relative success at delivering correctly spelled words of Microsoft’s autocorrect system. I performed three follow-up studies in which I attempted to isolate the effects of the interaction method from the different quality of service provided by the two spelling support systems. When these corrections were made to the systems, and the performance between autocorrect and spell-checking was comparable, user skill development was not significantly different between autocorrect and spell-checking (Studies 2, 3, and 4). This lack of significant difference between the conditions was apparent when the difficulty of the task was low (Study 2), when the difficulty of the task was high (Study 3), and when the spell-checking system presented the user with no ambiguity as to the correct selection (Study 4). From this, we can conclude that the more active, user-involved spell-checking process is no better at supporting the development of a user's spelling ability than autocorrect.

My last study presented here (Study 5) explores new means of interacting with spelling support systems. I introduce two new forms of user input for spell-checking systems, *part-word* and *full-word*, and a new form of autocorrect, *autohighlight*. The new spell-checking systems (part-word and full-word) aimed to create more user engagement by requiring users to either edit an incorrect word or type out the entirety of a correct word for themselves. The autohighlight system aimed to better inform users of changes made by the system and highlighted words that the system had corrected. Participants were asked to use these new tools, as well as traditional spell-checking and autocorrect systems, to correct passages containing misspelled words. The findings here indicate that only part-word system does more for creating useful cognitive residuals than conventional methods of spelling-assistant interaction. Beyond that, the affordances offered by the part-word system create a lasting performance enhancement better than if the user was provided the correct answers. The traditional means of interaction explored here were not able to accomplish this.

From these results, I argue that it may be possible to mitigate the skill loss observed in commonly used support systems by creating opportunities for task-relevant user interaction. User skill can be preserved without putting performance at risk, and it is possible to improve user skills while still providing the same level of support to the user. Rather than losing skills to our tools, it is entirely possible to imbue ourselves with skills through well-designed devices. I call this endeavor "Instructive Design."

Before I cover my findings in greater detail, I will now provide some background information regarding existing research on the effects of tool use on human cognition. I will begin with a survey of ways in which humans developed to be sensitive to information from our tools. After I have established why we are sensitive to tools, I will review cases in which our tools have incidentally impaired the development of our abilities. Finally, I will explore some of the existing explanations for why these detriments arise.

The Cognitive Consequences of Tool Use

The term "tool" will be used here to refer to a recognizable and usable object or class of objects that enhances or creates a capability that allows a user to pursue a goal. Successful tool use requires (in part) knowledge about how to co-ordinate with the tool in a manner that achieves the user's goals (Baber, 2006). We have evolved to better co-ordinate with an environment rich in tools, or potential tools.

Unlike our extant ape relatives, humans appear better able to more carefully control our motor systems to create finer edges in stone tools (Byrne, 2004). We can see that as the brain developed, so did the quality and specificity of our tools (Stout, Toth, Schick, & Chaminade, 2008). Our parietal cortex, compared to that of non-human primates, is particularly sensitive to the type of 3D rotation movements useful for manual tool manipulation (Orban et al., 2006), and object recognition

(Kastner, Chen, Jeong, & Mruczek, 2017). Our cognitive systems, both perceptual and motor, make rapid adjustments to help us engage with tools (Orban & Caruana, 2014). As a result, we come to conform to a set of behaviors and aims suited to the affordances presented by our environment (Pezzulo & Cisek; Noack 2012; de Wit, de Vries, van der Kamp, & Withagen, 2017).

The cognitive changes that we make while co-ordinating with tools leave a lasting imprint on us. Using tools cognitively changes the user, and in humans this effect can be seen in changes to our cognitive processes and capabilities. Brief usage of hand tools can modify our sense of self and can change our beliefs on what we are capable of ourselves (Cardinali et al., 2012). In developing motor routines and perceptual expectations for engaging hand tools, we develop a tool-person system (van Andel, Cole, & Pepping, 2017; D'Angelo et al., 2018). The cognitive elements that both inform and execute the actions necessary to work in coordination with the tool are known as schema, or schemes (Baber, 2006; Plant & Stanton, 2013). The schemes we develop are in turn, shaped by the demands placed on us by the tool, and features of the schemes change based on the tool used (Salmon et al., 2014; Baber, 2006). The schemes and other learned tendencies produced by interaction with a tool that are created not as a direct aim of the device in question are known as *cognitive residuals.*

Cognitive residuals are the lasting effects on cognitive processes *of* using a device (Salomon, Perkins, & Globerson, 1991). These lasting effects are distinct from the effects *with* technology, a term to distinguish the change in capabilities of users when they work with a device. While using a crutch to help one walk is an effect of working with the device, the opportunity to heal (or hurt, in the case of overreliance) through its use is an effect of using the device. The focus of this document will be on the cognitive impact of tool use, as the cognitive consequences of learning with a device are the focus of other research programs. The reason for this focus is found in the apparent consequences of ignoring these effects.

Cognitive residuals do not necessarily support the development of independent user skill. People who navigate with GPS tend to develop worse knowledge about the path they took (Ishikawa et al., 2008; Fenech et al., 2010). Similarly, people who use digital storage systems to either store information or as reference show reduced ability to recall the information stored on these devices (Sparrow et al., 2011; Dong & Potenza 2015). More generally, people who develop in a context that is rich with these digital tools exhibit breadth-based cognitive styles, and difficulty focusing in classroom environments or single tasks for extended periods of time (Dempsey, Lyons, & McCoy 2018; Kirschner & De Bruyckere, 2017; Loh, & Kanai 2016; Firth et al., 2019). The cognitive residuals that tools imbue us with are a subject worthy of further inspection.

These detriments appear to be caused by changes to how we interact with our tools. We are well equipped to discern technical knowledge from interactions with physical tools, but it is not clear if this the case for modern tools, whose actions are arbitrary relative to their outcomes (Osiurak, Navarro, and Reynaud 2018; Osiurak & Heinke, 2018). Similarly, we are more frequently relying on our tools to perform tasks for us, rather than using tools to assist us in performing these tasks for ourselves (Firth et. al, 2019; Hamilton & Benjamin, 2019; Heersmink & Sutton, 2018). Offloading responsibility for accomplishing these tasks can free up cognitive resources we need for learning (Paas, Renkl, & Sweller, 2004; Risko & Gilbert, 2016), but offloading must be carefully managed, or we risk further diminishing the opportunities for us to learn for ourselves (Jonassen, 1995; Van Merriënboer, Kirschner, & Kester, 2003; Ayres, 2006; Paas, Van Gog, & Sweller, 2010; Risko & Gilbert, 2016)

These detriments are not observed in all digital tools, however. Spelling support systems, or at least spell-check and autocorrect, appear to support the development of independent user spelling skill (Arif, Sylla, & Mazalek, 2016; Lin, Liu, & Paas, 2017). However, these tools differ with regards to their level of user engagement, and the amount of information they present that is relevant to the task. Thus, this difference will be the focus of inquiry across these five studies.

I hope to demonstrate in these studies that by considering the types of actions demanded by a tool's affordances, it is possible to construct tools that, in turn, create cognitive residuals that support tool-independent user skill. To that end I conducted five studies exploring the development of cognitive residuals during the use of autocorrect and spell-checking. A review of the findings of these studies will be conducted below.

Interactive Spell-Checker, Passive Autocorrect: Findings

My studies on spell-checking and auto-correct serve two purposes. The first is to explore the effects of existing spelling assistant technology on the creation of cognitive residuals (measured by user spelling skills). The second is to explore the effects on user skill development of engaging users in the editing process, or passively supporting them in the editing process. The results from these studies suggest that only certain types of engagement, providing them information while requiring them to complete the task for themselves, can assist a user more than passive assistance.

Spell-check and auto-correct were created to support the same task, creating correctly spelled words. While the aim of the tools is the same, the tools differ in the user roles they create. Spell checking systems require the user to identify the misspelled words and then select the correct spelling from a list of alternative spellings. In contrast, auto-correct systems perform the identification and selection task for the user, providing a correct spelling for any misspelled words entered by the user. This provides a natural contrast between devices, as they offer the same product (correctly spelled words) yet differ concerning the tasks they require the user to accomplish.

I explored the consequences of this difference on cognitive residuals across a set of five studies wherein I looked at the performance of users in a spelling task before and after using either a spell-checking or auto-correcting device. The results of these studies suggest that both of these devices can create useful cognitive residuals in their users, by merely presenting them with relevant information. While I observed enhanced performance in users provided access to spell-checking system in the first of the four studies, this difference did not survive subsequent studies in which stricter criteria for successful performance during the training trials were included. However, it is important to note that as the performance was constrained across these subsequent studies, the task the user was faced with also changed. A look at the tasks necessary to successfully correct a misspelled word show that as the options for selection were diminished and the difficulty in selecting the correct spelling was eliminated, the observed difference in effect size between spell-checking and auto-correct also diminished.

These studies appear to suggest that interaction by itself is not a crucial component for spelling-skill development, and that passive presentation of information is enough to improve spelling capabilities. Conversely, these results may also suggest that a diminished user role also carries with it diminished the task-relevant value for the cognitive residuals generated by the tool.

A New Way to Interact with Spell-Checking: Findings

In my last experiment, I compare the performance of spell-checking and autocorrect systems with the mere presentation of task-relevant information. I accomplished this by creating a new control called the *corrected* condition, in which participants were provided a correctly spelled document (rather than a document they are asked to correct). Further, in the interest of exploring the enhancement of cognitive residuals, I also explored the effects of introducing two new modes of user interaction for a spell-checking system and one new mode of user interaction for autocorrect. The new spell-checking systems were designed to ensure that they required that the user was taking physical actions relevant to correcting a misspelled word for themselves. The new autocorrect system was designed to ensure that participants were made aware of the changes made by the system.

The two new spell-checking systems still provided the basic highlighting and dictionary functions of a modern spell-checking system. In other words, misspelled words were still underlined in red, and right-clicking on these incorrectly spelled words would produce a list of adequately spelled correct words. In the interest of ensuring high-accuracy during the training trials, the list of words provided only contained one option for each misspelled word.

The first new system, referred to in this document as *part-word,* required that users perform the required edit themselves to correct the misspelled word. They were able to right-click the misspelled word, but rather than being able to click on the context menu to replace the misspelled word, the user would have to type in the correct word themselves manually. The second new system, referred to from here on as *full-word* required that users rewrite the entire misspelled word. When a misspelled word is right-clicked, the word was removed entirely from the textbox. The correct spelling was still provided in the context menu produced by the right-click, but the user had to re-enter the correct word into the document manually.

Both of these conditions feature a form of input that is relevant to the task of manually editing spelling in a document. They require that the user perform either part or the entirety of typing the correctly spelled word. In the case of the part-word system, the user is provided with ample opportunity to compare the misspelling in the original document to the correctly spelled word supplied by the spell-checker. The user of the full-word system was, in contrast, afforded less of an opportunity to compare the misspelling and the correct spelling. Instead they were required to practice typing out the entirety of the correctly spelled word.

The new autocorrect system, which I am calling *autohighlight*, provided users visual feedback of recently corrected items. This autocorrect system was modeled after the system used by Arif et al. (2017). After the autocorrect system replaced an incorrectly spelled word with a correctly spelled word, the corrected word was highlighted in the document. This highlighting remained present as the user worked on the rest of the document, allowing the user to review where the system had intervened.

To satisfy the question of whether or not spell-checking and autocorrect were any better at creating useful cognitive residuals than merely presenting task-relevant information, a new control was introduced in this study. This control consisted of a correctly spelled document and was referred to as the "corrected" condition. Participants were provided a correctly spelled document and were made aware that this document was correctly spelled. Through this, users would receive both exposure to the correct spellings, as well as practice typing in the correctly spelled words. This would provide a basis for comparison of cognitive residuals created through mere exposure to those produced through interactions with a spell-checking or auto-correct device.

Results from these studies support the hypothesis that traditional means of auto-correct and spell-checking do not create cognitive residuals that are observably more useful than the cognitive residuals generated from mere exposure to the correctly spelled words. However, this does not appear to be the case for only one of the new spelling support systems, part-word. The results discussed here will demonstrate that the part-word system provides observably more useful cognitive residuals than those created from mere exposure to the correctly spelled words.

Conclusion

This dissertation aims to demonstrate that useful cognitive residuals can be created in our interactions with technology. To begin, there will be review prior work that explores the cognitive effects of tool use, concerns regarding the development of detrimental cognitive residuals through tool use, and potential routes by which we can enhance our tools to assist the development of our skills instead. In the studies presented after the review, I will demonstrate what I have discovered regarding the development of spelling skills through the use of spelling support systems. I will show that the modern autocorrect and spell-checking systems are likely supporting the development of user spelling skills by providing relevant information to users and that learning new words can be supported by both of these modern spelling support systems about equally (though no better than if they had simply reviewed the proper spellings). Finally, I will demonstrate that there are ways that users can interact with their spelling support systems that are better for developing user skill than existing methods.

Chapter 2 - Tool Use and Human Cognition

Spell-checking and auto-correct represent two forms of spelling support system. These systems both deliver the same product to the user: a correctly spelled word. However, they differ in the required user action. Autocorrect passively provides a user with a correction to a mistake. The user enters a misspelled word, an autocorrect system attempts to determine the intended word, and then without further user intervention, replaces the misspelling with a correctly spelled word. Spell-checking requires a user to engage in the process of producing a correct spelling actively. When a misspelled word is entered, a spell-checking system underlines the misspelling until the user takes action (such a clicking a mouse button), at which point the user is presented with options for possible corrections to select, further engaging the user in the correction process.

This difference between autocorrect and spell-checking systems allowed me to explore the effect of a tool's interaction style on human cognitive development. In particular, the differences represent a means of identifying the role of active user engagement in the development of spelling skill. This chapter provides some background on the possible cognitive impacts of spelling support systems by making two key points. First, this review aims to establish that we can reasonably assume that human cognition is affected by our tools not just through intended use of educational systems, but from mere exposure to nearly all other devices as well. We usually understand humans as creators and users of tools. But a review of our evolutionary history and our particular cognitive quirks regarding tools will show that we are also a creation of our devices. The term “creation” is meant here in both a historically and individually meaningful sense. Humankind evolved alongside the tools they wielded, and this journey has made how we think intimately connected to the types of devices we use. As a result of this development, human cognition changes with respect to the tools we use. Through everyday interaction with tools, humans develop cognitive residuals that reflect task-relevant information garnered from those ordinary interactions (Salomon, Perkins, & Globerson, 1991). Specific tools or devices, such as GPS, digital reference systems, and even simple grasping tools, induce cognitive changes in their users that diminish or change user capabilities in tasks relevant to the types of interactions they have with these devices (Ishikawa et al., 2008; Fenech, Drews, and Bakdash, 2010; Sparrow, Liu, & Wegner 2011). Therefore, it seems reasonable to explore spell-checking and autocorrect as devices that may change user cognitive processes as well.

The second key point is that occlusion and diminished user responsibility are aspects of modern tools that impair user skill development. While humans are capable of deducing the functionally relevant aspects of mechanical tools through use, digital tools do not naturally afford such learning opportunities (Osiurak, Navarro, and Reynaud 2018; Osiurak & Heinke, 2018). The occlusion of task-relevant functional information is expected to limit the ability of users to learn from their devices. Autocorrect, compared to spell-checking, occludes the user from the spelling correction process by not alerting the user to an improper spelling. Identification of an incorrect spelling is no longer a process that involves the user, as it does with spell-checking. Thus, it is reasonable to explore the effects of autocorrect's increased occlusion of the spell correction process on user learning.

Diminished user responsibility also presents a problem for skill development. As tool designers automate tasks, they consequently also reduce the required capabilities of their users. Reducing the cognitive capacities needed of users in a task by modifying the environment is often referred to as a form of cognitive offloading (Risko & Gilbert, 2016). Cognitive offloading is the process in which the performance of a cognitive task is either assisted, enhanced, or replaced by an element of the environment. However when the burden of solving a task is removed from a user, this can interfere with the development of germane cognitive resources by removing the need for their growth in the user (Jonassen, 1995; Van Merriënboer, Kirschner, & Kester, 2003; Ayres, 2006; Paas, Van Gog, & Sweller, 2010; Risko & Gilbert, 2016). As germane resources represent the task-relevant capabilities of a user, the reduced user responsibilities present a threat to the ability of users to learn in the course of using a tool. Autocorrect, compared to spell-checking, creates a diminished user role. Thus, it is reasonable to explore the effects of this reduced role and enhanced occlusion on user learning.

Human Cognition is Affected by Our Tools

Tool use is intricately linked to cognitive development. Tools have been central to human cognitive development for as long as humans, and our close ancestors, have walked the earth. Tool-use is believed to have predated the emergence of modern humans, being a feature of social hominids that predate the first Homo sapiens sapiens fossils (Semaw et al., 2003).

Tool use appears to have given our ancestors distinct biological advantages, such as allowing access to new food sources (Power & Williams, 2018) and the ability to rapidly deploy new capabilities in the face of changing environments (Chase et al., 2018). To make better use of these capabilities, hominid bodies, in particular, hands, adapted across generations to make better use of tools (Williams-Hatala et al., 2018). We appear to have far greater fine-motor control than our closest relatives. This capability allows humans to execute more complex manual tasks and is believed to have assisted early hominids in the development of more regularly and finely made tools (Byrne, 2004). Hands, and hand control represent at least one way in which technology has modified humans.

Our ancestors necessarily must have changed cognitively as well. Modern humans also appear to be uniquely adapted to learning the processes for developing tools. We exhibit the capability of rapid acquisition of manufacturing techniques for ancient stone tools. This process is associated with activation of brain areas developed relatively late in the primate lineage (Stout et al., 2008). Compared to our close relative, modern chimpanzees, modern human brains show far higher sensitivity to tool properties (Stout et al., 2008; Kastner et al., 2017). While it is not the case that this establishes that it was the presence of tools in our ancestor’s lives that spurred the development of such sensitivity, it does confirm that the brain is at least highly responsive to tools.

Humans also have a far richer set of linguistic and social capabilities that assist in the integration and transmission of new tools (Orban et al., 2006; Orban &

Caruana, 2014). Further, changes in our dorsal-parietal visual pathway seem to support our ability to identify tool capabilities and generalize knowledge about these capabilities to new objects (Kastner et al., 2017). Across generations, humans have developed both behaviorally and physically in a manner that has made us better at using tools than seemingly any other organism on earth.

Cognitive Residuals and Tool-Shaped Interactions

The studies presented here focus on how tools change an individual. This section aims to establish that humans are prone to tool-influenced cognitive changes within short periods. Rapid adaptation to a diverse set of tools and potential applications requires that we cognitively change when we attempt to detect and engage the affordances in our environment (de Wit et al., 2017). For example, our concept of peripersonal space adjusts to the affordances of the tools we wield; Our sense of 'reachability' dynamically changes to match the reachable space accessible with the tools we have just used (Martel, Cardinali, Roy, & Farnè, 2016). Importantly, this phenomenon seems to arise only in two particular settings: when the perceptual feedback generated through tool use is task-relevant (van Andel, Cole, & Pepping, 2017), and when the user has some indication of agency in the process (D'Angelo, di Pellegrino, Seriani, Gallina, & Frassinetti, 2018).

These changes do not just affect tool-users as they are using tools; interactions with these tools appear to imbue users with observable cognitive residuals. Using hand tools for reaching affects our perception of our tool-less self as well. These effects are realized in two senses: (1) we see changes in our sense of the peripersonal space accessible by our hand as a result of interactions with tools, and (2) we see that our sense of peripersonal space remains changed even after we stop interacting with the device (Bourgeois, Farnè, & Coello, 2014).

Changes in post-tool cognition appear to be driven by motor rather than perceptual interactions with tools (Cardinali et al., 2012). Cardinali et al. (2012) show that after using a small device for gripping objects at a distance, users would grasp small blocks differently when only using their hands. These changes were not observed with participants that used the tool to estimate the size of the block. Using the tool to move the block proved to have a more significant impact on the user's un-assisted grip characteristics (grip strength, the velocity of reaching moment) than merely observing the block while holding the tool and assessing the possible affordances of the device. Manual, goal-relevant action appears to play an essential role in the cognitive residuals that tools imbue on us, at least in the cases described above.

The lasting cognitive effects of tool use can be considered a result *of* technology, rather than an effect rendered *with* technology (Salomon, 1990). This distinction helps us understand that there are two separate consequences of tool use for the tool user. The effects with technology refer to changes in capabilities that arise as a result of the user being engaged with technology. As in the case of the reaching tools described above, these would be the changes in future reaching actions created by having a longer reaching range. Effects with technology include applications that tool designers intended their tools to solve, as well as the affordances offered by the tool that is discovered by the user. Effects of technology include the *cognitive residuals*, lasting effects on cognition that persist even after the tool use has ceased, and that arise from interactions with devices. For instance, while one can produce notes of an event with either a pen and paper or a laptop, the effects of using a pen and paper provides users with greater memory of the contents of the notes (Smoker, Murphy, & Rockwell, 2009; Mueller, & Oppenheimer, 2014).

One well-known form of cognitive residual is the scheme (Plant, & Stanton, 2017). A scheme is a set of related concepts, action plans, and expectations that help organisms make use of environmental affordances. Schemes are generated through interaction with the environment and are in turn shaped with the elements of our environment that we either engage with, or are engaged by (Baber, 2006). As we learn about affordances and the success of our action sequences in utilizing them to achieve our goals, our actions (and thereby our schemes) adjust to better make use of the affordances presented to us.

This affordance-sensitive adjustment is to overcome the obstacles natural to self-conduct in a physical environment and achieve the goals we select for ourselves (Pezzulo, & Cisek, 2016). Pezzulo and Cisek (2016) propose that our sensitivity to affordances is created through neural systems devoted to the covert simulation of prior experience with overt action sequences in order to produce sensory expectations. Effectively, how we used tools in the past forms the basis of how we see objects as tools in the future. Pezzulo and Cisek further suggest that evidence of this supposedly central cognitive process should be found in residual products of experiencing specific action sequences. In other words, how we use the tools should change the way we think in observable ways.

One possible explanation for why humans appear to be so sensitive to affordances lies in the function of the frontal areas of the brain. Noack (2012) theorized that one of the defining characteristics of human cognition is rich feedback instituted and managed by the frontal cortices. At some point in primate evolution, the brain switched from a feed-forward dynamic in which information from sensory areas guided dynamics in the motor association areas, to a feedback dynamic of activation in which motor association areas govern sensory area dynamics. This feedback allows the creation of a category attractor-action scheme complex, a system by which the brain associates a variety of sensory inputs (a category) with relevant action sequences.

To summarize, human brains form memories of perceptual experiences as they relate to motor experiences in a manner respectful of the type of goals a human might seek (Pezzulo & Cisek, 2012; Noack, 2012). Our perceptual systems, and what we know about the world through them, are in turn shaped by the types of tools we use (Martel, Cardinali, Roy, & Farnè, 2016). Further, how we use these tools shapes how our cognitive systems are prepared to take actions in the future (Cardinali, Jacobs, Brozzoli, Frassinetti, Roy, & Farnè, 2012). The types of cognitive changes created by using a tool include changes that were not intended in the design, the effects of technology (Salomon, 1990). The tools that we use leave us with cognitive residuals relevant to their use.

Detrimental Cognitive Residuals

So far, I have discussed the biological consequences of humans developing in line with their tools, explaining why we expect humans to be cognitively sensitive to tools and demonstrating that indeed humans are cognitively sensitive to tools. Further, I have touched upon some of the existing explanations for that sensitivity and provided a name for this sensitivity, “cognitive residuals.” In this section, I show that cognitive residuals are a recognized consequence of tool use, and that cognitive residuals can inadvertently be harmful to the user.

Not all cognitive residuals are inherently valuable for users. Some frequently used modern devices appear to have detrimental effects on related user-skills. Google, the most commonly used search engine (Davies, 2018), has been shown to interfere with declarative memory capabilities (Sparrow, Liu, & Wegner, 2011). While such findings are limited regarding actual usage of Google, Sparrow et al.’s results primarily speak to the effects of interacting with a digital information storage system similar to a modern Windows or Apple Operating System. In their study, participants were offered the opportunity to read about trivia facts, type out passages about those facts, and then save their typed documents to a position in a digital folder system (e.g., “C:\Desktop\TriviaNotes”). If participants believed their typed passages were saved and remained retrievable in the system, their recall of the content of those passages was worse than for participants who thought that their messages had been erased. If a person believed the information was stored on a computer, they were less likely to recall the content of that information accurately. Sparrow et al. then demonstrate that this occurs within participants, on an interaction by interaction basis. For a given participant, if their notes concerning a fact were believed to be saved, that fact was less likely to be remembered, while if they were told their notes regarding a fact would be deleted that fact was more likely to be remembered. Effectively, the information people understood to be stored by the computer was more likely to be forgotten.

Similar results were found by Dong & Potenza (2015), who showed that participants that prepared for a memory task by using the internet for reference performed worse on a recall task than participants who prepared by using an encyclopedia. Participants exhibited lower confidence in known facts, and fewer remembered facts if they learned them through the internet. This was coupled with the observation that participants that had used the internet showed less activation in temporal-parietal-occipital associative network that “integrates information from different sensory areas and relate the information to past experiences.” Across both sets of participants, greater activation of this area was associated with better recall.

The loss of task-relevant memories also occurs when people use cameras to take pictures of their environment (Henkel, 2014). When participants were asked to take pictures on a museum tour, Henkel showed that their memory for the things that participants took a picture of was less accurate than the memory of the objects they were not instructed to photograph. It would appear that the act of creating a photograph impairs people’s memory of that photo’s subject. Interestingly this effect is very subject dependent. When participants were asked to zoom in on only a portion of the object before taking a photo, their memory for the whole object was comparable to baseline (no photo). Users did not experience memory impairment when they knew the image of the whole object would not be "stored" in the photo.

The creation of detrimental cognitive residuals is not a phenomenon limited to digital storage systems, as we can see similar detriments arise from users of GPS navigational systems. Ishikawa, Fujiwara, Imai, and Okabe (2008) showed that people using GPS for wayfinding while navigating a route for the first time performed more poorly on path recall than participants that were either given prior experience of the route (without GPS assistance) or provided a map to help them navigate the route for the first time. Participants were not instructed to memorize the route they walked, so GPS systems appear to have interfered with incidental learning processes in the user.

A failure to develop familiarity with routes also appears to happen when the GPS provides users with turn-by-turn directions. Fenech, Drews, and Bakdash (2010) explored a virtual driving scenario in which participants drove a specified route twice. The virtual city was organized such that streets in one direction were sequentially numbered (e.g. "2nd Street" is between "1st Street" and "3rd Street"), and cross streets were sequentially lettered (e.g. "B Street" is between "C Street" and "A Street"). Participants were then asked to navigate to a house on "E Street, between 3rd and 4th street". In the participant's first drive-through, they were either provided turn-by-turn directions as they would receive from most commercial GPS systems, or they were required to find the destination using street names alone. Participants that received turn-by-turn directions took significantly longer on their second drives, while participants that had to determine for themselves the proper path based on street names took less time on their second drives.

Fenech, Drews, and Bakdash (2010) attribute this difference in drive time to a failure of the GPS users to develop familiarity with the route. This failure to develop familiarity is further supported by their findings in a scene recall task. Participants in the driving task were asked to determine if a visual scene came from the route they had just driven. Participants that were provided turn-by-turn directions could not recall visual scenes from the route as accurately as participants that were not provided turn-by-turn directions. Fenech, Drews, and Bakdash suggest that this difference may be attributed to inattentional blindness created by being provided turn-by-turn directions. As will be discussed in later sections, the difference in accuracy may be attributed to differences in user demands, and the development of a transactional relationship between the participant and the turn-by-turn GPS.

More broadly, though, being immersed in digital technology appears to have lasting effects on cognition (Loh, & Kanai 2016; Firth et. al, 2019). "Digital Natives," a term used by Loh and Kanai (2016) to refer to people that develop in the presence of internet technologies from an early age, exhibit cognitive profiles distinct from "Digital Immigrants" (those who only began to use internet technologies in adulthood). For instance, children that own a smartphone early in life perform more poorly on standardized tests than children that received a smartphone later in life (Dempsey, Lyons, & McCoy 2018). "Digital Natives" exhibit less in-depth information processing and increased attempts at multitasking. They appear to adopt a breadth-biased attentional control style that is better suited towards the integration of multiple sources of information at the cost of long-term retention and greater susceptibility to distracting information.

However, it is important to note that operating like a "Digital Native" appears to be a consequence of familiarity with and usage of current digital technologies rather than a consequence of having developed in an environment rich with digital tools (Kirschner & De Bruyckere, 2017; Firth et. al, 2019). A digitally rich environment both lends itself to, and may even demand, multitasking. The concept of multitasking is constructed from the observation that people can perform two separate tasks in close temporal proximity, seemingly at the same time. This apparent dual-task performance belies the nature of the process, where scrutiny of users often reveals that they are not actually performing the tasks concurrently, but instead switching rapidly between the two tasks to manage load in various cognitive bottlenecks (Salvucci & Taatgen, 2008). Multitasking is frequently seen in the modern-day when people will opt to use social media while they perform other tasks. In the case of students, this decision often impairs their ability to learn (Rosen, Carrier, & Cheever 2013).

Explaining the Detriments

I have reviewed some cases in which the cognitive residuals of tool use were detrimental to users. It would appear that users of digital devices may occasionally, if not frequently, develop cognitive deficits as a consequence of using these devices. Identifying why exactly these tools may impair rather than enhance user skills is essential to inform effective tool design. I propose that digital devices may create cognitive deficits for two reasons: (1) occlusion of technical information relevant to how the device achieves its goal, and (2) displacement of the user as a task-responsible party (commonly referred to as *offloading*).

Technical Occlusion

One possible cause for the loss of particular skills may be a product of how we make our tools. Whereas physical tools have associations between their accessible properties (such as length, weight, color, or shape), modern devices hide how they accomplish their tasks. Hiding the solution limits what we can learn about an object, and by extension, the task it assists. Osiurak and Heinke (2018), in a tool cognition framework they have called Intooligence, distinguish three types of interactions we can have with our tools. Tools can use be assistive, arbitrary, or free. Assistive tools do not require users to be aware of their desire to use them. They are the kind of tools that can be created once and can be used again by others without the user conceiving them as a means to an end. Unlike deciding to use a television, a knife, or a car, we can use assistive tools without intending to make use of them. We do not need to decide to use a road, an awning, a wall, or even a support system like autocorrect. These systems can be used without the user intentionally employing them for a purpose.

Arbitrary tools must be intentionally selected (or "mentally made", to use the words of Osiurak & Heinke) by a user, but their proper use does not require an understanding of the technical processes by which they accomplish a goal. This category includes things like television remotes, calculators, and smartphones. These are objects with a limited set of interactive procedures for use. Only pressing on the buttons, and in a tool-arbitrary order, will allow the user to make use of the object. For many arbitrary devices, only knowledge of the correct procedure is necessary for successful usage.

Finally, free use tools require the user to envision how the tool will be used in order to accomplish a goal. One must know how to move a knife to cut a tomato, but one must also know the relationship between sharpness, pressure, the ripeness of the tomato, and the desired thickness of the slice. To use a knife, one must know something about how they (the tool and the user) interact with general physical principles about the world.

Osiurak and Heinke use physical tools as the primary example of what they mean by free tools. The relationship between the purpose of the device and how the tool solved a problem tended to be represented in the physical properties of our old tools, the identification of which we seem uniquely able to discern (Penn, Holyoak, & Povinelli, 2008). Our modern tools tend not to offer the same level of transparency regarding how they accomplish their functions (Osiurak, Navarro, & Reynaud, 2018). Instead, the association between the function and shape of the tool are divorced in modern devices. How the tool accomplishes, the task is hidden behind interfaces that reduce interactions into task-arbitrary actions.

Consider the act of typing on a keyboard. The development of useful routines for accurate and rapid use of a keyboard is unlikely to contribute to a deeper understanding of how a keyboard transforms key presses into letters on a screen. The inability to derive lessons about the function of a keyboard by using a keyboard stands in contrast to using a hammer. Learning to wield a hammer properly is quite likely to confer information on how mass and leverage help perform the hammer's task. The keyboard's task-arbitrary actions do not provide the information necessary to develop technical reasoning abilities. Instead, they support the development of procedural memories to assist the user in generating the proper motor sequences necessary to complete the task. Humans are generally quite adept at learning the mechanical principles behind the physical tools they use (Osiurak Navarro, & Reynaud, 2018). Tools that are built to be interacted with through an interface, not directly applied to physical problems, do not naturally afford the same learning opportunities.

This phenomenon is what I am referring to as *technical occlusion.* Technical occlusion is the degree to which a tool hides the process with which it solves a problem. Tools such as hammers and knives typically have low levels of technical occlusion, in that the information regarding how they assisted a task is apparent. It is through the application of force in a particular manner and the respective shapes of the tools. Tools such as smartphones have high levels of technical occlusion. They do not make the mechanical and physical processes by which information is displayed and transmitted available to the user; that information is hidden behind the interface. This hiding of the solution can also occur with other types of information relevant to solving a problem, not just the physical aspects. A map has a low degree of technical occlusion. In using a map, the user has access to location-relevant information for each point of the journey. They can discover how close destinations are by referring to the area between two points and can learn the names of intervening or surrounding streets for their destination. Receiving turn-by-turn GPS directions, however, reduces the information relevant to navigation to when and in which direction the user should turn. This form of navigation assistance would be said to exhibit a high level of technical occlusion.

User Responsibilities

Aside from the inscrutability of the mechanisms by which modern digital devices solve the problems that they do, modern digital technologies also suffer as a source of skill development in that automation does not naturally create learning opportunities for users. Automated functions aim to reduce the burden upon users, and successfully automated tasks require minimal user input. Consequently, when a task is automated correctly, the user never has the opportunity to experience how to perform the task themselves.

Whereas our old tools enhanced our capabilities, modern tools often provide us capabilities by automating the tasks that underlie them. The impact of this shift in task responsibility can be seen in the explanation for the effects of digital memory storage on user memory capabilities (Firth et al., 2019). The maintenance, storage, and organization of that information are entirely automated relative to the user, who only has to engage in the process of retrieving the information (assuming the source is trusted). Modern digital reference systems can be seen as part of our “transactive memory system” (Hamilton & Benjamin, 2019), the store of information we have access to through exchange with other entities, e.g., other people and digital reference systems. This “transactive memory” offers both information we could not possibly have generated ourselves, as well as resilient storage of what would otherwise have to be committed to rote memory.

Digital transactive memory may not be appropriately called a *transactive memory system* in the traditional sense (Heersmink & Sutton, 2018). In social contexts, where transactive memory systems were first described, responsibility for the information is often developed through explicit negotiation of which individuals are responsible for which information. In the case of the internet, or devices outside of the user’s designs (i.e., not created directly by the user), this relationship is frequently non-existent. The information stored on the internet that users have not created, endorsed the creation of, or agreed for others to create, dwarfs the information they have put on the internet themselves. Users have not agreed to delegate knowledge about the weather to www.weather.com, but nonetheless the website still generates, gathers, and stores information relevant to the weather.

The information exchange between the user and the internet is not quite like the user operating in a community of peers to accomplish a memory task jointly. It is more like the user operating in a community of potential experts who have already solved memory tasks (generate, gathering, and storing information) that the user can peruse at their discretion. Unlike traditional “Transactive Memory Systems,” where users store information for one another, the “Transactive Memory Systems” formed with the internet are frequently more one-way. For instance, the vast majority of people visiting www.weather.com are not doing so to store new information but to access what was already stored for them. Effectively these devices primarily form one-way information transactions, providing information to users without requiring any reciprocation.

Engaging with transactive memory systems, either in a traditional bidirectional exchange or in a one-way reception relation between user and tool, is referred to as a form of *cognitive offloading* (Risko & Gilbert, 2016). Cognitive offloading occurs when humans utilize environmental elements to ease or enhance the performance of cognitive tasks. If carefully managed, the process of offloading can benefit learning (Jonassen, 1995; Oviatt, 2006). A user that is not attempting to improve their skill in memorizing numbers but wishes to improve their ability with long division can keep track of the numbers they're practicing with on a sheet of paper. By offloading the storage of these numbers into the environment, users can free up resources for learning (Paas, Renkl, & Sweller, 2004).

However, when the system in question promotes the offloading of task-relevant information, learning outcomes can be impaired (Jonassen, 1995; Van Merriënboer, Kirschner, & Kester, 2003; Ayres, 2006; Paas, Van Gog, & Sweller, 2010). Risko and Gilbert (2016) suggest, citing the work of Sparrow et al., (2011), that this type of interaction trains users not to develop the same skills that the systems assist. Users cannot be expected to learn how to perform the tasks that the system accomplishes for them. Risko and Gilbert (2016) propose that this entails that designers should take particular care when constructing tools for educational purposes, as assistance in the task the user is being trained in can interfere with the user's learning.

Conclusion

Human cognition is oriented towards utilizing the environment as a tool (or rather, a set of tools). Perceptual and motor systems in the human brain are highly sensitive to the affordances offered by elements of the environment. The affordances of our devices can change the way we think, even after the user has disengaged with the part of the environment that was used as a tool. The residual effects on our cognition appear to be shaped by both what service the tools provide us and how they provide that service to us. Modern and emerging technologies are trending towards greater automation and less transparency in how they perform their functions.

Autocorrect represents a step towards greater automation and less transparency relative to spell-checking. The following set of studies will compare the cognitive residuals that arise from interacting with these devices. I expect that autocorrect, being both less transparent and involving greater automation, will prove less effective at increasing (if not detrimental to) a user's ability to correct a document's spelling than a spell-checking system.

Chapter 3 - Studies

Study 1 - A Comparison of Editing Skill Development Using Microsoft Word's Autocorrect and Spell-Checking Support Systems

Word processing is a real-life, everyday task for people all over the world. Modern word processing systems allow for document editing with a suite of support systems, such as those that assist with organization, typesetting, multi-user coordination, and spelling. These next few studies will focus on the cognitive effects of the sorts of spelling support systems that are included in many modern word processing programs, such as Microsoft Word. Of particular focus will be the spelling support systems that "autocorrect" misspelled words automatically and spelling support systems that do "spell-checking" to highlight misspelled words for user action.

As spelling support systems, both present users with roughly the same result: correctly spelled words. However, they differ significantly in how spelling support is delivered. Autocorrect systems may be thought of as user-passive and spell-checking systems may be thought of as user-active.

For autocorrect, once a misspelled word is entered, the identification of the misspelled word, the selection of the proper word, and the replacement of the misspelled word with the properly spelled word all occur without any further user input. No information is provided to the user outside of the corrected word, and occasionally a brief visual indication, such as an underline of the corrected word that fades over time. The user remains passive in the editing process.

For spell-checking, upon the entry of a misspelled word, a spell-checking system will identify the misspelled word for the user and underline or otherwise draw attention to the possible misspelling. It is up to the user to determine if the highlighted word is actually different from what was intended, and further, it is up to the user to engage a selection process to find the correct spelling, for instance, by bringing up a context menu with suggested corrections for the highlighted word. Once the user identifies a candidate alternative spelling, the misspelled word can then be replaced by clicking on the candidate. The multiple points in this process that require user input means that the user is far more active in the editing process compared to autocorrect. A comparison of user-active and user-inactive systems for editing spelling allows us to explore the importance of user-active participation in the development of useful cognitive residuals.

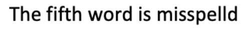
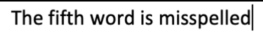
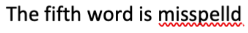
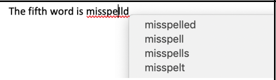
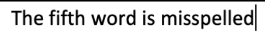
Prior research into the effects of spell-checking systems shows that spell-checkers can assist users in learning the proper spelling of English words through the act of editing a document with the spell-checker (Lin, Liu, & Paas 2017). As the goal before students was to use the device to correct the spelling, rather than to learn new words, we can say that learning the words is a cognitive residual (Lin et al. refer to the effects as a product of incidental learning) of having used the spell-checker. However, this study did not look at the effects of autocorrect, a relatively new method of spelling assistance.

Conceivably both of these systems should produce cognitive residuals, as both provide the user the opportunity to see correctly spelled words. However, the user-passive technique presented by autocorrect creates the sort of one-way information interactions that Heersmink and Sutton (2018) claim characterizes interactions with modern internet devices. In the parlance of Osiurak and Heinke (2018), this also shifts autocorrect to an almost purely "assistive tool," having required a massive amount of planning and design on the part of the creator, but little to no effort on the part of the user to adequately use the tool. When it functions properly, one does not have to go out of the way to engage the autocorrection function.

Spell-checking, in contrast, would fall into the category of "arbitrary" tool-use (Osiurak & Heinke, 2018). It would constitute an "arbitrary tool" as it requires specific but task-arbitrary user inputs to function. "Arbitrary tools" require users to develop procedural memories unique to the process of using the device. To know how to use a light-switch, an "arbitrary tool," one needs to know that certain actions will be successful while others will not. Only manipulating the up and down position of the switch will allow the user to make use of the primary affordances provided by the device while pressing, pushing, pulling, or smashing the device will not produce any useful effects.

Whereas it is conceivable that autocorrect and spell-checking systems both create cognitive residuals, it is reasonable to assume that the cognitive residuals created by each are different, as the two differ in the need for user input and in the demand that need places on the user to create procedural memories necessary to deliver that input. As a result of this extra need for procedural memory, I hypothesize that current spell-check systems will produce a better incidental learning experience than current autocorrect systems. This hypothesis can be tested by comparing the success of users on a spelling task, before and after using a spell-check or autocorrect support system.

Table 1 - Phases of the Spelling Assistance Process

System	Identification	Selection	Execution
Autocorrect	 The fifth word is misspelld	No new information is presented to the user for this step	 The fifth word is misspelled
Spell-Checking	 The fifth word is misspelld	 The fifth word is misspelld misspelled misspell misspells misspelt	 The fifth word is misspelled

This table summarizes the steps the user takes in correcting a spelling error with either Microsoft Word's autocorrect, or spell-checking assistant. Of particular note are the differences in the identification and selection steps. In the identification step, autocorrect provides no indication of a misspelling. Spell-check, on the other hand, draws the user's attention to the misspelling with a red underline. In the selection step, autocorrect handles the process of selecting a properly spelled word without prompting the user in any manner. Spell-check requires that the user perform the selection process (in part) themselves. Spell-check offers the user candidate properly spelled words, and the user must select the proper spelling from that list. In both cases the final step is replacing the misspelling with a properly spelled word.

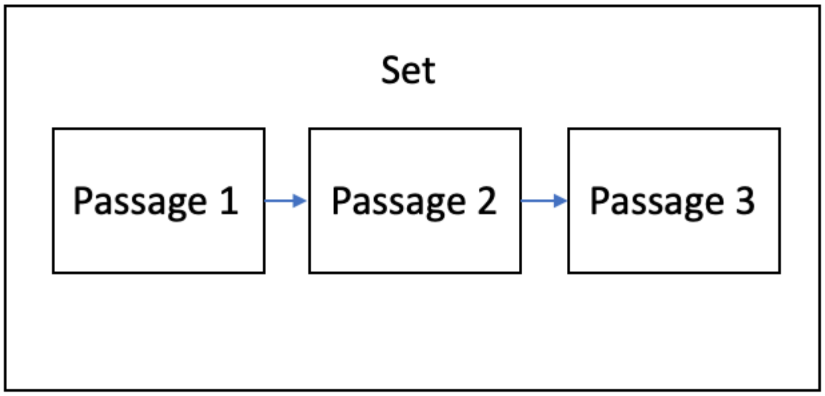
Methods

Data Collection

To test the hypothesis that cognitive residuals will differ between the different sorts of spelling support systems, I conducted a study comparing the incidentally learned spelling skills of users who used either an autocorrect or a spell-checking system to correct the spelling of a set of misspelled words in a short document. A total of 18 participants were recruited from the UC Merced subject pool through the Sona Experiment Management System and participated in exchange for class credit. Participants were randomly assigned to either the spell-check or autocorrect conditions (i.e., a between-subjects design).

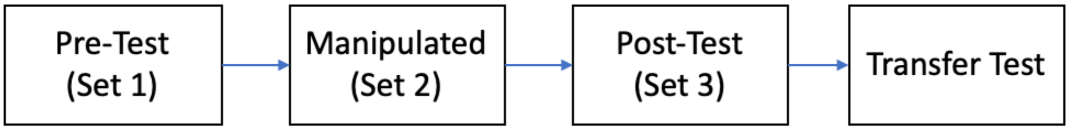
In this study, participants were asked to copy three short passages from a paper document into a Microsoft Word (Microsoft, 2018) document and correct the spelling mistakes the passages contained. These three passages composed a set (see Appendix A.1.2-A.1.4), and this set of three passages was repeated three times per participant. The first instance of the set was called the pre-test period. This set was used to establish the participant's base rate of word correction without any support system present. The second instance of the set was in an experimentally manipulated condition in which participants used either the autocorrect system or the spell-checking system native to Microsoft Word (Microsoft, 2018). The third instance of the set was called the post-test period and was used to establish whether any spellings were learned. In addition, after this post-test, participants also copied one additional passage that contained all the misspelled words in the set but in a different order and context than in the original passages (see Appendix A.1.5), constituting a transfer test.

Figure 1 - Diagram of a Set



Each set was composed of three trials. Each trial consisted of one passage containing misspelled words. Participants were instructed to correct the spelling in each of these passages. Passages 1, 2, and 3 all contain different misspelled words and are reproduced in Appendix A.1.2.2-A.1.2.4. Each set contained the same three passages, presented in the same order.

Figure 2 - Diagram of Experiment Flow



Participants saw the same set of passages (described above) three times. The first set constituted a "Pre-Test", in which base-line spelling performance was established by having participants correct the spelling in the passages without assistance. The second set constituted a "Manipulated" condition, in which participants corrected the same three passages they saw in the pre-test. The difference between the "Manipulated" condition and the "Pre-Test" condition was that participants were provided a spelling assistant (either autocorrect or spell-check) while they corrected the spelling in the passages. The third instance of the set was a "Post-Test", where once again participants were asked to correct the spelling in the passage without a spelling assistant. Finally, participants were asked to complete a "Transfer Test", in which participants were again asked to correct the spelling of a passage, though this time it was a passage they had not seen before. The transfer test passage contained the same target words, but in a different order within a new passage (passage is reproduced in Appendix A.1.2.5).

The passages were constructed from a Wikipedia list of commonly misspelled English words (Wikipedia contributors, 2017), and are reproduced in Appendix A. A total of 27 words were used, with eleven misspellings in the first passage, seven misspellings in the second passage, and nine misspellings in the last passage. A custom dictionary was constructed for the Word document to ensure that the misspellings produced spell-checking options that contained the correct spelling of the target word (Microsoft, 2018). Participants were allotted 3 minutes for each trial. If the trial ended before the 3 minutes, the experiment proceeded anyway. Any student unable to complete all trials was not included in this study.

Assessment

Scores were calculated for each set as the number of spelling errors remaining in the documents submitted across all three trials. This means there was a pre-test score, "manipulated" score (from the trials with either autocorrect or spell-checking available), and a post-test score. Both misspellings (e.g. "misspellng") and substitutions (e.g., "misstopping") were considered errors. Thus, a lower score indicated better performance. Scores on the pre-test and post-test were compared to assess how much participants improved across trials and between conditions. A generalized Poisson mixed model was fit to the data, and the relative rate ratio for the number of errors on the post-test was estimated for spell-checking compared to auto-correct.

Scores on the set with a support system (the "manipulated" set) were compared between support systems (spell-check, autocorrect) to assess how effective these support systems were in providing assistance. To ensure that performance was comparable between autocorrect and spell-check, a Kruskal-Wallis rank sum test was applied to participant's scores on the "manipulated" set. If performance was comparable between the two conditions, we would expect a non-significant value for the Kruskal-Wallis test statistic.

Regression

The statistical software R (R Core Team, 2018) was used to construct and test a generalized Poisson mixed model, using the *lme4* package (Bates et al., 2015). The response variable for this model was the total number of errors remaining in the post-test set. The model used the sum of the total number of errors made on the pre-test set, as well as the condition as independent variables. An observation-level random effect (OLRE) was also included in the model to account for the over-dispersion common to count data (Harrison, 2014).

To assess the impact of spell-checking on editing skills, a Poisson mixed model was constructed to predict performance on an editing task after using a spell-checker or an autocorrect system. The Poisson distribution is suggested for handling count data (Coxe, West, & Aiken, 2009). Because there were only 27 target words, the response variable violated the Poisson model assumption of a theoretically infinite possible count (Ferrari, & Comelli, 2016), but both quasi-Poisson and generalized linear mixed models have shown to be robust in cases where this assumption is violated (Lazic, 2015). Poissonality of the observed data was assessed by plotting estimated Poisson quantiles of the post-test scores against observed quantiles, with a 95% confidence band. All values fell within the bands, indicating that a Poisson distribution is an appropriate model for the data (see Appendix A.2.1.2).

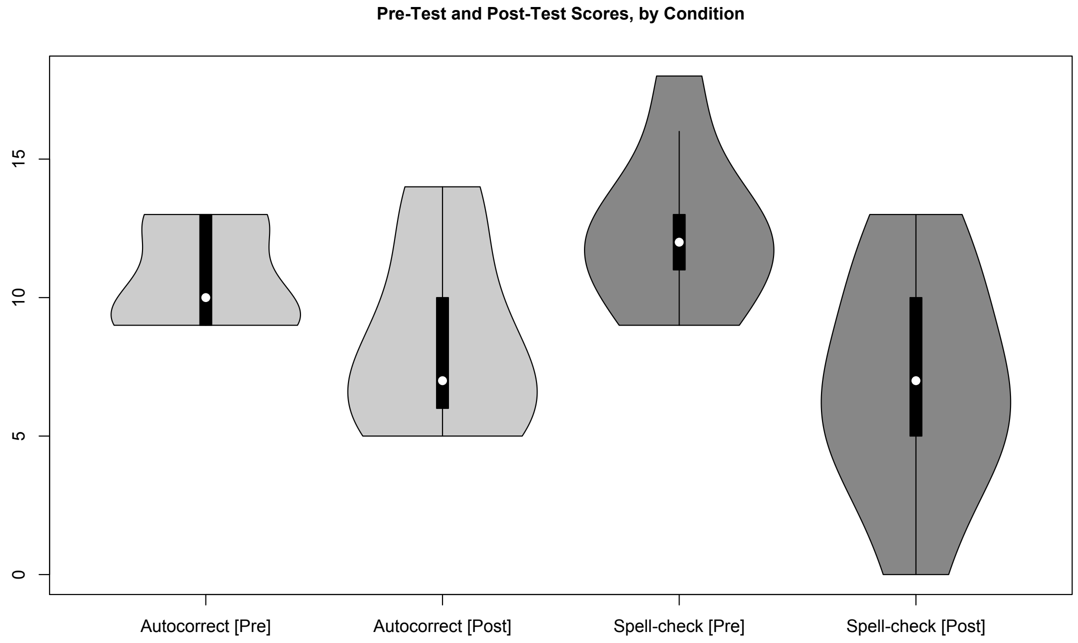
The error structure offered by generalized linear mixed models (GLMM) allows us to model count data where the mean is not exactly equal to the variance of the data, a core assumption of the Poisson distribution (Harrison, 2014). Due to the presence of singularity while fitting the generalized linear mixed model with an OLRE, a second generalized linear model with a quasi-Poisson link function and no OLRE was fit as well. This produced no errors, and the results were compared. No concerning differences were observed between the estimates, and the results from these fits are reproduced in Appendix A.2.3.

The predictor variables included *Pre-Test Score* (a count of the number of errors made on the pre-test) and *Condition* (a two-level categorical variable representing either the use of auto-correct or spell-checking during the middle trials). No interaction terms were included, as data exploration determined no need for an interaction term (Appendix A.2.1.1). To account for over-dispersion an OLRE term, *ID,* was included.

Verification of the model was done by plotting Pearson residuals against fitted values, all covariates included in the model, and one covariate not included in the model (Zuur & Ieno, 2016). These plots can be found in Appendix A.2.2.2. No issues were found in these plots. A Levene's test of the assumption that variance was homogenous across grouping levels was also performed, and also indicated that there was no significant violation of this assumption. Results are available in Appendix A.2.2.1. R^2 was calculated using the *r.squaredGLMM* in the *MuMIn* package (Bartoń, 2019), and output is reported in Appendix A.2.3. The OLRE was estimated to have a mean less than 0.001 and a standard deviation of less than 0.001. The results from the tri-gamma estimation of R^2 will be reported, per the recommendation of Nakagawa & Schielzeth (2017).

Results and Discussion

Figure 3 -Pre-Test and Post-Test Scores, by Condition



This is a violin plot of the distributions of pre-test and post-test scores, grouped by condition. Lower scores indicate fewer errors. Participants in the spell-checking condition, on the post-test, produced fewer errors than participants in the autocorrect condition. This was despite having produced more errors in the pre-test. The difference between pre-test and post-test scores varied significantly across conditions, with participants in the spell-checking condition producing fewer errors in the post-test ($\beta_{Condition}$ = -0.461 ,p<0.05).

Participants in the spell-checking condition outperformed the participants in the autocorrect condition. The hypothesis was that the number of errors on the post-test would be lower for participants that were in the spell-checking condition than those in the autocorrect condition, in a manner that differed relative to their initial ability as measured by the number of errors they had on the pre-test. Results support this hypothesis.

The model accounted for 43.8% more variance than the null model (tri-gamma R^2c = 0.4383). Both the score on the pre-test (*Pre-Test*) and the tool (*Condition*) were significant predictors in the model ($\beta_{Pre\text{-}Test}$ = 0.131, p<0.05; $\beta_{Condition}$ = -0.461 ,p<0.05). According to the model, we should expect that participants in *spell-checking* conditions to produce 36.9% fewer errors on the post-test than participants in an *autocorrect* condition. The lower bound of this estimate is 9.5%, and the higher bound is 56.5%.

The transfer test model accounted for 31.5% more variance than the null model (tri-gamma $R^2c = 0.3150$). The score on the pre-test (*Pre-Test*), but not the tool (*Condition*), was a significant predictor in the model ($\beta_{Pre\text{-}Test} = 0.110$, $p<0.05$; $\beta_{Condition} = -0.240$,$p = 0.2062$).

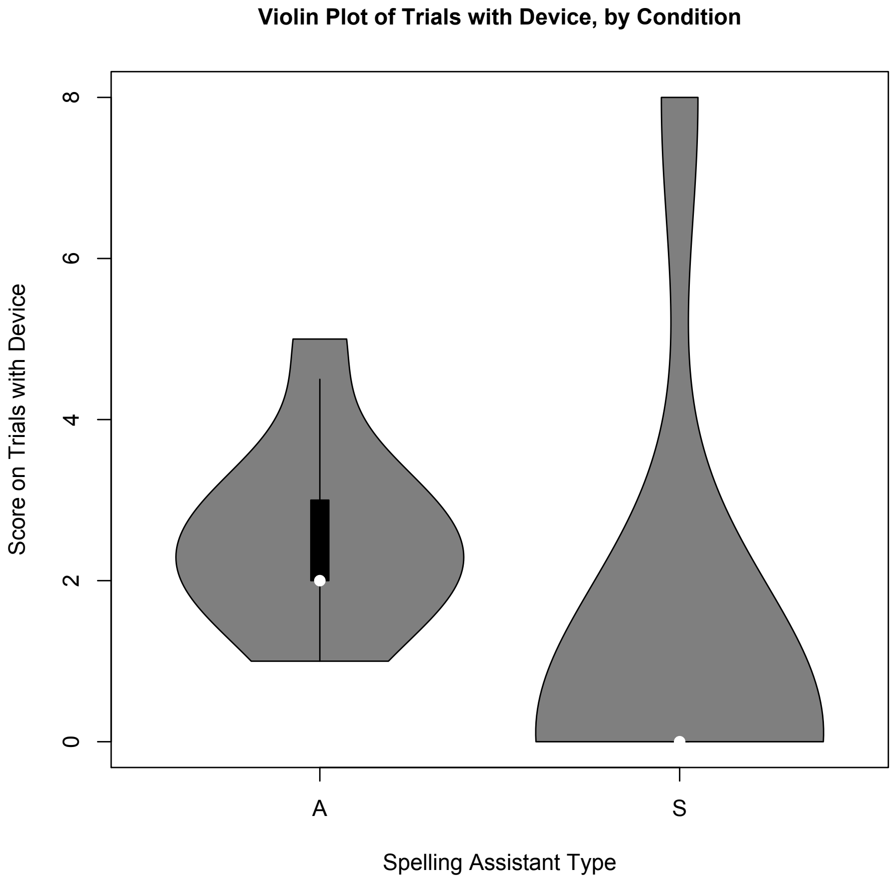
Table 2 - Table of Regression Coefficients for Study 1

Variable Name	Estimate	Std. error	z-value	*P*-value	Rate Ratio	2.5% C.I. RR	97.5% C.I. RR
Intercept	0.707	0.412	1.719	0.086	2.027	0.906	4.527
Pre-Test	0.131	0.036	3.678	<0.05	1.140	1.063	1.223
Condition: Spell-Check	-0.461	0.190	-2.431	<0.05	0.631	0.435	0.915

Alpha level for all tests presented here was $\alpha = 0.05$. The rate ratio was calculated as the exponentiated regression coefficient. Upper and lower bounds confidence interval bounds were calculated using the 'emmeans' package from R (Lenth, 2019).

The use of the two spelling support systems appeared to have different rates of success during use. All of the autocorrect users had at least two errors during the trials that they had access to an autocorrect system, while only 22.2% of participants in the spell-checking condition had one or more errors. This difference was detected by the Kruskal-Wallis one-way analysis of variance $\chi^2(1, N=18) = 8.086$, $p < 0.005$. Density plot and results from the Kruskal-Wallis test can be found in Appendix A.2.1.1.

Figure 4 - Distribution of Scores (Number of Errors) on Manipulated Trials



This is a violin plot depicting the distribution of errors (score) in the trials where participants were using either Microsoft's spell-checking or autocorrect support system to correct errors in the passage. The width of each 'violin' represents the number of observations at that level. A wider point in the violin corresponds to more observations, while a narrower point in the violin corresponds to fewer observations. Participants in the autocorrect condition (A) never achieved perfect performance (a score of zero), unlike the majority of the participants (seven of ten) in the spell-checking condition (S). Spell-checking had the highest single score of four errors (the passages contained 27 spelling errors to begin with).

Conclusions for Study 1

Results indicate that users supported by spell-checking performed better than users supported by autocorrect on an unsupported editing task after using the support system. However, the size of this effect (i.e., the 95% confidence interval) ranges widely, from about 10% to 33% improvement. In addition, spell-check and autocorrect also produced different error rates for the trials where participants had access to one of these support systems. To ensure improvement observed in participants in the spellchecking condition is attributable to the difference in how the tool is used (either actively with spell-checking, or passively with autocorrect) rather than in the relative success of the spelling support systems, user performance during these trials should be strictly controlled.

Study 2 – Interaction Styles and Word Learning: Comparing Autocorrect and Spell-Check When Support System Performance is Controlled

In the prior study, participants that used a spell-checker performed better on the post-test than participants that used an autocorrect device. This suggests that spell-checking is a better system for developing spelling skills than autocorrect. However, the first study does not demonstrate that this difference is due to how the user interacts with the two systems. Participants that used a spell-checking device also saw more correctly spelled words during the training trials than participants in the autocorrect condition. In order to explore the effects of the method of interaction on word learning, rather than the relative success of each device in assisting the user with spelling, performance during training trials should be better controlled. If both sets of participants see the same set of properly spelled words, differences in performance on the post-test can be more reasonably attributed to differences in delivery.

In this study, we controlled performance during the trials where participants had access to a spelling support system. We controlled for performance to shift focus from the question of "how do these devices perform," to "how does the method of interaction impact what the user learns". Assuming that participants see roughly the same number of words in either condition, then differences in score on the post-test should be the result of how those words were delivered by each system. Either participants are witnessing these words as a product of their intentional selection in a spell-checking interface, or they are witnessing these words as a product of the passive intervention offered by an auto-correct interface. If we control for the different levels of performance observed in Study 1, we can interrogate the method of interaction as a possible source of differences in post-test scores.

As one further step of control, scores for this study were counted as the number of correctly spelled target words. This is technically a different measure than that of Study 1, where the number of errors were counted instead. Study 1 aimed to explore the effects of different commercially available devices on user editing skills. Here the aim is to see if people will differently integrate the information provided to them by either spell-checking or autocorrect style feedback.

This study was undertaken to test the following hypothesis:

Interaction with a spell-checking system will improve participants' post-test scores more than interaction with an auto-correct system.

Methods

Performance Control

To control for the differences in performance during the training phase of Study 1, I performed a web-based study that used a JavaScript auto-correct system. It is based off the *npm* javascript package 'autocorrect' created by Yefim (2016) (https://www.npmjs.com/package/autocorrect). The system interfaces with standard HTML textbox objects. This allows it to be deployed and used in most modern browsers. In order to emulate the function of modern autocorrect systems, when a user presses space while typing into a textbox this autocorrect system 'checks' the last set of letters entered after the previous space (the most recent word entered). This 'check' constitutes a comparison between the word entered, and the words contained in the system's internal dictionary. In all cases the word entered is finally replaced with the closest dictionary match, as measured by Levenshtein distance.

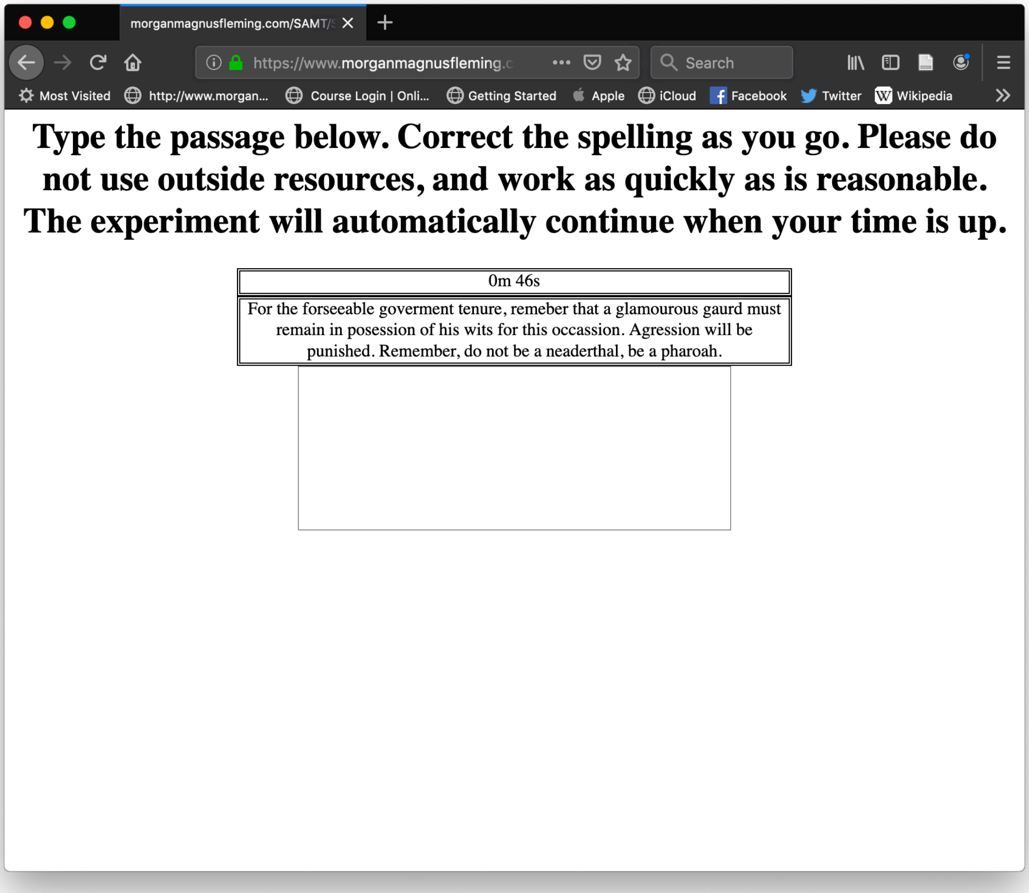
This autocorrect system replaces every word that a participant enters with a correctly spelled word. This aims to avoid the issue observed in the first study where participants using an autocorrect system saw fewer target words during the training trial. This new system also featured a customizable internal dictionary, and by ensuring that the autocorrect system had a limited dictionary the novel misspellings entered by users (i.e. misspellings that were different from those in the passage) had fewer non-target words from which to select corrections. These two features, ubiquitous replacement and limited selection for corrections, helped ensure that autocorrect users are far more likely to produce the target words than they were in Study 1.

Further steps were also taken to control for performance on trials with a support system present. First, a performance criterion was set at 90% of the target words spelled correctly. If participants spelled fewer than 25 out of 27 words spelled correctly on the trials with a support system present, they would be rejected from further participation in the trial.

The misspelling for each target word was selected both to ensure that the first option presented in the context menu by the browser's spell-checking system was the target word associated with that misspelling (e.g. "freind" if right-clicked while spell-checking was enabled would produce "friend" as the top result). This was done to ensure participants in the spell-checking condition could reliably achieve the strict performance criterion set. Participants in the spell-checking condition were made aware of these affordances, and were told that if they entered the misspelling into the text box they could be certain that the first option presented by the spell-checking system would be the correct one.

Data Collection

Figure 5 - Screenshot of Website Used to Gather Data



Type the passage below. Correct the spelling as you go. Please do not use outside resources, and work as quickly as is reasonable. The experiment will automatically continue when your time is up.

0m 46s

For the forseeable goverment tenure, remeber that a glamourous gaurd must remain in posession of his wits for this occassion. Agression will be punished. Remember, do not be a neaderthal, be a pharoah.

A total of 22 participants were recruited from the Amazon Mechanical Turk system and participated in exchange for $7.51 for 30-45 minutes of work. Participants were part of the master class of mTurk Workers (Barr, 2018), and were all responding from IP addresses within the US.

The structure of the experiment remained the same to that of Study 1, with the exception that the administration of the task was managed via website rather than administered in person. Participants were asked to type out three short passages that were displayed in their browser and correct any spelling mistakes these passages may contain. Participants were allotted 3 minutes to complete each trial. Responses were submitted through a textbox on the webpage.

The three passages composed a set, and this set was repeated three times. The first instance was a pre-test period to establish their base rate of word correction without any support system offered in the webpage. The second instance was an experimentally manipulated condition referred to as the training trials. In this instance participants were able to use either the autocorrect system described above, or the spell-checking system native to their browser. Participants were randomly assigned to either of these conditions. The third instance was a post-test period in the same unsupported condition as the pre-test. There was no transfer test included to limit the cost of data collection.

The same passages were used as the ones used in Study 1. Passages are fully reproduced in Appendix B.1.2.2-B.1.2.4. The transfer test was removed.

Assessment and Regression

Scores were calculated for each set of trials as the number of target words spelled correctly across all three trials. Thus, a higher score indicated better performance on the task. A Poisson mixed model was fit and a Kruskal-Wallis goodness-of-fit test was run using the same steps described in Study 1.

Poissonality of the observed data was assessed by plotting estimated Poisson quantiles of the post-test scores against observed quantiles, with a 95% confidence band. All values except for two fell within the bands, indicating that a Poisson distribution is an appropriate model for the data. These plots can be found in Appendix B.2.1.2.

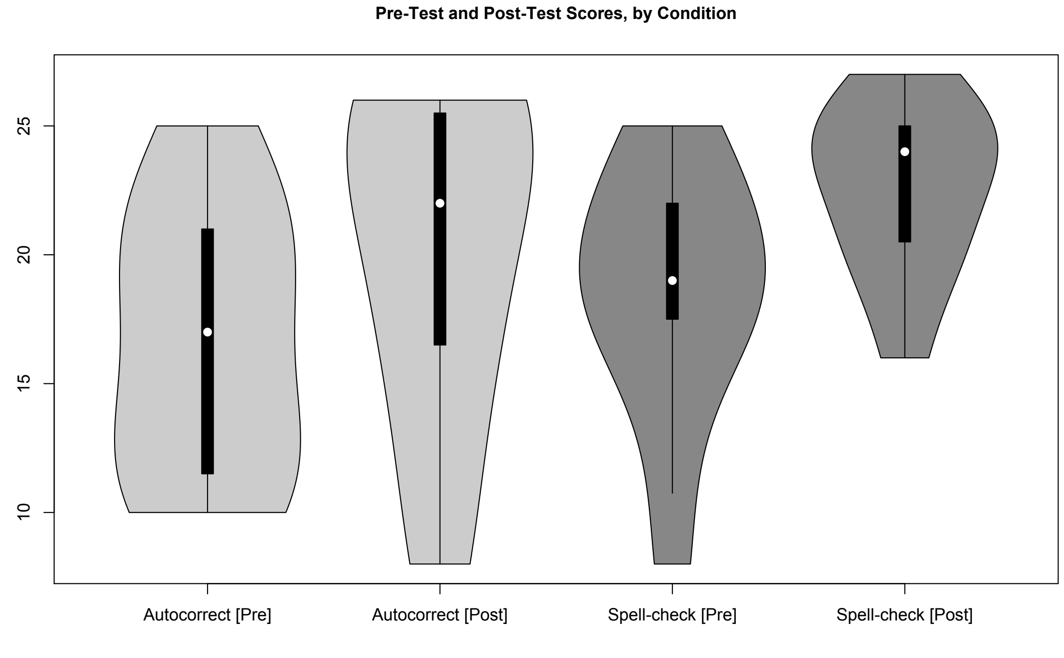
The predictor variables included *Pre-Test Score* (a count of the number of errors made on the pre-test) and *Condition* (a two-level categorical variable representing either the use of auto-correct or spell-checking during the training trials). No interaction terms were included, as data exploration determined no need for an interaction term (Appendix B.2.1.1). To account for over-dispersion an OLRE term, *ID,* was included.

Because there were only 27 target words, the response variable necessarily violates the Poisson model assumption of a theoretically infinite possible count (Ferrari, A., & Comelli, M. 2016). However, both quasi-Poisson and generalized linear mixed models have shown to be robust in cases where this assumption is violated (Lazic, 2015). For this model the response variable is the proportion of words spelled correctly over the total number of target words.

Verification of the model was performed by plotting Pearson residuals against fitted values and all covariates included in the model (Zuur & Ieno, 2016). These plots can be found in Appendix B.2.2.2. No issues were found in these plots. A Levene's test of the assumption that variance was homogenous across grouping levels was also performed, and also indicated that there was no significant violation of this assumption. Results are available in Appendix B.2.2.1. R^2 was calculated using the *r.squaredGLMM* in the *MuMIn* package (Bartoń, 2019), and output is reported in Appendix B.2.3. The OLRE was estimated to have a mean less than 0.001 and a standard deviation of less than 0.001.

Results and Discussion

Figure 6 -Pre-test and Post-test Scores, grouped by Condition



This is a violin plot of pre-test and post-test scores, grouped by condition. Pre-test and post-test differences were not significantly different between the conditions. Participants in spell-

checking were observed correctly spelling more words on the post-test than participants in the autocorrect condition, but this difference was not significantly different.

Results indicate that there was not a significant difference in post-test scores between the participants that had access to spell-checking software and the participants that had access to autocorrect software. The model accounted for 47.6% of more variance than the null model (tri-gamma $R^2c = 0.476$). The score on the pre-test (*Pre-Test*) was a significant predictor in the model ($\beta_{Pre\text{-}Test} = 0.041$, $p<0.05$) while the tool they used (*Device*) was not significant ($\beta_{Condition} = 0.038$, $p = 0.685$). According to the model, we observed that participants in *spell-checking* conditions to produce 3.9% more correctly spelled words on the post-test than participants in an *autocorrec*t condition. The lower bound of this estimate is 13.6% fewer correctly spelled words, and the higher bound is 25% more correctly spelled words. This range includes zero, suggesting that this difference is not reliable.

The performance controls were successful in ensuring that participants had similar performance during the training trials. The use of the two devices did not appear to have different rates of success during the training trials. The group that used a spell-checking system and the group that used an auto-correct system had an identical distribution. This difference was not detected by the Kruskal-Wallis one-way analysis of variance. Density plot and results from the Kruskal-Wallis test can be found in Appendix B.2.1.1.

Table 3 - Table of Regression Coefficients for Study 2

Variable Name	Estimate	Std. error	z value	*P*-value	Rate Ratio	2.5% C.I. RR	97.5% C.I. RR
Intercept	2.308	0.188	12.274	<0.05	10.054	6.909	14.445
Pre-Test	0.041	0.010	4.058	<0.05	1.041	1.021	1.062
Condition: Spell-Check	0.038	0.094	0.406	0.685	1.039	0.864	1.250

Alpha level for all tests presented here was $\alpha = 0.05$. The rate ratio was calculated as the exponentiated regression coefficient. Upper and lower bounds confidence interval bounds were calculated using the 'emmeans' package from R (Lenth, 2019).

Conclusions for Study 2

The new auto-correct system proved successful in limiting the number of errors during usage, which helped avoid the differences in performance during the training trials observed in Study 1. Further, spell-checking did not appear to significantly assist users in learning more words beyond what they would have learned with an auto-correct system. This may have been an issue with the difficulty of the stimuli. On the post-test a number of participants achieved perfect performance, raising the possibility that the words in this study were too easy for participants to learn.

Study 3 - Interaction Styles and Word Learning: Comparing Spell-Check to Autocorrect with Difficult Words

Study 2 tells us that with regards to users correcting familiar words, users of either spell-checking or autocorrect learn roughly the same amount. This, however, may have been due to the apparent ceiling effect observed in Study 2. Furthermore, there has yet to be a comparison of auto-correct and spell-checking with unsupported conditions. This third study aimed to answer both of the following questions. First, autocorrect and spell-checking provide similar skill improvement when users are simply learning to detect errors in familiar words. Are there observable differences that arise when users are faced with far fewer familiar words? Second, do either of these devices actually create useful cognitive residuals?

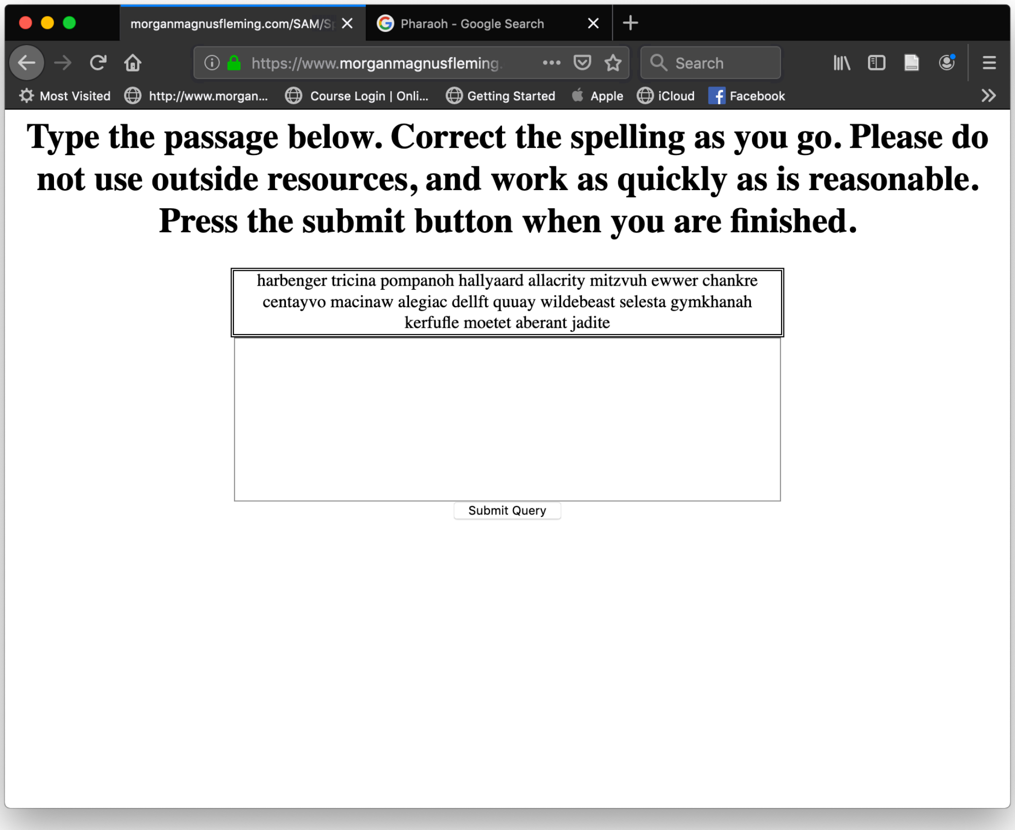
To answer the first question, this study used new words and new passages. These words were selected from a list of words used for adult spelling bees (The National Senior Spelling Bee, 2018.) Adult spelling bees are contests that hope to challenge adults on their ability to recognize and spell difficult words. The passages contained 20 target words each. The hypothesis is that these more difficult words will help avoid the ceiling effect in the previous study, and that spell-checking systems will have a significantly greater impact on post-test scores than auto-correct systems.

To answer the second question, this study employed a control condition. This control condition featured training trials in which participants did not have access to a spelling support system. This offers an opportunity to explore how participant spelling skill compares between supported and unsupported conditions. If these systems create useful cognitive residuals, then we would anticipate that both spell-checking and auto-correct systems will positively impact user post-test scores when compared to the unsupported (control) condition.

Methods

A total of 56 participants were recruited from the Amazon Mechanical Turk and participated in exchange for $7.51 for 30-45 minutes of work. Participants were part of the master class of mTurk Workers (Barr, 2018), and were all responding from IP addresses within the US.

Figure 7- Screenshot of Website Used to Gather Data



Type the passage below. Correct the spelling as you go. Please do not use outside resources, and work as quickly as is reasonable. Press the submit button when you are finished.

harbenger tricina pompanoh hallyaard allacrity mitzvuh ewwer chankre centayvo macinaw alegiac dellft quuay wildebeast selesta gymkhanah kerfufle moetet aberant jadite

The structure of the experiment remained the same to that of Study 2, with the exception of an additional control condition and the removal of a time limit. In the control condition participants all saw the same sets of trials (Pre-Test, Manipulated, Post-Test) as the other conditions. However, during their manipulated trials, participants received no support system. They were asked to try their best to correct the words and, as was the case for all the other conditions, refrain from using outside resources to assist them. Performance was also controlled for by setting a performance criterion of 90% for the trials with a spelling support system, as was done in Study 2. This criterion was not enforced for participants in the control condition, given that they did not have access to any form of assistance.

The time limit was removed with the intention of allowing participants as much time as they needed to get through all 60 misspellings. Participants were able to freely submit their work by pressing a "Submit" button at the bottom of the textbox. Participants were only provided pay if they completed all trials, and no participants that failed to complete all the trials were included in the analysis.

The passages consisted entirely of words taken from the 2018 National Senior Spelling Bee (The National Senior Spelling Bee, 2018). They are reproduced in Appendix A. A total of 60 words were used, with twenty misspellings presented for each trial. Passages are fully reproduced in Appendix C.

To ensure consistent performance between the autocorrect and spell-checking conditions, the words and the misspellings were selected such that upon entry of the misspelling, the target word would be the first option presented by the browser's spell-checking system. This was confirmed in Google Chrome, Mozilla Firefox, and Safari browsers. A custom dictionary was also constructed for the auto-correct system to ensure that the misspellings produced the correct target word from the passage.

Assessment

Scores were calculated for each set of trials as the number of target words spelled correctly across all three trials. Thus, a higher score indicated better performance on the task. Scores on the trials with a support system were compared between support systems (spell-check, autocorrect) to assess how effective these support systems were in providing assistance. To ensure that performance was comparable between autocorrect and spell-check a Kruskal-Wallis rank sum test was applied to participant's scores on the supported trials. Scores on the pre-test and post-test were compared to assess how much participants improved across trials and between conditions. A generalized linear mixed model in the Poisson family was fit to the data, and the relative rate ratio for the number of correctly spelled words on the post-test was estimated for both spell-checking and auto-correct conditions.

Regression

The statistical software R (R Core Team, 2018) was used to construct and test a generalized mixed Poisson model, using the *lme4* package (Bates et al., 2015). The response variable for this model was the total number of words correctly spelled in the post-test set. All models used the sum of the total number of words correctly spelled on the pre-test set, as well as the device participants used during the trials as independent variables. An observation-level random effect (OLRE) was also included in the generalized linear mixed Poisson model, to account for the over-dispersion common to count data (Harrison, 2014).

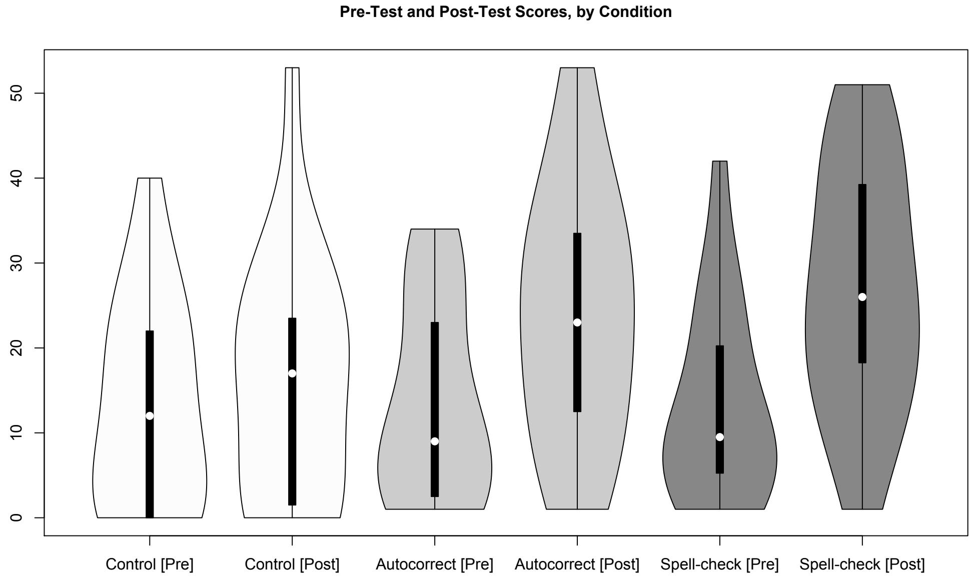
In order to assess the impact of spell-checking on editing skills, a Poisson mixed model was constructed to predict performance on an editing task after using a spell-checker or an autocorrect system. The Poisson distribution is suggested for handling count data (Coxe, West, & Aiken, 2009). Because there were only 60 target words, the response variable necessarily violated the Poisson model assumption of a theoretically infinite possible count (Ferrari & Comelli, 2016). However, both quasi-Poisson and generalized linear mixed models have shown to be robust in cases where this assumption is violated (Lazic, 2015). Poissonality of the observed data was assessed by plotting estimated Poisson quantiles of the post-test scores against observed quantiles, with a 95% confidence band. All values fell within the bands, indicating that a Poisson distribution is an appropriate model for the data. These plots can be found in Appendix C.2.2.

The predictor variables included *Pre-Test Score* (a count of the number of errors made on the pre-test) and *Condition* (a two-level categorical variable representing either the use of auto-correct or spell-checking during the training trials). No interaction terms were included, as data exploration determined no need for an interaction term (Appendix D.2.1). To account for over-dispersion an OLRE term, *ID,* was included. The OLRE was estimated to have a mean less than 0.3157 and a standard deviation of 0.5619.

If autocorrect and spell-checking systems create useful cognitive residuals, we anticipated that regression coefficient for "Condition" to be significantly positive for both of these systems. To test the hypothesis that spell-checking systems create more useful cognitive residuals than autocorrect systems, the *glht* function of the *multicomp* package will be used to compare the regression coefficient for spell-checking and autocorrect (Hothorn, Bretz, & Westfall, 2008). If spell-checking produces more useful cognitive residuals than autocorrect, then the regression coefficient should be significantly larger for spell-checking than autocorrect. Results from this test can be found in Appendix C.3.3.

The error structure offered by generalized linear mixed models (GLMM) allows us to model count data where the mean is not exactly equal to the variance of the data, a core assumption of the Poisson distribution (Harrison, 2014). Due to the presence of singularity while fitting the generalized linear mixed model with an OLRE, a second generalized linear model of the quasi-Poisson family (without an OLRE) was fit as well. This produced no errors, and the results were compared. No concerning differences were observed between the estimates, and the results from these fits are reproduced in Appendix C.3.3.

Figure 8 - Scores on the Pre-Test and Post-Test, Grouped by Condition



This is a violin plot of user pre-test and post-test scores, grouped by condition. Participants in the control condition primarily produced the same number of correctly spelled words in the post-test as they did in the pre-test. Only one participant in the control condition had a score in the post-test above 40 correctly spelled words. Both autocorrect and spell-check conditions were associated with significantly greater differences between the pre-test and post-test scores ($\beta_{autocorrect} = 0.807$, $p<0.05$ and $\beta_{spell\text{-}check} = 0.599$, $p<0.05$). Spell-check and autocorrect were not significantly different from one another, with regards to user's improvement on spelling.

Verification of the model was performed by plotting Pearson residuals against fitted values and all covariates included in the model (Zuur & Ieno, 2016). These plots can be found in Appendix C.3.2. No issues were found in these plots. A Levene's test of the assumption that variance was homogenous across grouping levels was also performed, and also indicated that there was no significant violation of this assumption. Results are available in Appendix C.3.1. R^2 was calculated using the *r.squaredGLMM* in the *MuMIn* package (Bartoń, 2019), and output is reported in Appendix C.3.3. The results from the tri-gamma estimation of R^2 will be reported, per the recommendation of Nakagawa & Schielzeth (2017).

Results and Discussion

The model accounted for 62.5% more variance than the null model (tri-gamma $R^2c = 0.625$). The hypothesis was that the number of correctly spelled words on the post-test would be higher for participants that were in the spell-checking condition than those in the autocorrect condition, in a manner that differed relative to their initial ability as measured by the number of correctly spelled words they had on the pre-test. The score on the pre-test (*Pre-Test*) was a significant predictor in the model ($\beta_{Pre\text{-}Test} = 0.058$, $p<0.05$) while both spell-checking ($\beta_{spell\text{-}check} = 0.599$, $p<0.05$) and auto-correct ($\beta_{autocorrect} = 0.807$, $p<0.05$) significantly predicted higher scores on the post-test compared to the control. According to the model, we should expect that participants in *spell-checking* conditions to produce 124.1% more correctly spelled words on the post-test than participants in the control condition. The lower bound of this estimate is 48.7% more correctly spelled words, and the higher bound is 250.4% more correctly spelled words. For participants in the *autocorrect* condition, we should expect that participants to produce 82.2% more words than participants in the control condition. The lower bound of this estimate is 21% more correctly spelled words, and the higher bound is 182.9% more correctly spelled words. Despite this difference, results from the planned post-hoc comparison of coefficient regressions between auto-correct and spell-checking did not show a significant difference between the two conditions.

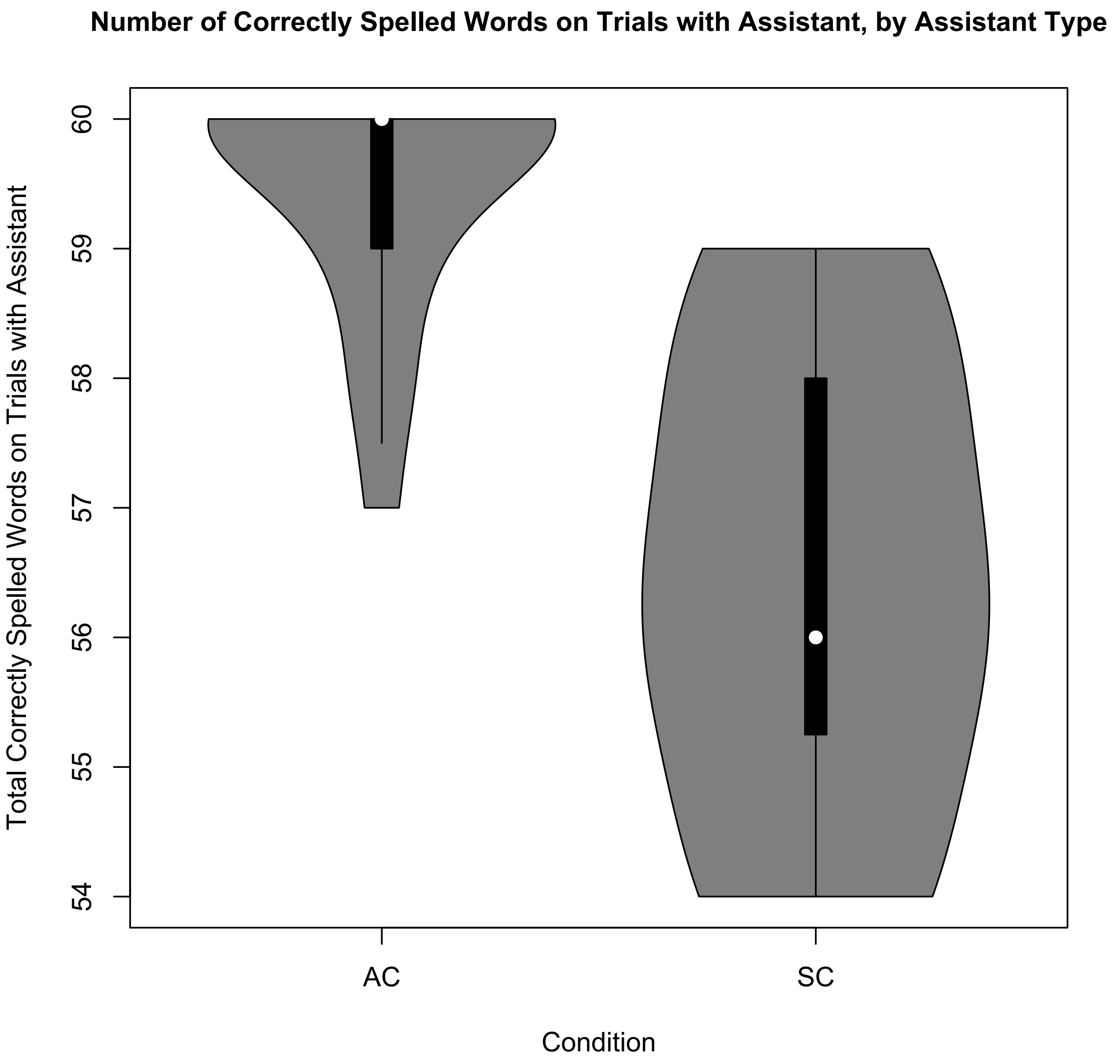
Table 4 - Table of Regression Coefficients for Study 3

Variable Name	Estimate	Std. error	z value	*P*-value	Rate Ratio	2.5% C.I. RR	97.5% C.I. RR
Intercept	1.526	0.203	7.524	<0.05	4.600	2.950	6.693
Pre-Test	0.058	0.008	7.776	<0.05	1.060	1.045	1.078
Condition: AutoCorrect	0.600	0.210	2.864	<0.05	1.822	1.210	2.829
Condition: Spell-Check	0.807	0.212	3.812	<0.05	2.241	1.487	3.504

Alpha level for all tests presented here was α = 0.05. The rate ratio was calculated as the exponentiated regression coefficient. Upper and lower bounds confidence interval bounds were calculated using the 'emmeans' package from R (Lenth, 2019).

The use of the two devices appeared to have different rates of success during usage. The group that used a spell-checking system never achieved a perfect score in the trials where they were allowed to use a spell-checking system, while the group that used an auto-correct system primarily achieved perfect scores on the same trials. This difference was detected by a Kruskal-Wallis rank sum test $\chi^2(1, N= 56)$ $= 22.101$, $p < 0.0001$.

Figure 9-Number of Correctly Spelled Words on Trials with Assistant, by Assistant Type



The violin plot indicates that participants in the autocorrect condition saw at least 57 correctly spelled words, while the participants in the spell-checking condition saw at least 54 correctly spelled words and at most 59 correctly spelled words.

Conclusions for Study 3

It appears that both spell-checking and autocorrect create useful cognitive residuals above and beyond what can be achieved without them. It does not appear that spell-checking significantly assisted users more than autocorrect. This appears to remain true even when dealing with words that are considered difficult to spell for adults.

Unfortunately, the Kruskal-Wallis rank-sum test indicated that performances were different in the training trials across autocorrect and spell-checking conditions. This appears to be a limitation of spell-checking systems in scenarios where the software has a high degree of certainty in the selection. Even if a fully automated system knows the proper spelling, users are still able to select non-target words. Unfortunately, this interferes with inferring the impact that interaction has on the development of useful cognitive residuals

Study 4- Interaction Styles and Word Learning: A Comparison of Selecting a Word with Receiving a Word

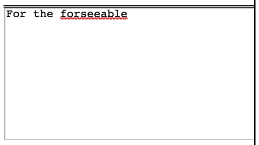
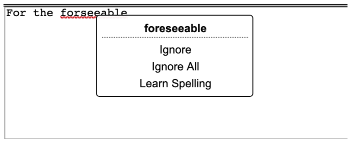
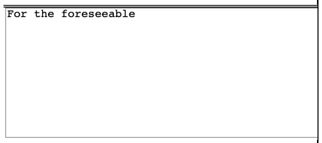
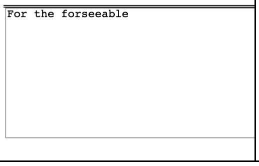
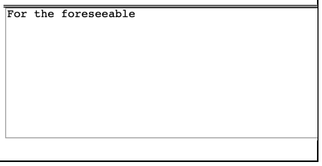
Spell-checking requires the user to select an alternative spelling, whereas autocorrect passively provides alternative spellings to the user. This study aimed to test if this act of selection required in using a spell-checking system is more useful in the creation of spelling-relevant cognitive residuals compared to autocorrect. This may be affected by the difficulty of the words, so the effects should be surveyed across both difficult and easy words.

Further, the commercially available spell-checking software embedded into modern browsers presents users with the opportunity to select incorrect words. While this feature is important in assessing the application of existing spell-checking software as a tool to assist in learning spelling, this feature interferes with the ability to assess the impact of user-interaction on the development of cognitive residuals. To correct for this issue, this study used a java-script spell-checking system with a custom dictionary. This device functioned like most other spell-checking systems by underlining the misspelled word and allowing the user to select a correctly spelled alternative from a list. The important feature is that both the dictionary and the software remained the same across all systems. Misspellings and words were selected for the purpose of creating only a single option for each misspelled word.

Methods

Performance Control

Table 5 - Software Assistance Table

Spelling Support System	Identification Assistance	Selection Assistance	Execution
Spell-Checking	 For the forseeable	 For the forseeable 	 For the foreseeable
Autocorrect	 For the forseeable	This step is handled by the support system	 For the foreseeable

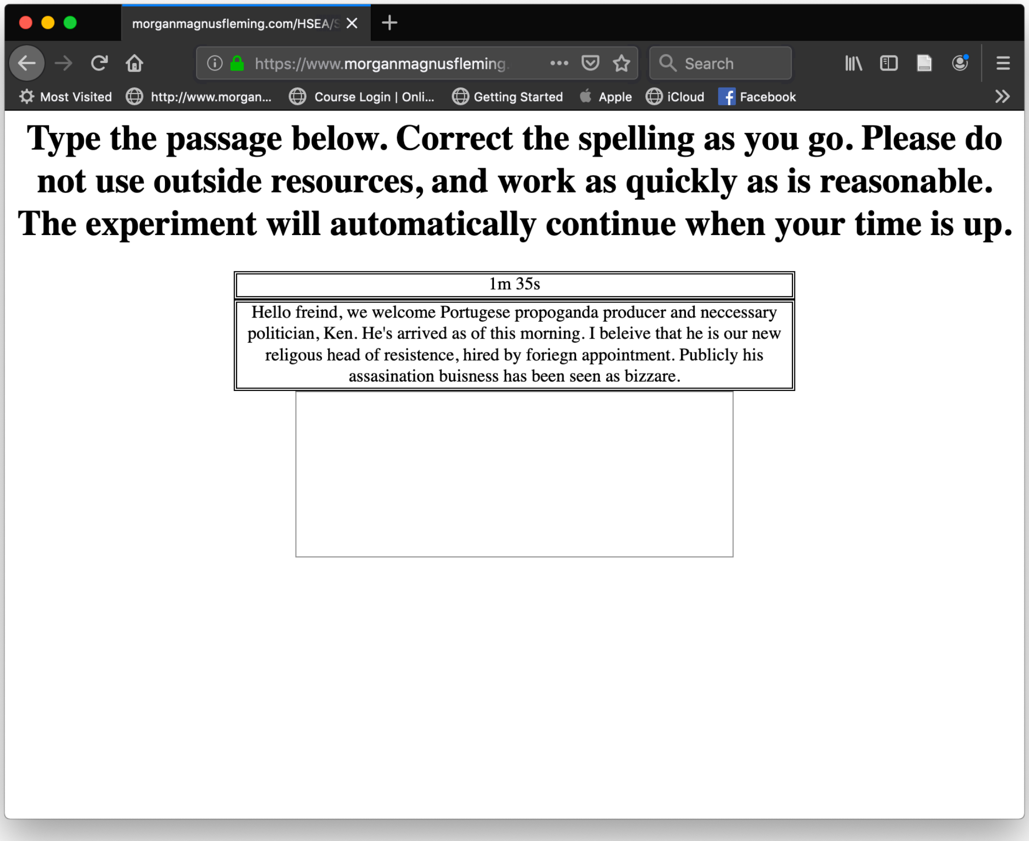
Visual depictions of each phase of the usage procedure for each device. Selection assistance is not displayed for the autocorrect support system, as selection is performed without any visible change in state.

The same measures as Study 2 were taken here to control for performance. This means that all trials with a device had a performance criterion of 90%, all words were selected such at their misspellings aligned with target words, and participants in the autocorrect condition used the javascript autocorrect system described in Study 2.

In addition to the above controls, a javascript spell-checking system was implemented. This was based off of the "Javascript SpellChecker" software provided by Nanospell (Nanospell, 2018). Their product provides the option to create a custom dictionary for a client-side spell-checking system. This software functioned much like a browser or Windows Word spell-checking system. When a user entered a misspelled word, the system would underline this misspelling in red. If the user clicks on the misspelling, a context menu containing a properly spelled alternative word would appear. The user can then select that word by clicking on the context menu button, and the correctly spelled alternative will replace the red underlined word. The custom dictionary feature allowed us to limit the number of possible correctly spelled alternatives to only the words contained within the passage. This helped ensure that users were not selecting non-target words, as they did in Study 3.

Data Collection

Figure 10 – Screenshot of Webpage Used for Data Collection



Type the passage below. Correct the spelling as you go. Please do not use outside resources, and work as quickly as is reasonable. The experiment will automatically continue when your time is up.

1m 35s

Hello freind, we welcome Portugese propoganda producer and neccessary politician, Ken. He's arrived as of this morning. I beleive that he is our new religous head of resistence, hired by foriegn appointment. Publicly his assasination buisness has been seen as bizzare.

This is a screenshot taken from the first pre-test trial. The directions are repeated at the top of the screen for users, they are provided a timer to display their remaining time, the trial passage they are to copy is in the textbox below the timer, and finally the box in which the user types the passage is located just below the passage textbox. Trials proceeded automatically once the timer reached zero.

A total of 36 participants were gathered using Amazon's Mechanical Turk. Participants were asked to copy three short passages and correct any spelling mistakes they may contain. They had 3 minutes to copy each passage before the system would move on to the next trial. They repeated these three times. The first instance was a pre-test period to establish their base rate of word correction without any support system present. The second instance was an experimentally manipulated condition where participants were able to use either an autocorrection system or a "right-click on the underlined word" spell-checking system. The third instance was a post-test period in the same non-supported condition as the first period.

Participants then repeated the above steps, with a separate set of words and the device that they had not used in the previous manipulated condition. Participants saw two different sets of pre-train-post tests: One containing difficult words, and another containing easy words. If they had used spell-checking with hard words in the first pre-con-post set, then they would use autocorrect with easy words in the second pre-con-post set. Easy and hard words were selected from two separate collections, with difficulty being assessed from the source context. Easy words were collected from a Wikipedia list on commonly misspelled English words (Wikipedia contributors, 2017), and the hard words were taken from the 2018 National Senior Spelling Bee (The National Senior Spelling Bee, 2018). Passages are available in Appendix D.1.2.

Table 6 - Target Words

Easy Misspellings	Easy Proper Spelling	Hard Misspellings	Hard Proper Spelling
freind	friend	brasero	bracero
Portugese	Portuguese	vellar	Velar
propoganda	propaganda	darma	dharma
neccessary	necessary	trychina	trichina
religous	religious	pompeno	pompano
resistence	resistance	scanscion	scansion
foriegn	foreign	faillet	faille
beleive	believe	bateste	batiste
assasination	assassination	hallyard	halyard
buisness	business	alacrety	alacrity
bizzare	bizarre	mitsvah	mitzvah
calender	calendar	aballone	abalone
collegue	colleague	chrore	crore
Carribbean	Caribbean	ewwer	ewer
concious	conscious	chanchre	chancre
tendancy	tendency	sashey	sashay
forseeable	foreseeable	centavvo	centavo
goverment	government	mackenaw	mackinaw
remeber	remember	paruke	peruke
gaurd	guard	dellft	delft
posession	possession	kuay	quay
occassion	occasion	wildebeast	wildebeest
agression	aggression	selesta	celesta
neaderthal	Neanderthal	eloadea	elodea
pharoah	Pharaoh	ayuah	ayah
humourous	humorous	xeolite	zeolite
chauffer	chauffeur	gymkana	gymkhana

27 "easy" words were selected from Wikipedia list on commonly misspelled English words (Wikipedia contributors, 2017), and 27 "hard" words were taken from the 2018 National Senior Spelling Bee (The National Senior Spelling Bee, 2018). All misspellings were generated with the intention of ensuring that both the autocorrect and spell-checking systems were able to generate their corresponding proper spellings.

Scores were calculated for each set of trials as the number of target words spelled correctly across all three trials. The statistical software R (R Core Team, 2018) was used to construct and test a generalized mixed Poisson model, using the *lme4* package (Bates et al., 2015). The response variable for this model was the total number of words correctly spelled in the post-test set. The model used the sum of the total number of words correctly spelled on the pre-test set, as well as the device used as independent variables. An observation-level random effect was also included in the model, to account for the over-dispersion common to count data (Harrison, 2014).

Regression

The statistical software R (R Core Team, 2018) was used to construct and test a generalized mixed Poisson model, using the *lme4* package (Bates et al., 2015). The response variable for this model was the total number of words correctly spelled in the post-test set. Difficult and easy words were analyzed separately, using their respective pre-test and post-test scores. All models used the sum of the total number of words correctly spelled on the pre-test set, as well as the device participants used during the trials as independent variables. An observation-level random effect (OLRE) was also included in the generalized linear mixed Poisson model, to account for the over-dispersion common to count data (Harrison, 2014).

In order to assess the impact of spell-checking on editing skills, a Poisson mixed model was constructed to predict performance on an editing task after using a spell-checker or an autocorrect system. The Poisson distribution is suggested for handling count data (Coxe, West, & Aiken, 2009). Because there were only 27 target words for each set, the response variable necessarily violated the Poisson model assumption of a theoretically infinite possible count (Ferrari & Comelli, 2016). However, both quasi-Poisson and generalized linear mixed models have shown to be robust in cases where this assumption is violated (Lazic, 2015). Poissonality of the observed data was assessed by plotting estimated Poisson quantiles of the post-test scores against observed quantiles, with a 95% confidence band. For difficult words, only 3 values fell outside the 95% confidence band. For easy words, 5 values fell outside the 95% confidence band, with the majority being at higher values of the distribution. These were deemed acceptable errors.

The predictor variables for both models included *Pre-Test Score* (a count of the number of errors made on the pre-test) and *Condition* (a two-level categorical variable representing either the use of auto-correct or spell-checking during the training trials). No interaction terms were included, as data exploration determined no need for an interaction term (Appendix D.2.1). To account for over-dispersion an OLRE term, *ID,* was included. For the easy word set the OLRE was estimated to have a mean less than 0.0001 and a standard deviation less than 0.0001. For difficult words the OLRE was estimated to have a mean of 0.126 and a standard deviation of 0.354.

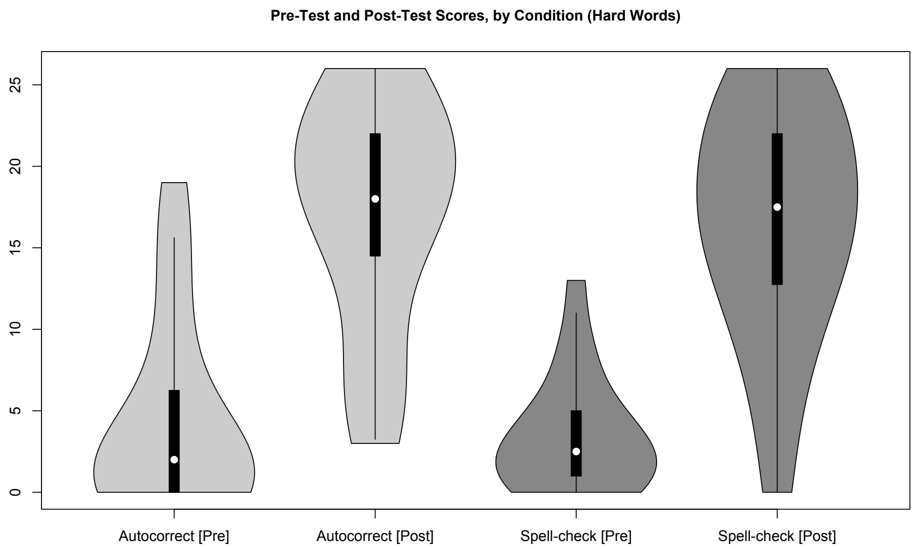
Because there were only 27 target words for each set, the response variable necessarily violates the Poisson model assumption of a theoretically infinite possible count (Ferrari, A., & Comelli, M. 2016). However, both quasi-Poisson and generalized linear mixed models have shown to be robust in cases where this assumption is violated (Lazic, 2015). For this model the response variable is the proportion of words spelled correctly over the total number of target words.

The error structure offered by generalized linear mixed models (GLMM) allows us to model count data where the mean is not exactly equal to the variance of the data, a core assumption of the Poisson distribution (Harrison, 2014). Due to the presence of singularity while fitting the generalized linear mixed model with an OLRE for easy words, a second generalized linear model of the quasi-Poisson family (without an OLRE) was fit for easy words as well. This produced no errors, and the results were compared. No concerning differences were observed between the estimates, and the results from these fits are reproduced in Appendix D.3.3.

Verification of the model was performed by plotting Pearson residuals against fitted values and all covariates included in the model (Zuur & Ieno, 2016). These plots can be found in Appendix D.3.2. No issues were found in these plots. A Levene's test of the assumption that variance was homogenous across grouping levels was also performed and indicated that there was no significant violation of this assumption for the difficult word set. There was, however, excess heterogeneity of variance detected in the easy set $F(1,34)=6.8749$, p=0.013. Results are available in Appendix D.3.1. R^2 was calculated using the *r.squaredGLMM* in the *MuMIn* package (Bartoń, 2019), and output is reported in Appendix D.3.3. The results from the tri-gamma estimation of R^2 will be reported, per the recommendation of Nakagawa and Schielzeth (2017).

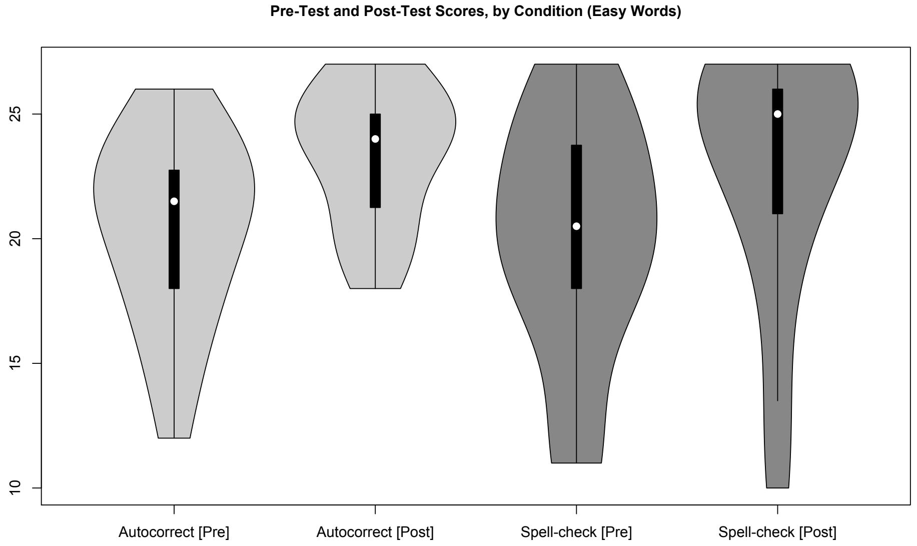
Results and Discussion

Figure 11 - Pre-Test and Post-Test Scores, by Condition (Hard Words)



This is a violin plot depicting the distributions for both pre-test and post-test scores for trials containing the hard words, grouped by condition. Distributions appear very similar across conditions, as was confirmed by the model ($\beta_{Condition\text{-}Hard}$ = -0.004, p = 0.764) .

Figure 12 - Pre-Test and Post-Test Scores, by Condition (Easy Words)



This is a violin plot depicting the distributions for both pre-test and post-test scores for trials containing the easy words, grouped by condition. Distributions appear very similar across conditions, as was confirmed by the model ($\beta_{Condition\text{-}Easy}$= -0.0125, p = 0.980).

The hypothesis was that the number of correctly spelled words on the post-test would be higher for participants that were in the spell-checking condition than those in the autocorrect condition, in a manner that differed relative to their initial ability as measured by the number of correctly spelled words they had on the pre-test. This hypothesis was not confirmed by this study. There were not significant differences observed between participants across the conditions. This was true of both participants working with hard words, and participants working with easy words.

The model for the set with easy words accounted for 31.0% more variance than the null model (tri-gamma R^2c = 0.310). For the model built for the easy set of words, the score on the pre-test (*Pre-Test*) was a significant predictor in the model ($\beta_{Pre\text{-}Test}$ = 0.033, $p<0.05$) while the device used during the training trials was not significant ($\beta_{Condition\text{-}Easy}$= -0.0125, p = 0.980). According to the model, we should expect that participants in *spell-checking* conditions to produce 1.2% fewer correctly spelled words on the post-test than participants in the autocorrect condition. The lower bound of this estimate is 13.8% fewer correctly spelled words, and the higher bound is 13.1% more correctly spelled words (than participants in the autocorrect condition).

The model for the set with hard words accounted for 73.0% more variance than the null model (tri-gamma R^{2}c = 0.730). For the model built for the hard set of words, the score on the pre-test (*Pre-Test*) was a significant predictor in the model ($\beta_{Pre\text{-}Test} = 0.040$, $p<0.05$) while the device used during the training trials was not significant ($\beta_{Condition\text{-}Hard} = -0.004$, $p = 0.764$). For participants in the *spell-checking* condition, we should expect that participants to produce 0.4% fewer words than participants in the autocorrect condition. The lower bound of this estimate is 26.2% fewer correctly spelled words, and the higher bound is 34.2% more correctly spelled words.

Table 7 - Table of Coefficients, Easy Words in Study 4

Variable Name	Estimate	Std. error	z value	*P*-value	Rate Ratio	2.5% C.I. RR	97.5% C.I. RR
Intercept	2.458	0.112	13.078	<0.05	11.691	8.050	16.823
Pre-Test	0.033	0.005	3.846	<0.05	1.034	1.017	1.052
Condition: Spell-Check	0.005	0.042	-0.076	0.940	0.995	0.867	1.142

Alpha level for all tests presented here was α = 0.05. The rate ratio was calculated as the exponentiated regression coefficient. Upper and lower bounds confidence interval bounds were calculated using the 'emmeans' package from R (Lenth, 2019).

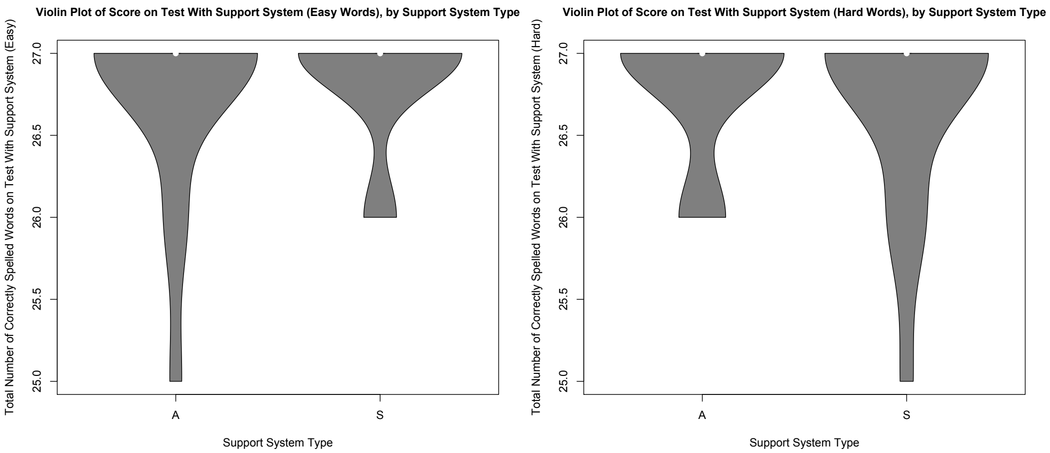
Table 8 - Table of Coefficients, Hard Words in Study 4

Variable Name	Estimate	Std. error	z value	*P*-value	Rate Ratio	2.5% C.I. RR	97.5% C.I. RR
Intercept	2.586	0.128	20.284	<0.05	11.691	10.108	17.015
Pre-Test	0.041	0.015	2.722	<0.05	1.034	1.011	1.075
Condition: Spell-Check	-0.014	0.151	-0.093	0.926	0.995	0.722	1.342

Alpha level for all tests presented here was α = 0.05. The rate ratio was calculated as the exponentiated regression coefficient. Upper and lower bounds confidence interval bounds were calculated using the 'emmeans' package from R (Lenth, 2019).

The use of the two devices did not appear to have different rates of success during usage. The group that used a spell-checking system and the group that used an auto-correct system had an approximately identical distribution. The differences between the two conditions were not detected by the Kruskal-Wallis rank sum test (Hard Words χ^2 (1, N= 36)=0.0327, p = 0.8565; Easy Words χ^2 (1, N = 36) = 0.0229, p = 0.8798). Density plot and results from the Kruskal-Wallis test can be found in Appendix D.2.1.1.

Figure 13- Violin Plot of Scores on Trials with Support Systems Enabled



Participants in the spell-checking condition properly spelled between 25 and 27 of the 27 target words. Participants in the autocorrect condition correctly spelled between 26 and 27 of the 27 target words. This difference was not found to be significant (Appendix D.2.1.1.)

Conclusions for Study 4

Both devices appear to have given users comparable exposure to correctly spelled words, as is indicated by the results of the Kruskal-Wallis rank sum test. These devices did not appear to differ in their ability to create useful cognitive residuals. Consequently, it may be reasonable to conclude that the mere act of selecting a word from a context menu does not create a more useful cognitive residual than passive correction (autocorrect). This appears to be true whether or not users are dealing with common or uncommon words.

Study 5 - Building a Spell-Check That Teaches: Instructive Design Through Procedure or Perception

Study 1 showed that spelling performance on a post-test differs across commercially available spell-checking and autocorrect systems. Study 3 showed that both autocorrect and spell-checking systems can create useful cognitive residuals. Yet taken together, the findings of Studies 2, 3, and 4 show that if each system provides equal success in finding the correct spelling of a word, spell-checking and autocorrect systems do not create comparatively more or less word learning. Unlike in Studies 2 and 4, the scores on training trials in Study 1 differed significantly, meaning differences on the post-test could be attributed to a different number of properly spelled words seen during the training trials. If performance correlated with exposure to properly spelled words, the best way to instruct users in spelling may be to reserve spelling instruction to intervals separate from on-line editing.

To determine if the interaction with these systems is useful for learning, rather than if the systems are only useful in that they provide us proper spellings, we can compare the performance development experienced by spelling support system users to that of participants exposed only to properly spelled words. Users that have to copy a passage that already contains correctly spelled words can provide such a comparison. If participants are only asked to copy a properly spelled passage, then they will have witnessed and interacted with the proper spellings independent of a spelling support system. If we compare their performance on the post-test to the performance of spelling support system users, we can get an impression of how the user's mode of interaction (supported or unsupported) affects the development of useful cognitive residuals.

After the previous studies, I do not believe that either of these systems provide users the basis necessary to develop cognitive residuals beyond what could be achieved by copying the correctly spelled words. My reasoning follows. From Studies 2, 3, and 4, participants that were asked to select properly spelled words from a list (spell-checking) and participants that had proper spelling provided to them automatically (autocorrect) performed roughly equivalently with regards to improving their post-test scores. Compared to the act of copying a properly spelled word, receiving a correction from an autocorrect system is a seemingly information impoverished learning opportunity. Users of an autocorrect system are required to take no action beyond making a mistake. Neither of these systems create a chance for the user to either practice typing out the word, nor do they encourage the user to consider the proper spelling once it is provided.

Consequently, it seems reasonable to believe that users of either system would perform (with regards to spelling skill development), at best, equivalently with users who simply copied a fully corrected passage. Copying a corrected passage requires users to both practice seeing and typing the correctly spelled word. Both spell-checking and autocorrect systems require a user to make a mistake, and in that act, they must practice typing an incorrect spelling. The most consideration a user needs to place in correcting this misspelling is with a spell-checking system. Even then though, the user needs only to identify if the proper spelling offered by the system closely resembles the word they are correcting. They do not need to consider the differences in spelling between the proper spelling and the misspelling, nor the details of the proper spelling.

However, there are more ways to interact with spell-checking systems rather than just list selection and automatic replacement. Selecting a properly spelled word from a list is the method of input provided by most modern spell-checking systems. Modern autocorrect systems automatically replace misspelled words. But more actions exist for users to participate in the assisted editing process than are represented by these two systems.

Of particular note are methods of user interaction that align well with the act of properly spelling words. This includes typing out the properly spelled word, as one might do if they had genuine prior knowledge of the proper spelling as well as the act of correcting a misspelled word by modifying an existing misspelling. Rather than typing out the entire correction, users that discover a misspelled word in their work can often opt to modify the existing misspelling into the properly spelled word (e.g. “helllo” becomes “hello” by simply deleting the extra “l”). Spell-checking systems can be modified to require that users take these actions, either by typing out the entire word or by modifying an existing misspelling. This study included spell-checking software that made such requirements, in order to explore the potential of new methods of user interaction in creating useful cognitive residuals.

Similarly, autocorrect may be attentionally enhanced to assist users in learning from their mistakes. For instance, the act of correcting a mistake provides a moment in which the system is capable of spelling a word that the user has not. In addition to simply providing the correction, the system can draw the attention of the user to the correction. Some existing autocorrect software has this function and provides the user a fleeting visual cue that a given word has been corrected. Additionally, color highlighting of corrections in an autocorrect system has been explored as a means of enhancing spelling learning in children (Arif, Sylla, & Mazalek, 2016). Highlighting misspelled words that the autocorrect system assisted the user with can potentially enhance the opportunity for users to practice seeing properly spelled words.

New Devices

To explore these methods of spell-checking and autocorrecting as they relate to both copying properly spelled words and unassisted performance, I constructed the following seven conditions for this study. Participants were placed in one of seven conditions. In the trials between the pre-test and post-test, referred to as the training trials, participants were either provided one of the five support systems created for the study or they were required to copy the passage without any assistance. Screenshots of the support systems in action can be found in Appendix F.1.3.

The *control* condition provided users no spelling support systems and asked them to attempt to correct the spelling in the passages. These trials were identical to those in the pre- and post- tests. Additionally, participants in the *control* condition did not have to satisfy the 90% performance criterion. This condition aimed to represent user spelling ability development in the absence of a support system.

The *corrected* condition formed a second control condition. In this condition users were provided passages containing the correctly spelled target words. The passages did not require further editing, and participants were made aware of this fact. Participants were instructed to copy these passages rather than copy and correct them. Participants in the *corrected* condition were required to spell at least 90% of the target words correctly. This condition aimed to represent user spelling ability development through exposure (seeing the correctly spelled word) and practice (typing the correctly spelled word).

The *spell-checking* condition used the javascript spell-checking system that was introduced in Study 4. This system functioned similarly to modern spell-checking devices found in Microsoft Word and modern browsers. The system indicated misspelled words for users by underlining the misspelling with a red line. Right clicking this misspelling opened a context menu with properly spelled words as potential substitution candidates. The user was then able to select one of these words and the system would replace the underlined misspelling with the selected word. This condition aimed to represent user spelling ability development through use of commercially available spell-checking systems.

The *full-word* condition used a modified version of the javascript spell-checking system introduced in Study 4. The system was designed with the intention to ensure that users practiced typing out the correctly spelled word. Users who entered a misspelling would still receive the underline characteristic of traditional spell-checking systems. The selection step was modified to delete the word upon opening the context menu. The execution step was modified to prevent replacement of the misspelled word with the target word. By deleting the word and removing the usual replacement function, the only way for the user to make use of the proper spelling was to type out the word themselves. Thus, in order to use the device participants would need to practice typing out the properly spelled word.

The *part-word* condition also used a modified version of the javascript spell-checking system introduced in Study 4. This system was designed with the intention to ensure that users practiced editing an incorrectly spelled word. Only the execution step was modified to prevent replacement of the misspelled word with the target word. This ensured that participants would have to modify the misspelling themselves in order to create a properly spelled word. Thus, in order to use the device successfully participants would need to edit the misspelled word themselves.

The *autocorrect* condition used the same autocorrection system first used in Study 2. This system replaces misspelled words as the user types. Whenever the user pressed the space bar, the last word in the text they had written would be substituted with a word from the system's dictionary. This was done automatically and without indication to the user that it had been performed. This condition aimed to represent user spelling ability development through use of commercially available autocorrect systems.

The *autohighlight* condition used the same autocorrection system as the *autocorrect* condition, with one modification. Target words were highlighted in light blue. This change was made with the intention of giving users a durable indication of which words the autocorrect system had to correct.

I designed this study to address the following theoretical questions:

1. Are selecting from a list and automatic replacement, the methods of text-interaction provided by spell-checking and autocorrect respectively, able to teach users how to correct spelling errors? Are they able to teach users better than if those users had practiced seeing and typing the correct word without the support system?
2. Are there other ways to interact with spelling support systems that can passively enhance the user's own spelling capabilities? If so, are they able to teach users how to correct spelling errors better than if they practiced seeing and typing the correct word without the support system?

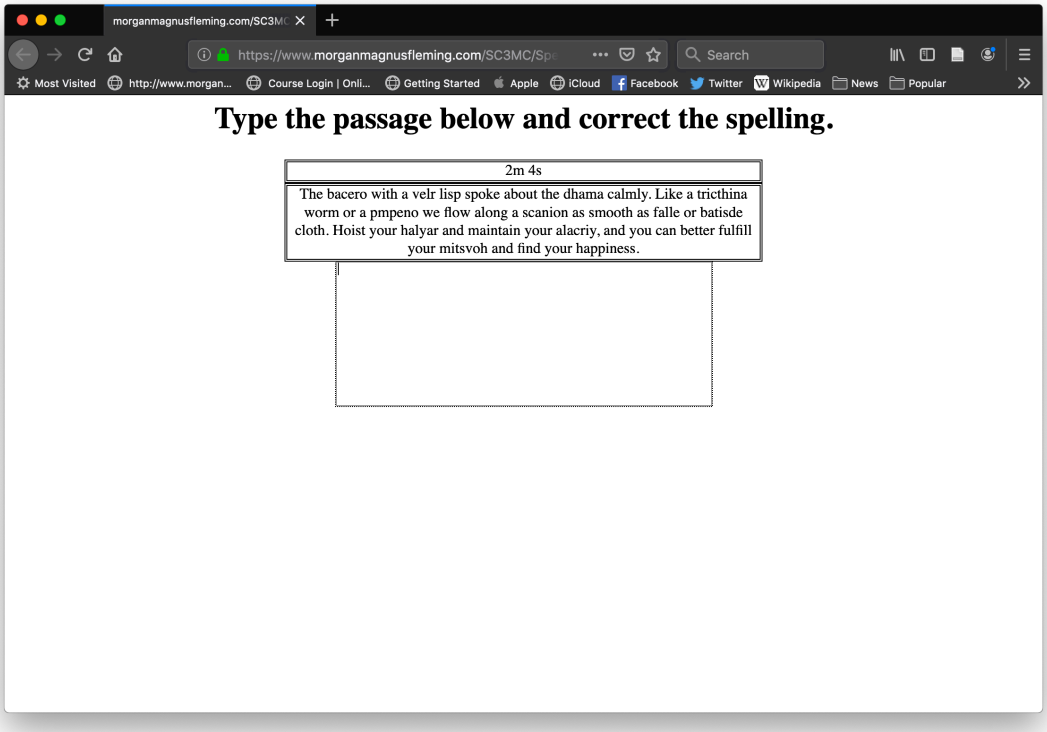
The new spelling support systems described above (and in more detail below) introduce new ways for users to interact with the spell-checking process. The *corrected* condition provided a standard for assessing the effects of how the user interacted with the spell-checking process, by showing the effects of simply having the solutions provided. This allowed me to test the corresponding hypotheses:

1. The *autocorrect* and *spell-checking* systems do not have a greater positive impact on post-test scores than copying the passages in the *corrected* condition, but they do have a greater positive impact on post-test scores than copying the passages in the *control* condition.
2. The new spelling support systems (*part-word, full-word, autohighlight*) have a greater positive impact on post-test scores than simply copying the passages in the *corrected* or *control* condition.

Methods

Data Collection

Figure 14 - Screenshot of the Website Used to Gather Data



Type the passage below and correct the spelling.

2m 4s

The bacero with a velr lisp spoke about the dhama calmly. Like a tricthina worm or a pmpeno we flow along a scanion as smooth as falle or batisde cloth. Hoist your halyar and maintain your alacriy, and you can better fulfill your mitsvoh and find your happiness.

A total of 126 participants were gathered using both the UC Merced SONA system(N=62) and Amazon's Mechanical Turk (N=64). Participants were asked to copy three short passages and correct any spelling mistakes they may contain. They repeated these three times. The first instance was a pre-test period to establish their base rate of word correction without any support system present. The second instance was an experimentally manipulated condition where participants were able to use a spelling support system or were placed in one of two control conditions. Participants in the control conditions were either provided a misspelled passage during these trials and asked to correct the passage, or they were provided a properly spelled passage and asked to copy the passage into the text-box. Participants that were provided a spelling support system were instructed to correct the spelling in a misspelled passage and were required to correctly spell at least 90% of the words in these trials correctly. The third instance was a post-test period in the same non-supported condition as the first period. This matches the pre-training-post-transfer structure of Study 1.

Participants saw two different sets of pre-train-post tests: One containing difficult words, and another containing easy words. Easy and hard words were selected from two separate collections, with difficulty being assessed from the context of the collection. Easy words were collected from a Wikipedia list on commonly misspelled English words (Wikipedia contributors, 2017), and the hard words were taken from the 2018 National Senior Spelling Bee (The National Senior Spelling Bee, 2018).

All data was collected with a website in the style of the one presented in Study 2. Both UC Merced SONA students and Amazon mTurk workers used this same website to participate in the study. The only modification made between the two communities were instructions relating to their SONA and Mechanical Turk IDs.

Scores were calculated for each set of trials as the number of target words spelled correctly across all three trials. The score for the transfer test was the total number of target words spelled correctly during the transfer test. The statistical software R (R Core Team, 2018) was used to construct and test a generalized mixed Poisson model, using the *lme4* package (Bates et al., 2015). The response variable for this model was the total number of words correctly spelled in the post-test set. The model used the sum of the total number of words correctly spelled on the pre-test set, as well as the condition as independent variables. An observation-level random effect was also included in the model, to account for the over-dispersion common to count data (Harrison, 2014).

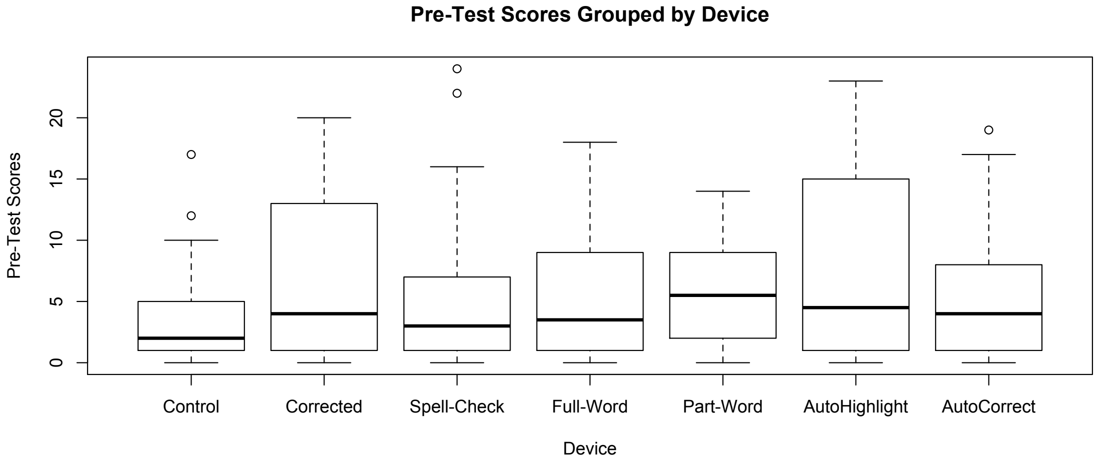
Regression

In order to assess the impact of the different forms of spelling assistance on editing skills, a Poisson mixed model was constructed to predict performance on an editing task. This model was constructed using the same process described in Study 2 and 3. Poissonality of the observed data was assessed by plotting estimated Poisson quantiles of the post-test scores against observed quantiles, with a 95% confidence band. 12 observations fell outside these bands. Given the size of the sample (n=126), this was deemed an acceptable deviation and these observations were used in the model.

The predictor variables for the model included *Pre-Test Score* (a count of the number of errors made on the pre-test) and *Device* (a seven-level categorical variable representing the use of one of the seven support systems described above). No interaction terms were included, as data exploration determined no need for an interaction term (Appendix F.2.1). To account for over-dispersion an OLRE term, *ID,* was included. For the easy word set the OLRE was estimated to have a mean less than 0.102 and a standard deviation less than 0.319.

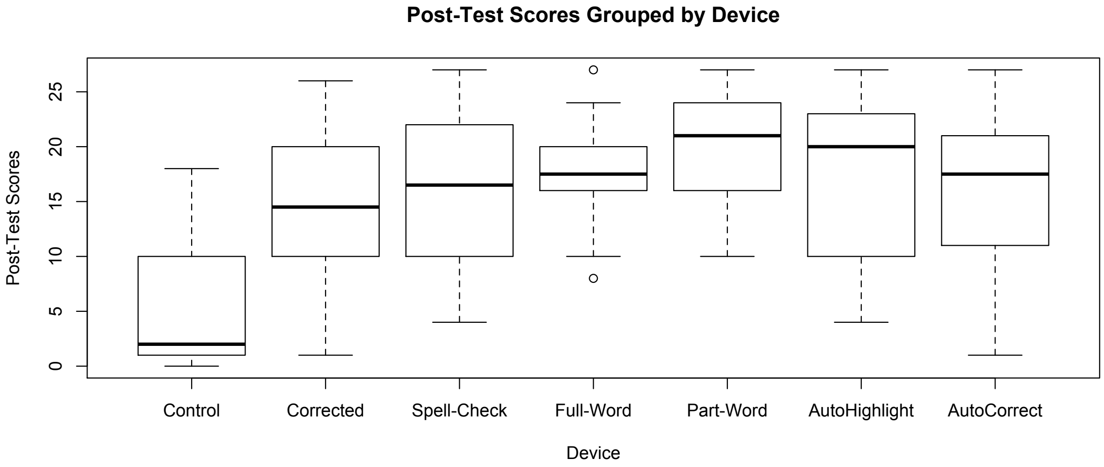
Because both hypotheses depend on specific levels of the variable *Device*, a comparison of the regression coefficients for all levels will be made to all other levels using the multiple comparisons function *glht* in the R package *emmeans* (Hothorn, Bretz, & Westfall, 2008). If the first hypothesis is correct, then estimates for the regression coefficients for both autocorrect and spell-checking should be significantly larger than the regression coefficient for participants in the *control* condition. Additionally, estimates for the regression coefficients for both *autocorrect* and *spell-checking* conditions should not be significantly larger than the regression coefficient for participants in the *corrected* condition. If the second hypothesis is correct, then estimates for the regression coefficients for *part-word*, *full-word*, and *autohighlight* should all be significantly greater than the regression coefficients for participants in both the *corrected* and *control* conditions.

Figure 15 - Pre-Test Scores Grouped by Device



These are the distributions of scores from the pre-test (Pre-Test Scores), grouped by the type of spelling support system provided in the training trials (Device).

Figure 16 - Post-Test Scores Grouped by Device



These are the distributions of scores on the post-test, grouped by the type of spelling support system provided in the training trials (Device).

Results and Discussion

The results confirm the hypothesis that *autocorrect* and *spell-checking* systems do not have a greater positive impact on post-test scores than copying the passages in the *corrected* condition. Autocorrect and spell-check, however, do have a greater positive impact on post-test scores than copying passages in the *control* condition. The model these conclusions are based on accounted for 85.3% more variance than the null model (tri-gamma $R^2c = 0.8533$). Both *autocorrect* ($\beta_{autocorrect} = 1.079$, $p<0.05$) and *spell-checking* ($\beta_{spell\text{-}checking} = 1.032$, $p<0.05$) conditions had significantly greater regression coefficients than the *control* condition. They both also had greater regression coefficients than the *corrected* condition ($\beta_{corrected}= 0.851$, $p<0.05$), but this difference was not significant ($\text{Ratio}_{Autocorrect/Corrected} = 0.2281$, z = 1.631, p =0.6598; $\text{Ratio}_{Corrected/Spell\text{-}Check} = 0.1812$, z = 1.292, p = 0.8547).

Table 9 - Table of Regression Coefficients

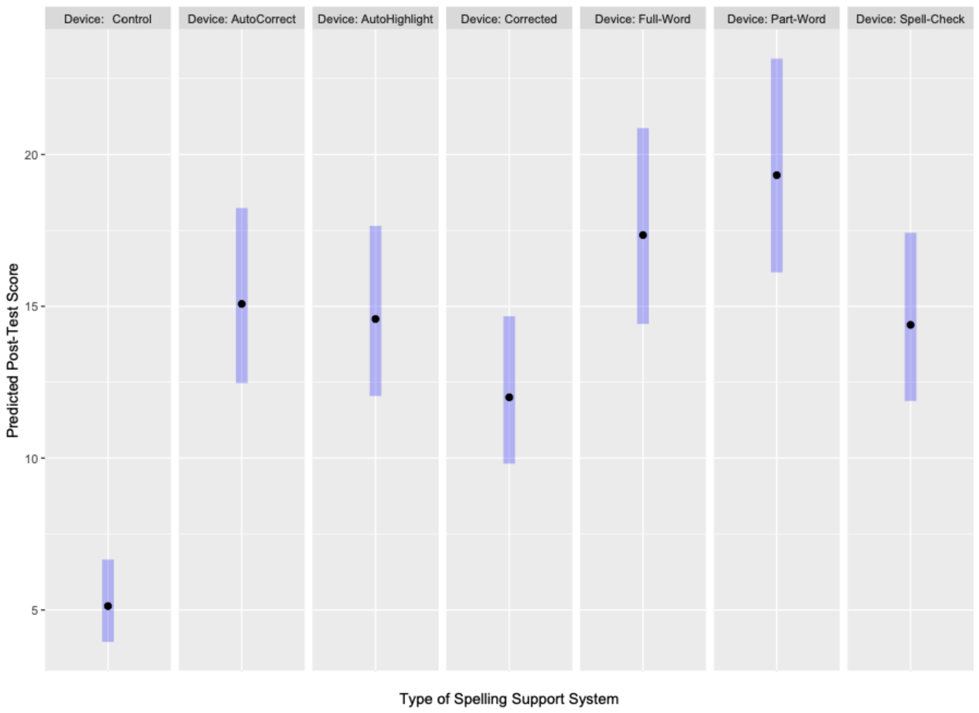
Variable Name	Estimate	Std. error	z value	*P*- value	Rate Ratio	2.5% C.I. RR	97.5% C.I. RR
Intercept	1.350	0.139	9.678	<0.05	3.857	2.886	5.015
Pre-Test	0.049	0.006	7.978	<0.05	1.050	1.038	1.063
Corrected	0.851	0.164	5.157	<0.05	2.341	1.699	3.269
Spell-Check	1.032	0.164	6.291	<0.05	2.806	2.046	3.922
Full-Word	1.219	0.163	7.485	<0.05	3.383	2.475	4.724
Part-Word	1.327	0.162	8.201	<0.05	3.768	2.763	5.252
Autocorrect	1.079	0.163	6.604	<0.05	2.941	2.146	4.104
AutoHighlight	1.045	0.163	6.395	<0.05	2.844	2.074	3.968

Alpha level for all tests presented here was $\alpha = 0.05$. The rate ratio was calculated as the exponentiated regression coefficient. Upper and lower bounds confidence interval bounds were calculated using the 'emmeans' package from R (Lenth, 2019).

Of the new spelling support systems (*part-word, full-word, autohighlight*), only part-word showed a greater positive impact on post-test scores than simply copying the passages in the *corrected* or *control* condition. The new spelling support systems *part-word* ($\beta_{part\text{-}word} = 1.327$, p<0.05)*, full-word* ($\beta_{full\text{-}word} = 1.219$, p<0.05)*,* and *autohighlight* ($\beta_{autohighlight}$ = 1.045, p<0.05) had a greater positive impact on post-test scores than not having access to a support system (the control condition). However, *full-word* and *autohighlight* had regression coefficients that were not significantly different from the *corrected* condition ($\text{Ratio}_{Corrected/Full\text{-}Word}$= 0.3682, z = 2.650, p = 0.11; $\text{Ratio}_{Autohighligh/\ Correctedt}$= 0.1945, z = 1.396, p = 0.8024). Only the regression coefficient associated with *part-word* was significantly greater than the regression coefficient associated with the *corrected* condition($\text{Ratio}_{Part\text{-}Word/Corrected}$ = 0.4759, z = 3.460, p < 0.05).

Results for the transfer test contained similar patterns. Due to a singularity in fitting the Poisson GLM with an OLRE, results reported for this section regard an equivalent Quasi-Poisson model. The model accounted for 49.4% more variance than the null model (tri-gamma R^2c = 0.4938). All systems tested, as well as the *corrected* condition, produced users with higher transfer test scores than the control condition($\beta_{full\text{-}word}$ = 1.0308, p <0.05; $\beta_{part\text{-}word}$ = 1.0862, p <0.05; $\beta_{spell\text{-}check}$ = 0.7677, p <0.05; $\beta_{autohighlight}$= 0.8818, p <0.05; $\beta_{autocorrect}$ = 0.8900, p <0.05; $\beta_{corrected}$ = 0.6574, p <0.05). Only the part-word condition outperformed the corrected condition ($\text{Ratio}_{Part\text{-}Word/Corrected}$= 0.4288, z = 3.211, p <0.05).

Figure 17 - Predicted Post-Test Score Grouped by Type of Spelling Support System Used During Training Trials



The estimated marginal means for post-test score is represented by the point, and the corresponding 95% confidence interval in that estimation is represented by the purple bar. Estimated marginal means were computed using the emmeans package in R (Lenth, 2019). The part-word spelling support system outperformed all other conditions, including significantly surpassing the post-test scores from participants in the corrected condition.

Conclusions for Study 5

All support systems tested here had a significant impact on the spelling abilities of users. Compared to the control condition, having access to any of the spelling support systems explored here increased the post-test scores of participants. Both traditional autocorrect and spell-checking increased post-scores higher than had participants not had access to any spelling support system. From this, it would appear that yes, both of these devices can teach their users how to correct spelling errors in a document.

However, neither of these spelling support systems significantly outperformed users who practiced seeing and typing the correctly spelled words. While participants that were asked to copy a correctly spelled passage did perform worse than participants that used either spell-checking or autocorrect, this was not significantly so. It does not appear that autocorrect or spell-checking improves spelling beyond what could be achieved by practicing typing and seeing the properly spelled words.

Only participants in the *part-word* condition significantly exceeded the performance of the participants in the *corrected* condition. It appears that neither drawing their attention to automatically provided proper spellings (as in the *autohighlight* condition) nor requiring the user to practice typing out the proper spelling (as in the *full-word* condition) exceeded the performance improvements observed when participants simply saw and copied properly spelled passages. However, allowing participants to edit existing misspellings while providing them the proper spelling did appear to raise their post-test scores beyond those who saw and copied properly spelled passages.

The results suggest that modern autocorrect and spell-checking systems provide users the means to improve their spelling abilities. Compared to the absence of any spelling support system, participants that use a spelling support system are provided both a means of identifying misspellings and a means of acquiring a properly spelled word. And they are able to take advantage of these affordances in order to learn to identify and correct misspellings.

As shown above, traditional spell-checking and autocorrect are not the only ways spelling support systems could provide users this information. With regard to skill development, they are also not the best ways for users to interact with spelling support systems. Providing users with a properly spelled passage and asking them to copy it enhanced their ability to spell the words about as well as either of these systems. The only system that exceeded the learning experienced by participants in the *corrected condition* was the *part-word* system. The *part-word* system required that users make the edit themselves. This encouraged them to more deeply consider what was wrong about the initial misspelling, and what constituted a proper spelling of that word. It appears that this created cognitive residuals that were useful to the task at hand.

Chapter 4 - Conclusion

This project began with the aim of comparing different methods of interaction for existing spelling support systems offered by Microsoft to explore their ability to foster user spelling skill development. I demonstrated that within the domain of Microsoft Word, users can expect to get better at editing for spelling if they use Microsoft's spell-checker rather than Microsoft's autocorrect. However, further study suggests that this result is not tied to a difference in interactive affordances. By taking steps to control for performance between autocorrect and spell-checking systems, it became apparent that selecting an adequately spelled word from a context menu, as one does with spell-checking, does not confer any additional benefit beyond what is offered by autocorrect. That said, both systems did help users improve their spelling abilities.

The opportunities for autocorrect and spell-checking to create incidental spelling skill improvement appear to lie mostly in their provision of correctly spelled words. Traditional spell-checking and autocorrect systems do not improve users' spelling skills any more than providing users correctly spelled words would. Neither do autocorrect systems that highlight corrections, nor spell-checking systems that require users to type out the correctly spelled word. Only a spelling support system that requires users to make edits themselves significantly outperformed participants that were provided a correctly spelled document.

What follows is a review of the studies that established these conclusions. After that, a brief discussion of the shortcomings of this study. Finally, a mention of possible future directions.

Review

I conducted five studies to explore the impact of different ways of interacting with a spelling device on the development of user skill. The first study established that for users of Microsoft Word, their ability to correct errors was better improved by spell-checking rather than autocorrect. Errors on the post-test were 36.9% lower for participants that had previously been allowed access to Microsoft's spell-checking function rather than Microsoft's autocorrect capability. I observed this difference in conjunction with slightly lower scores for autocorrect users during the supported trials. Participants that had access to the autocorrect support system had more errors in the trials in which they had access to autocorrect, compared to participants that had access to the spell-checking system. It may be the case that post-test score differences arose not from how users interacted with these tools, but the relative success of these tools at delivering correctly spelled words. Regardless of the cause, it would appear that users that wish to improve their editing ability would derive more benefit from using Microsoft's spell-checking system rather than Microsoft's autocorrect system.

The second study aimed to explore further the differences in skill development driven by the method of interaction, rather than the capabilities of the spelling support system to render the correct, correctly spelled word. Here I introduced a limited-dictionary autocorrect system and was able to create comparable spelling correction performance between autocorrect and spell-checking systems. Users scored roughly the same on trials supported by either autocorrect or spell-checking systems. With this performance during supported trials controlled, no significant differences between conditions were observed in the post-test scores.

The third study explored if increasing the difficulty of the target words would expose differences in post-test scores between participants in the spell-checking and autocorrect conditions. It also aimed to establish whether or not either system conferred useful cognitive residuals to their users. Again, as was found in the second study, no reliable differences were observed between participants that had access to spell-check or autocorrect with regards to their post-test scores. However, participants that were provided either support system outperformed the participants who were never provided a support system. From this study, we can conclude that even if the user is dealing with words that are considered challenging to spell, autocorrect and spell-checking offer similar beneficial cognitive residuals.

The fourth study aimed to explore more specific aspects of interaction with spell-checking devices. In particular, the object of inquiry was the act of selecting a correctly spelled word. I introduced a spell-checking system in which the user was presented with a single option and compared their performance on a post-test with users that were provided an autocorrect support system. This comparison was made with two sets of words, both challenging and easy, to ensure that the results of this test did not vary with difficulty. The results indicated that there was no reliable difference in post-test performance between users that selected a word offered by spell-check, and those that had their corrections automatically handled by autocorrect.

With the fifth study, I explored alternative designs for spell-checking and autocorrect support systems, and if the type of interaction they provide is doing more for users than merely presenting them with valuable information. I introduced three new designs for spelling support systems: part-word, full-word, and autohighlight. The first two systems, part-word, and full-word, aimed to develop new ways to engage users in the editing process by requiring users to make the corrections themselves. The third system, autohighlight, introduced automatic highlighting of corrected words as a means for notifying users about the corrections that the autocorrect system. These three systems, as well as the traditional autocorrect and spell-checking systems, were compared to a control in which participants practiced typing an already corrected set of passages. This control ensured that participants received all the same information they would have if the spelling support systems duly delivered it. After correcting for differences in user's capabilities, it was only the part-word edit that appeared to outperform the new control condition. It was the only system providing the user more than just the information regarding the corrected word.

It is interesting that when users are presented with a challenge to spell words correctly, neither users of autocorrect nor spell-checking exhibit a loss of relevant skills. This lack of skill-loss would distinguish editing a document with a spell-checker from navigating with GPS or using digital storage for reference (Ishikawa et al., 2008; Fenech et al., 2010; Sparrow et al., 2011; Dong & Potenza, 2015). GPS, certain forms of digital storage, and spelling support systems all provide users information relevant to the solution of their respective task domains. GPS provides a route, digital storage provides a reference, and spelling support systems provide a proper spelling. Unlike GPS and digital storage systems, both forms of spelling support systems benefitted the development of independent spelling skills in users. This improvement in skill agrees with the findings of Lin et al. (2017) and Arif et al. (2016), who showed that both spell-checking and autocorrect (respectively) could improve user's abilities to spell in educational contexts. This point warrants further exploration, as it is still not clear why spelling support systems are not prone to the same sorts of detrimental cognitive residuals observed in these other systems.

The incidental learning opportunities that allow users to develop useful cognitive residuals while using autocorrect and spell-checking do not appear to be unique to these systems' methods of interaction. As was shown in Study 5, the user skill improvement created by traditional autocorrect and spell-checking support systems was directly comparable to the user skill improvement created by exposing users to correctly spelled words. As further evidence against the importance of the method of interaction, when performance between the devices was carefully controlled (Studies 2, 3, and 4), differences in user skill-development between autocorrect and spell-check became insignificant. It does not appear that the greater user responsibility and exposure to the correction process confer more useful cognitive residuals than the passive provision of correct words.

Luckily these are not the only ways of interacting with spelling support systems. The part-word support system significantly outperformed directly exposing users to correctly spelled words (Study 5). Interestingly, neither the full-word support system nor the autohighlight support system was able to outperform simple exposure. This failure of all other tested interaction methods to exceed simple exposure suggests that there was something unique to the part-word support system. The part-word support system made available information regarding the solution, which was present in all conditions, but it also allowed prolonged user exposure to the error they were attempting to correct as well. The full-word support system accomplished the task of requiring users to input the entire correct spelling by deleting the misspelled word. The autohighlight support system automatically eliminated the misspelled word by replacing it with the correctly spelled word. The part-word support system, however, kept the misspelling on the screen while the user edited the word. Given that spelling correction can only occur once an error is detected, this may have offered users a better chance to become acquainted with the nature of the error.

Methodological Issues

The studies presented here suffered from several methodological issues. These studies aimed to explore incidental spelling skill development that arose from using autocorrect and spell-checking devices. Consequently, the use of a pre-test phase may have interfered with this goal. By providing users an opportunity to compare their performance between the unsupported pre-test and the supported set of trials, it is possible they were incentivized to pursue learning. If this were the case, it would be difficult to describe this as a form of incidental learning. However, Sparrow et al. (2011) noted that the memory deficits they observed with digital reference systems were present regardless of whether or not participants were explicitly instructed to remember the facts they were trained on or not.

These studies also used the same misspellings in all sets of trials. “Improvements in spelling ability” may, therefore, be a misnomer, as there is no evidence to show that participants' overall ability to spell words improved. Instead, these studies used “the ability to spell the tested words” as a stand-in for overall “spelling ability”. In one sense, this is reasonable, as overall spelling ability can only be measured in the number of words that participants can correctly spell. If one can spell more words, one can say that they have improved their spelling abilities. The issue arises when looking at adult spelling skills as a whole. While orthographic memorization is fundamental to spelling skill in English, spelling skill also involves understanding phonetic/orthographic relationships, spelling rules, and morphological units (Holmes & Malone, 2004). These elements are outside of the scope of this study. They also are not directly relevant to the functions of either of the primary spelling support systems explored here. They are, however, interesting in their own right. The cognitive residuals conferred by either autocorrect and spell-checking may be relevant to some of these more in-depth concepts.

Future Directions

This study does not explore the longitudinal effects of the regular use of either spell-checking or autocorrect devices. Due to the recognized long-term effects of using digital devices (Loh, & Kanai 2016; Firth et al., 2019; Kirschner & De Bruyckere, 2017), it seems like it will be necessary to try to extend these results into a study of spelling skill as it develops across time. Early adopters of spelling support systems may have different spelling skill trajectories than people who adopt the technology early.

Another element worth exploring is the importance of user motivation in incidental skill development. This study was structured with the intention of exploring incidental, or rather non-intentional, learning in users. Findings regarding the actual “incidental-ness” of the learning in this paradigm, such as by asking participants if they were intentionally trying to learn from the systems, could help explore if the effects observed here are incidental or driven by the intentions of the user.

Additionally, the findings here are limited to the realm of spell-checking. It would be interesting to see if similar approaches to deepening user involvement can assist in skill development in fields such as navigation or memory. The results here suggest that if users are required to participate in solving a problem while using a digital assistant, they may better develop their own independent abilities. It is conceivable to consider similar actions being worked into GPS and storage systems, or that such affordances may already exist. Modifying the act of looking something up that is stored in digital memory, such as requiring users to recall exact passages from the material (assuming the information being referred to is in the form of written material), may assist with recall. Similarly, asking users to identify landmarks while navigating with GPS may enhance a user’s ability to recall the route they’ve taken. This study has not confirmed that deepening user involvement extends beyond word learning from spelling support systems, though it appears that it may be a fruitful space to explore.

It also seems imperative to explore precisely why autocorrect and spell-check do not induce the type of skill loss associated with other digital devices (Ishikawa et al., 2008; Fenech et al., 2010; Sparrow et al., 2011; Henkel, 2014; Dong & Potenza 2015). Given the relatively diminished role of interaction in how these spelling support systems foster user skill development, we must look to other aspects of either their design, use, or domain of application (assisting spelling) that could explain why they create (rather than impair) user skill development. The results here would suggest that this reason lies in their domain of application. Varying the design and method of use for these tools never created a situation in which the devices created a loss of user skill. Further, directly providing users proper spellings appeared to increase user spelling skills as well. It seems likely that spelling is a task that is well suited to incidental learning through exposure to proper solutions, and digital devices (such as GPS and reference systems) appear well equipped to expose users to agreed-upon solutions (such as properly spelled words).

Summary

Spelling support systems can create useful cognitive residuals for their users. Assuming that both systems are comparably successful in providing proper spellings, their ability to support user skill development is not significantly different from one another. Further, the ability to create useful cognitive residuals does not appear to extend beyond what can be achieved by exposing users to correctly spelled words. However, the incidental learning benefits of spelling support systems can still be enhanced by requiring the user to make the necessary edits themselves.

References

Arif, A. S., Sylla, C., & Mazalek, A. (2016, November). Learning new words and spelling with autocorrections. In Proceedings of the 2016 ACM International Conference on Interactive Surfaces and Spaces (pp. 409-414). ACM.

Ayres, P. (2006). Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *20*(3), 287-298.

Baber, C. (2006). Cognitive aspects of tool use. *Applied ergonomics*, *37*(1), 3-15.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

Barr, J., (2018, March 20). Get Better Results with Amazon Mechanical Turk Masters: Amazon Web Services. Retrieved from https://aws.amazon.com/blogs/aws/amazon-mechanical-turk-master-workers/

Bourgeois, J., Farnè, A., & Coello, Y. (2014). Costs and benefits of tool-use on the perception of reachable space. *Acta Psychologica*, *148*, 91-95.

Byrne, R. W. (2004). The manual skills and cognition that lie behind hominid tool use. The evolution of thought: Evolutionary origins of great ape intelligence.

Cardinali, L., Jacobs, S., Brozzoli, C., Frassinetti, F., Roy, A. C., & Farnè, A. (2012). Grab an object with a tool and change your body: tool-use-dependent changes of body representation for action. *Experimental Brain Research*, *218*(2), 259-271.

Chase, B. M., Faith, J. T., Mackay, A., Chevalier, M., Carr, A. S., Boom, A., ... & Reimer, P. J. (2018). Climatic controls on Later Stone Age human adaptation in Africa's southern Cape. *Journal of human evolution*, *114*, 35-44.

Coxe, S., West, S. G., & Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of personality assessment*, *91*(2), 121-136.

D'Angelo, M., di Pellegrino, G., Seriani, S., Gallina, P., & Frassinetti, F. (2018). The sense of agency shapes body schema and peripersonal space. Scientific reports, 8.

Davies, D. (2018, June 22). The 7 Most Popular Search Engines in the World - SEO 101. Retrieved from https://www.searchenginejournal.com/seo-101/meet-search-engines/

Dempsey, S., Lyons, S., & McCoy, S. (2018). Later is better: mobile phone ownership and child academic development, evidence from a longitudinal study. *Economics of Innovation and New Technology*, 1-18.

Dong, G., & Potenza, M. N. (2015). Behavioural and brain responses related to Internet search and memory. *European Journal of Neuroscience*, *42*(8), 2546-2554.

Fox, J., & Weisberg, S., (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Fenech, E. P., Drews, F. A., & Bakdash, J. Z. (2010, September). The effects of acoustic turn-by-turn navigation on wayfinding. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 54, No. 23, pp. 1926-1930). SAGE Publications.

Firth, J., Torous, J., Stubbs, B., Firth, J.A., Steiner, G.Z., Smith, L., Alvarez-Jimenez, M., Gleeson, J., Vancampfort, D., Armitage, C.J. and Sarris, J. (2019). The "online brain": how the Internet may be changing our cognition. *World Psychiatry*, *18*(2), 119-129.

de Wit, M. M., de Vries, S., van der Kamp, J., & Withagen, R. (2017). Affordances and neuroscience: Steps towards a successful marriage. Neuroscience & Biobehavioral Reviews, 80, 622-629.

Hamilton, K. A., & Benjamin, A. S. (2019). The human-machine extended organism: New roles and responsibilities of human cognition in a digital ecology. *Journal of Applied Research in Memory and Cognition*, *8*(1), 40-45.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, *2*, e616.

Henkel, L. A. (2014). Point-and-shoot memories: The influence of taking photos on memory for a museum tour. *Psychological science*, *25*(2), 396-402.

Heersmink, R., & Sutton, J. (2018). Cognition and the Web: Extended, Transactive, or Scaffolded?. *Erkenntnis*, 1-26.

Holmes, V. M., & Malone, N. (2004). Adult spelling strategies. *Reading and Writing*, *17*(6), 537-566.

Hothorn, T., Bretz, F., & Westfall, P., (2008). Simultaneous Inference in General Parametric Models. Biometrical Journal 50(3), 346--363.

Ishikawa, T., Fujiwara, H., Imai, O., & Okabe, A. (2008). Wayfinding with a GPS-based mobile navigation system: A comparison with maps and direct experience. *Journal of Environmental Psychology*, *28*(1), 74-82.

Bartoń, K. (2019). MuMIn: Multi-Model Inference. R package version 1.43.6. https://CRAN.R-project.org/package=MuMIn

Kastner, S., Chen, Q., Jeong, S. K., & Mruczek, R. E. B. (2017). A brief comparative review of primate posterior parietal cortex: a novel hypothesis on the human toolmaker. *Neuropsychologia*, *105*, 123-134.

Kirschner, P. A., & De Bruyckere, P. (2017). The myths of the digital native and the multitasker. *Teaching and Teacher Education*, *67*, 135-142.

Lazic, S. E. (2015). Analytical strategies for the marble burying test: avoiding impossible predictions and invalid p-values. *BMC research notes*, *8*(1), 141.

Lenth, R. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.https://CRAN.R-project.org/package=emmeans

Loh, K. K., & Kanai, R. (2016). How has the Internet reshaped human cognition?. *The Neuroscientist*, *22*(5), 506-520.

Lin, P. H., Liu, T. C., & Paas, F. (2017). Effects of spell checkers on English as a second language students' incidental spelling learning: a cognitive load perspective. *Reading and Writing*, *30*(7), 1501-1525.

Nanospell (2018). JavaScript SpellCheck. Retrieved from https://www.javascriptspellcheck.com/.

Jonassen, D. H. (1995). Computers as cognitive tools: Learning with technology, not from technology. *Journal of Computing in Higher Education*, *6*(2), 40.

Martel, M., Cardinali, L., Roy, A. C., & Farnè, A. (2016). Tool-use: An open window into body representation and its plasticity. *Cognitive neuropsychology*, *33*(1-2), 82-101.

Microsoft (2018). Microsoft Word for Mac (Version 16.19 (181109)) [Computer software]. Microsoft.

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological science*, *25*(6), 1159-1168.

The National Senior Spelling Bee. (2018). Master List - 2018. Retrieved from http://www.nationalseniorspellingbee.com/pdfs/Master_List_2018.pdf

Nakagawa, S., & Schielzeth, H. (2017). Extending R2 and Intra-class Correlation Coefficient from Generalized Linear Mixed-effects Models: Capturing and Characterizing Biological Variation. *BioRxiv beta: the preprint server for Biology http://biorxiv. org/content/early/2017/04/16/095851. article. info*, *10*, 095851.

Noack, R. A. (2012). Solving the "human problem": The frontal feedback model. Consciousness and cognition, 21(2), 1043-1067.

Orban, G. A., & Caruana, F. (2014). The neural basis of human tool use. Frontiers in psychology, 5, 310.

Orban, G. A., Claeys, K., Nelissen, K., Smans, R., Sunaert, S., Todd, J. T., ... & Vanduffel, W. (2006). Mapping the parietal cortex of human and non-human primates. *Neuropsychologia*, *44*(13), 2647-2667.

Oviatt, S. (2006, October). Human-centered design meets cognitive load theory: designing interfaces that help people think. In *Proceedings of the 14th ACM international conference on Multimedia* (pp. 871-880). ACM.

Osiurak, F., Navarro, J., & Reynaud, E. (2018). How our cognition shapes and is shaped by technology: a common framework for understanding human tool-use interactions in the past, present, and future. *Frontiers in psychology*, *9*, 293.

Osiurak, F., & Heinke, D. (2018). Looking for intoolligence: A unified framework for the cognitive study of human tool use and technology. *American Psychologist*, *73*(2), 169.

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive load theory: Instructional implications of the interaction between information structures and cognitive architecture. *Instructional science*, *32*(1), 1-8.

Paas, F., Van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational psychology review*, *22*(2), 115-121.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake:

Explaining the discontinuity between human and nonhuman minds. *BEHAVIORAL AND BRAIN SCIENCES*, *31*, 109-178.

Pezzulo, G., & Cisek, P. (2016). Navigating the affordance landscape: feedback control as a process model of behavior and cognition. *Trends in cognitive sciences*, *20*(6), 414-424.

Power, R. C., & Williams, F. L. E. (2018). Evidence of increasing intensity of food processing during the Upper Paleolithic of Western Eurasia. *Journal of Paleolithic Archaeology*, *1*(4), 281-301.

Plant, K. L., & Stanton, N. A. (2013). The explanatory power of Schema Theory: theoretical foundations and future applications in Ergonomics. *Ergonomics*, *56*(1), 1-15.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosen, L. D., Carrier, L. M., & Cheever, N. A. (2013). Facebook and texting made me do it: Media-induced task-switching while studying. *Computers in Human Behavior*, *29*(3), 948-958.

Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological review*, *115*(1), 101.

Salomon, G. (1990). Cognitive effects with and of computer technology. Communication research, 17(1), 26-44.

Salmon, P. M., Lenné, M. G., Walker, G. H., Stanton, N. A., & Filtness, A. (2014). Exploring schema-driven differences in situation awareness between road users: an on-road study of driver, cyclist and motorcyclist situation awareness. *Ergonomics*, *57*(2), 191-209.

Salomon, G., Perkins, D., & Globerson, T. (1991). Partners in Cognition: Extending Human Intelligence with Intelligent Technologies. *Educational Researcher, 20*(3), 2-9. Retrieved from http://www.jstor.org/stable/1177234

Semaw, S., Rogers, M. J., Quade, J., Renne, P. R., Butler, R. F., Dominguez-Rodrigo, M., ... & Simpson, S. W. (2003). 2.6-Million-year-old stone tools and associated bones from OGS-6 and OGS-7, Gona, Afar, Ethiopia. *Journal of Human Evolution*, *45*(2), 169-177.

Smoker, T. J., Murphy, C. E., & Rockwell, A. K. (2009, October). Comparing memory for handwriting versus typing. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 53, No. 22, pp. 1744-1747). Sage CA: Los Angeles, CA: SAGE Publications.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *science*, *333*(6043), 776-778.

Stout, D., Toth, N., Schick, K., & Chaminade, T. (2008). Neural correlates of Early Stone Age toolmaking: technology, language and cognition in human evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1499), 1939-1949.

van Andel, S., Cole, M. H., & Pepping, G. J. (2017). A systematic review on perceptual-motor calibration to changes in action capabilities. *Human movement science*, *51*, 59-71.

Van Merriënboer, J. J., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. *Educational psychologist*, *38*(1), 5-13.

Wikipedia contributors. (2017, January 4). Commonly misspelled English words. In Wikipedia, The Free Encyclopedia. Retrieved 05:53, January 4, 2017, from https://en.wikipedia.org/w/index.php?title=Commonly_misspelled_English_words&oldid=876715287

Williams-Hatala, E. M., Hatala, K. G., Gordon, M., Key, A., Kasper, M., & Kivell, T. L. (2018). The manual pressures of stone tool behaviors and their implications for the evolution of the human hand. *Journal of human evolution*, *119*, 14-26.

Yefim (2016). autocorrect. Retrieved from https://www.npmjs.com/package/autocorrect

Zuur, A. F., & Ieno, E. N. (2016). A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, *7*(6), 636-645.

Appendix A

A.1 Materials

A.1.1 Stimuli

A.1.1.1 Target Words

Table 10 - Target Words

Misspelling	Proper Spelling	Misspelling	Proper Spelling	Misspelling	Proper Spelling
freind	friend	buisness	business	forseeable	foreseeable
Portugese	Portuguese	bizzare	bizarre	goverment	government
propoganda	propaganda	calender	calendar	remeber	remember
neccessary	necessary	collegue	colleague	gaurd	guard
religous	religious	Carribbean	Caribbean	posession	possession
resistence	resistance	chauffer	chauffeur	occassion	occasion
foriegn	foreign	concious	conscious	agression	aggression
beleive	believe	tendancy	tendency	neaderthal	Neanderthal
assasination	assassination	humourous	humorous	pharoah	Pharaoh

This is a list of the target words and the associated misspellings.

A.1.1.2 Trial 1 Passage

“Hello freind, we welcome Portugese propoganda producer and neccessary politician, Ken. Religous head of resistence by foriegn appointment, I beleive. Publicly his assasination buisness has been seen as bizzare.”

A.1.1.3 Trial 2 Passage

“A calender filled with meetings with a specific collegue in the Carribbean, concious avoidance of his chauffer, and his tendancy for humourous rants gave him away.”

A.1.1.4 Trial 3 Passage

“For the forseeable goverment tenure, remeber that a glamourous gaurd must remain in posession of his wits for this occassion. Agression will be punished. Remember, do not be a neaderthal, be a pharoah.”

A.1.1.5 Transfer Test Passage

"Calender in some goverment resistence," he told the Carribean assasination team. Buisness posession was still the strategy of the Portugese, or so his collegue told him. Their tendancy toward the publicly bizzare was neccessary to not be caught off gaurd. "Stay conscious in the forseeable future, chauffer the propoganda team, and manage my agression" he thought to himself. While humourous, his freind would have to wait. "A pharoah and a glamourous religous leader" was the pass phrase he has to remeber. Beleive it or not this occassion was foriegn to him and his Netherdal brain.

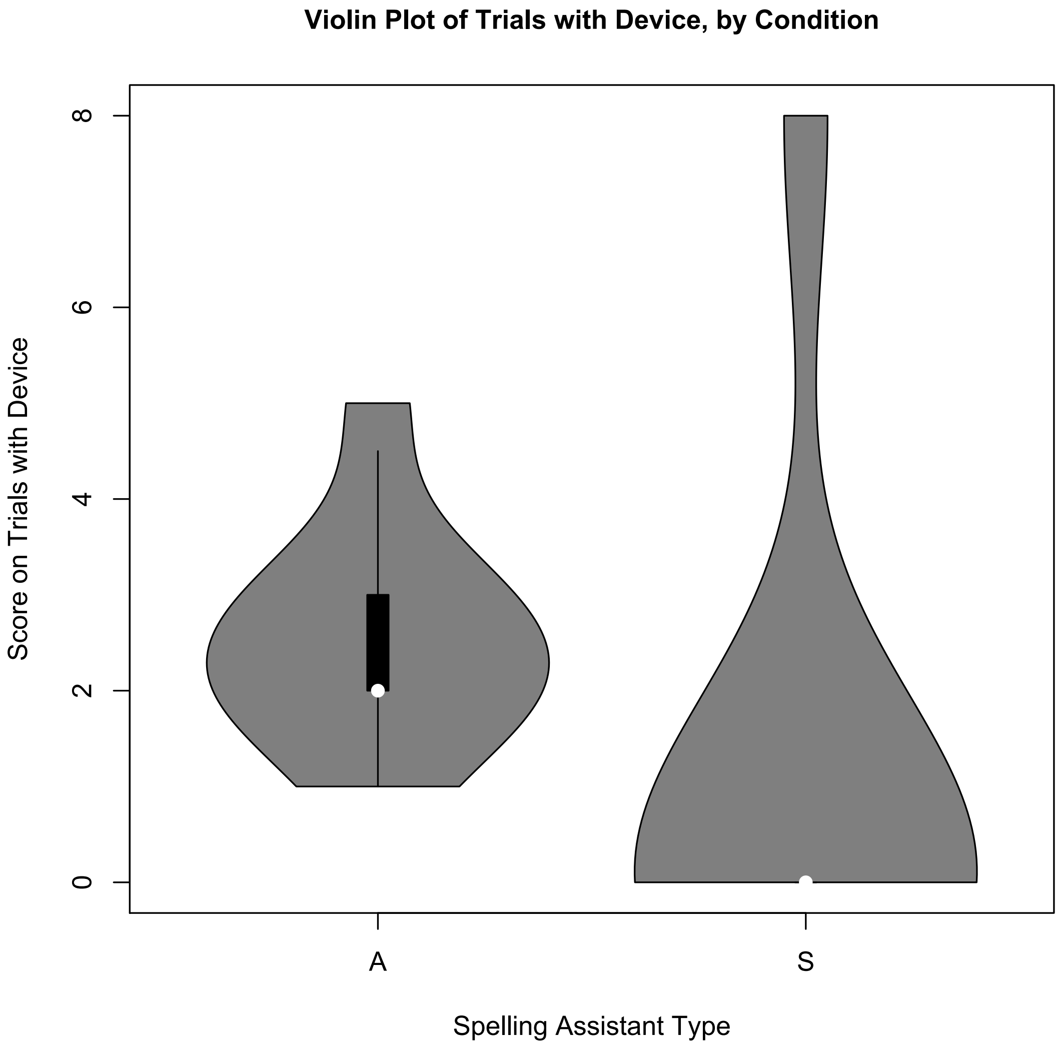
A.2 Model Verification

A.2.1 Data Exploration

A.2.1.1 Distributions

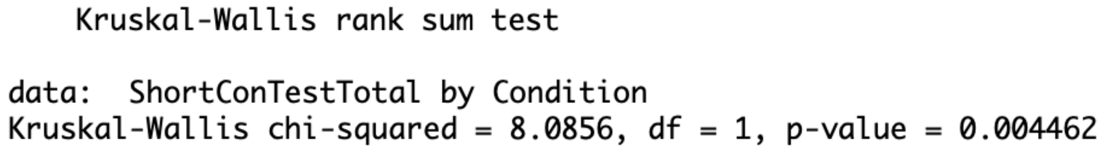
Assessment of Device

Figure 18 - Distribution of Scores (Number of Errors) on Manipulated Trials



This is a violin plot depicting the distribution of errors (score) in the trials where participants were using either Microsoft's spell-checking or autocorrect support system to correct errors in the passage. The width of each 'violin' represents the number of observations at that level. A wider point in the violin corresponds to more observations, while a narrower point in the violin corresponds to fewer observations. Participants in the autocorrect condition (A) never achieved perfect performance (a score of zero), unlike the majority of the participants (seven *of ten) in the spell-checking condition (S). Spell-checking had the highest single score of four errors (the passages contained 27 spelling errors to begin with).*

Figure 19 - Results from the Kruskal-Wallis Rank Sum Test

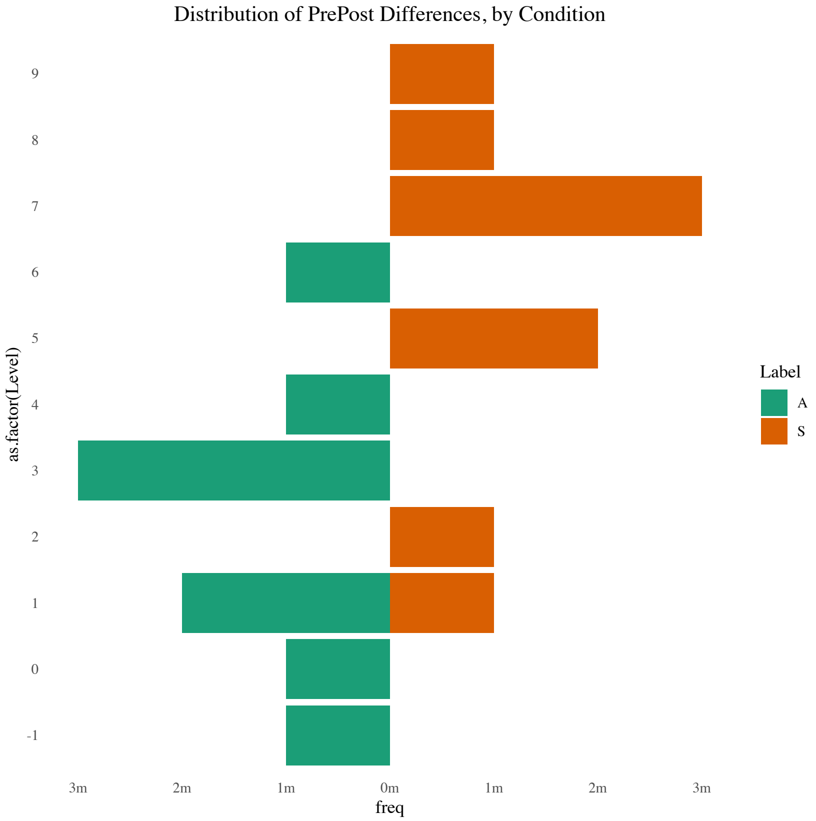


```
	Kruskal-Wallis rank sum test

data:  ShortConTestTotal by Condition
Kruskal-Wallis chi-squared = 8.0856, df = 1, p-value = 0.004462
```

Results indicate that participants in the autocorrect condition, during the trials where they used a spelling support system, had a significantly different distribution of errors from the participants in the spell-checking condition. Overall, participants that were using Microsoft's autocorrect corrected fewer errors than the participants using Microsoft's spell-checking feature.

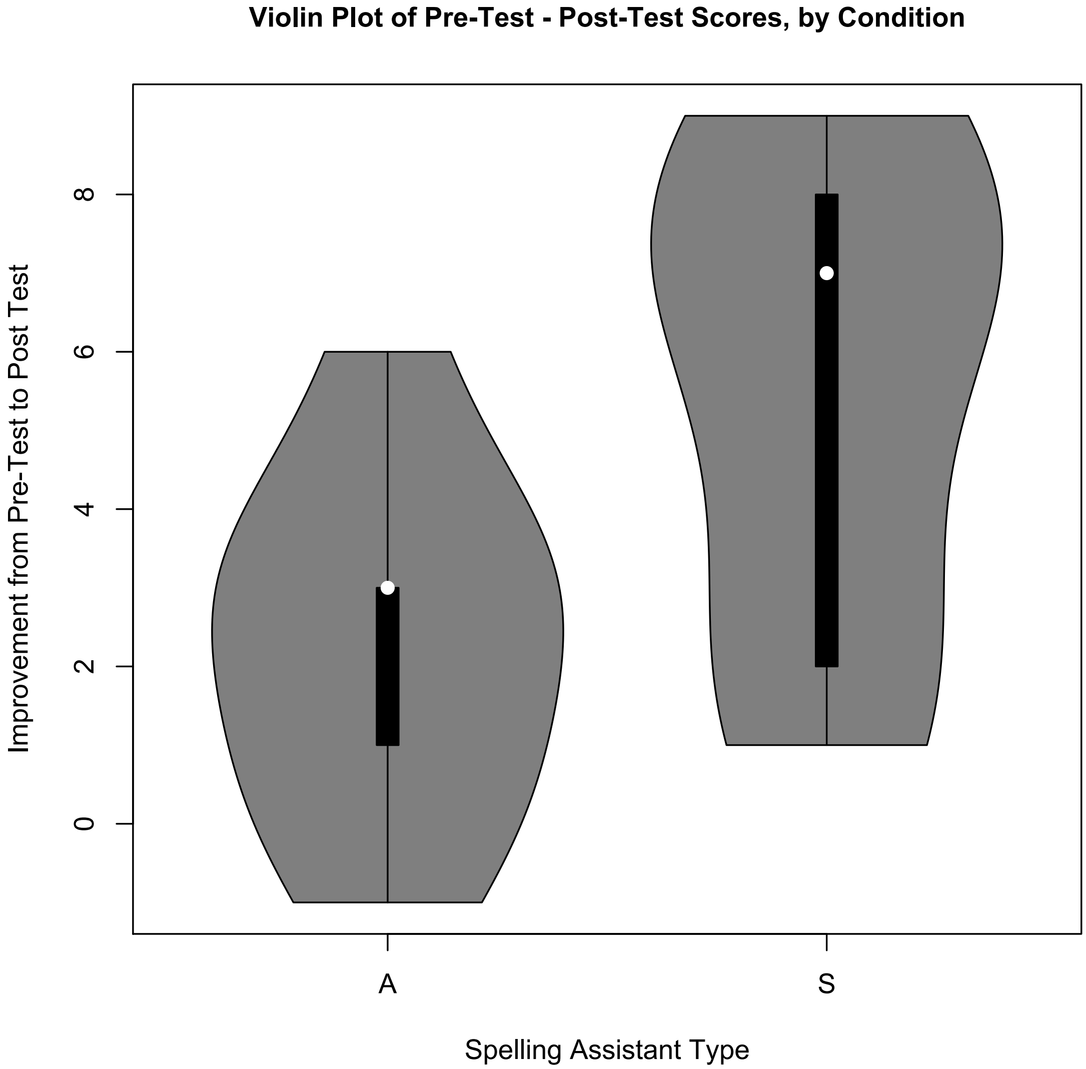
Assessment of Residuals

Figure 20 - Population chart of Pre-Post Score Differences



Here are the distributions of pre-post differences in scores, grouped by condition ("A": Autocorrect; "S": Spell-check). Differences in scores were overall higher for participants in *the spell-checking condition, with the modal difference in score being close to 7 fewer errors on the post-test than the pre-test.*

Figure 21 - Violin Plot of Pre-Test



Post-Test Scores, by Condition – This diagram depicts the difference between pre-test and post-test scores for each condition ("A": Autocorrect; "S": Spell-check). The top performers in the spell-checking condition outperformed the top performers in the autocorrect condition. Additionally, the lowest performers in the spell-checking condition also outperformed the lowest performers in the autocorrect condition.

Figure 22 - Pre-Test vs Post-Test Score, by Condition

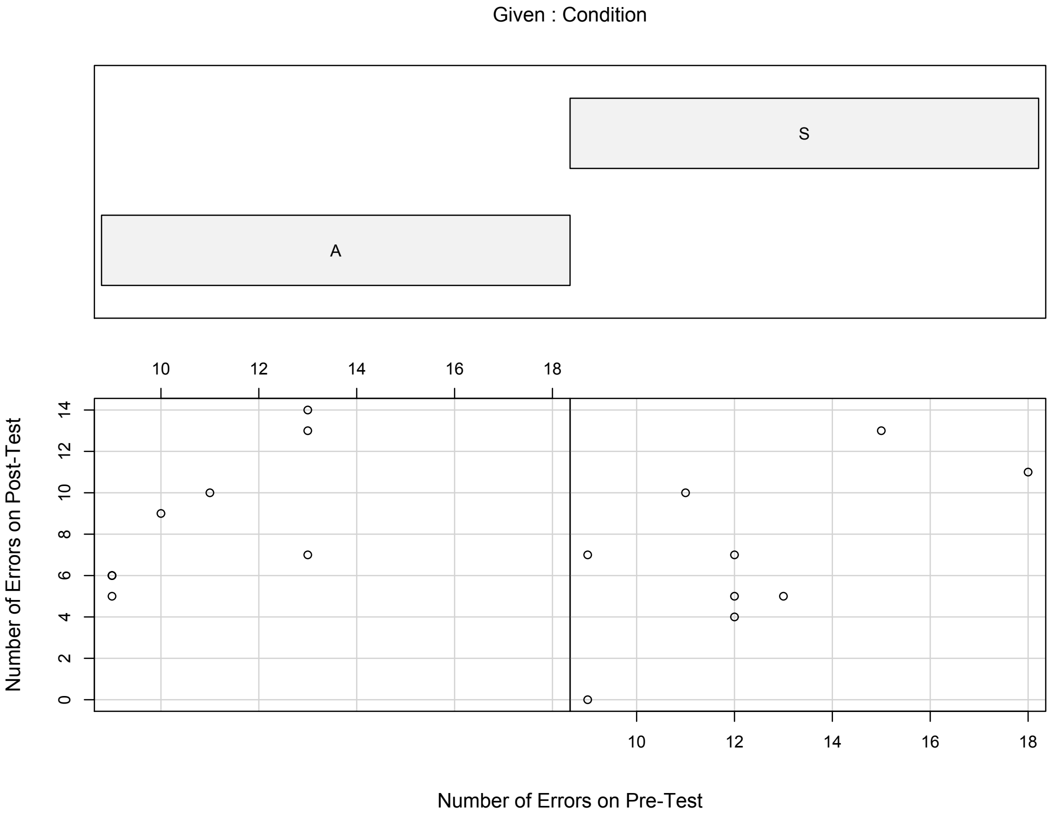
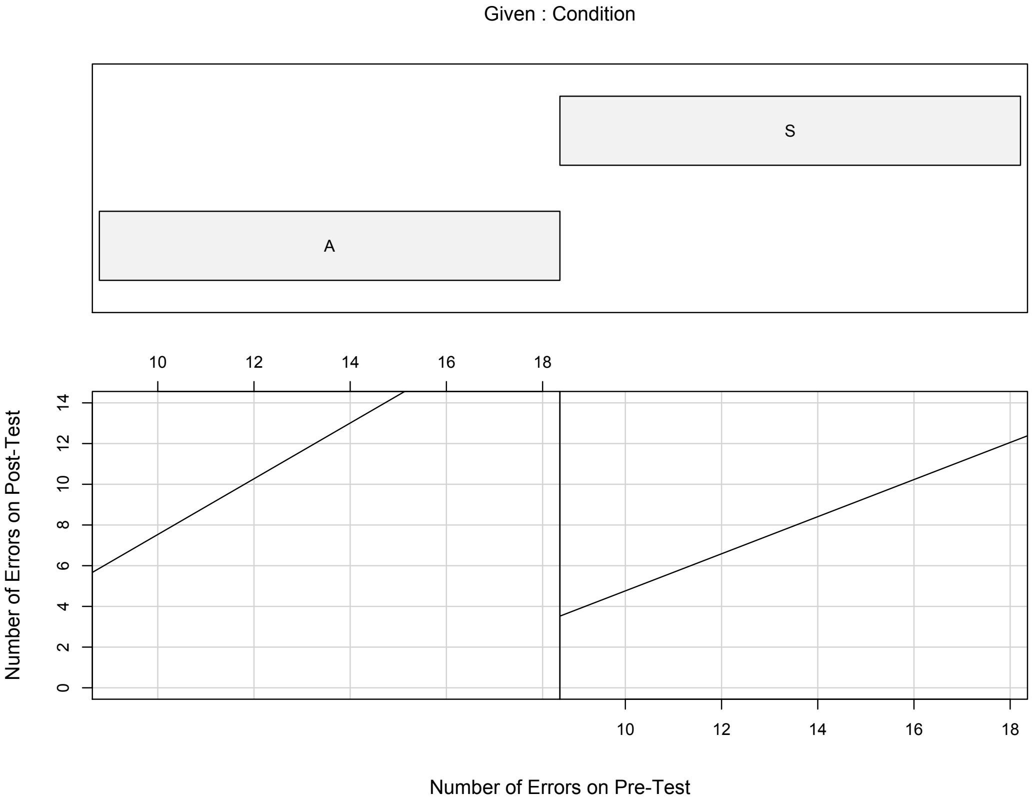
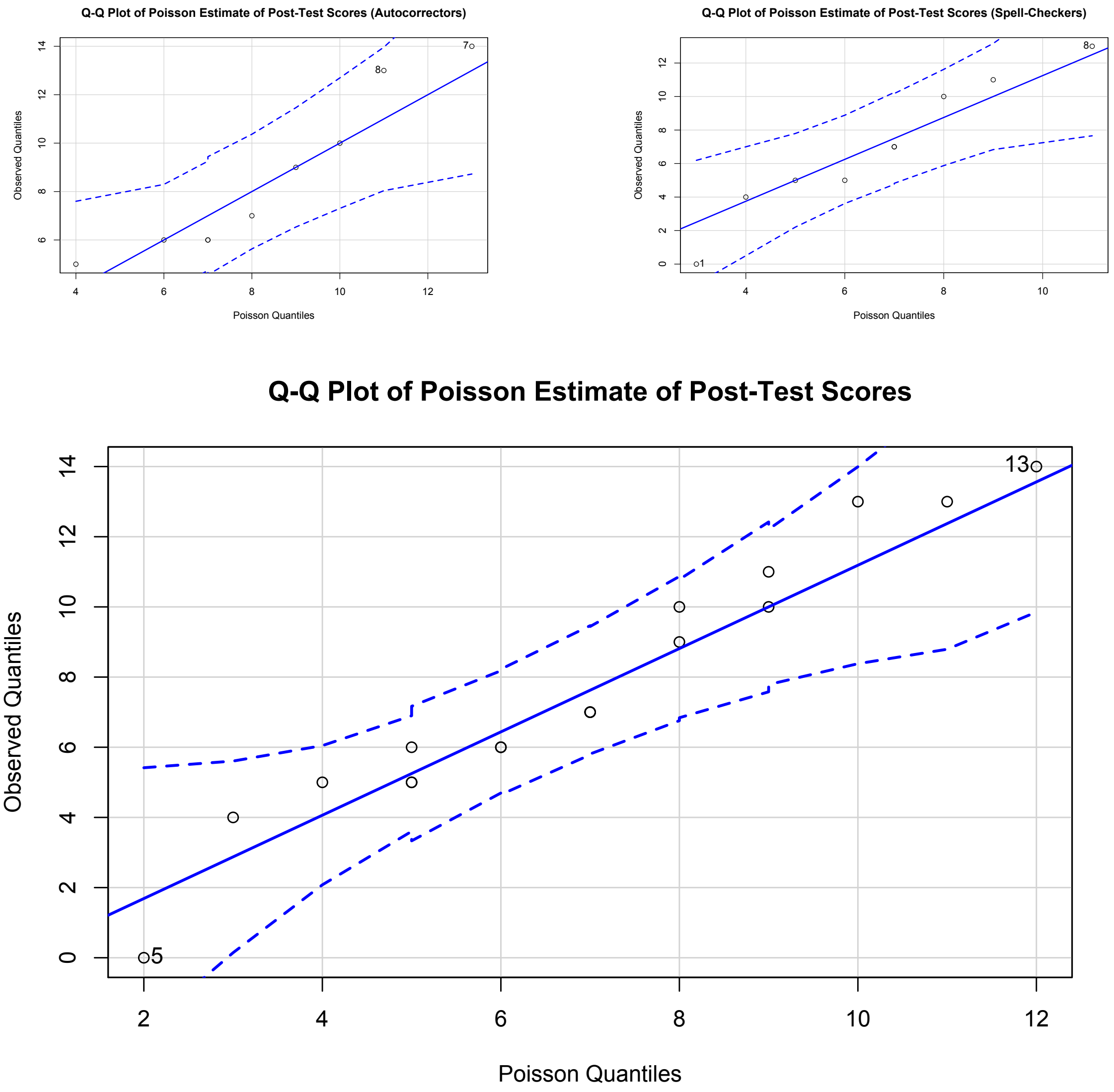


Figure 23 - Pre-Test vs Post-Test Score, by Condition (With Guiding Line)



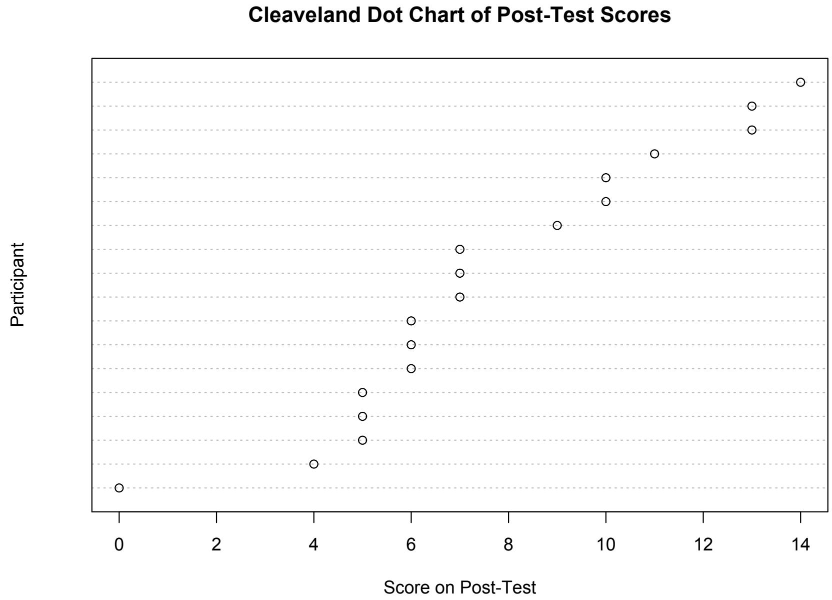
A.2.1.2 Poissonality

Figure 24- Measure of Poissonality of Response Variable



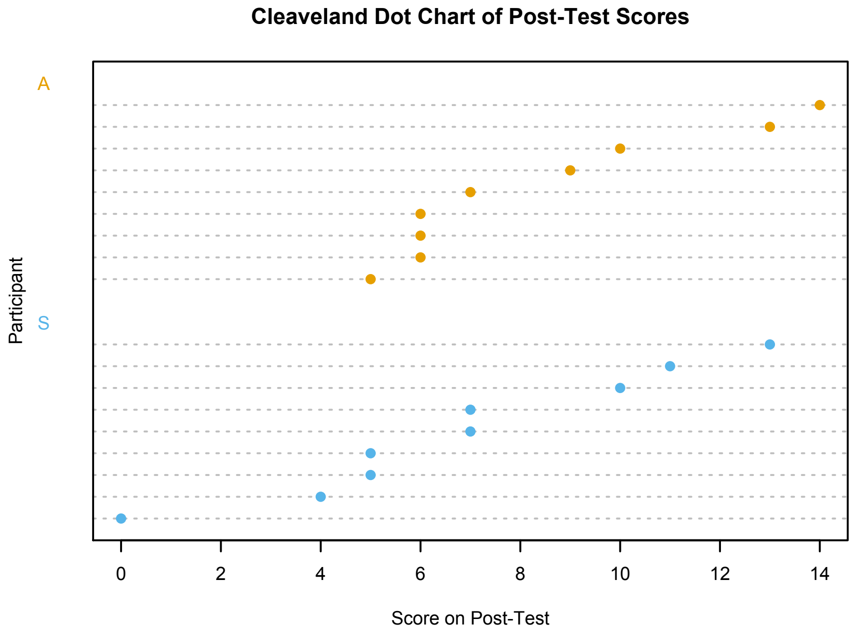
A.2.1.3 Outliers

Figure 25 - Cleaveland Dot Chart of Scores on Post-Test



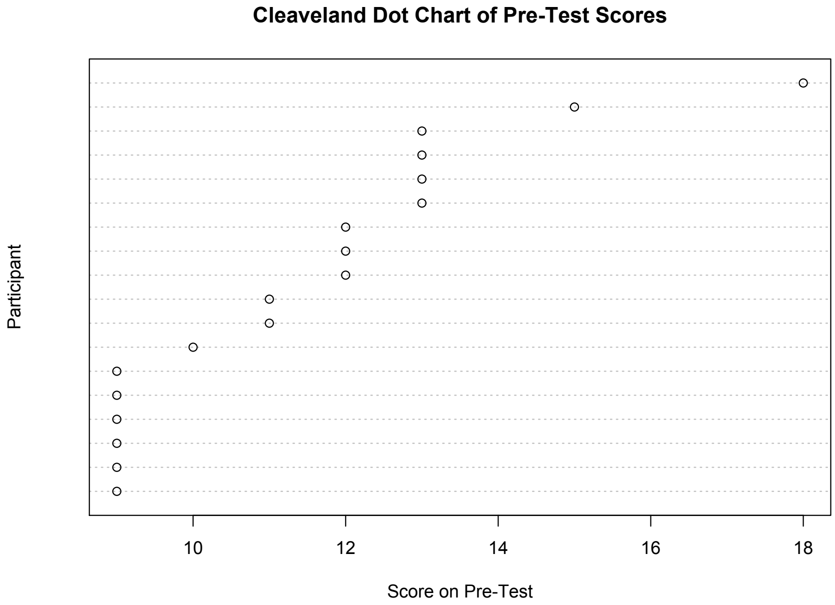
A plot of the number of errors participants made on the post-test. Each line represents a participant, and the dot on each line represents their score on the post-test. Plot indicates that most of the participants scored four our higher, with only one participant achieving zero spelling errors on the post-test.

Figure 26 - Cleaveland Dot Chart of Scores on Post-Test, Grouped by Condition



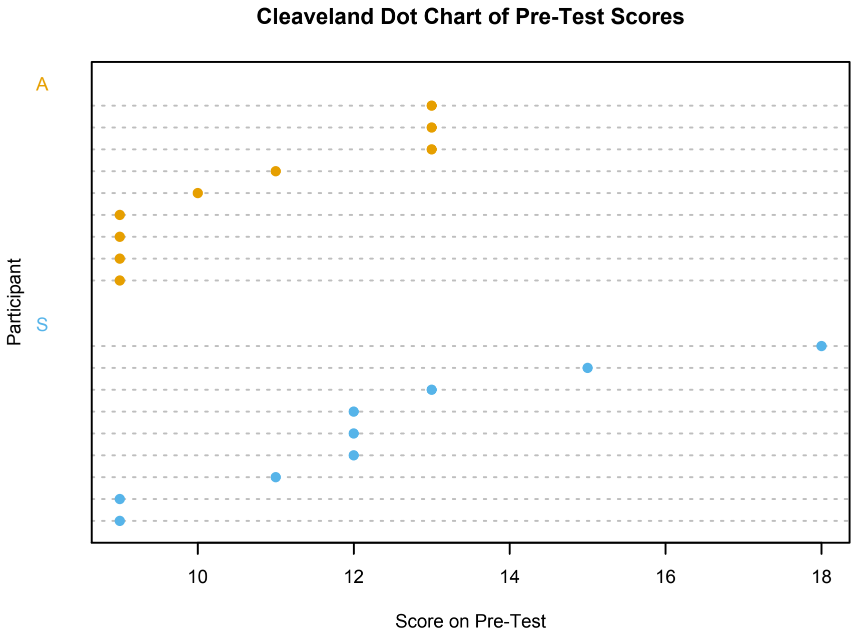
This presents the same information as above, but with points in gold, labelled "A" for "Autocorrect", representing participants in the "autocorrect" condition and points in blue, labelled "S" for "Spell-checker". Performance between the two conditions appears similar, with the participant at 0 still appearing to be the most distant from the group.

Figure 27 -Cleaveland Dot Chart of Pre-Test Scores



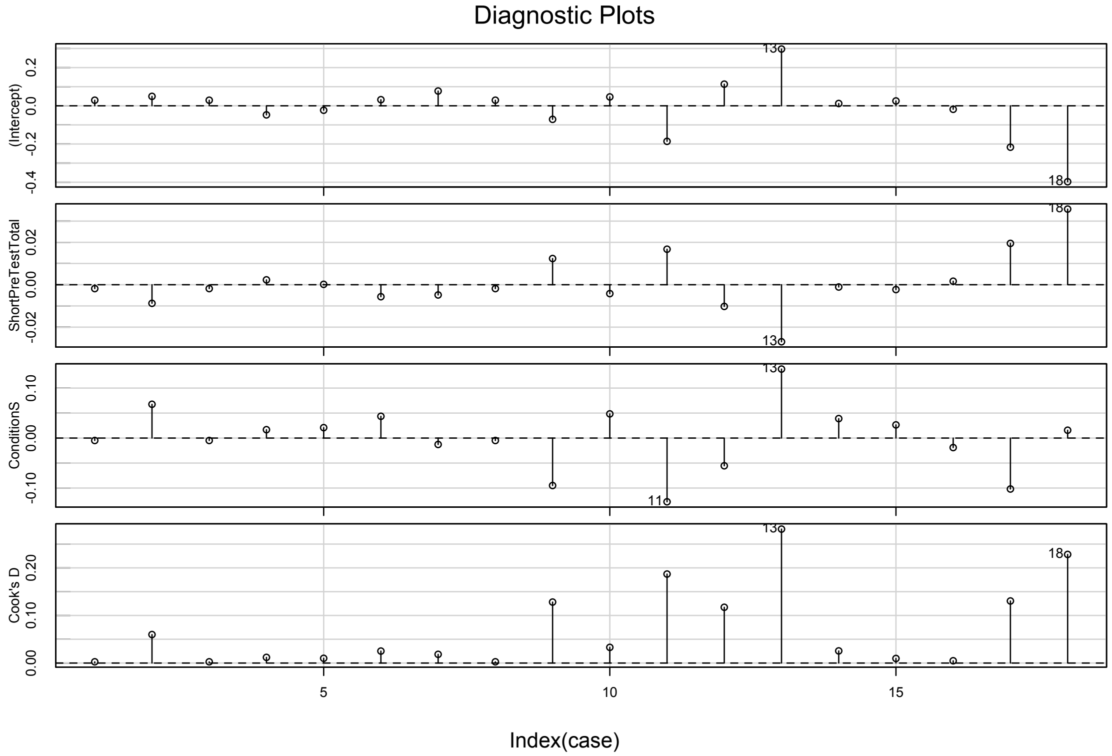
Each line on the chart represents an individual participant, and the dot represents the number of spelling errors in their pre-test.

Figure 28 - Cleaveland Dot Chart of Pre-Test Scores, by Condition



Each line on the chart represents an individual participant, and the dot represents the number of spelling errors in their pre-test. Participants are grouped by Condition, where "A" are participants in the "Autocorrect" condition and "S" in the "Spell-Checking" condition.

Figure 29 - Influential Observation Diagnostic Plot



Produced using the 'infIndexPlot' function of the R package 'car' (Fox & Weisberg, 2011). Influence diagnostics for generalized linear mixed models were calculated with the 'influence.lme4' function of the 'lme4' package for R (Bates, Maechler, Bolker, & Walker, 2015). A large number of observations, namely observations 9,11,12,13, 17 and 18. Follow-up analysis included looking at how estimated parameters of the change as a result the removal of these observations.

Table 11 - Parameter Estimations with Influential Observations Removed

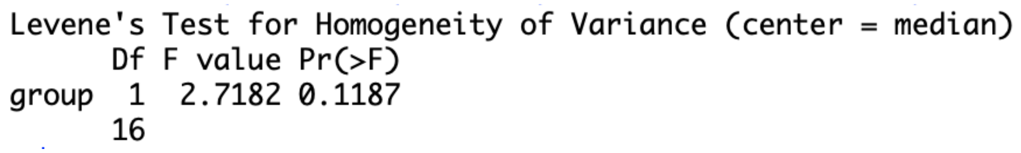
Observation Index	***Intercept***	***Pre-Test***	***Condition***
9	*0.6366*	*0.1435*	*-0.5559*
11	*0.5214*	*0.1479*	*-0.5884*
12	*0.8207*	*0.1209*	*-0.5164*
13	*1.0045*	*0.1042*	*-0.3229*
17	*0.4907*	*0.1506*	*-0.5629*
18	*0.3098*	*0.1668*	*-0.4452*
Full-Model	*0.7075*	*0.1311*	*-0.4609*

Each Observation Index corresponds to a given participant that was observed with a Cook's D above 0.2 (see above). Model parameters are re-estimated for the model via the 'influence.lme4' function of the 'lme4' package for R (Bates et al., 2015). Estimates for the intercept varied the most across all observation removals, and the influence of pre-test varied the least. Estimates for condition remain negative, indicating that estimates for participant errors in the post-test are consistently lower for participants that were in the spell-checking condition.

A.2.2 Model Validation

A.2.2.1 Heterogeneity

Figure 30-Output from Levene Test of Homogeneity of Variance on Pearson residuals, by Condition

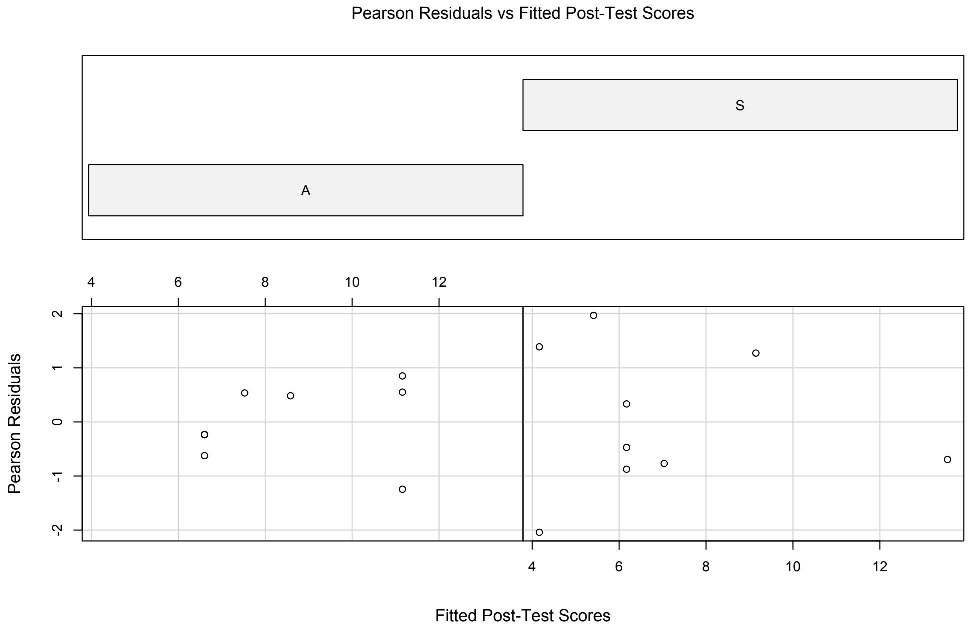


```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  2.7182 0.1187
      16
```

Results indicate that observed heterogeneity does not exceed the assumption of homogenous variance between conditions.

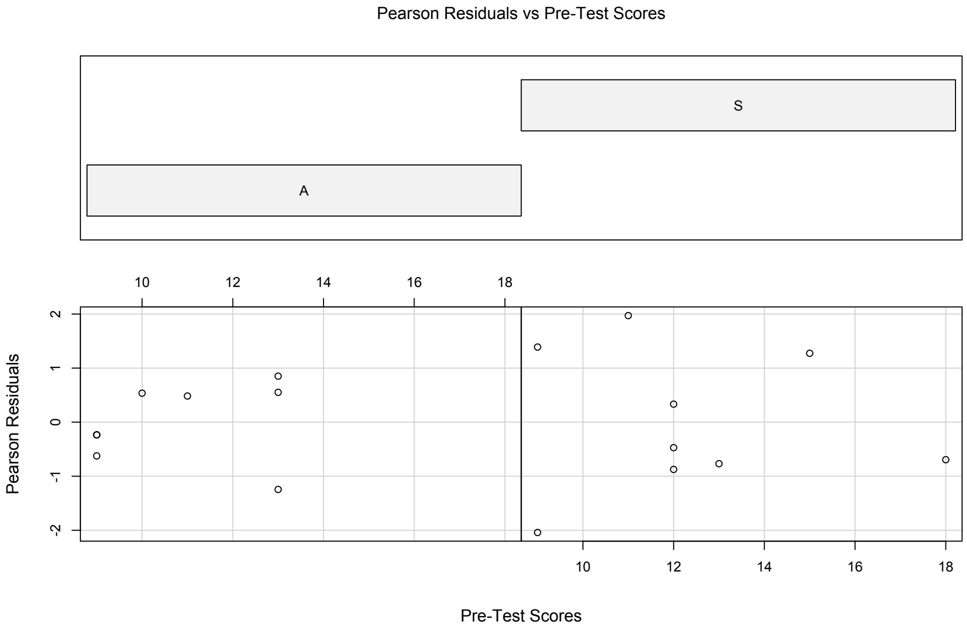
A.2.2.2 Residual Plots

Figure 31 - Pearson Residuals vs Fitted Post-Test Scores



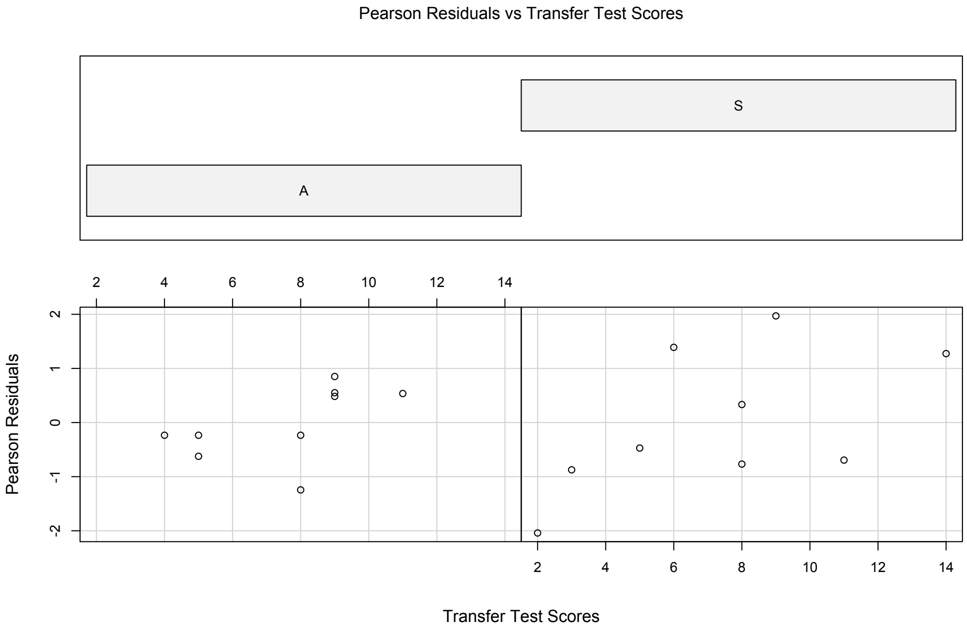
Pearson residuals were calculated and plotted against the model-predicted post-test scores. Distribution appears centered around 0 in both the auto-correct and spell-checking condition. Heterogeneity across Condition is further explored in section B.3.3.2

Figure 32 - Pearson Residuals vs Pre-Test Scores



Residuals appear evenly distributed around zero. No issues were noted.

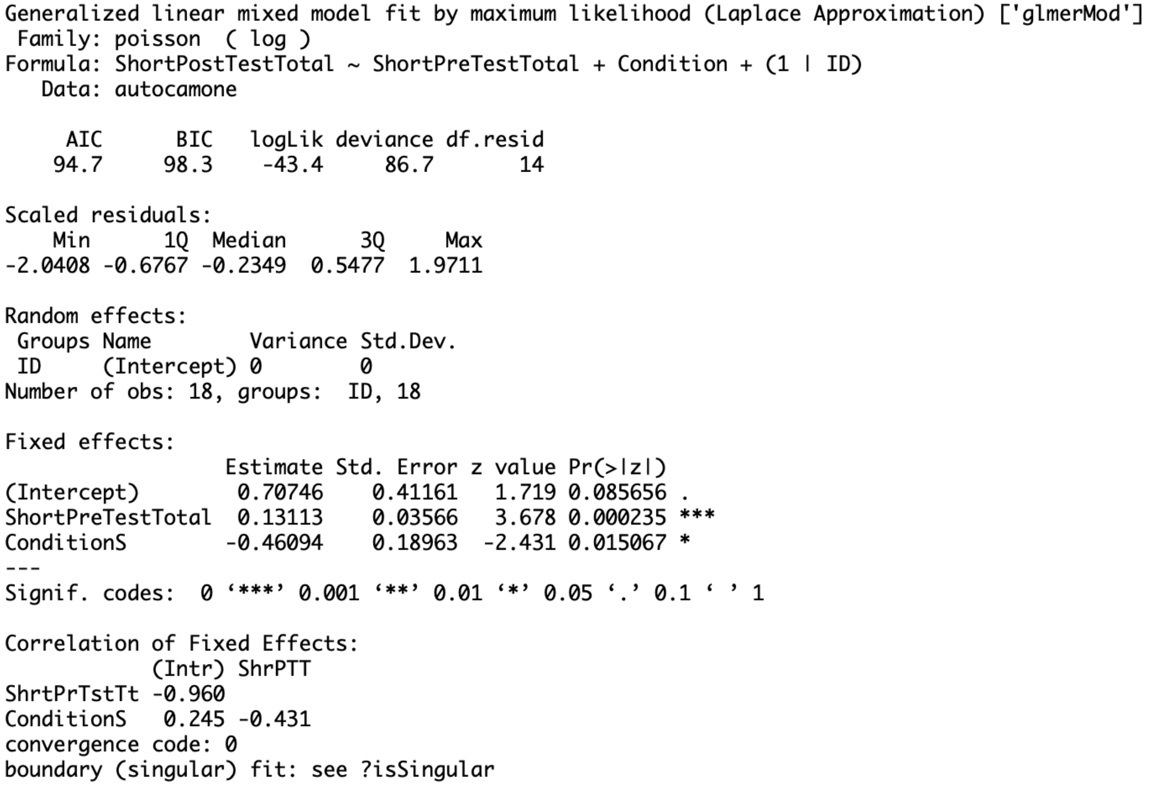
Figure 33 -Pearson Residuals vs Transfer Test Scores



To determine if the data is overfit to the response variable, a non-model covariate can be plotted against residuals. These residuals should be evenly distributed around 0. Transfer Test was a covariate of Post-Test score that was not a term included in the model. Residuals appear evenly distributed around 0, indicating no issue.

A.2.3 Model Fit

Figure 34 - Results of GLMM



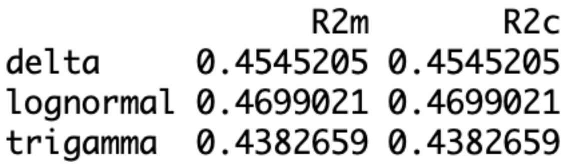
```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: ShortPostTestTotal ~ ShortPreTestTotal + Condition + (1 | ID)
   Data: autocamone

     AIC      BIC   logLik deviance df.resid
    94.7     98.3    -43.4     86.7       14

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0408 -0.6767 -0.2349  0.5477  1.9711

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 0        0
Number of obs: 18, groups:  ID, 18

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.70746    0.41161   1.719 0.085656 .
ShortPreTestTotal  0.13113    0.03566   3.678 0.000235 ***
ConditionS        -0.46094    0.18963  -2.431 0.015067 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) ShrPTT
ShrtPrTstTt -0.960
ConditionS   0.245 -0.431
convergence code: 0
boundary (singular) fit: see ?isSingular
```

Results of a generalized linear mixed model fit with the 'lme4' package in 'R' (Bates et al., 2015)., using the log-link function in the poisson family. An OLRE was included to capture overdispersion frequently found in count data. The fit was determined to be singular and the estimate for the variance and standard deviation of the OLRE were both estimated to be zero. To determine if this was detrimental to the estimate of the fixed effect co-efficients, a GLM was also fit without the OLRE. Results without the ORLE are described below.

Table 12 - Table of Generalized Poisson Mixed Mode Regression Coefficients

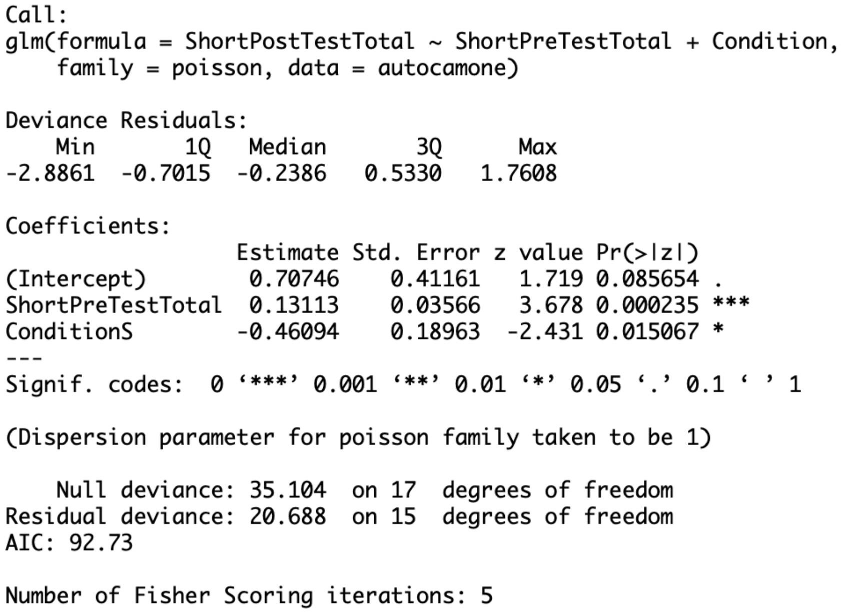
	Estimate	**Std. Error**	**z value**	***P*-value**
Intercept	0.70746	0.42261	1.719	0.085656
Pre-Test Score	0.13113	0.03566	3.678	0.000235*
Spell-Checking	-0.46094	0.18963	-2.431	0.015067*

Figure 35 - R^2 of GLMM



```
                R2m       R2c
delta     0.4545205 0.4545205
lognormal 0.4699021 0.4699021
trigamma  0.4382659 0.4382659
```

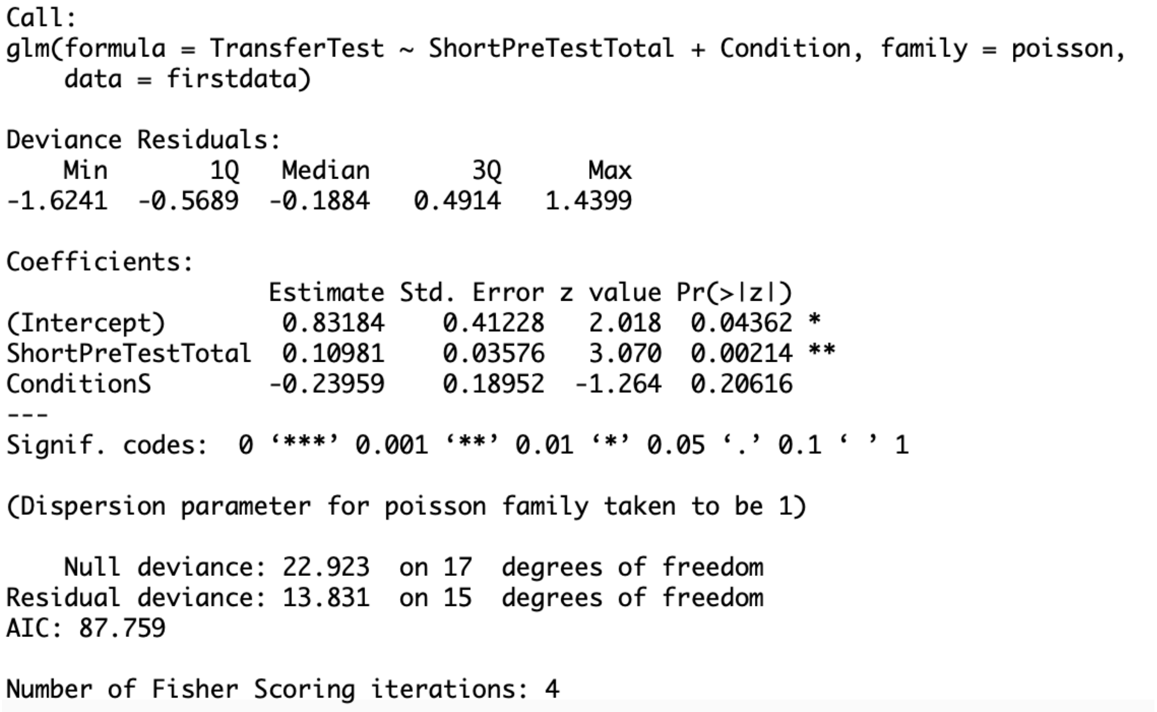
Figure 36 - Result of GLM



```
Call:
glm(formula = ShortPostTestTotal ~ ShortPreTestTotal + Condition,
    family = poisson, data = autocamone)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8861  -0.7015  -0.2386   0.5330   1.7608

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.70746    0.41161   1.719 0.085654 .
ShortPreTestTotal  0.13113    0.03566   3.678 0.000235 ***
ConditionS        -0.46094    0.18963  -2.431 0.015067 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 35.104  on 17  degrees of freedom
Residual deviance: 20.688  on 15  degrees of freedom
AIC: 92.73

Number of Fisher Scoring iterations: 5
```

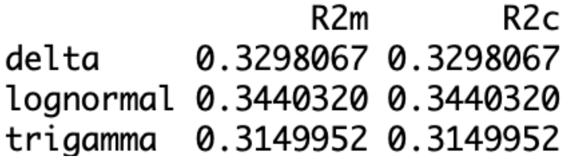
Removal of the OLRE resolved the singularity. Estimates for the fixed-effect co-efficients remained the same after removal of the OLRE.

Figure 37 - GLM for Transfer Test



```
Call:
glm(formula = TransferTest ~ ShortPreTestTotal + Condition, family = poisson,
    data = firstdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6241  -0.5689  -0.1884   0.4914   1.4399

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.83184    0.41228   2.018  0.04362 *
ShortPreTestTotal  0.10981    0.03576   3.070  0.00214 **
ConditionS        -0.23959    0.18952  -1.264  0.20616
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 22.923  on 17  degrees of freedom
Residual deviance: 13.831  on 15  degrees of freedom
AIC: 87.759

Number of Fisher Scoring iterations: 4
```

Figure 38 – R^2 for Transfer Test Model



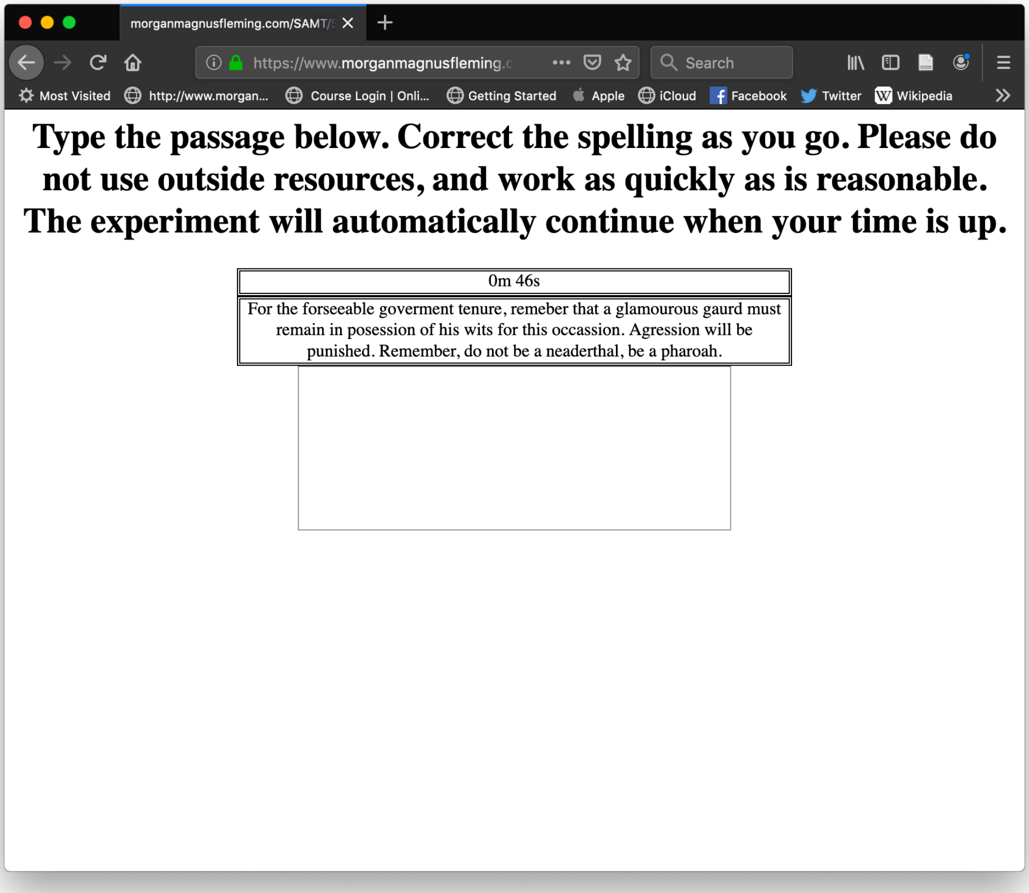
```
                R2m       R2c
delta     0.3298067 0.3298067
lognormal 0.3440320 0.3440320
trigamma  0.3149952 0.3149952
```

Appendix B

B.1 Materials

B.1.1 Website and Support Systems

Figure 39 - Screenshot of Website Used to Gather Data



Type the passage below. Correct the spelling as you go. Please do not use outside resources, and work as quickly as is reasonable. The experiment will automatically continue when your time is up.

0m 46s

For the forseeable goverment tenure, remeber that a glamourous gaurd must remain in posession of his wits for this occassion. Agression will be punished. Remember, do not be a neaderthal, be a pharoah.

B.1.2 Stimuli

B.1.2.1 Target Words

Table 13 - Target Words

Misspelling	Proper Spelling	Misspelling	Proper Spelling	Misspelling	Proper Spelling
freind	friend	buisness	business	forseeable	foreseeable
Portugese	Portuguese	bizzare	bizarre	goverment	government
propoganda	propaganda	calender	calendar	remeber	remember
neccessary	necessary	collegue	colleague	gaurd	guard
religous	religious	Carribbean	Caribbean	posession	possession
resistence	resistance	chauffer	chauffeur	occassion	occasion
foriegn	foreign	concious	conscious	agression	aggression
beleive	believe	tendancy	tendency	neaderthal	Neanderthal
assasination	assassination	humourous	Humorous	pharoah	Pharaoh

These are the target words and their associated misspellings used in Study 2. They are the same words as those used in Study 1.

B.1.2.2 Passage 1

"Hello freind, we welcome Portugese propoganda producer and neccessary politician, Ken. Religous head of resistence by foriegn appointment, I beleive. Publicly his assasination buisness has been seen as bizzare."

B.1.2.3 Passage 2

"A calender filled with meetings with a specific collegue in the Carribbean, concious avoidance of his chauffer, and his tendancy for humourous rants gave him away."

B.1.2.4 Passage 3

"For the forseeable goverment tenure, remeber that a glamourous gaurd must remain in posession of his wits for this occassion. Agression will be punished. Remember, do not be a neaderthal, be a pharoah."

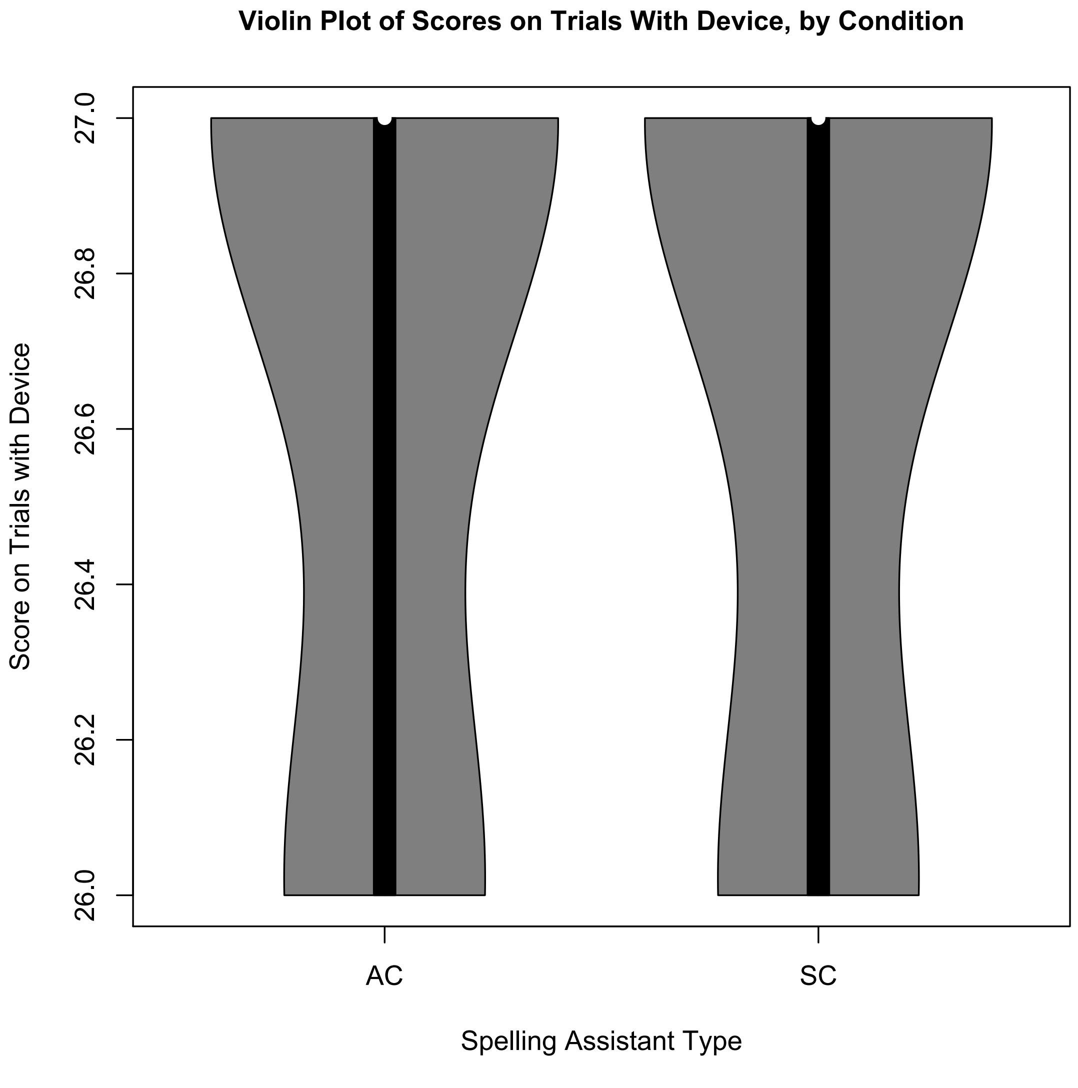
B.2 Model Verification

B.2.1 Data Exploration

B.2.1.1 Distributions

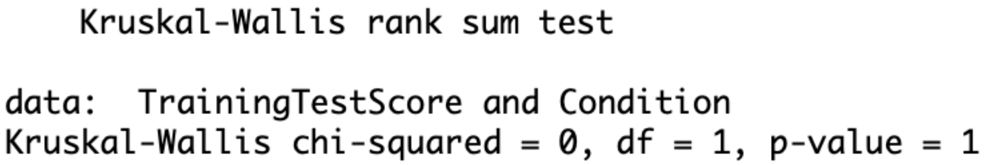
Assessment of Device

Figure 40- Violin Plot of Scores on Trials with Device, by Condition



The distribution of scores on the trials with a spelling assistant were identical. For each group exactly four participants scored 26, and the other seven scored 27.

Figure 41 - Kruskal-Wallis Rank Sum Test Between Conditions



```
        Kruskal-Wallis rank sum test

data:  TrainingTestScore and Condition
Kruskal-Wallis chi-squared = 0, df = 1, p-value = 1
```

Score distributions on these trials were identical. For each group, exactly four people correctly spelled 26 words while exactly seven people spelled 27 words correctly.

Figure 42 - Score on Pre-Test vs Score on Post Test, Grouped by Condition

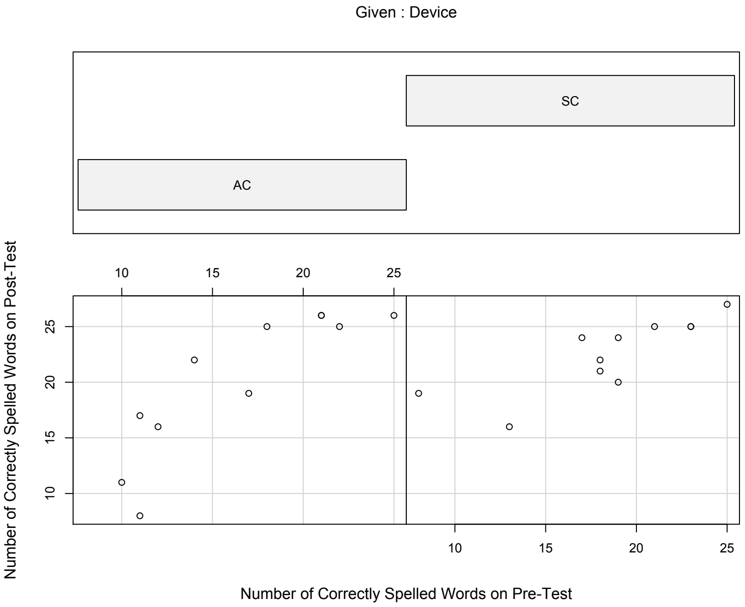
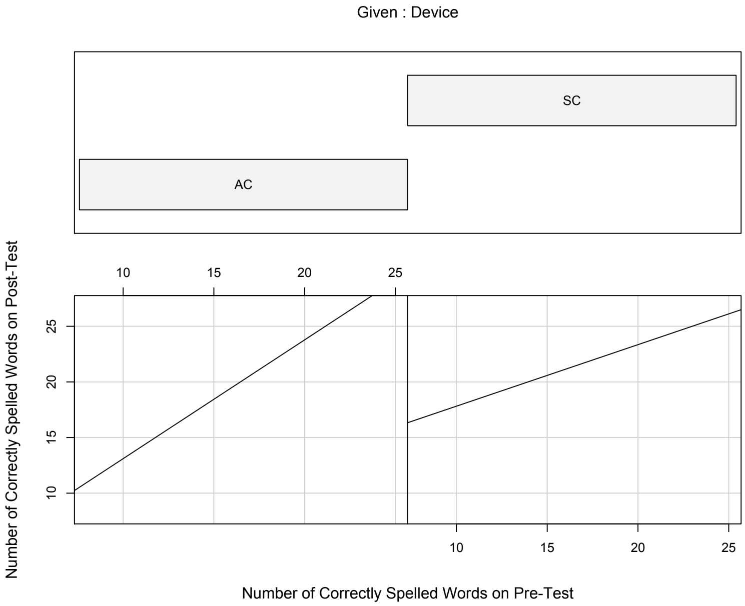
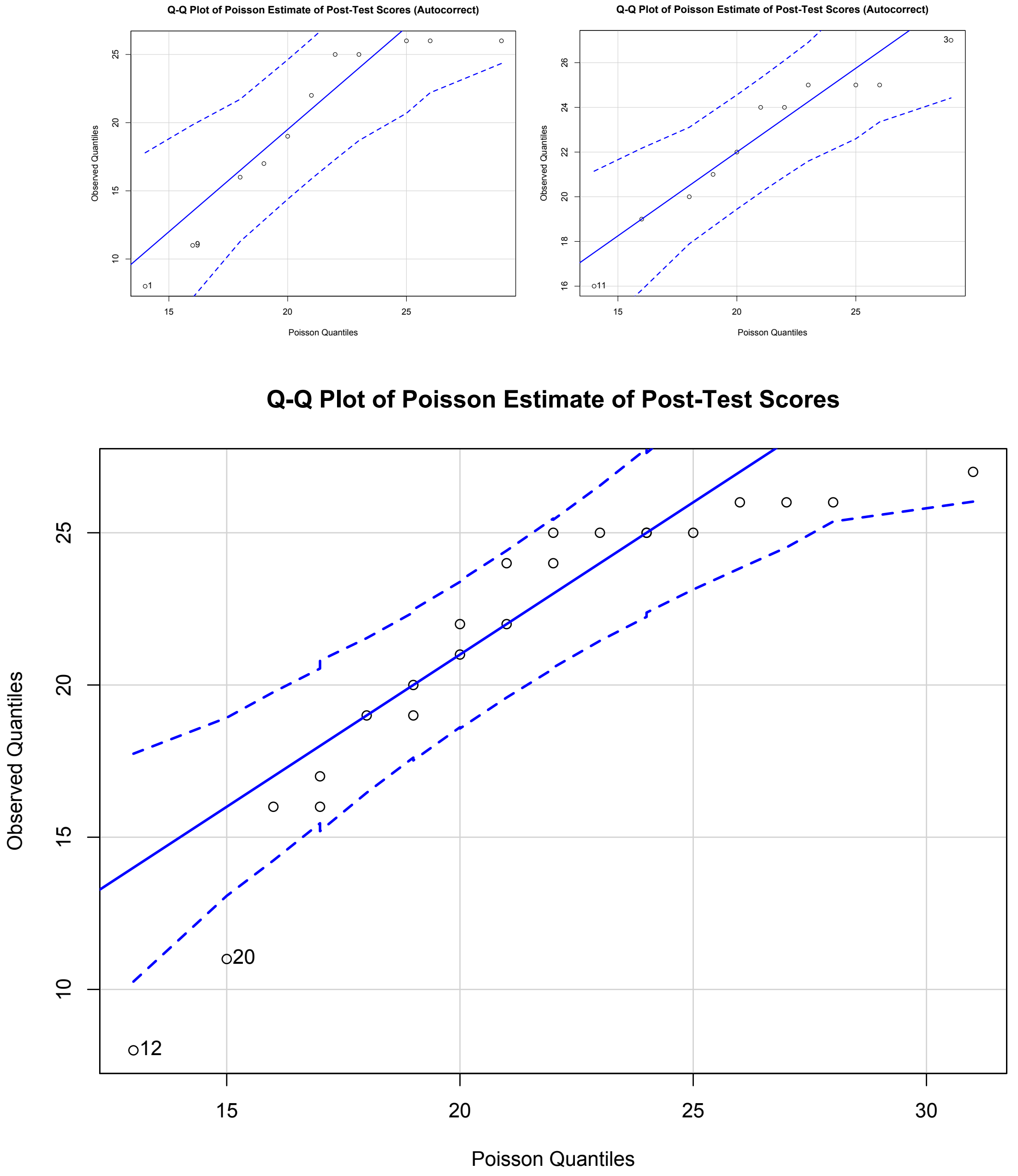


Figure 43 - Score on Pre-Test vs Score on Post-Test, Grouped by Condition



B.2.1.2 Poissonality

Figure 44 - Q-Q Plot of Estimated Poisson Quantiles and Observed Scores on Post-Test



B.2.1.3 Outliers

Figure 45 - Cleaveland Plot of Pre-Test Scores

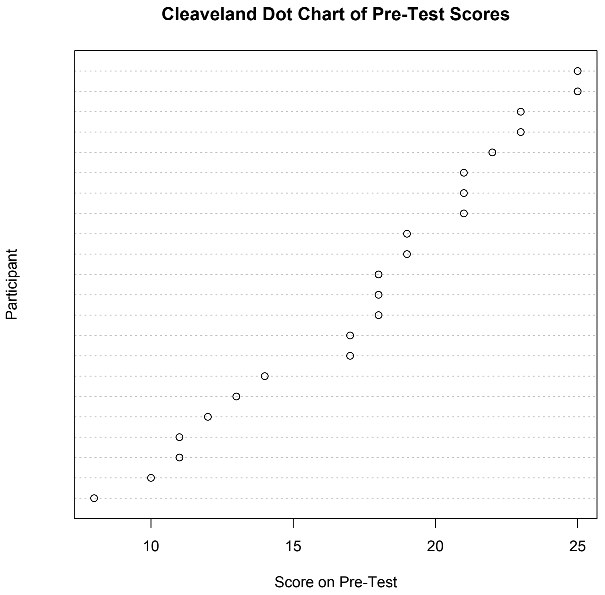


Figure 46 - Cleaveland Plot of Post-Test Scores

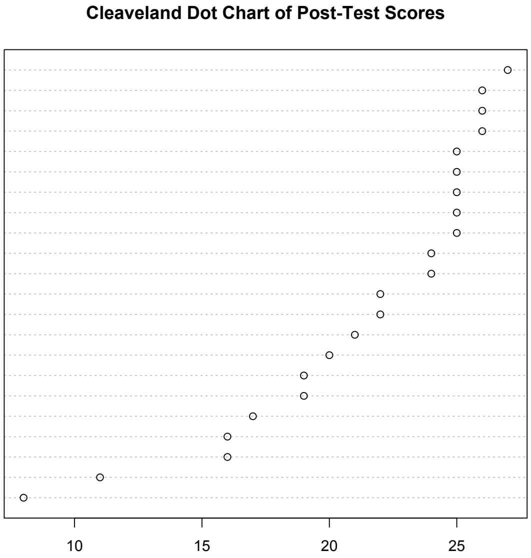


Figure 47 - Cleaveland Plot of Pre-Test Scores, Grouped by Condition

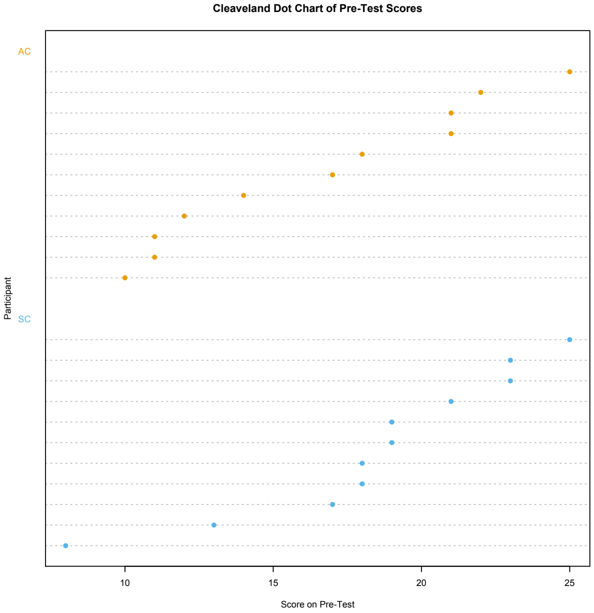


Figure 48 - Cleaveland Plot of Post-Test Scores, Grouped by Condition

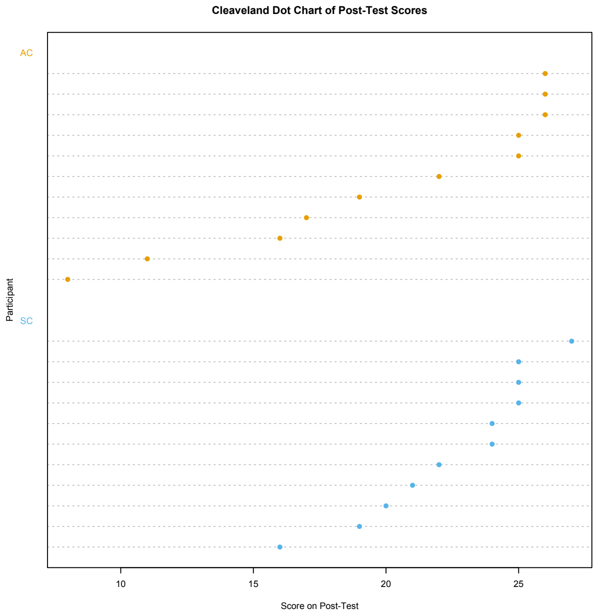


Figure 49 - Outlier Diagnostic Charts for Poisson Model with OLRE

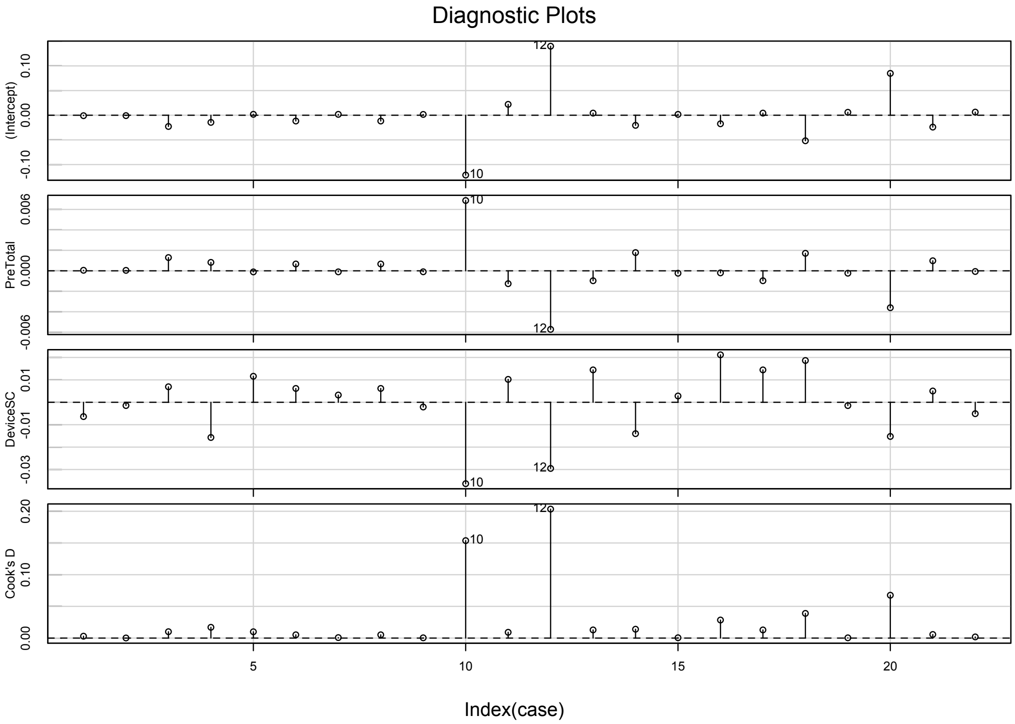


Table 14- Table of Regression Coefficients After Outlier Removal

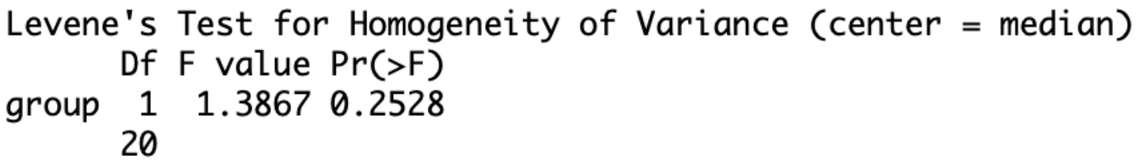
Observation Index	***Intercept***	***Pre-Test***	***Condition***
10	*2.1869*	*0.0474*	*0.0019*
12	*2.4482*	*0.0349*	*0.0087*
20	*2.3931*	*0.0370*	*0.0230*
Full-Model	*2.3083*	*0.0406*	*0.0383*

This table summarizes the new regression coefficients when the corresponding outlier is removed. Observations 10 and 12 appear to have the largest impact on the non-significant differences observed between conditions.

B.2.2 Model Validation

B.2.2.1 Heterogeneity

Figure 50- Levene's Test of Homogeneity of Variance for Poisson Regression with OLRE



```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  1.3867 0.2528
      20
```

Results indicate no violation of the assumption of homogeneity for the model.

B.2.2.2 Residual Plots

Figure 51 - Pearson Residuals from Poisson Regression with OLRE vs Observed Score on Pre-Test

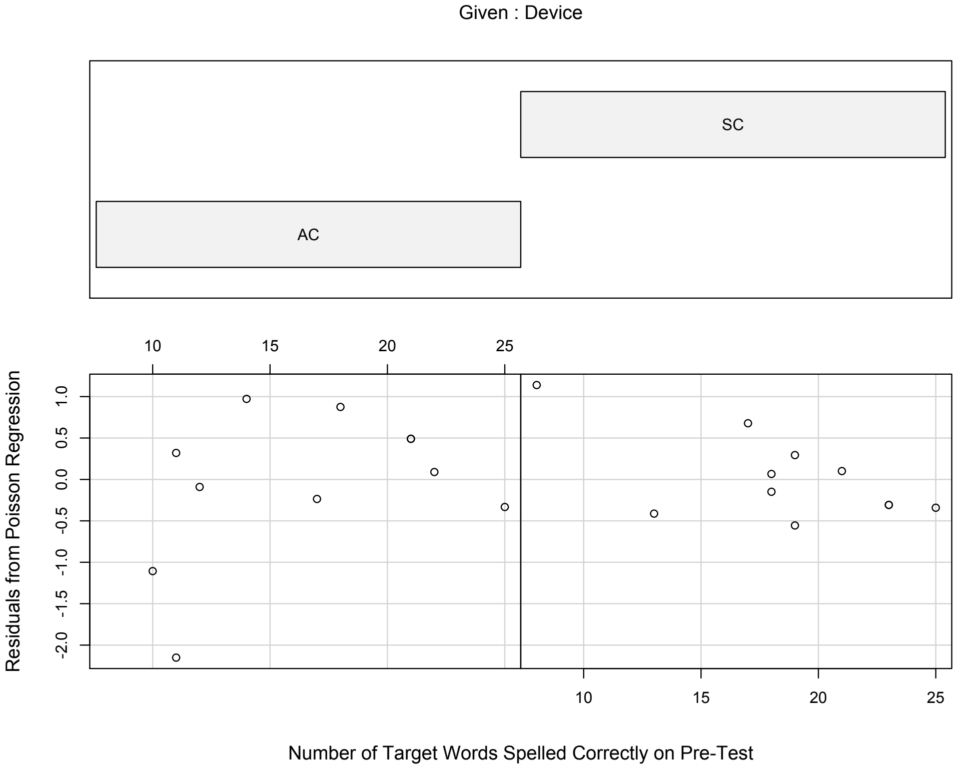
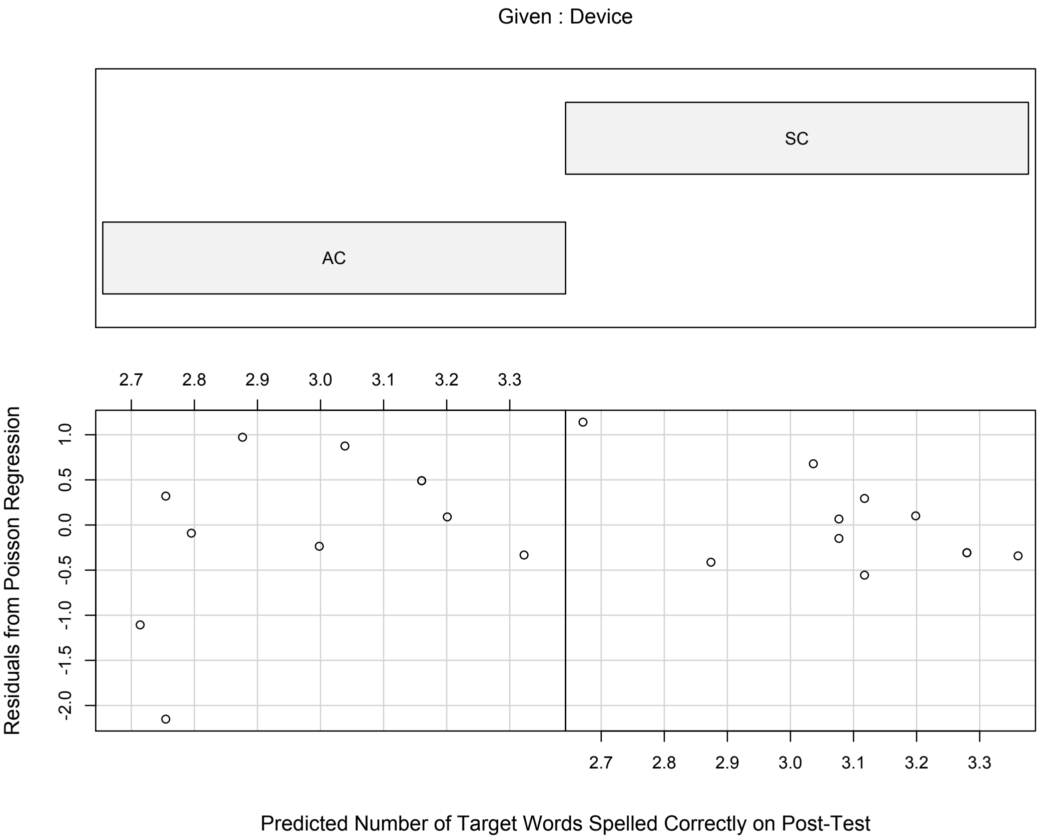
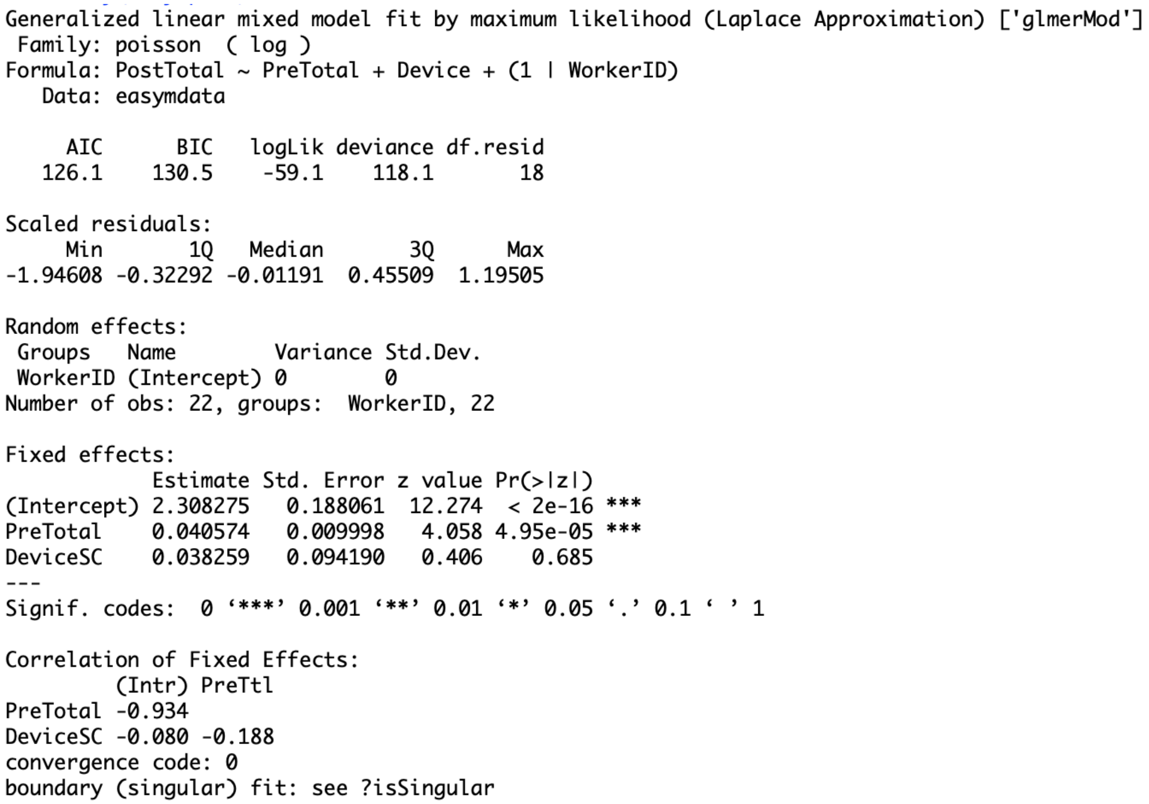


Figure 52 - Pearson Residuals from Poisson Regression with OLRE vs Predicted Score on Post-Test



B.2.3 Model Fit

Figure 53 - Results from GLMM (Poisson) Regression with OLRE

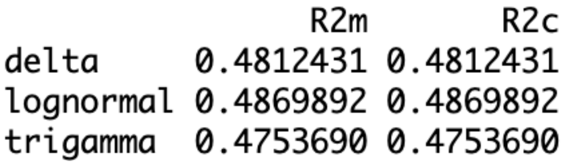


```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: PostTotal ~ PreTotal + Device + (1 | WorkerID)
   Data: easymdata

     AIC      BIC   logLik deviance df.resid
   126.1    130.5    -59.1    118.1       18

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.94608 -0.32292 -0.01191  0.45509  1.19505

Random effects:
 Groups   Name        Variance Std.Dev.
 WorkerID (Intercept) 0        0
Number of obs: 22, groups:  WorkerID, 22

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.308275   0.188061  12.274  < 2e-16 ***
PreTotal    0.040574   0.009998   4.058 4.95e-05 ***
DeviceSC    0.038259   0.094190   0.406    0.685
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
         (Intr) PreTtl
PreTotal -0.934
DeviceSC -0.080 -0.188
convergence code: 0
boundary (singular) fit: see ?isSingular
```

Table 15- Table of GLMM Coefficients

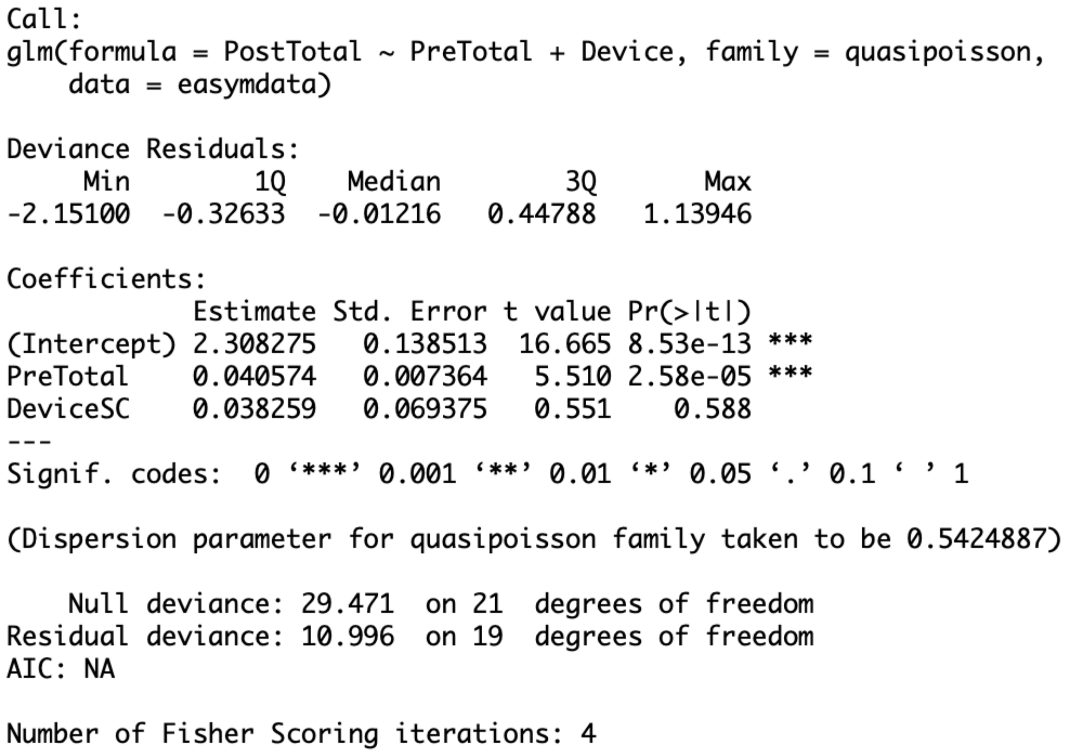
	Estimate	**Std. Error**	**z value**	***P*-value**
Intercept	2.3083	0.18806	12.274	<2e-16
Pre-Test Score	0.0406	0.0010	4.058	4.95e-05
Spell-Checking	0.0383	0.0942	0.406	0.685

Figure 54 -R^2 of GLMM



```
                R2m       R2c
delta     0.4812431 0.4812431
lognormal 0.4869892 0.4869892
trigamma  0.4753690 0.4753690
```

Figure 55 - Results from Quasipoisson Regression



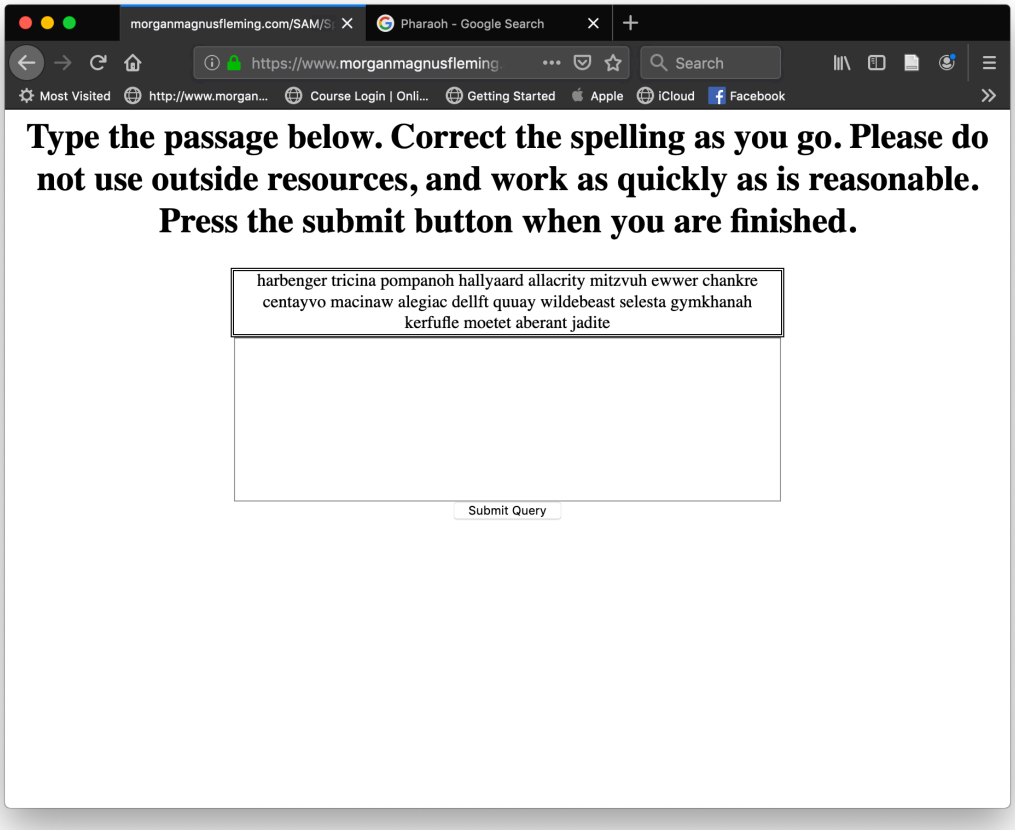
```
Call:
glm(formula = PostTotal ~ PreTotal + Device, family = quasipoisson,
    data = easymdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.15100  -0.32633  -0.01216   0.44788   1.13946

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.308275   0.138513  16.665 8.53e-13 ***
PreTotal    0.040574   0.007364   5.510 2.58e-05 ***
DeviceSC    0.038259   0.069375   0.551    0.588
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 0.5424887)

    Null deviance: 29.471  on 21  degrees of freedom
Residual deviance: 10.996  on 19  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Appendix C

C.1 Materials

C.1.1 Website and Support Systems

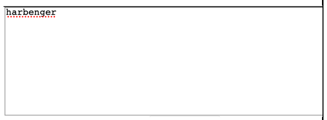
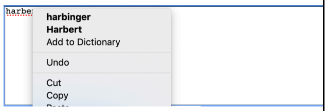
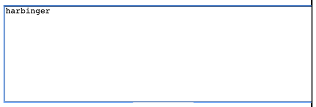
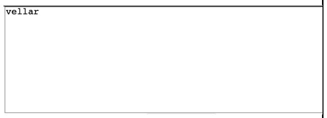
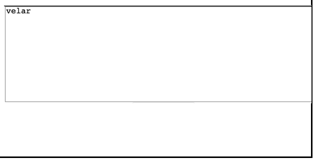
Figure 56- Screenshot of Website Used to Gather Data



Type the passage below. Correct the spelling as you go. Please do not use outside resources, and work as quickly as is reasonable. Press the submit button when you are finished.

harbenger tricina pompanoh hallyaard allacrity mitzvuh ewwer chankre centayvo macinaw alegiac dellft quuay wildebeast selesta gymkhanah kerfufle moetet aberant jadite

Table 16 - Table of Use Steps for Spelling Support Systems

Spelling Support System	Identification Assistance	Selection Assistance	Execution
Spell-Checking	 harbenger	 harbe harbinger Harbert Add to Dictionary Undo Cut Copy	 harbinger
Autocorrect	 vellar	Selection is handled by system	 velar

This table shows what the user saw during each step of using their respective support system. Selection is not depicted for the autocorrect, as the selection step is handled by the support system, without any user prompting

C.1.2 Stimuli

C.1.2.1 Target Words

Table 17 - Target Words

Misspelling	Correct Spelling	Misspelling	Correct Spelling	Misspelling	Correct Spelling
harbenger	harbinger	preogue	prorogue	kehpi	kepi
tricina	trichina	zhloty	zloty	treuculent	truculent
pompanoh	pompano	piaza	piazza	pheaton	phaeton
hallyard	halyard	aggar	agar	cenobyte	cenobite
allacrity	alacrity	zerography	xerography	yttreaum	yttrium
mitzvuh	mitzvah	Dirck	dirk	vectual	victual
ewwer	ewer	Rollic	rollick	schemma	schema
chankre	chancre	zephear	zephyr	upsillion	upsilon
centayvo	centavo	cayene	cayenne	bouffahnt	bouffant
macinaw	mackinaw	illeum	ileum	scherso	scherzo
alegiac	elegiac	raffea	raffia	exturpate	extirpate
dellft	delft	sacharin	saccharin	rheeum	rheum
quuay	quay	panashe	panache	oshiose	otiose
wildbeast	wildebeest	wapeeti	wapiti	bouillubaisse	bouillabaisse
selesta	celesta	fisile	fissile	vallise	valise
gymkhanah	gymkhana	ginko	ginkgo	ententae	entente
kerfufle	kerfuffle	grohsbeak	grosbeak	raymin	ramie
moetet	motet	bracerro	bracero	braem	bream
aberant	aberrant	toowhee	towhee	ginghum	gingham
jadite	jadeite	bragadocio	braggadocio	sepiya	sepia

These are the target words, and associated misspellings, used in Study 3

C.1.2.2 Trial 1

“harbenger tricina pompanoh hallyaard allacrity mitzvuh ewwer chankre centayvo macinaw alegiac dellft quuay wildebeast selesta gymkhanah kerfufle moetet aberant jadite”

C.1.2.3 Trial 2

“prerogue zhloty piaza aggar zerography dirck rollic zephear cayene illeum raffea sacharin panashe wapeeti fisile ginko grohsbeak bracerro toowhee bragadocio”
.”

C.1.2.4 Trial 3

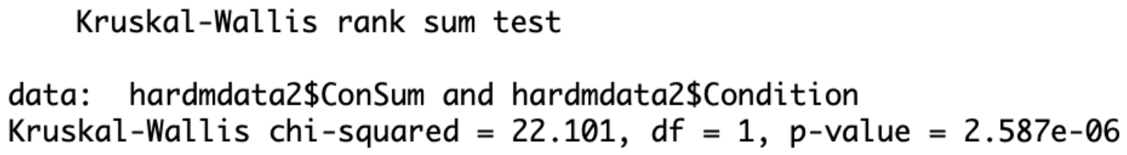
"kehpi treuculent pheaton cenobyte yttreaum vectual schemma upsillon bouffahnt scherso exturpate rheeum oshiose bouillubaisse vallise ententae raymie braem ginghum sepiya."

C.2 Model Verification

C.2.1 Data Exploration

C.2.1.1 Assessment of Device

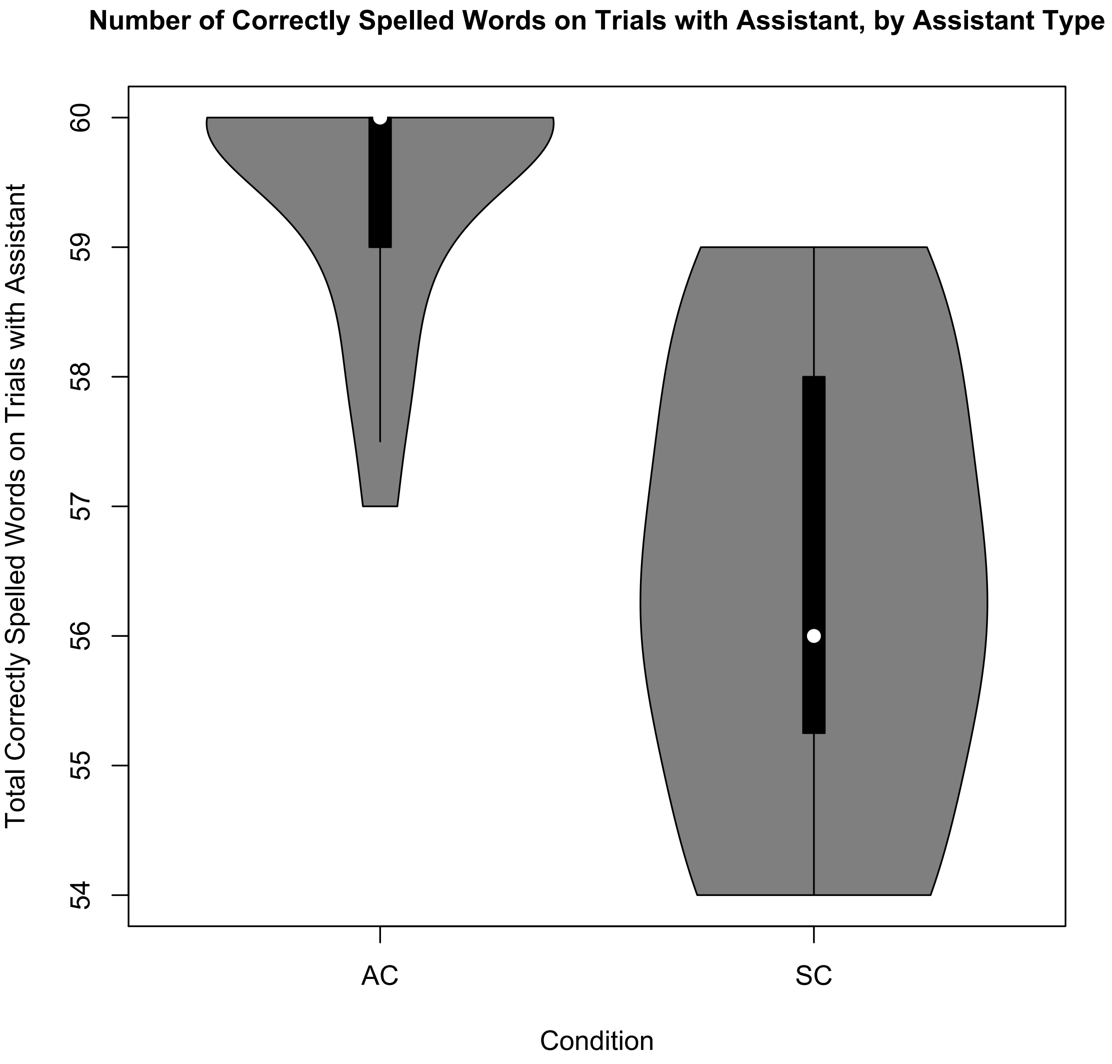
Figure 57 - Kruskal-Wallis Rank Sum Test



```
	Kruskal-Wallis rank sum test

data:  hardmdata2$ConSum and hardmdata2$Condition
Kruskal-Wallis chi-squared = 22.101, df = 1, p-value = 2.587e-06
```

Results indicate that for difficult words, users that had access to the browser based spell-checking systems performed significantly differently from users of the autocorrect system. This was for trials where they had access to the support systems.

Figure 58-Number of Correctly Spelled Words on Trials with Assistant, by Assistant Type



The violin plot indicates that participants in the autocorrect condition saw at least 57 correctly spelled words, while the participants in the spell-checking condition only saw at least 54 correctly spelled words and at most 59 correctly spelled words.

C.2.1.2 Post-Test Performance

Figure 59 - Post-Test Scores vs Pre-Test Scores, Grouped by Condition

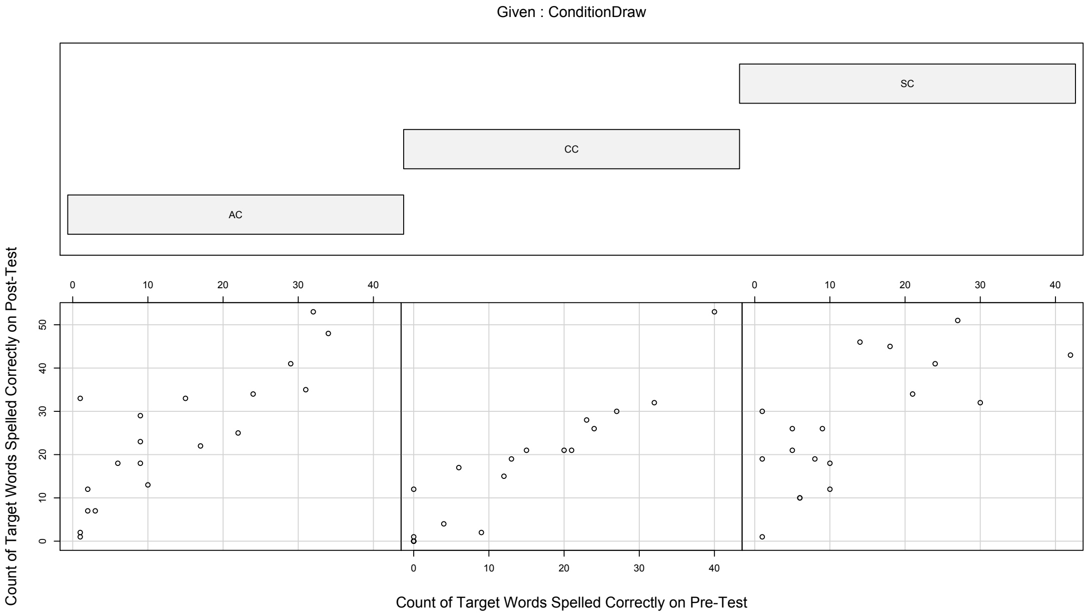
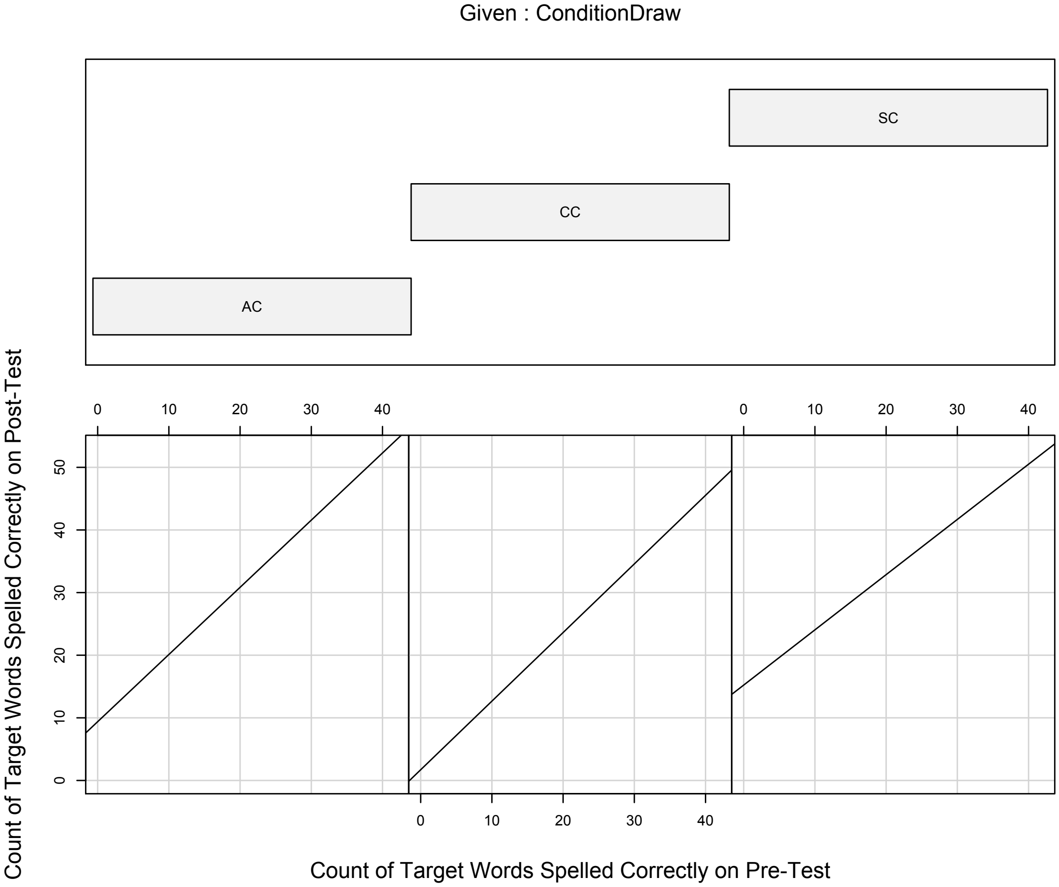
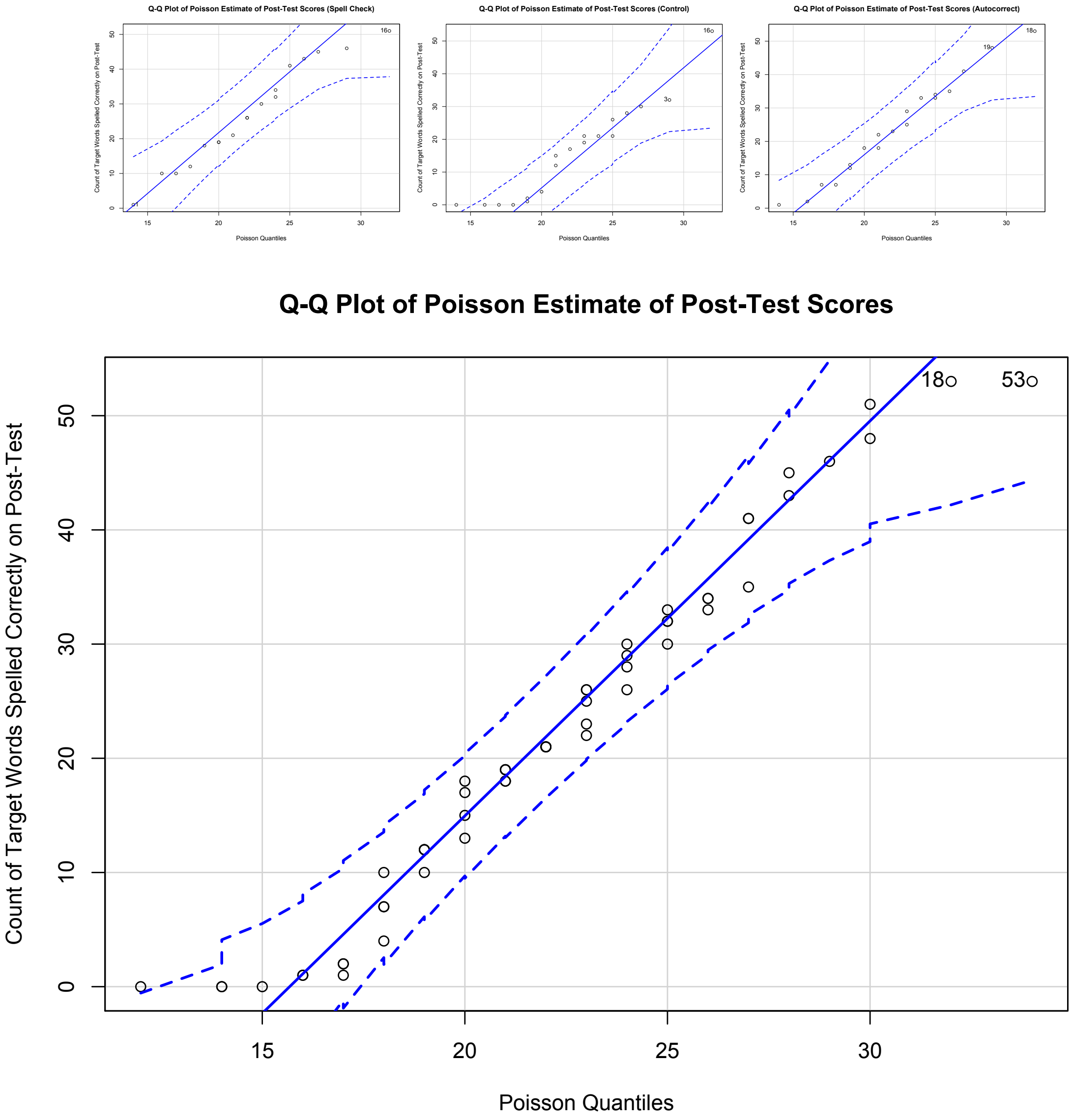


Figure 60 - Post-Test Scores vs Pre-Test Scores, Grouped by Condition (line)



C.2.1.3 Poissonality

Figure 61 - Q-Q Plot of Estimated Poisson Quantiles vs Observed Score on Post-Test



C.2.1.4 Outliers

Figure 62-Cleaveland Plot of Post-Test Scores

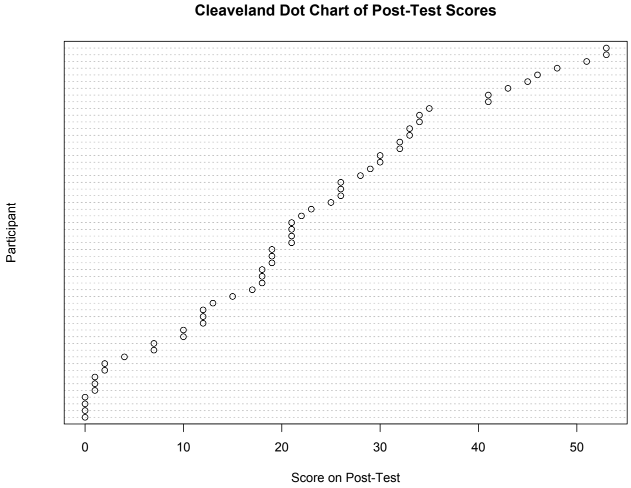


Figure 63- Cleaveland Plot of Post-Test Scores, Grouped by Condition

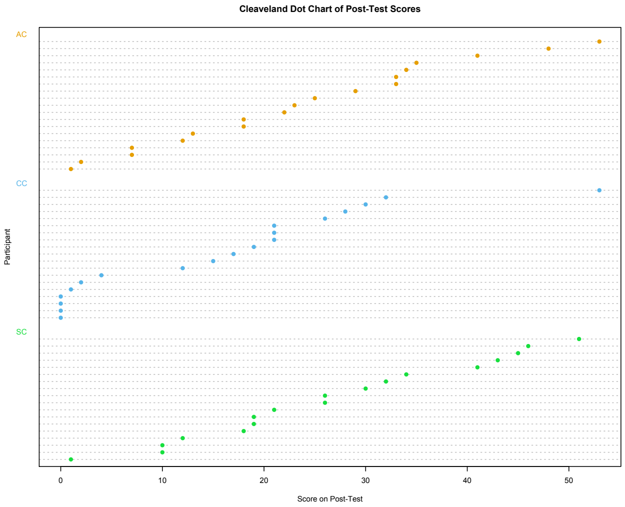


Figure 64 - Cleaveland Plot of Pre-Test Scores

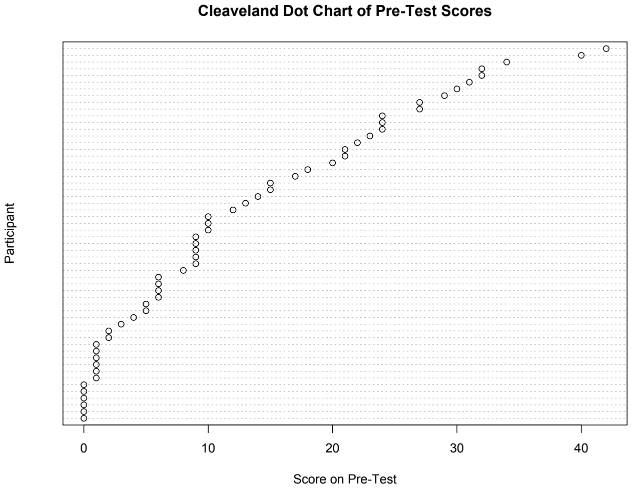


Figure 65 - Cleaveland Plot of Pre-Test Scores by Condition

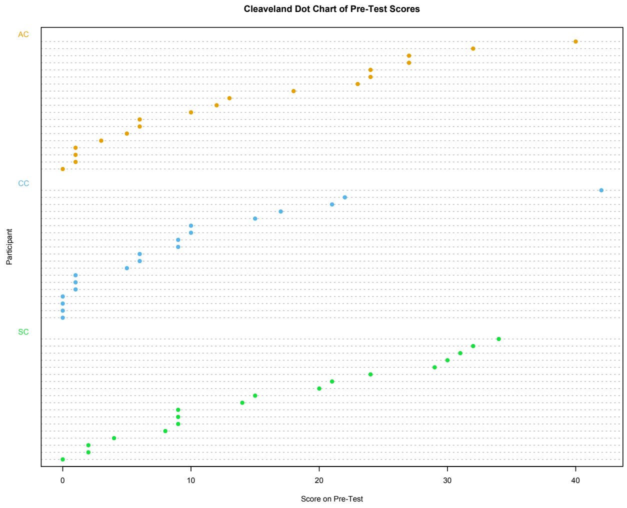
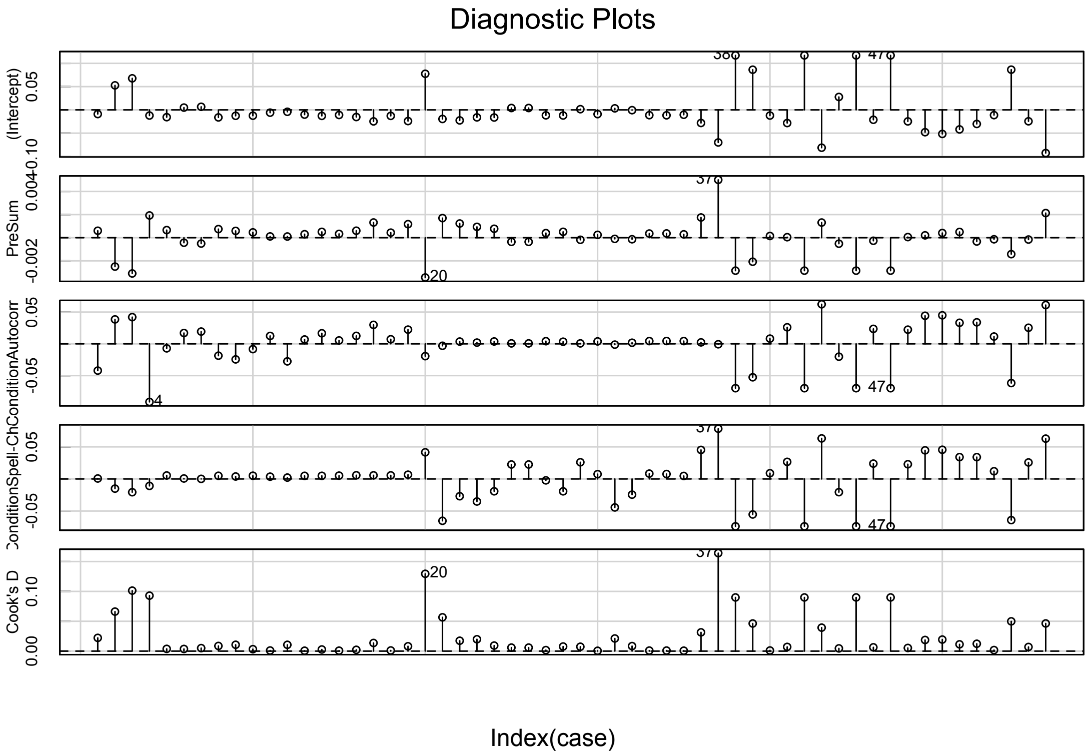


Figure 66-Outlier Diagnostics for Poisson Model with OLRE



Outliers were selected as being 2 standard deviations away from the average Cook's D.

Table 18- Table of Regression Coefficients After Removal of Outliers

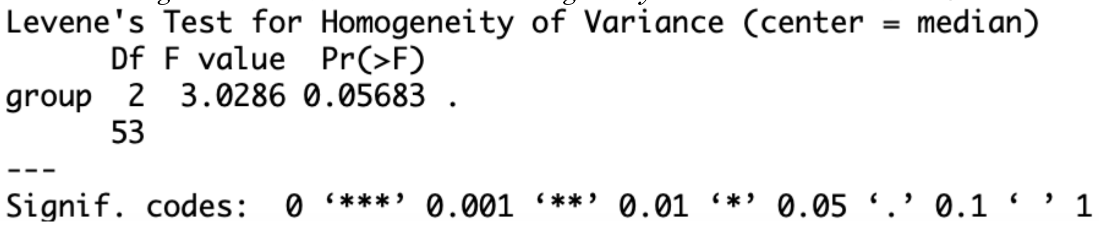
Observation Index	***Intercept***	***Pre-Test***	***Autocorrect***	***Spell Check***
3	*1.5171*	*0.0554*	*0.6420*	*0.7862*
20	*1.6037*	*0.0550*	*0.5806*	*0.8484*
37	*1.4561*	*0.0635*	*0.5993*	*0.8852*
Full-Model	*1.5261*	*0.0585*	*0.6000*	*0.8069*

This table summarizes the regression coefficients estimated by the model when the corresponding observed outlier is removed. Removing observation 3 lowers the estimated effect of spell-check by 2.09%, while removing observation 37 increases the estimated effect of spell-check by 8.14%.

C.2.2 Model Validation

C.2.2.1 Heterogeneity

Figure 67 - Levene's Test for Homogeneity in Poisson Model with OLRE



```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  2  3.0286 0.05683 .
      53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Results indicate no violation of the assumption of homogeneity of variance between levels.

C.2.2.2 Residual Plots

Figure 68-Pearson Residuals from Poisson Regression vs Predicted Number of Target Words Spelled Correctly on Post-Test

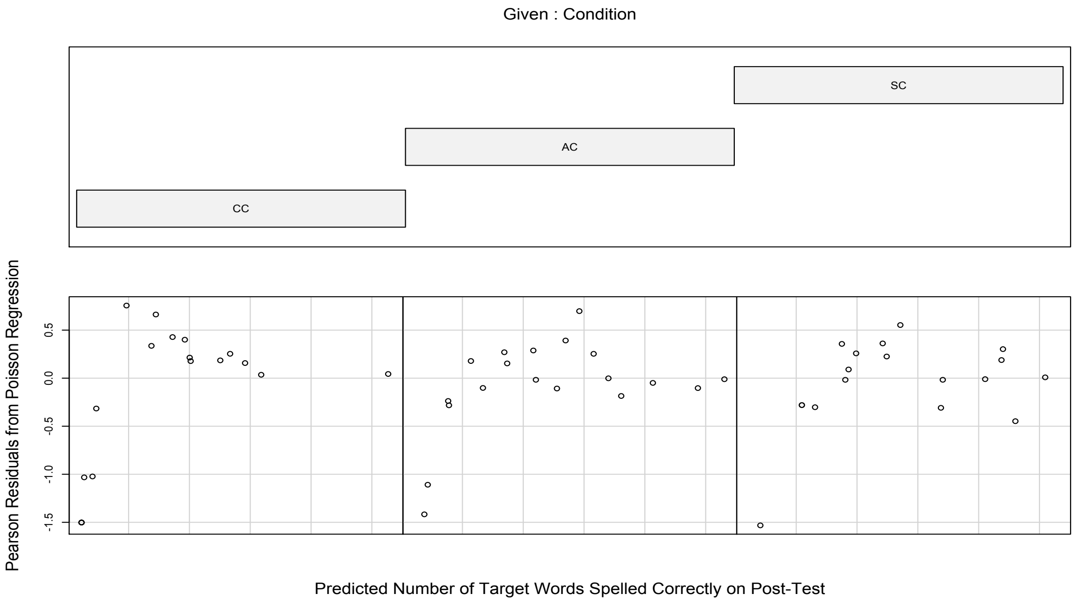
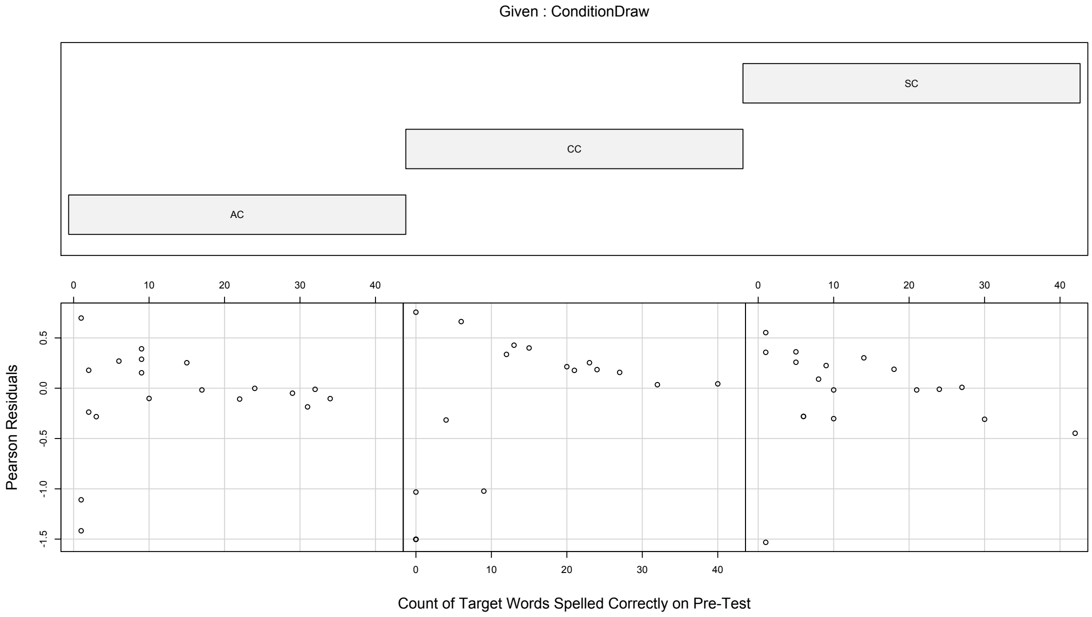
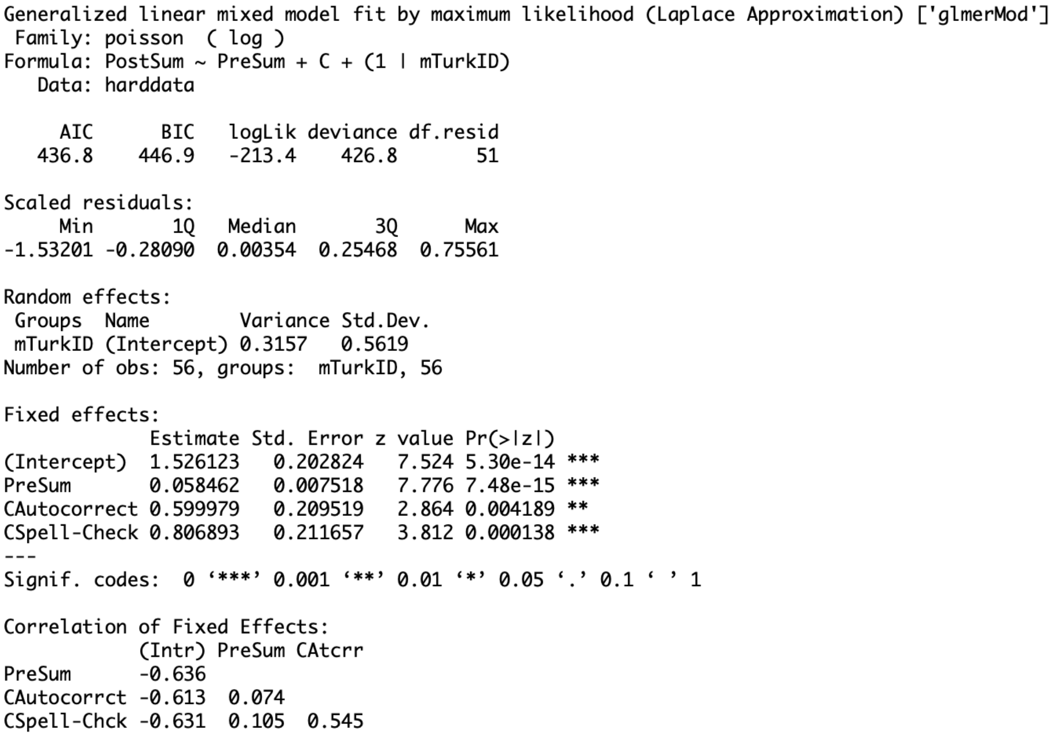


Figure 69- Pearson Residuals for Model vs Score on Pre-Test, Grouped by Condition



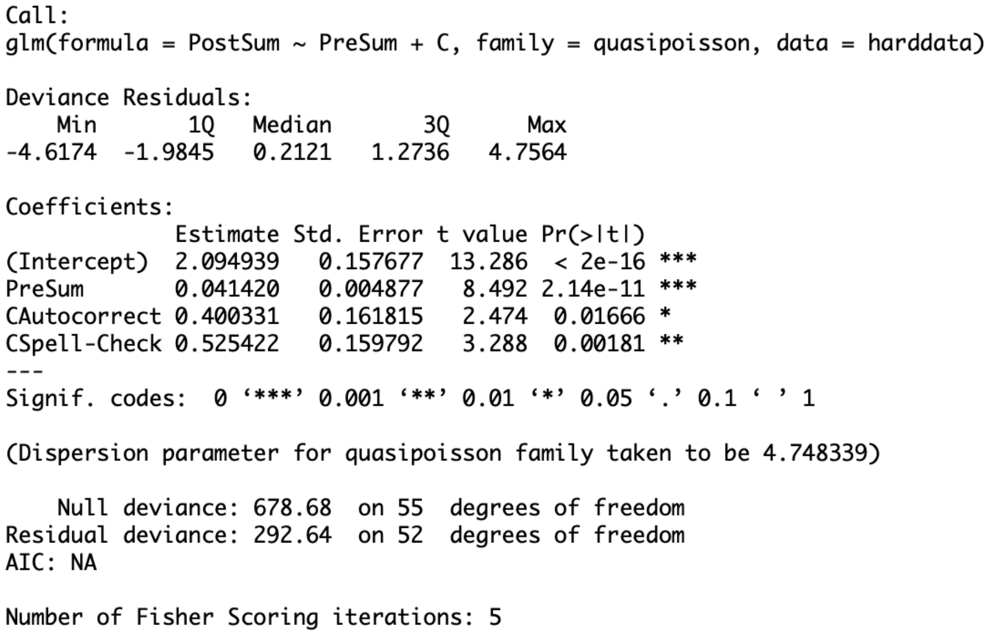
C.2.3 Model Fit

Figure 70 - Results from Poisson Regression with OLRE



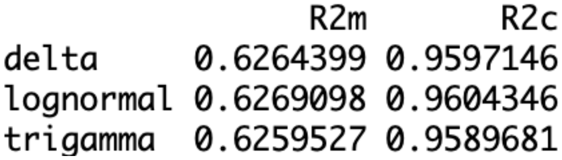
```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: PostSum ~ PreSum + C + (1 | mTurkID)
   Data: harddata

     AIC      BIC   logLik deviance df.resid
   436.8    446.9   -213.4    426.8       51

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.53201 -0.28090  0.00354  0.25468  0.75561

Random effects:
 Groups  Name        Variance Std.Dev.
 mTurkID (Intercept) 0.3157   0.5619
Number of obs: 56, groups:  mTurkID, 56

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.526123   0.202824   7.524 5.30e-14 ***
PreSum       0.058462   0.007518   7.776 7.48e-15 ***
CAutocorrect 0.599979   0.209519   2.864 0.004189 **
CSpell-Check 0.806893   0.211657   3.812 0.000138 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) PreSum CAtcrr
PreSum      -0.636
CAutocorrct -0.613  0.074
CSpell-Chck -0.631  0.105  0.545
```

Figure 71- Results from Quasipoisson Regression



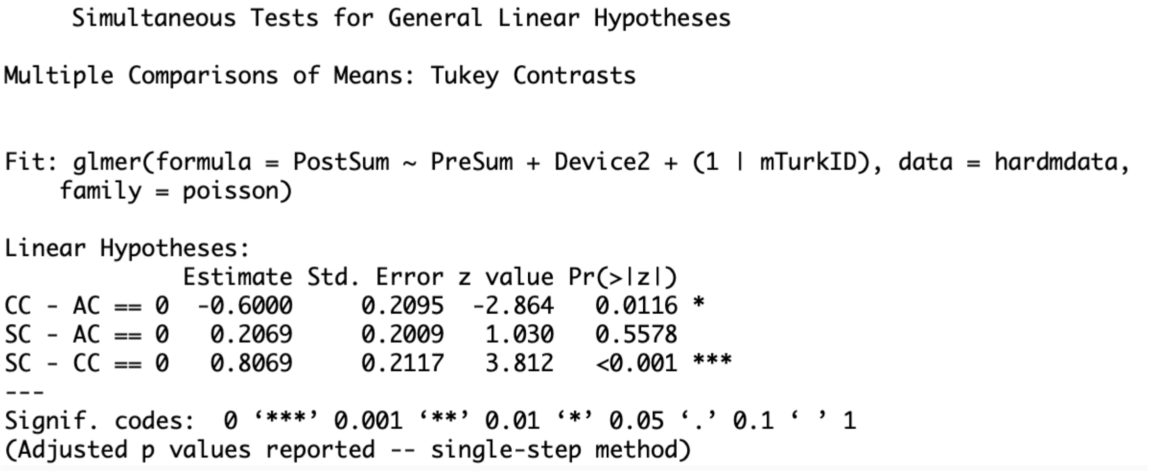
```
Call:
glm(formula = PostSum ~ PreSum + C, family = quasipoisson, data = harddata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6174  -1.9845   0.2121   1.2736   4.7564

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.094939   0.157677  13.286  < 2e-16 ***
PreSum       0.041420   0.004877   8.492 2.14e-11 ***
CAutocorrect 0.400331   0.161815   2.474  0.01666 *
CSpell-Check 0.525422   0.159792   3.288  0.00181 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 4.748339)

    Null deviance: 678.68  on 55  degrees of freedom
Residual deviance: 292.64  on 52  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure 72 -R^2 for GLMM



```
                 R2m       R2c
delta     0.6264399 0.9597146
lognormal 0.6269098 0.9604346
trigamma  0.6259527 0.9589681
```

Figure 73 - Multiple Comparisons Between Conditions,Output From 'glht' (Hothorn, Bretz, & Westfall, 2008).



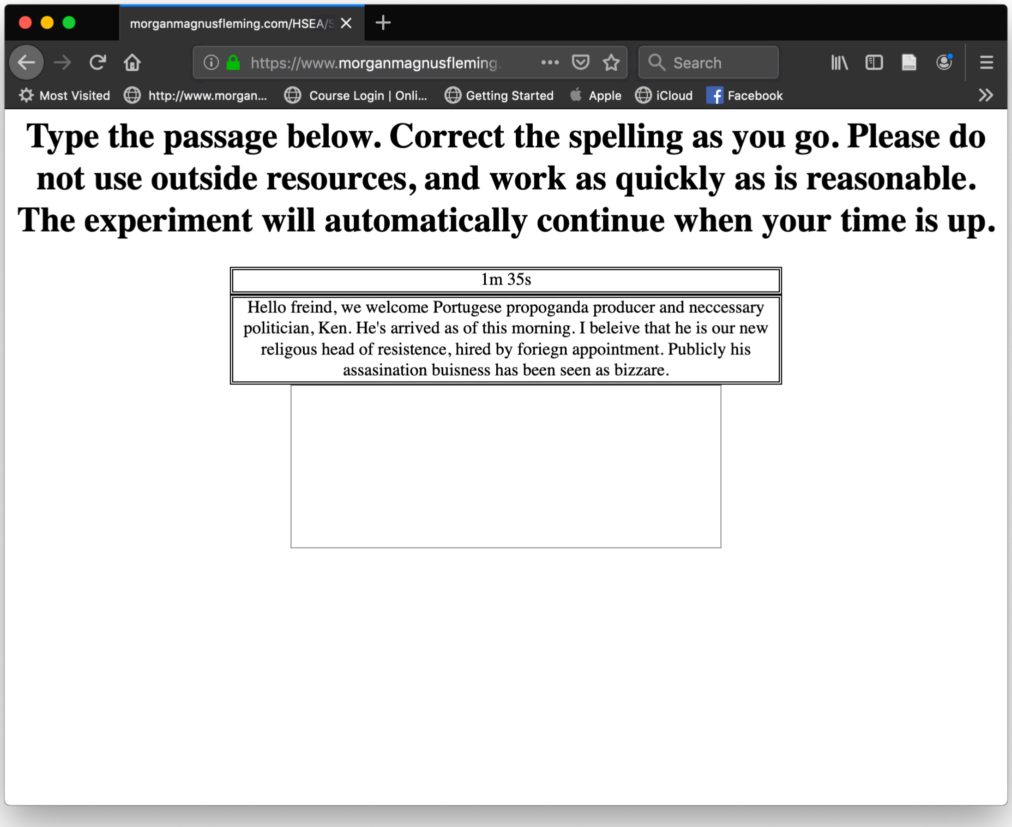
```
	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: glmer(formula = PostSum ~ PreSum + Device2 + (1 | mTurkID), data = hardmdata, 
    family = poisson)

Linear Hypotheses:
             Estimate Std. Error z value Pr(>|z|)    
CC - AC == 0  -0.6000     0.2095  -2.864   0.0116 *  
SC - AC == 0   0.2069     0.2009   1.030   0.5578    
SC - CC == 0   0.8069     0.2117   3.812   <0.001 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Appendix D

D.1 Materials

D.1.1 Website and Devices

Figure 74 - Screenshot of Website Used to Collect Data

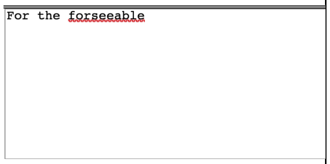
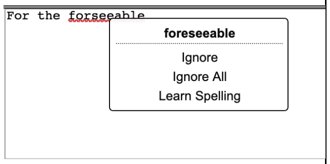
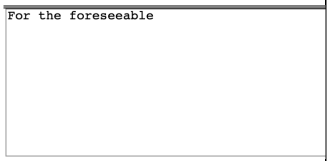
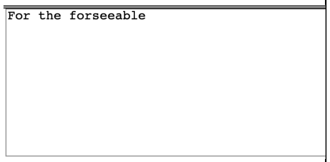
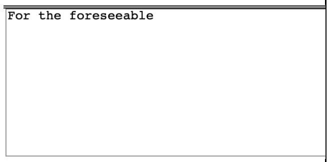


Type the passage below. Correct the spelling as you go. Please do not use outside resources, and work as quickly as is reasonable. The experiment will automatically continue when your time is up.

1m 35s

Hello freind, we welcome Portugese propoganda producer and neccessary politician, Ken. He's arrived as of this morning. I beleive that he is our new religous head of resistence, hired by foriegn appointment. Publicly his assasination buisness has been seen as bizzare.

Table 19- Phases of the Spelling Assistance Process

Spelling Support System	Identification Assistance	Selection Assistance	Execution
Spell-Checking	 For the forseeable	 For the forseeable	 For the foreseeable
Autocorrect	 For the forseeable	All words are spelled correctly	 For the foreseeable

This table summarizes the steps the user takes in correcting a spelling error with either the autocorrect, or spell-checking assistant created for this study. Of particular note are the differences in the identification and selection steps. In the identification step, autocorrect provides no indication of a misspelling. Spell-check, on the other hand, draws the user's attention to the misspelling with a red underline. In the selection step, autocorrect handles the process of selecting a properly spelled word without prompting the user in any manner. Spell-check requires that the user perform the selection process (in part) themselves. Spell-check offers the user candidate properly spelled words, and the user must select the proper spelling from that list. In both cases the final step is replacing the misspelling with a properly spelled word.

D.1.2 Stimuli

D.1.2.1 Target Words

Table 20 - Table of Target Words for Study 4

Easy Misspellings	Easy Proper Spelling	Hard Misspellings	Hard Proper Spelling
freind	friend	brasero	bracero
Portugese	Portuguese	vellar	velar
propoganda	propaganda	darma	dharma
neccessary	necessary	trychina	trichina
religous	religious	pompeno	pompano
resistence	resistance	scanscion	scansion
foriegn	foreign	faillet	faille
beleive	believe	bateste	batiste
assasination	assassination	hallyard	halyard
buisness	business	alacrety	alacrity
bizzare	bizarre	mitsvah	mitzvah
calender	calendar	aballone	abalone
collegue	colleague	chrore	crore
Carribbean	Caribbean	ewwer	ewer
concious	conscious	chanchre	chancre
tendancy	tendency	sashey	sashay
forseeable	foreseeable	centavvo	centavo
goverment	government	mackenaw	mackinaw
remeber	remember	paruke	peruke
gaurd	guard	dellft	delft
posession	possession	kuay	quay
occassion	occasion	wildebeast	wildebeest
agression	aggression	selesta	celesta
neaderthal	Neanderthal	eloadea	elodea
pharoah	Pharaoh	ayuah	ayah
humourous	humorous	xeolite	zeolite
chauffer	chauffeur	gymkana	gymkhana

Two lists of words were used in the study, categorized as 'Easy' and 'Hard'. These lists come from Studies 1 and 3 respectively.

D.1.2.2 Easy Passage 1

“Hello freind, we welcome Portugese propoganda producer and neccessary politician, Ken. Religous head of resistence by foriegn appointment, I beleive. Publicly his assasination buisness has been seen as bizzare.”

D.1.2.3 Easy Passage 2

“A calender filled with meetings with a specific collegue in the Carribbean, concious avoidance of his chauffer, and his tendancy for humourous rants gave him away.”

D.1.2.4 Easy Passage 3

“For the forseeable goverment tenure, remeber that a glamourous gaurd must remain in posession of his wits for this occassion. Agression will be punished. Remember, do not be a neaderthal, be a pharoah.”

D.1.2.5 Hard Passage 1

“The brasero with a vellar lisp spoke about the darma calmly. Like a trychina worm or a pompeno we flow along a scanscion as smooth as faillet or bateste cloth. Hoist your hallyard and maintain your alacrety, and you can better fulfill your mitsvah and find your happiness.”

D.1.2.6 Hard Passage 2

“He scooped out the aballone, now worth a chrore and a half. The medicine in the ewwer eased the chanchre plaguing the patient. Sashey for a centavvo was too low of a price. Upon the patient’s front was their mackenaw. “

D.1.2.7 Hard Passage 3

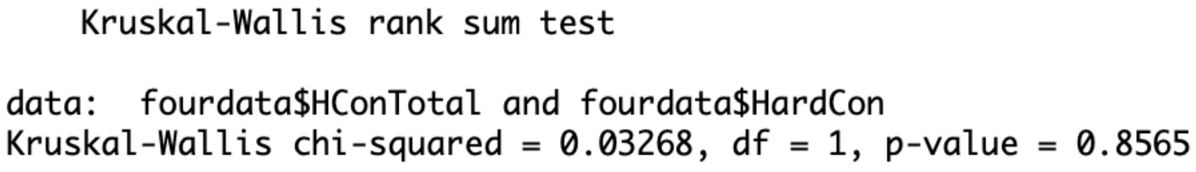
“His paruke toppled into the dellft bowl on the ground, just above the kuay below. A wildebeast herd outside stamped about like a selesta, and eloadea sat just under the surface of the owater. The ayuah went to the shelf to grab the xeolite to the tune of that gymkana from nature. “

D.2 Model Verification

D.2.1 Data Exploration

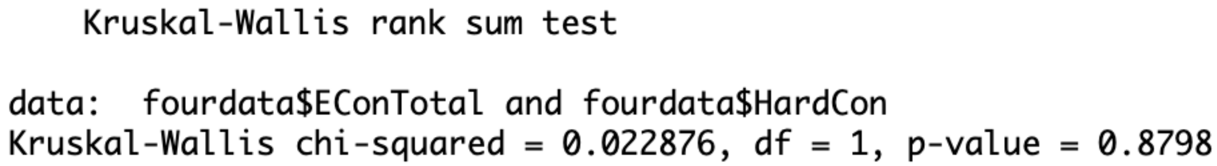
D.2.1.1 Device Assessment

Figure 75 - Rank Sum Test (Hard Words)



```
	Kruskal-Wallis rank sum test

data:  fourdata$HConTotal and fourdata$HardCon
Kruskal-Wallis chi-squared = 0.03268, df = 1, p-value = 0.8565
```

Figure 76- Rank Sum Test (Easy Words)



```
        Kruskal-Wallis rank sum test

data:  fourdata$EConTotal and fourdata$HardCon
Kruskal-Wallis chi-squared = 0.022876, df = 1, p-value = 0.8798
```

Figure 77 - Violin Plot of Score on Test with Support System (Easy Words)

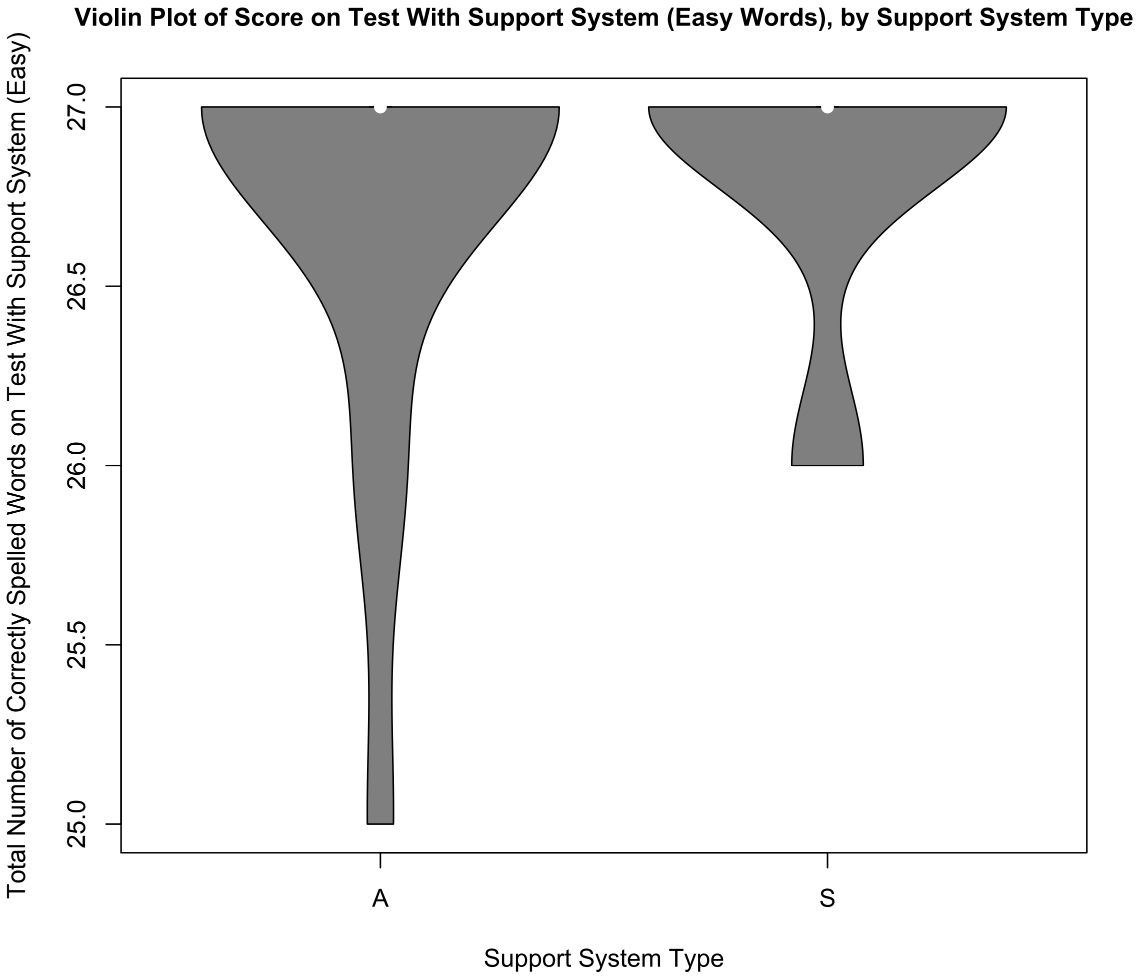
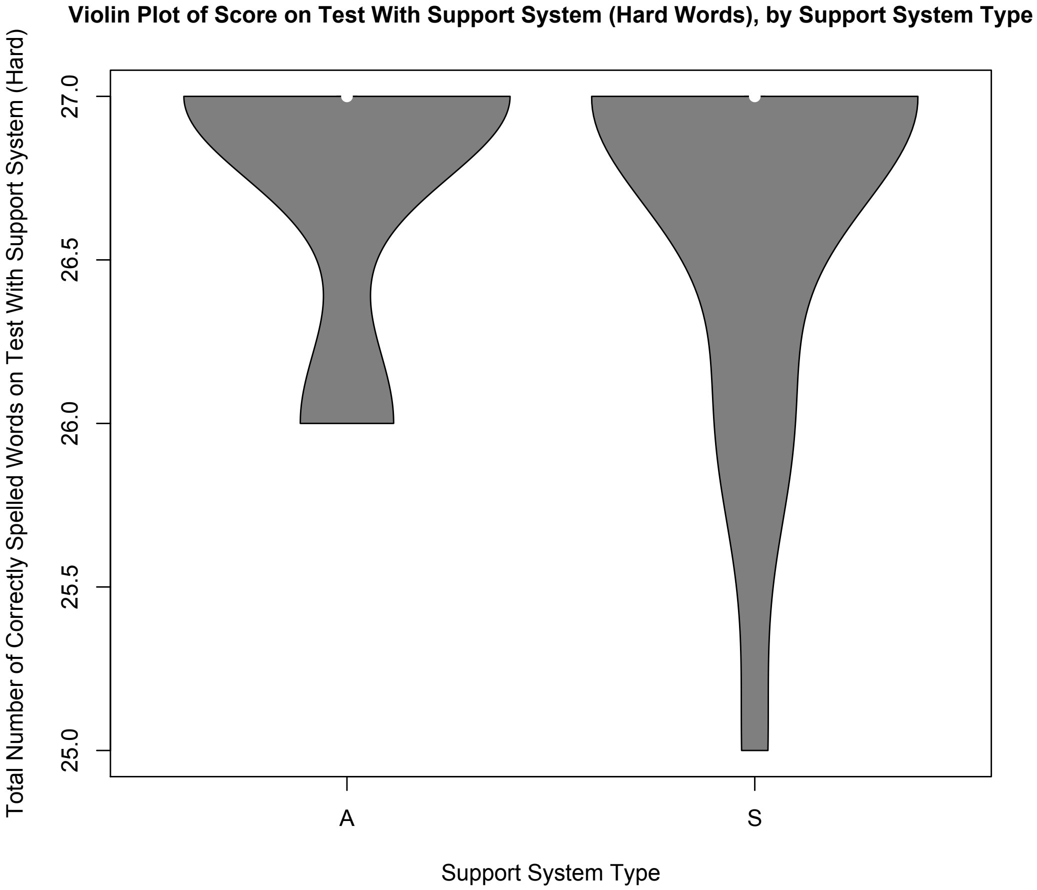


Figure 78- Violin Plot of Score on Test with Support System (Hard Words)



D.2.1.2 Distributions

Figure 79 - Plot of Post-Test Scores for Hard Words vs Pre-Test Scores for Hard Words, Grouped by Condition

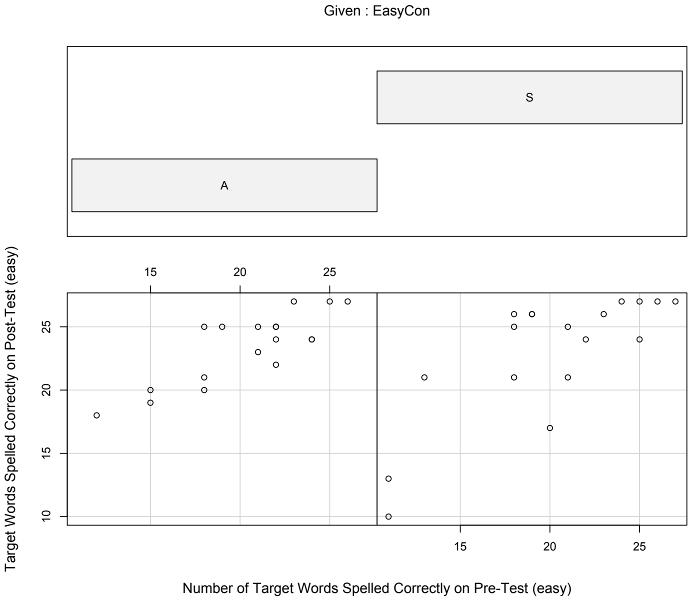
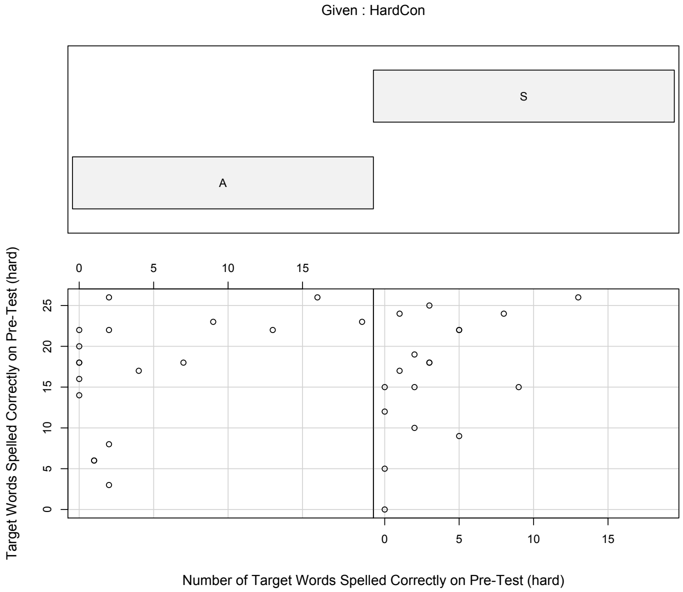


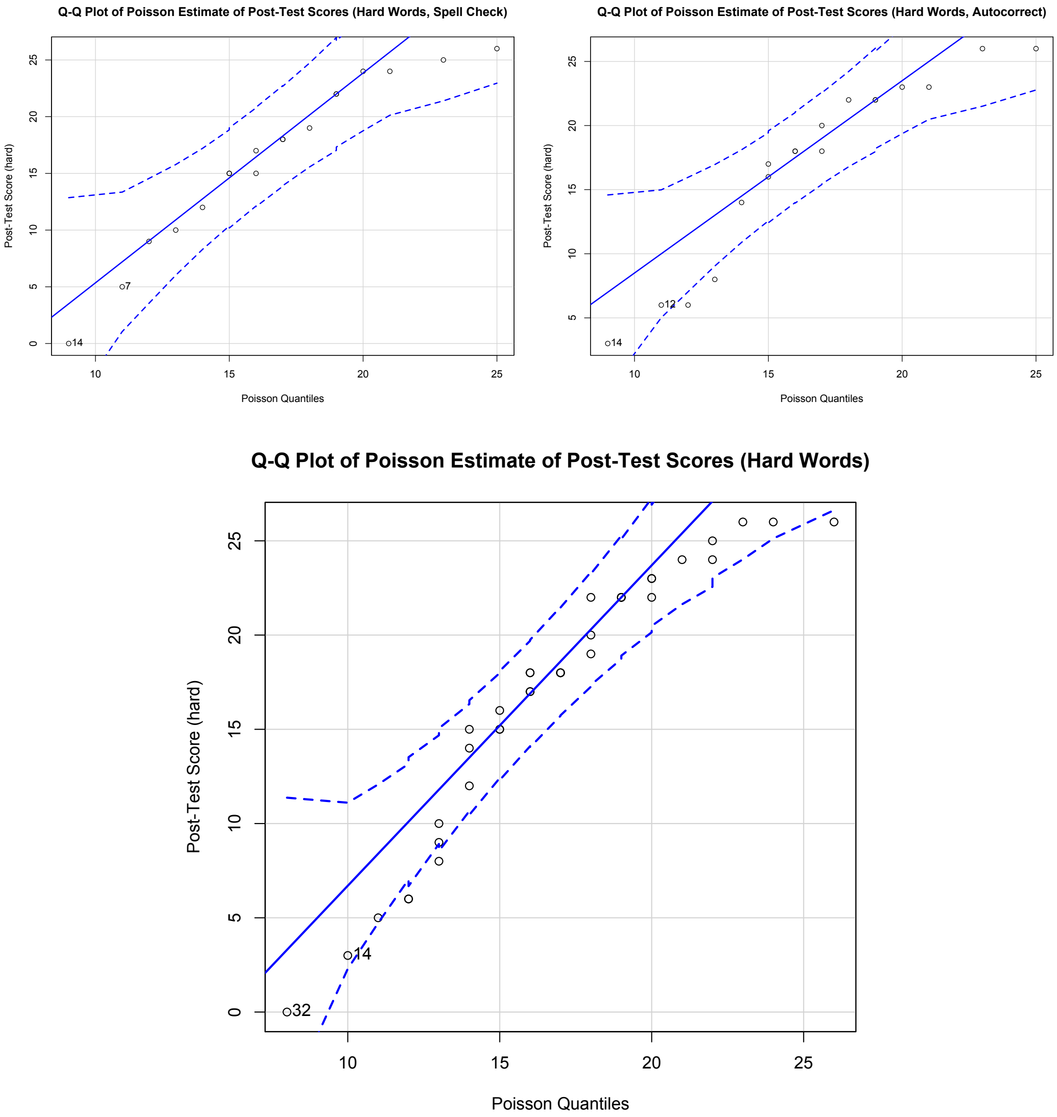
Figure 80 - Plot of Post-Test Scores for Easy Words vs Pre-Test Scores for Easy Words, Grouped by Condition



D.2.1.3 Poissonality

D.2.1.3.a Q-Q Plots for Hard Words

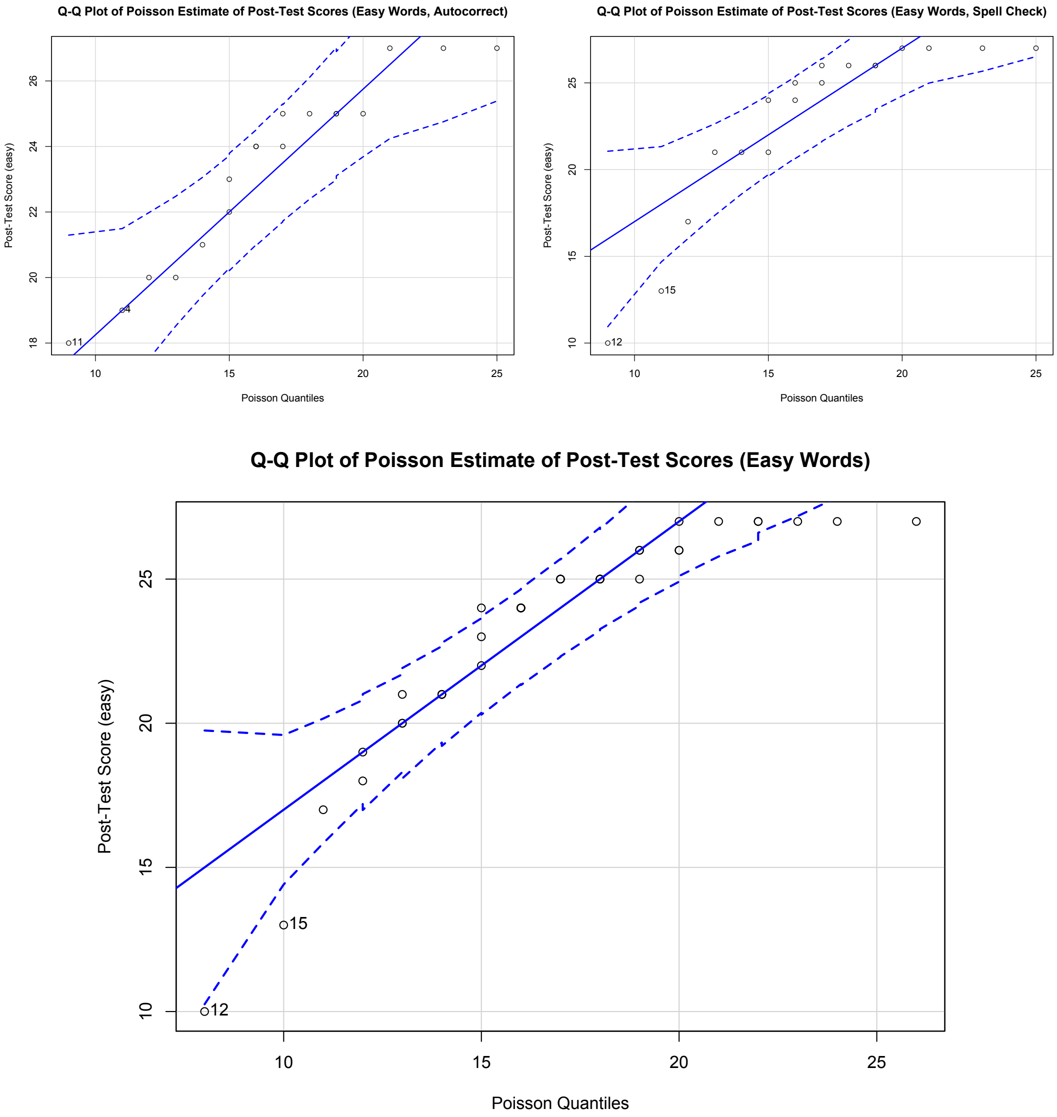
Figure 81 - Q-Q Plots of Poisson Quantile Estimates of Post-Test Scores vs Observed Post-Test Scores with Hard Words



Only three values fall outside of the 95% confidence interval band. Issues with poissonality appear to be mostly the result of over estimation of post-test scores at lower values.

D.2.1.3.b Q-Q Plots for Easy Words

Figure 82 - Q-Q Plots of Poisson Quantile Estimates of Post-Test Scores vs Observed Post-Test Scores With Easy Words



6 values fall outside the 95% confidence interval bands. This appears to have been the result of over estimation of lower post-test scores, and under estimation of three data points at the top end of the distribution.

D.2.1.4 Outliers

D.2.1.4.a Outliers - Hard Words

Figure 83-Cleaveland Plot of Post-Test Scores With Hard Words-

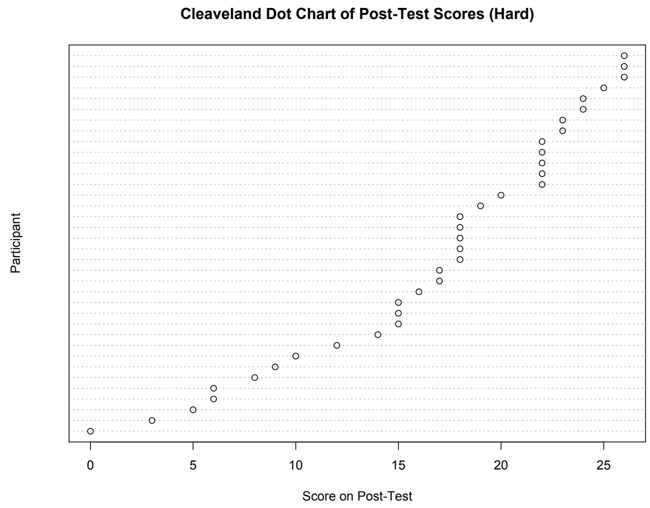


Figure 84-Cleaveland Plot of Post-Test Scores with Hard Words, Grouped by Condition

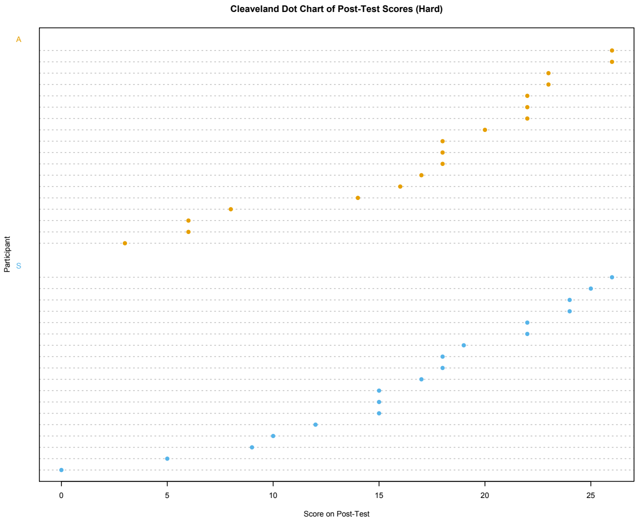


Figure 85 - Diagnostic Plot for Outliers, Hard Words

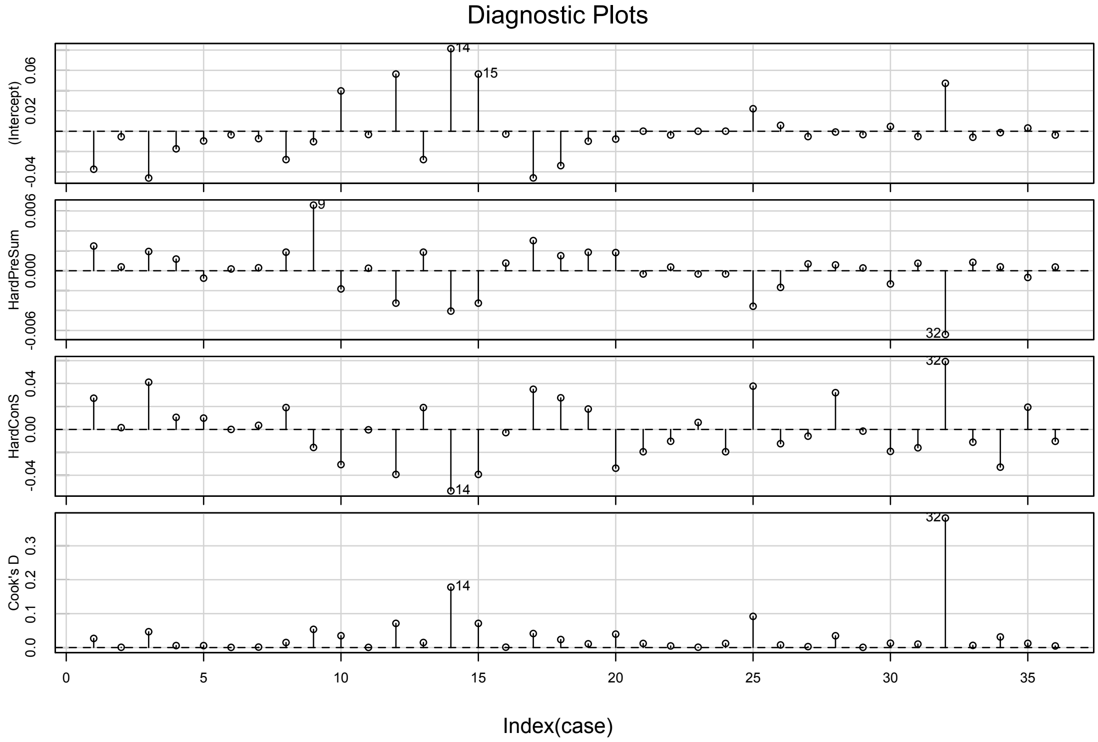


Table 21 - Table of Regression Coefficients After Removal of Outliers, Hard Words

Observation Index	***Intercept***	***Pre-Test***	***Spell Check***
32	*2.6374*	*0.0340*	*0.0555*
Full-Model	*2.5902*	*0.0404*	*-0.0037*

This table summarizes the regression coefficients estimated by the model when the corresponding observed outlier is removed.

D.2.1.4.b Outliers - Easy Words

Figure 86- Cleaveland Plot of Post-Test Scores with Easy Words

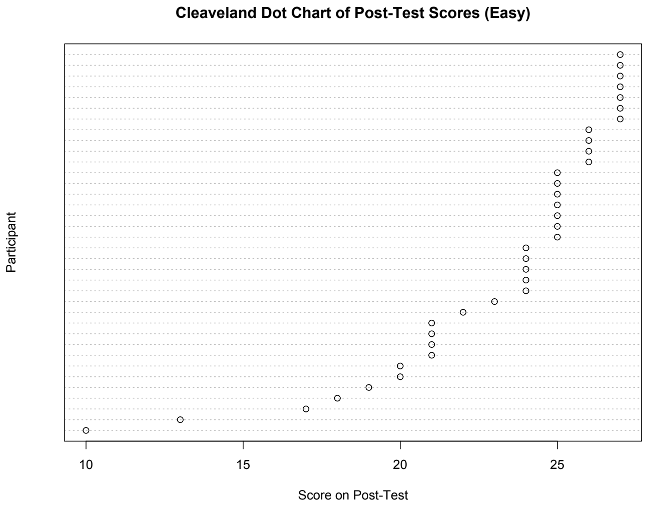


Figure 87- Cleaveland Plot of Post-Test Scores with Easy Words, Grouped by Condition

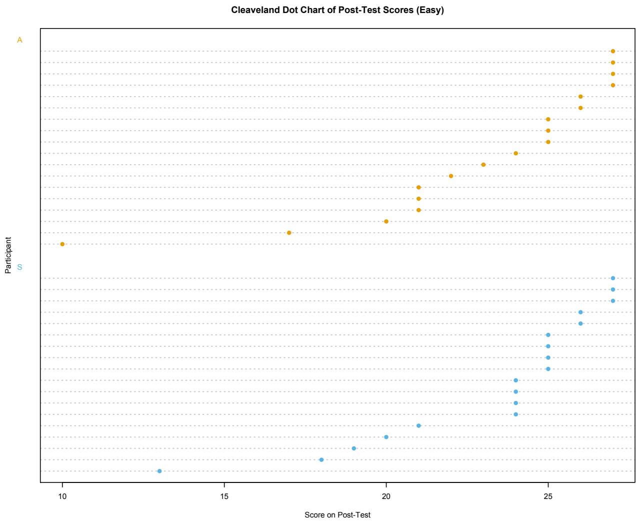


Figure 88 - Diagnostic Outlier Plot, Easy Words

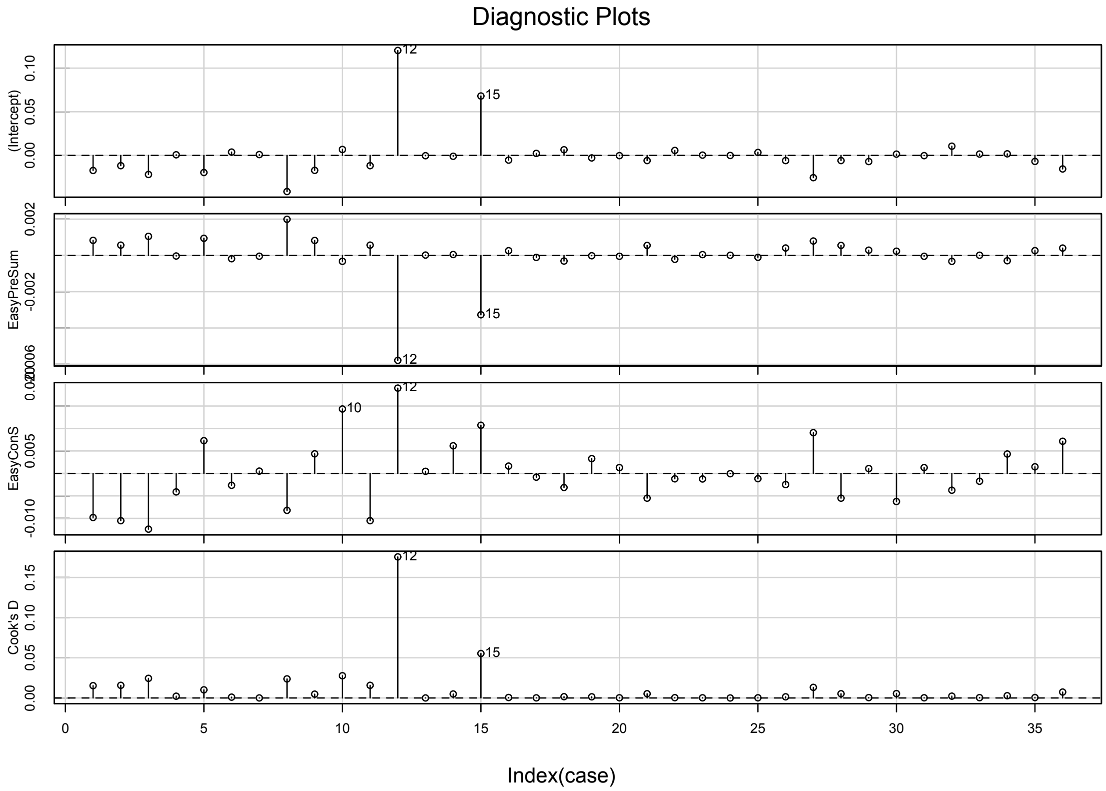


Table 22 - Table of Regression Coefficients After Removal of Outliers, Easy Words

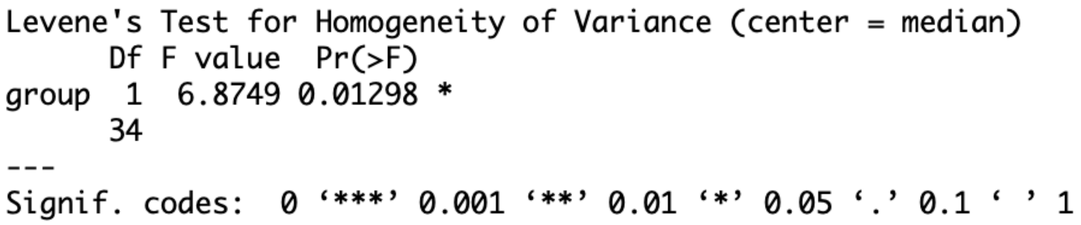
Observation Index	***Intercept***	***Pre-Test***	***Spell Check***
12	*2.5946*	*0.0271*	*0.0065*
Full-Model	*2.4744*	*0.0329*	*-0.0125*

This table summarizes the regression coefficients estimated by the model when the corresponding observed outlier is removed.

D.2.2 Model Validation

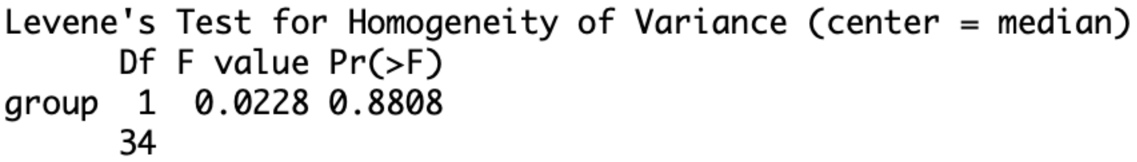
D.2.2.1 Heterogeneity

Figure 89- Levene's Test, Easy Words



```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)  
group  1  6.8749 0.01298 *
      34                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 90- Levene's Test, Hard Words



```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1  0.0228 0.8808
      34               
```

D.2.2.2 Residual Plots

Figure 91- Pearson Residuals from Easy Model vs Predictor Variable: Pre-Test

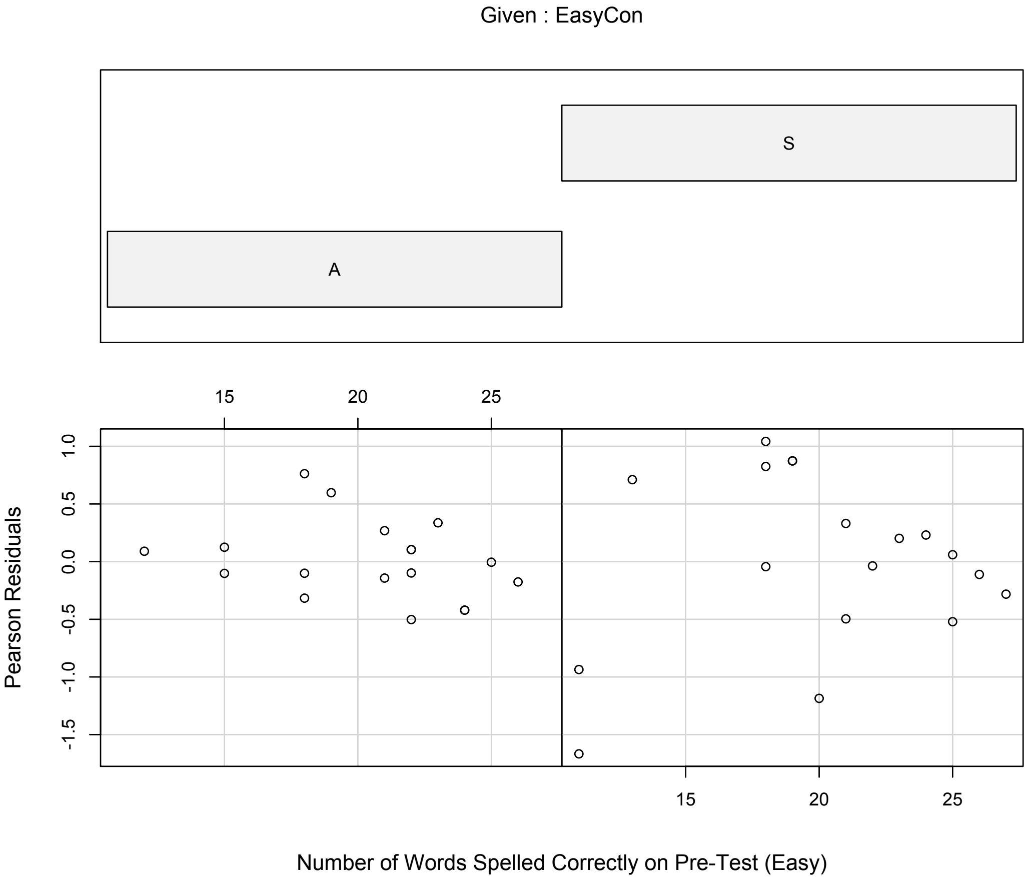


Figure 92- Pearson Residuals from Easy Model vs Predicted Post-Test Scores

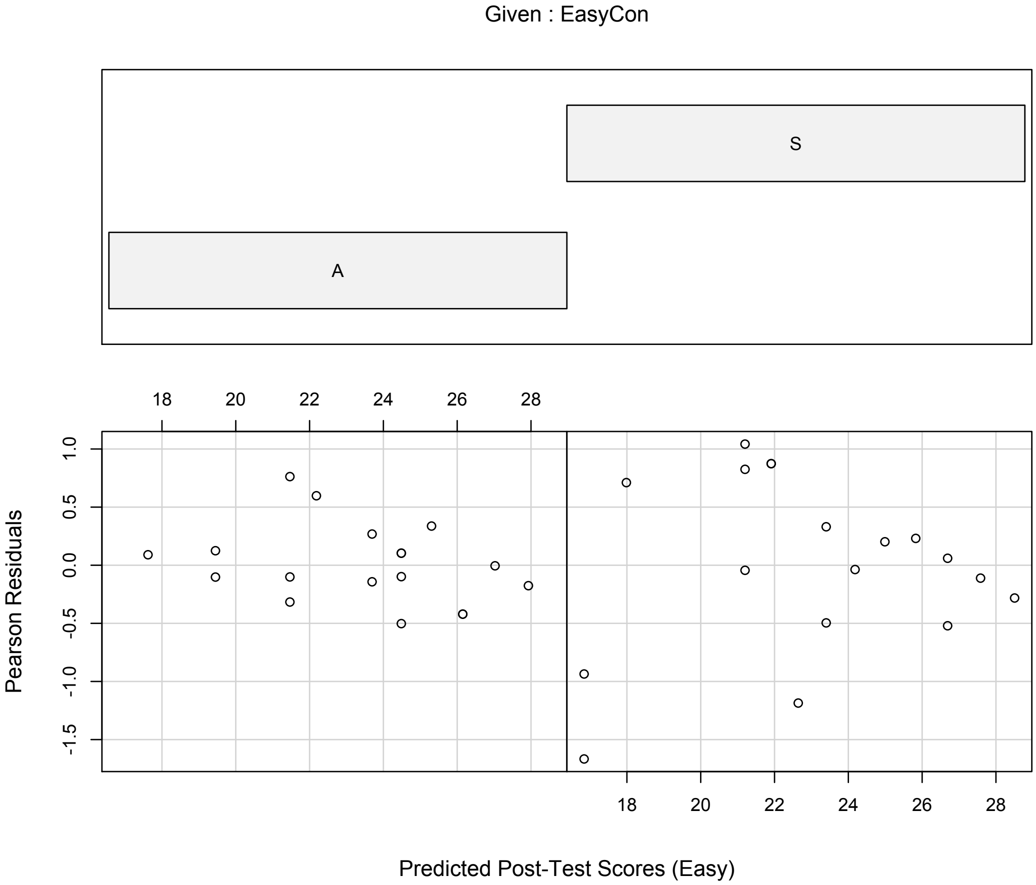


Figure 93 - Pearson Residuals from Hard Model vs Predicted Post-Test Scores

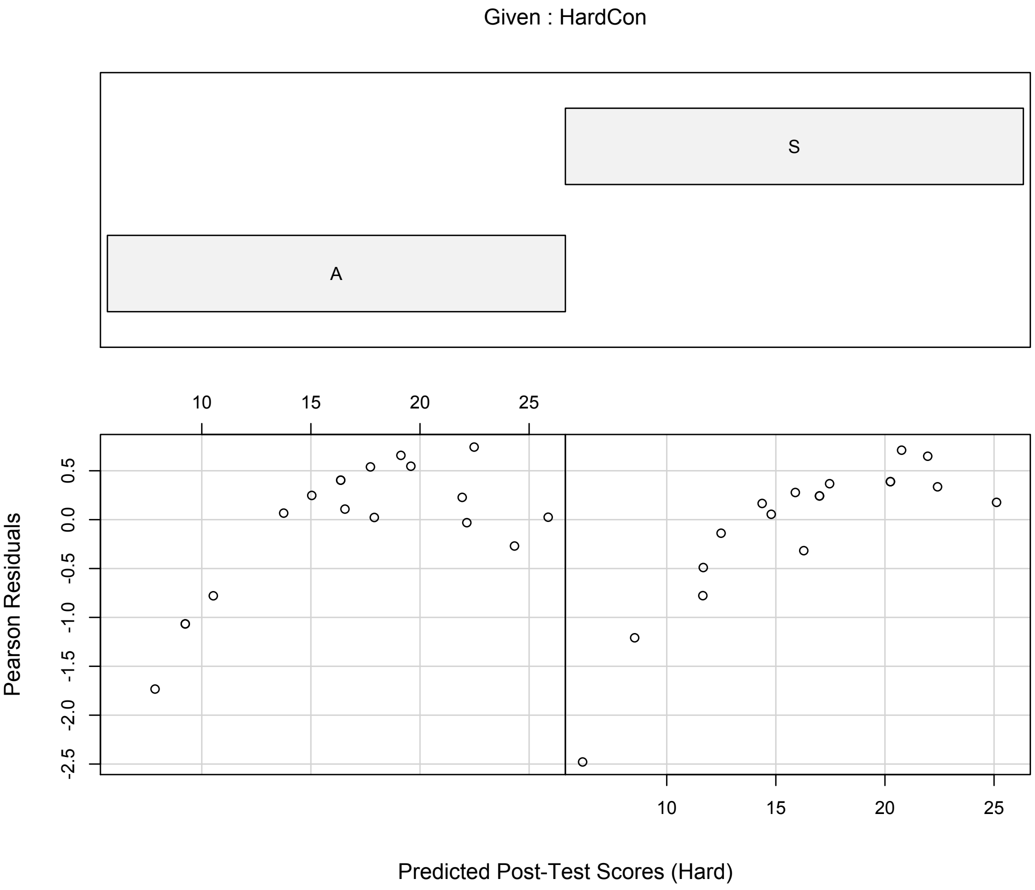

Given : HardCon
S
A
10
15
20
25
Pearson Residuals
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
10
15
20
25
Predicted Post-Test Scores (Hard)

Figure 94 - Pearson Residuals from Hard Model vs Predictor: Pre-Test

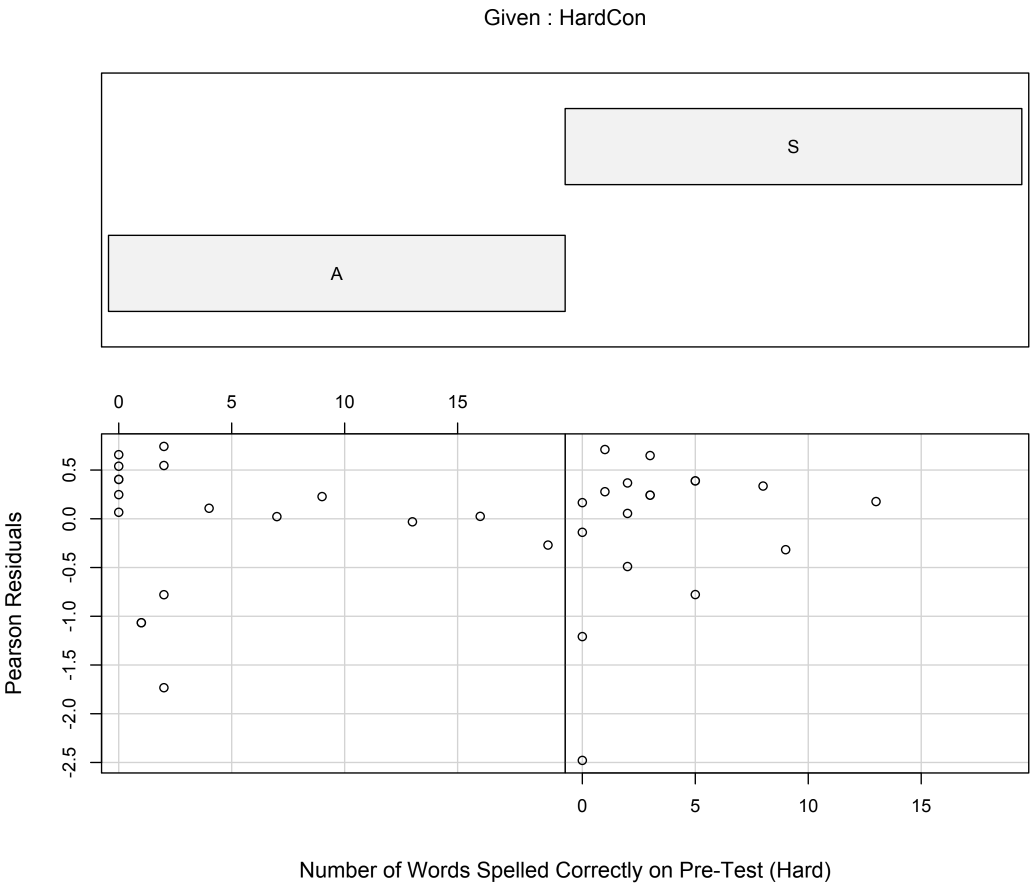


Figure 95 - Pearson Residuals from Easy Model vs Non-Model Co-Variate

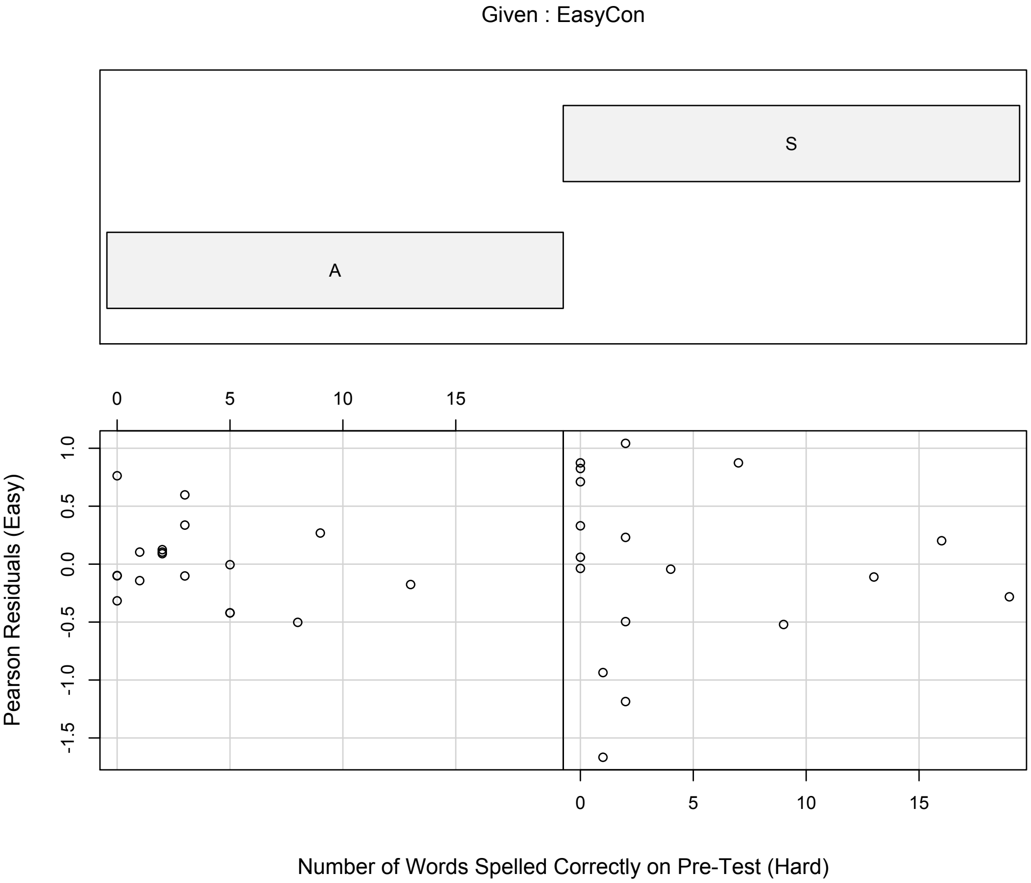
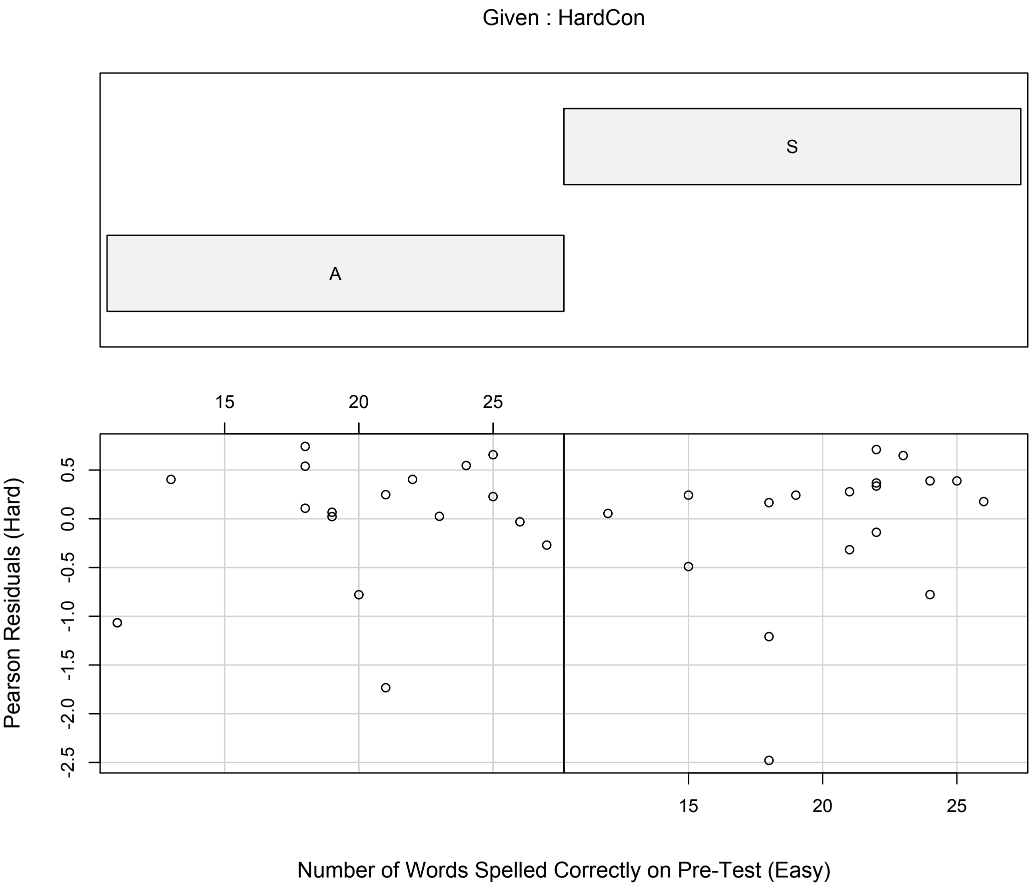
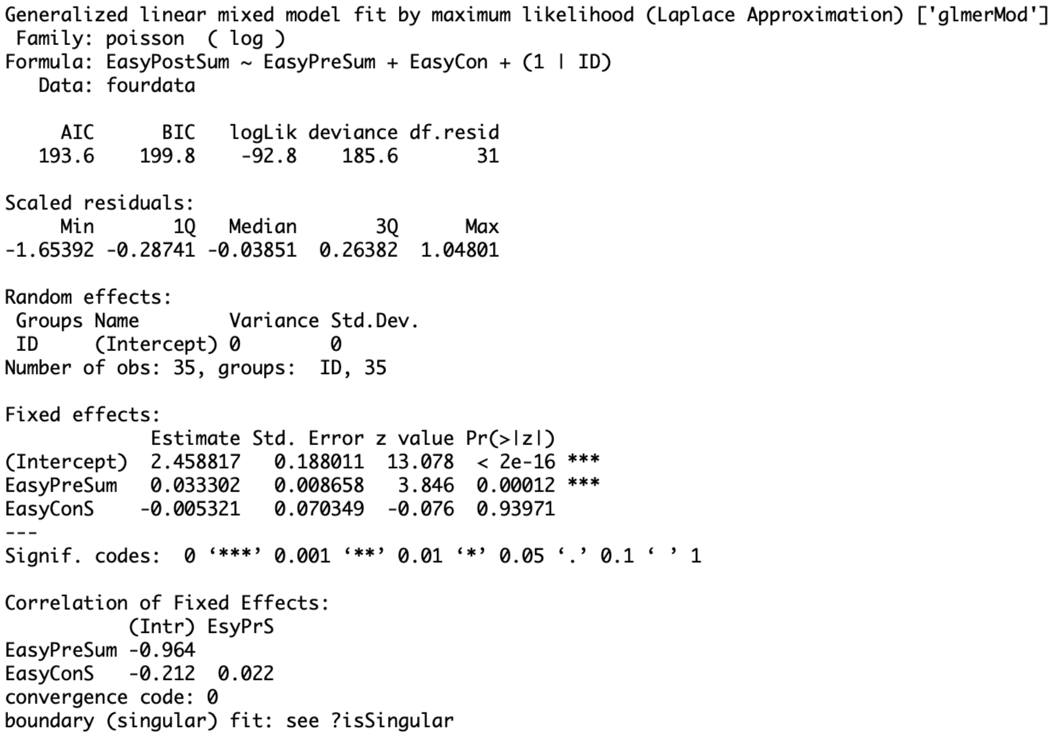


Figure 96 - Pearson Residuals from Hard Model vs Non-Model Co-Variate



D.2.3 Model Fit

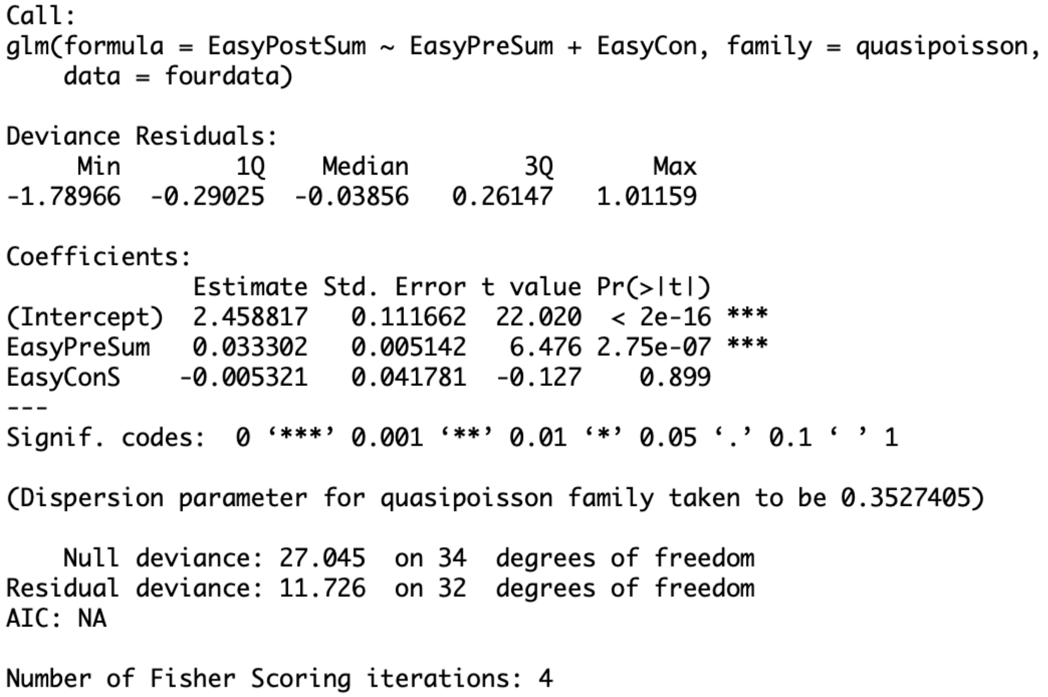
Figure 97- Poisson Regression with ORLE for Easy Words



```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: EasyPostSum ~ EasyPreSum + EasyCon + (1 | ID)
   Data: fourdata

     AIC      BIC   logLik deviance df.resid
   193.6    199.8    -92.8    185.6       31

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.65392 -0.28741 -0.03851  0.26382  1.04801

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 0        0
Number of obs: 35, groups:  ID, 35

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.458817   0.188011  13.078  < 2e-16 ***
EasyPreSum   0.033302   0.008658   3.846  0.00012 ***
EasyConS    -0.005321   0.070349  -0.076  0.93971
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr) EsyPrS
EasyPreSum -0.964
EasyConS   -0.212  0.022
convergence code: 0
boundary (singular) fit: see ?isSingular
```

Results from the regression indicate a singular fit. This is likely from the near zero value for the random effect.

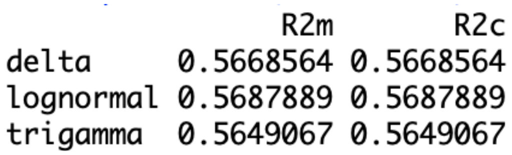
Figure 98 - QuasiPoisson Fit for Easy Words



```
Call:
glm(formula = EasyPostSum ~ EasyPreSum + EasyCon, family = quasipoisson,
    data = fourdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.78966  -0.29025  -0.03856   0.26147   1.01159

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.458817   0.111662  22.020  < 2e-16 ***
EasyPreSum   0.033302   0.005142   6.476 2.75e-07 ***
EasyConS    -0.005321   0.041781  -0.127    0.899
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 0.3527405)

    Null deviance: 27.045  on 34  degrees of freedom
Residual deviance: 11.726  on 32  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

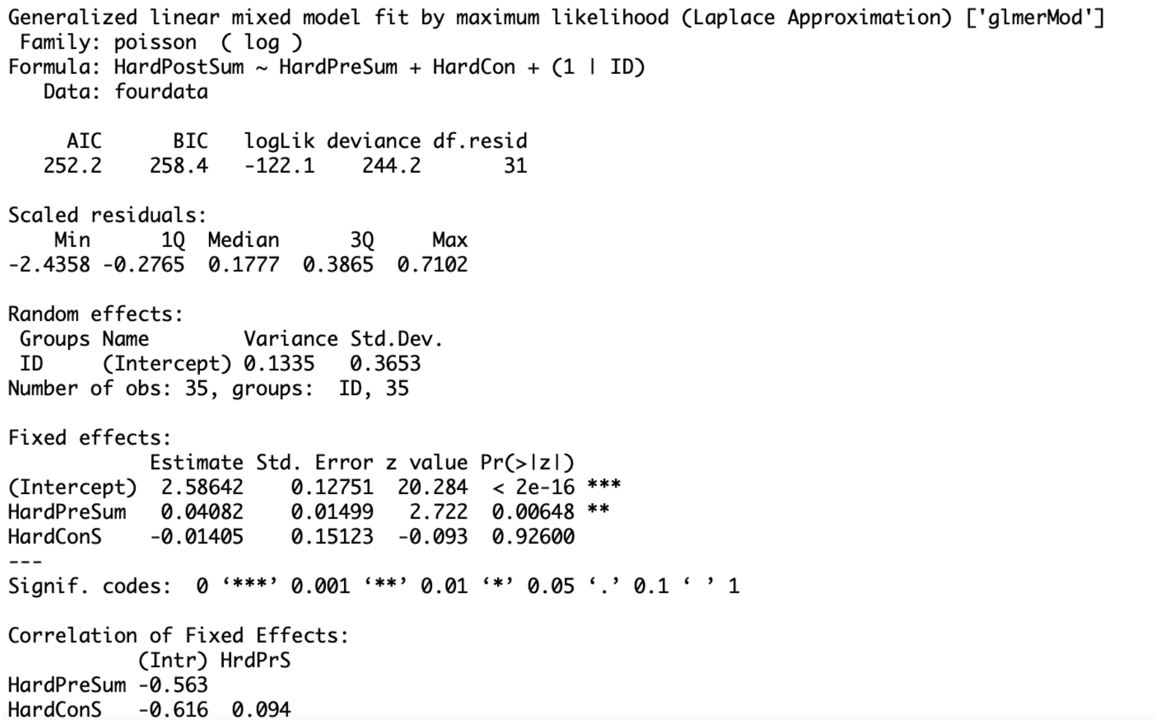
Fit after the detection of a singularity in the GLMM with OLRE for the set with easy words. Regression coefficients remain about the same. Pre-test score (EasyPreSum) significantly correlates with post-test scores. The type of device used during experimentally manipulated trials (EasyCon) does not significantly correlate with post-test scores.

Figure 99 - R^2 for Model of Easy Words



```
                R2m       R2c
delta     0.5668564 0.5668564
lognormal 0.5687889 0.5687889
trigamma  0.5649067 0.5649067
```

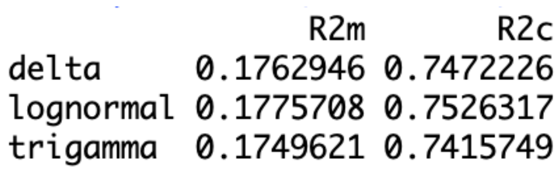
Figure 100- Poisson Regression with OLRE for Hard Words



```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: HardPostSum ~ HardPreSum + HardCon + (1 | ID)
   Data: fourdata

     AIC      BIC   logLik deviance df.resid
   252.2    258.4   -122.1    244.2       31

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4358 -0.2765  0.1777  0.3865  0.7102

Random effects:
 Groups Name        Variance Std.Dev.
 ID     (Intercept) 0.1335   0.3653
Number of obs: 35, groups:  ID, 35

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.58642    0.12751  20.284  < 2e-16 ***
HardPreSum   0.04082    0.01499   2.722  0.00648 **
HardConS    -0.01405    0.15123  -0.093  0.92600
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
           (Intr) HrdPrS
HardPreSum -0.563
HardConS   -0.616  0.094
```

Fit indicates no errors. For the set that includes difficult words, pre-test scores (HardPreSum) significantly correlated with post-test scores. This was not the case for the type of device (HardCon), which was not significantly correlated with the post-test score.

Figure 101 - R^2 for Hard Word Model



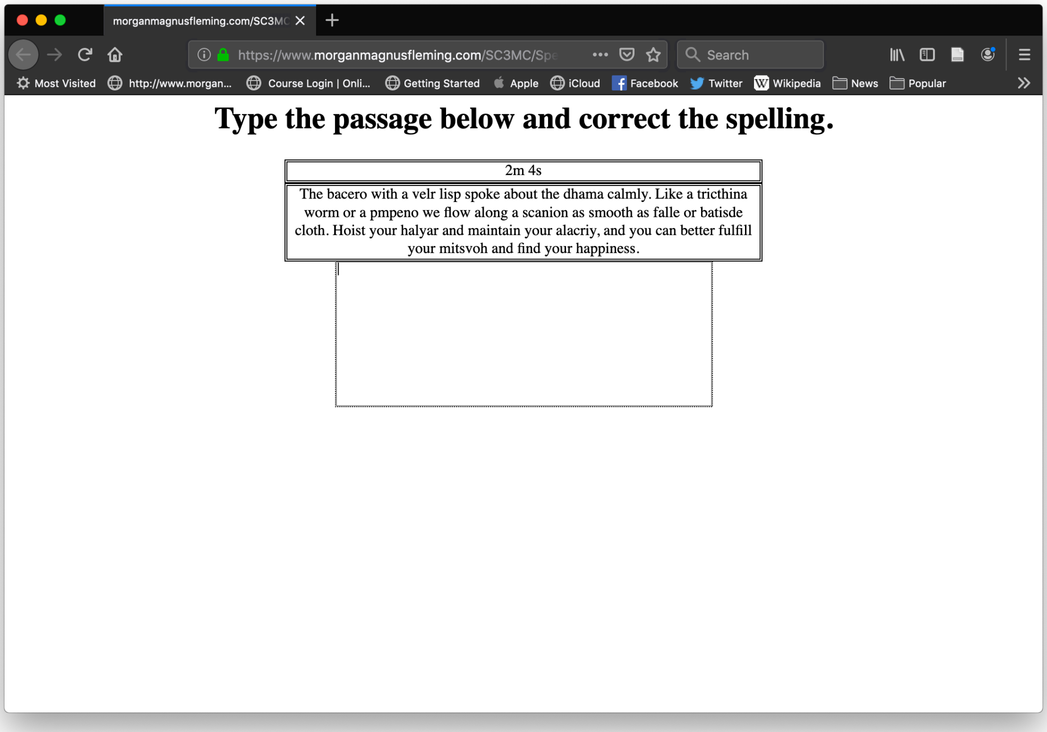
```
                R2m       R2c
delta     0.1762946 0.7472226
lognormal 0.1775708 0.7526317
trigamma  0.1749621 0.7415749
```

Appendix F

F.1 Materials

F.1.1 Website and Devices

Figure 102 - Screenshot of the Website Used to Gather Data

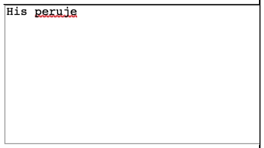
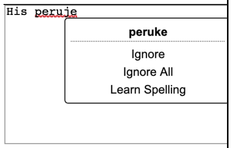
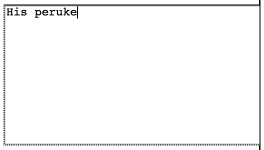
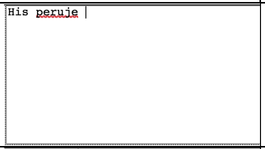
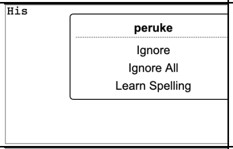
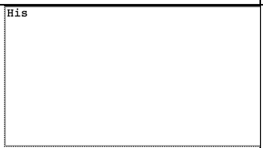
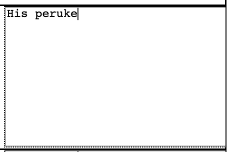
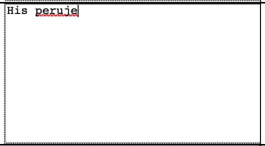
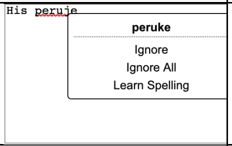
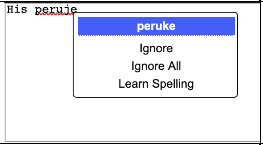
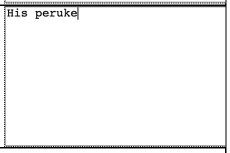
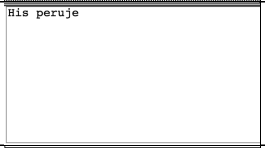
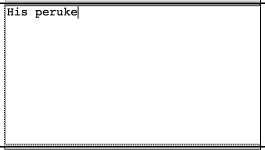
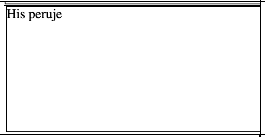
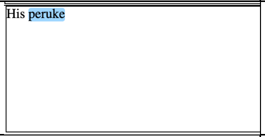


Type the passage below and correct the spelling.

2m 4s

The bacero with a velr lisp spoke about the dhama calmly. Like a tricthina worm or a pmpeno we flow along a scanion as smooth as falle or batisde cloth. Hoist your halyar and maintain your alacriy, and you can better fulfill your mitsvoh and find your happiness.

Table 23 - Phases of the Spelling Assistance Process

Condition	Identification Assistance	Selection Assistance	Execution	Extra Keyboard Entry
Spell-Check	 His peruje	 His peruje	 His peruke	None
Full-Word	 His peruje	 His	 His	 His peruke
Part-Word	 His peruje	 His peruje	 His peruje	 His peruke
Autocorrect	 His peruje	Handled by System	 His peruke	None
Autohighlight	 His peruje	Handled by System	 His peruke	None
Control	None	None	None	None
Complete	All words are spelled correctly	All words are spelled correctly	All words are spelled correctly	None

This table summarizes the steps the user takes in correcting a spelling error with the five support systems (traditional spell-check, full-word, part-word, autocorrect, autohighlight) or in either of the control conditions (control, complete). Below is a description of the steps for the spell-checking systems, a description of the steps for the autocorrect systems, and finally, a description of the steps for the 'control' and 'complete' conditions.

In the identification step, the traditional spell-check, part-word, and full-word systems all provide a red underline to help the user identify misspelled words. The user can then access the selection assistance features by right-clicking on a word with a red underline. When an underlined word is right-clicked by the user, the spell-check, part-word, and full-word systems all provide the user with a context menu containing correctly spelled words. The full-word system additionally deletes the word that the user right-clicked on. This extra step was taken in order to ensure that users have to enter the correctly spelled word for themselves in the execution step. In the traditional spell-checking system, the selection process can be completed by selecting a proper spelling from the context menu.

The execution step proceeds directly from the selection step. Upon selecting a proper spelling in the traditional spell-checking system, the edit is executed automatically for the *user. For both full-word and part-word systems, selecting a word from the context menu produces no change in the system. Instead, the user can only visually select a proper spelling.*

For the autocorrect and autohighlight systems, the majority of steps are handled by the system. The systems complete the spelling process with no prompting or necessary interaction to proceed through the steps. The system identifies the misspelling, selects a candidate word, and then executes the edit for the user. A critical difference between autocorrect and autohighlight lies in the execution step. Upon replacing a misspelled word, the autohighlight system highlights the word that was replaced. This highlight remains on-screen for the remainder of the trial.

In the 'control' and 'complete' conditions, participants were not offered any support systems. They were also asked not to use outside resources of any sort. The passages in the control condition were identical to the passages in the traditional spell-check, full-word, part-word, autocorrect, and autohighlight conditions. The passages in the complete condition only contained correctly spelled words. This difference in the passages distinguishes the control and complete conditions across the identification, selection, and execution steps. Participants in the control condition had to identify the misspelled words, select proper spellings from the proper spellings they knew of, and then correctly spell the words themselves. Participants in the 'complete' condition were provided proper spellings, ensuring that there were no misspellings that they had to identify. They did not need to select proper spellings for any misspellings, and correctly spelling the words consisted of merely copying what was already contained in the passage.

F.1.2 Stimuli

F.1.2.1 Target Words

Table 24 - Target Words

Misspellings	Proper Spelling
brasero	bracero
vellar	velar
darma	dharma
trychina	trichina
pompeno	pompano
scanscion	scansion
faillet	faille
bateste	batiste
hallyard	halyard
alacrety	alacrity
mitsvah	mitzvah
aballone	abalone
chrore	crore
ewwer	ewer
chanchre	chancre
sashey	sashay
centavvo	centavo
mackenaw	mackinaw
paruke	peruke
dellft	delft
kuay	quay
wildebeast	wildebeest
selesta	celesta
eloadea	elodea
ayuah	ayah
xeolite	zeolite
gymkana	gymkhana

F.1.2.2 Passage 1

"The brasero with a vellar lisp spoke about the darma calmly. Like a trychina worm or a pompeno we flow along a scanscion as smooth as faillet or bateste cloth. Hoist your hallyard and maintain your alacrety, and you can better fulfill your mitsvah and find your happiness."

F.1.2.3 Passage 2

"He scooped out the aballone, now worth a chrore and a half. The medicine in the ewwer eased the chanchre plaguing the patient. Sashey for a centavvo was too low of a price. Upon the patient's front was their mackenaw. "

F.1.2.4 Passage 3

“His paruke toppled into the dellft bowl on the ground, just above the kuay below. A wildebeast herd outside stamped about like a selesta, and eloadea sat just under the surface of the water. The ayuah went to the shelf to grab the xeolite to the tune of that gymkana from nature. “

F.1.2.5 Transfer Test Passage

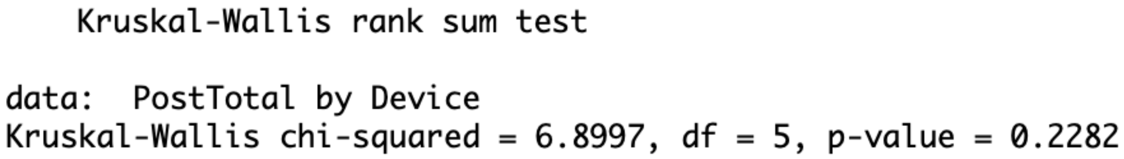
“The paruke upon the ayuah hid a chanchre shaped like an aballone. Under the injury, thoughts of the year's pompeno harvest flowed freely. They sat above a mouth known for its vellar annunciations that were as smooth as bateste cloth. The darma that guided this soul played the bones and tendons like a beautiful selesta when pleased, and like a raucous gymkana when not. Below the soul swept vast fields of eloadea, hidden under the dellft like reflections of the sea. The soul took a swig from their ewwer with an alacrety that spoke to their thirst. The faillet hallyard was drawn tight, like a mackenaw wrapped around a particularly rotund individual. The wildebeast meat would have been a nice accouterment, but that was lost to the kuay the soul had launched from. The soul's brasero companion whittled at a xeolite figure he'd been carving. It was his mitsvah, he claimed. More like the goading of some trychina meme, scoffed the soul. At the sound of a loud bang, both parties surveyed the surroundings, like a scanscion in preparation for the crescendo. "I bet you a centavvo you don't see what I see," the soul said with a sashey. "Make it a chrore, and you're on," replied the companion.”

F.2 Model Verification

F.2.1 Data Exploration

F.2.1.1 Device Assessment

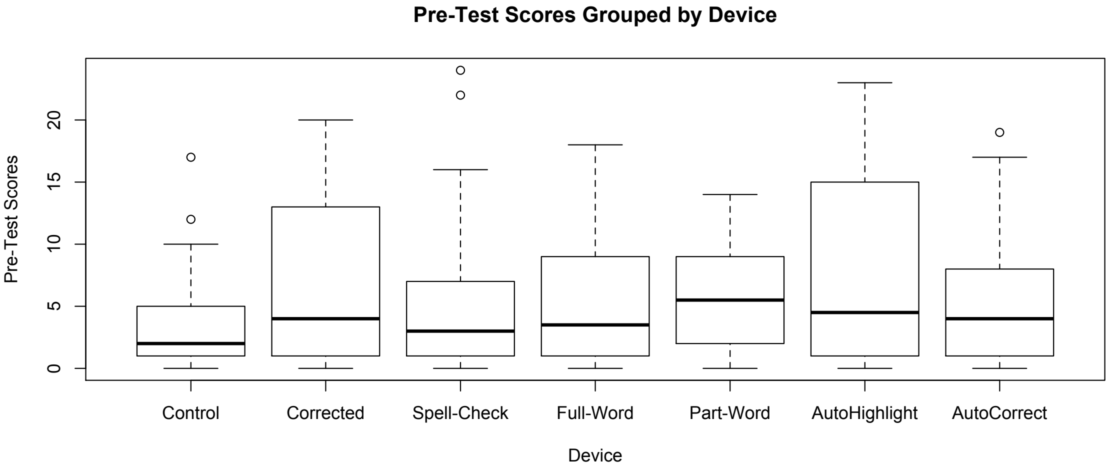
Figure 103 – Rank Sum Test (includes Corrected Control)



```
	Kruskal-Wallis rank sum test

data:  PostTotal by Device
Kruskal-Wallis chi-squared = 6.8997, df = 5, p-value = 0.2282
```

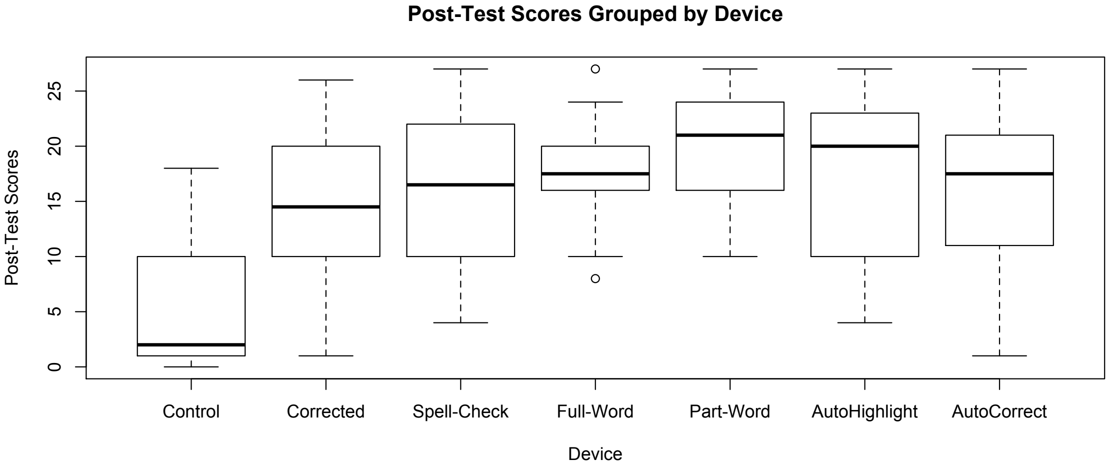
F.2.1.2 Distributions

Figure 104 - Pre-Test Scores Grouped by Device



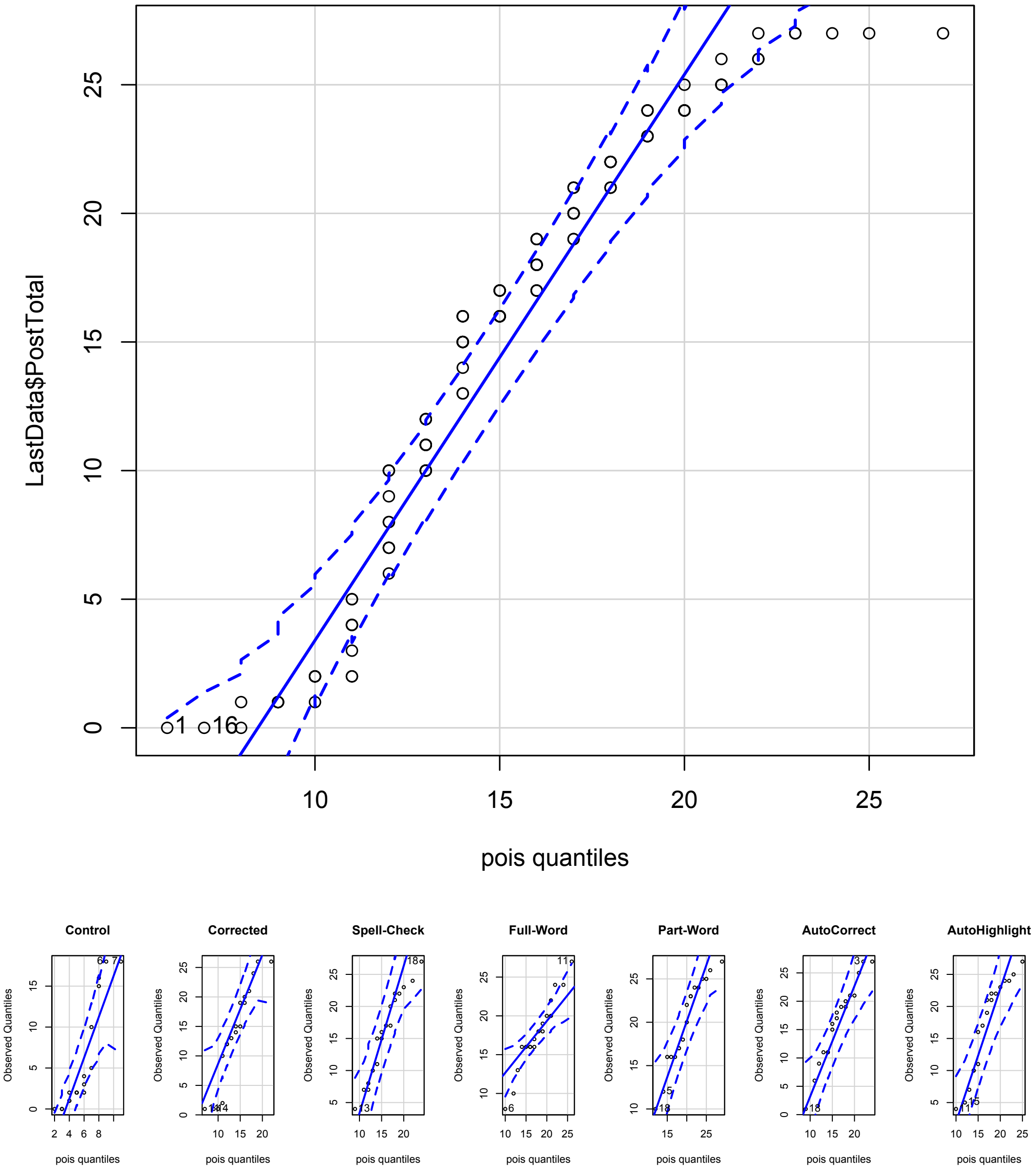
This is a diagram of pre-test scores for each condition. Participants in the autohighlight condition had the greatest variation in pretest scores, though traditional spell-check contained the highest performer overall.

Figure 105 - Post-Test Scores Grouped by Device



This is a diagram of post-test scores for each condition. Participants in the part-word condition showed the highest mean score on the post-test trials

Figure 106 – Q-Q Plot of Estimated Poisson Distribution and Observed Values, by Condition



Scores across all conditions (pictured: top) approximate the theoretical poisson distribution. Exceptions include an abundance of

F.2.1.3 Outliers

Figure 107- Cleaveland Dot Chart of Pre-Test Scores

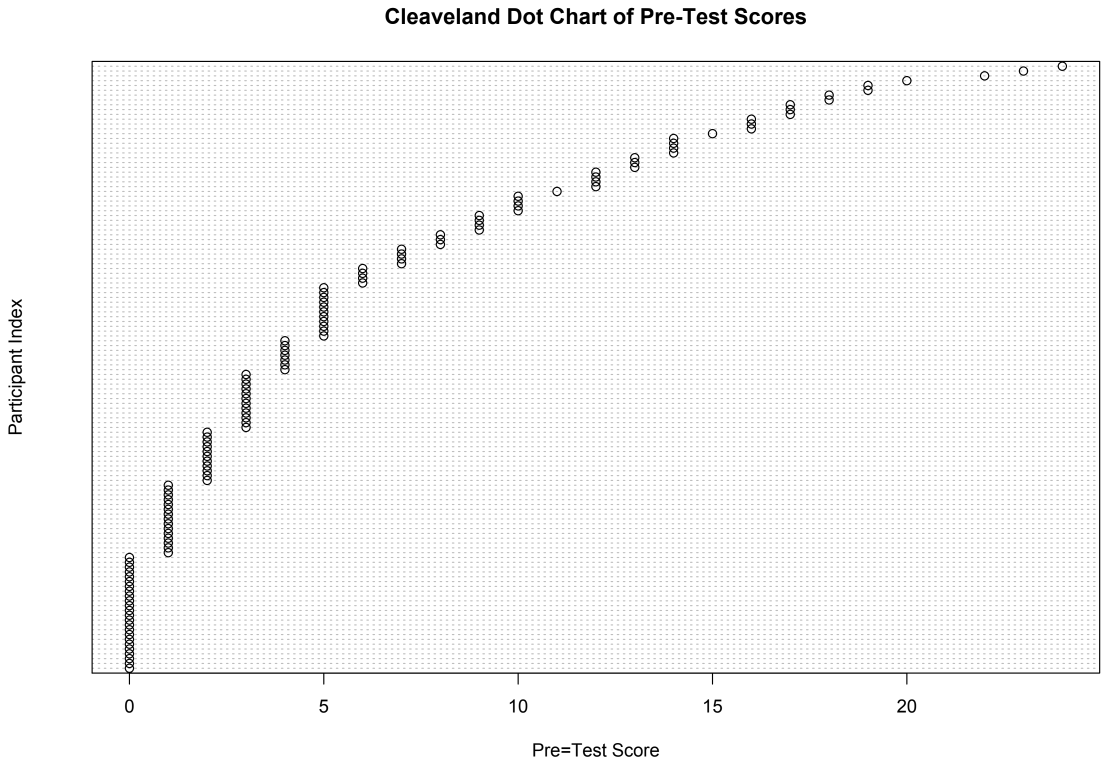


Figure 108 - Cleaveland Dot Chart of Pre-Test Scores by Condition

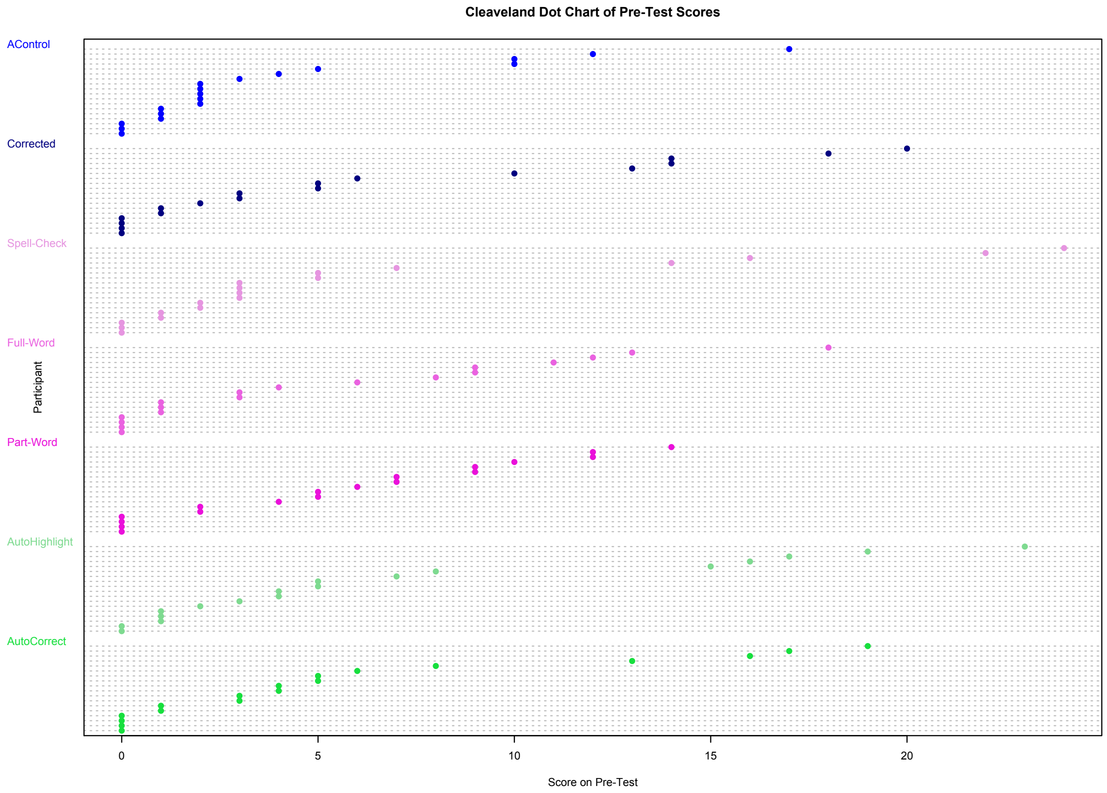


Figure 109- Cleaveland Dot Chart of Post-Test Scores by Condition

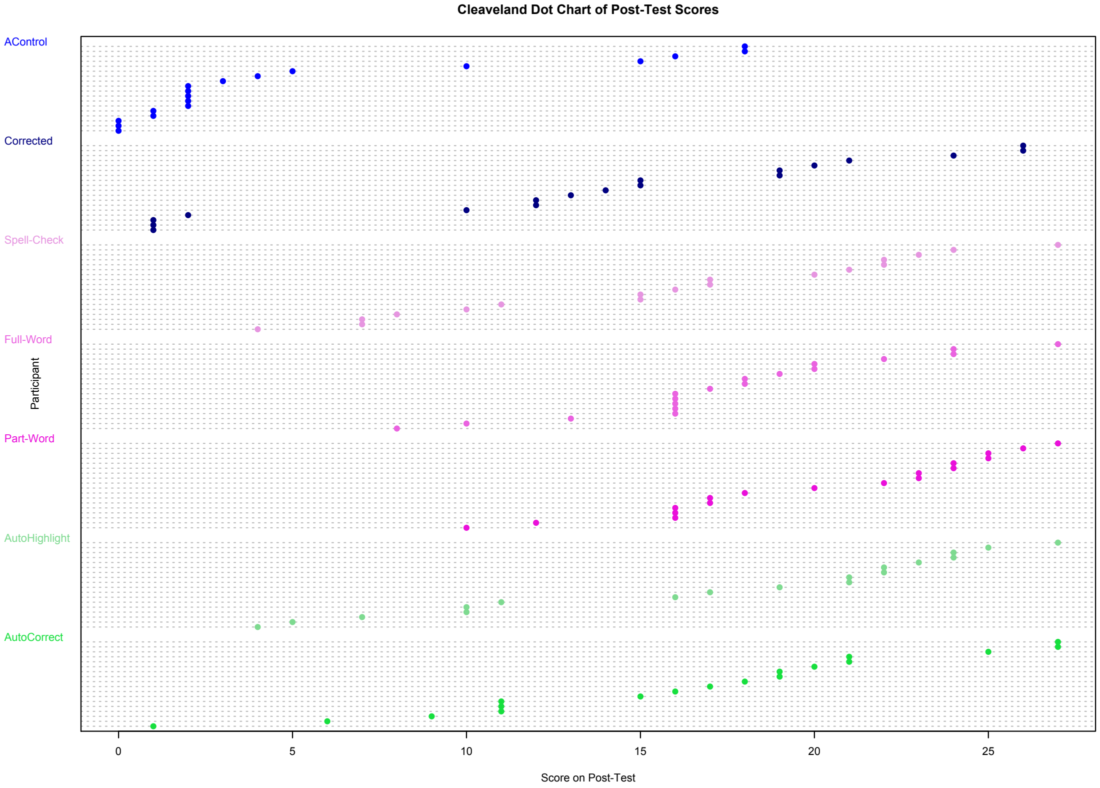


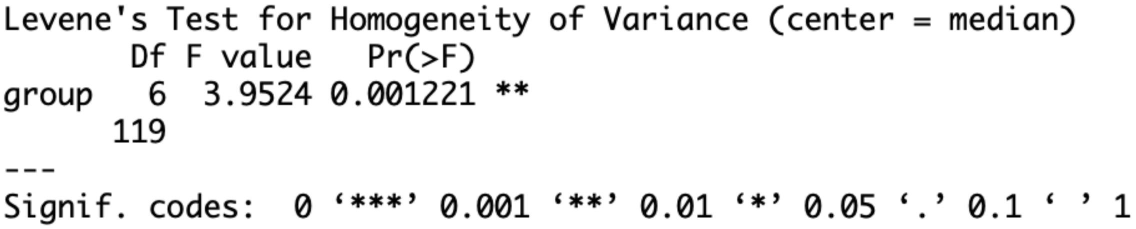
Table 25 - Parameter Estimates with Outliers Removed

Obs. Index	Intercept	PreTotal	Corrected	Spell-Check	Full-Word	Part-Word	Auto Highlight	Auto Correct
6	1 .2211	0.0486	0.9860	1.1649	1.3489	1.4557	1.1771	1.2136
7	1.2529	0.0471	0.9634	1.1409	1.3252	1.4333	1.1563	1.1877
17	1.2805	0.0474	0.9299	1.1098	1.2952	1.4032	1.1246	1.1563
29	1.3487	0.0485	0.8630	1.0329	1.2199	1.3277	1.0462	1.0797
32	1.3731	0.0467	0.8960	1.0234	1.2072	1.3153	1.0394	1.0701
36	1.3732	0.0468	0.8971	1.0227	1.2065	1.3146	1.0385	1.0695
124	1.3771	0.0466	0.8455	1.0211	1.2044	1.3125	1.0373	1.1211
Model	1.3498	0.0485	0.8507	1.0319	1.2189	1.3266	1.0452	1.0788

F.2.2 Model Validation

F.2.2.1 Heterogeneity

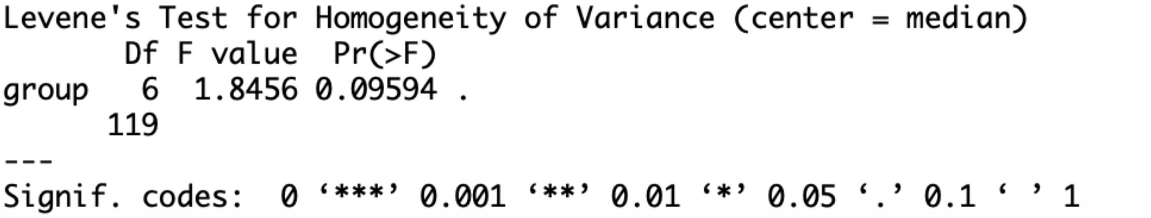
Figure 110 - Levene's Test for Homogeneity of Variance for GLMM



```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   6  3.9524 0.001221 **
      119
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Heterogenity was detected in this Levene Test of the residuals of the GLMM. Results indicate that the variance between estimated scores on the post-test did not remain the same across participants who used different tools (variable "Device" in the over-all model).

Figure 111 - Levene's Test for Homogeneity of Variance for Quasipoisson GLM

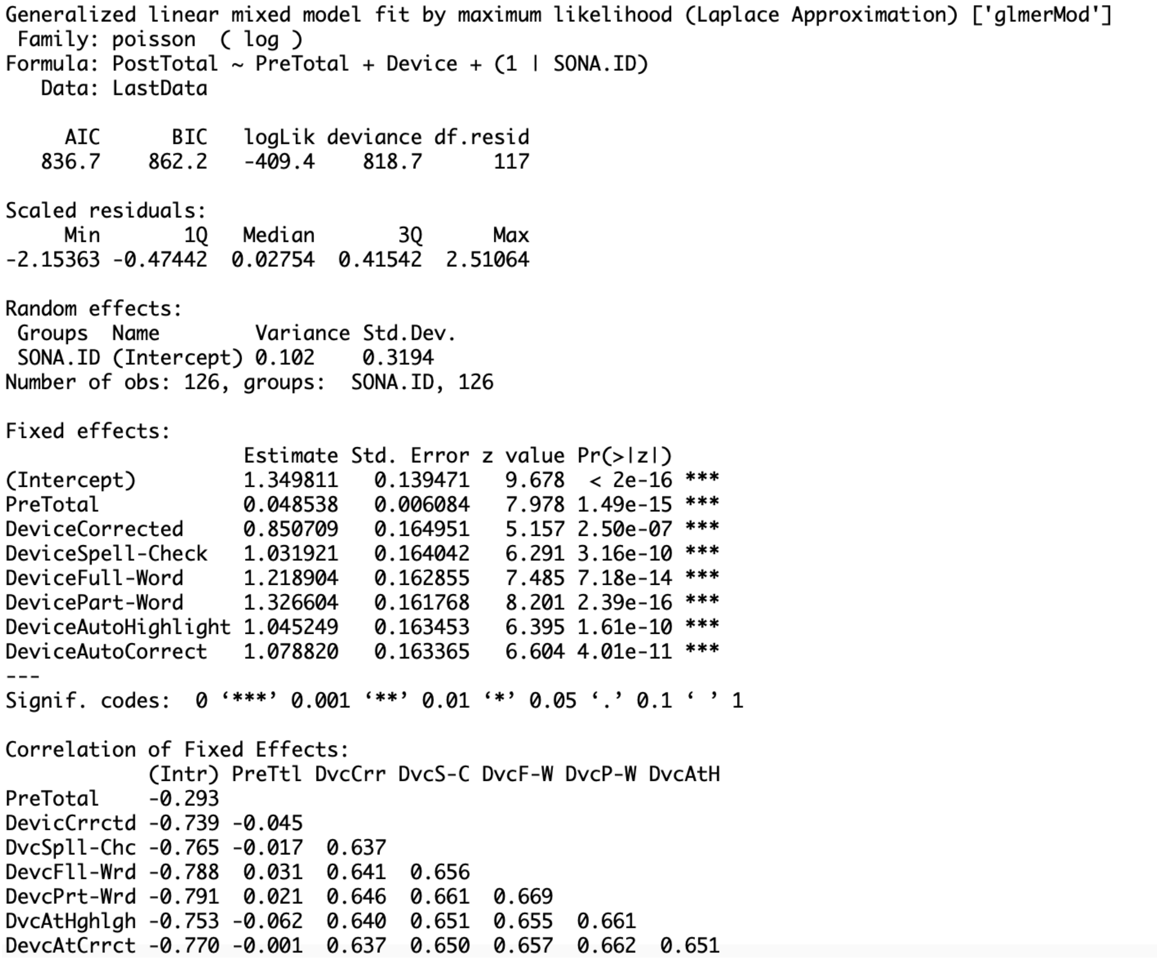


```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   6  1.8456 0.09594 .
      119
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In order to ensure that the results of the GLMM were not a spurious result of the GLMM's violation of homogeneity across different levels of "Device", a Quasipoisson GLM without an ORLE term was fit and the results were compared between the Quasipoisson GLM and the GLMM. Results indicate that the Quasipoisson GLM did not violate the assumption of homogeneity, and results from the GLM and GLMM will be compared below.

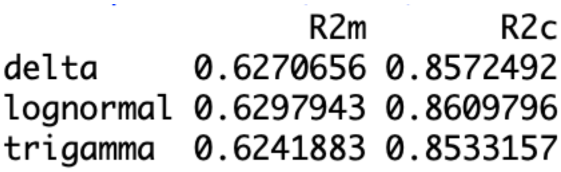
F.2.3 Model Fit

Figure 112- Generalized Poisson Mixed Model for Post-Test Scores



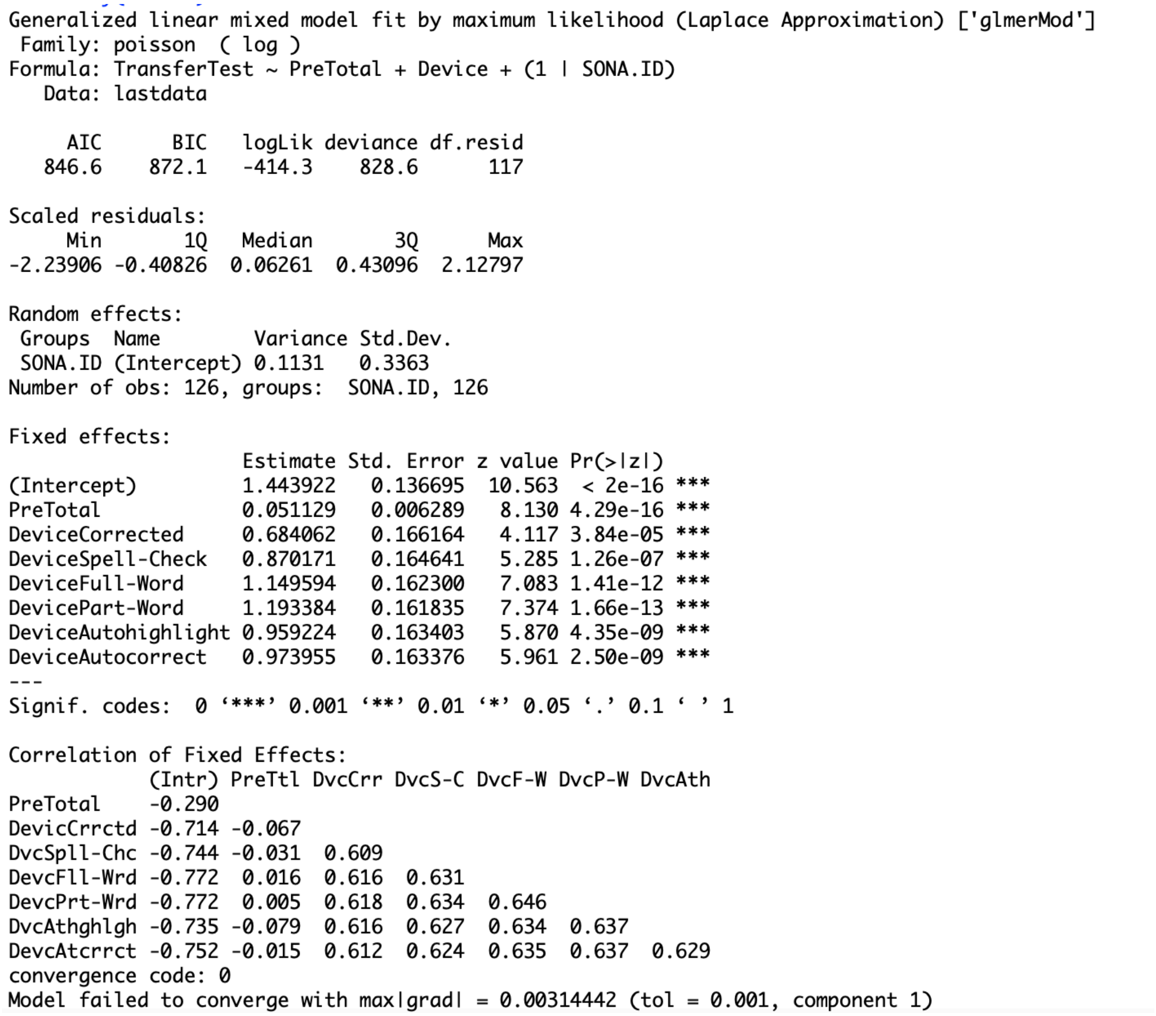
```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: PostTotal ~ PreTotal + Device + (1 | SONA.ID)
   Data: LastData

     AIC      BIC   logLik deviance df.resid
   836.7    862.2   -409.4    818.7      117

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.15363 -0.47442  0.02754  0.41542  2.51064

Random effects:
 Groups  Name        Variance Std.Dev.
 SONA.ID (Intercept) 0.102    0.3194
Number of obs: 126, groups:  SONA.ID, 126

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.349811   0.139471   9.678  < 2e-16 ***
PreTotal            0.048538   0.006084   7.978 1.49e-15 ***
DeviceCorrected     0.850709   0.164951   5.157 2.50e-07 ***
DeviceSpell-Check   1.031921   0.164042   6.291 3.16e-10 ***
DeviceFull-Word     1.218904   0.162855   7.485 7.18e-14 ***
DevicePart-Word     1.326604   0.161768   8.201 2.39e-16 ***
DeviceAutoHighlight 1.045249   0.163453   6.395 1.61e-10 ***
DeviceAutoCorrect   1.078820   0.163365   6.604 4.01e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) PreTtl DvcCrr DvcS-C DvcF-W DvcP-W DvcAtH
PreTotal    -0.293
DevicCrrctd -0.739 -0.045
DvcSpll-Chc -0.765 -0.017  0.637
DevcFll-Wrd -0.788  0.031  0.641  0.656
DevcPrt-Wrd -0.791  0.021  0.646  0.661  0.669
DvcAtHghlgh -0.753 -0.062  0.640  0.651  0.655  0.661
DevcAtCrrct -0.770 -0.001  0.637  0.650  0.657  0.662  0.651
```

Figure 113 - R^2 for Generalized Poisson Mixed Model (Post-Test Scores)



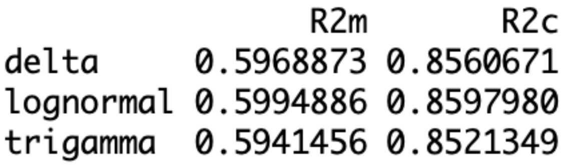
```
                R2m       R2c
delta     0.6270656 0.8572492
lognormal 0.6297943 0.8609796
trigamma  0.6241883 0.8533157
```

Figure 114 – Generalized Poisson Mixed Model for Transfer Test Scores



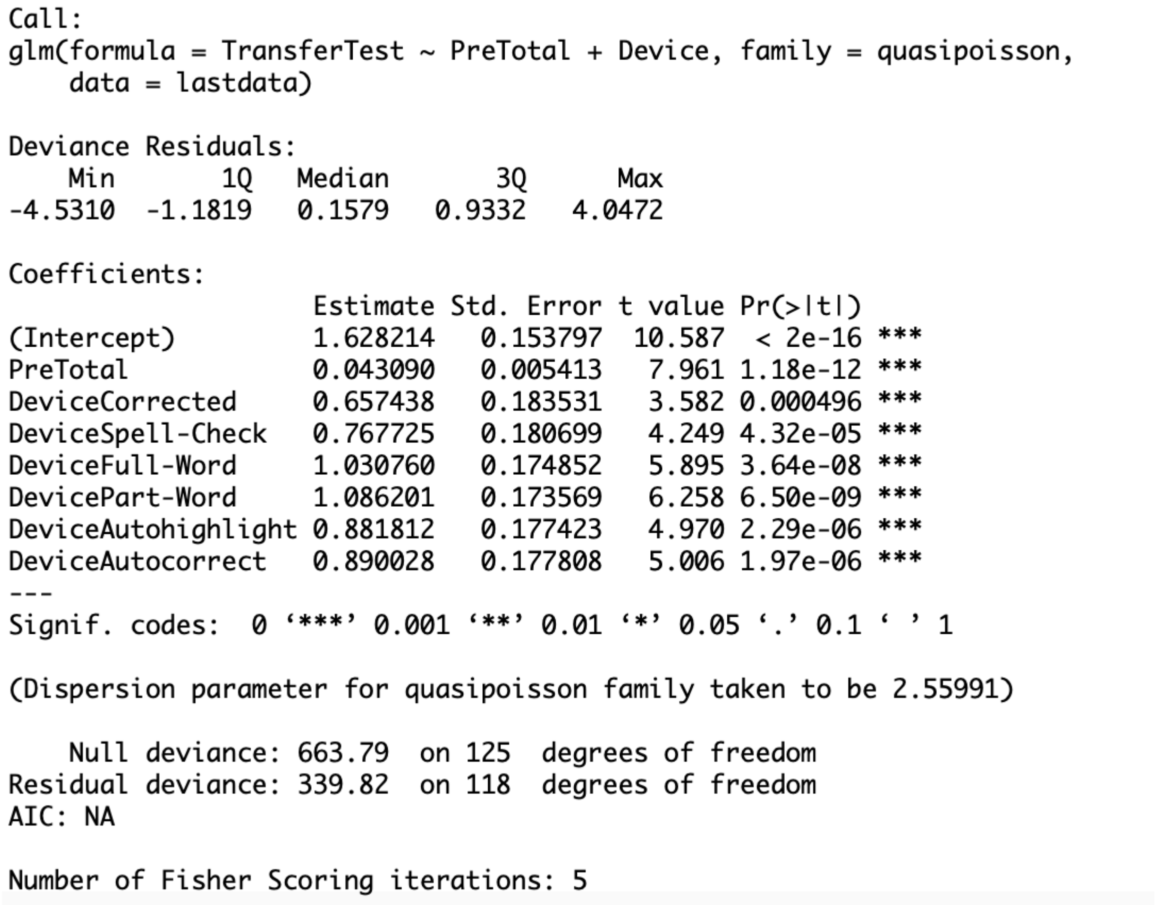
```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: TransferTest ~ PreTotal + Device + (1 | SONA.ID)
   Data: lastdata

     AIC      BIC   logLik deviance df.resid
   846.6    872.1   -414.3    828.6      117

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.23906 -0.40826  0.06261  0.43096  2.12797

Random effects:
 Groups  Name        Variance Std.Dev.
 SONA.ID (Intercept) 0.1131   0.3363
Number of obs: 126, groups:  SONA.ID, 126

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.443922   0.136695  10.563  < 2e-16 ***
PreTotal            0.051129   0.006289   8.130 4.29e-16 ***
DeviceCorrected     0.684062   0.166164   4.117 3.84e-05 ***
DeviceSpell-Check   0.870171   0.164641   5.285 1.26e-07 ***
DeviceFull-Word     1.149594   0.162300   7.083 1.41e-12 ***
DevicePart-Word     1.193384   0.161835   7.374 1.66e-13 ***
DeviceAutohighlight 0.959224   0.163403   5.870 4.35e-09 ***
DeviceAutocorrect   0.973955   0.163376   5.961 2.50e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) PreTtl DvcCrr DvcS-C DvcF-W DvcP-W DvcAth
PreTotal    -0.290
DevicCrrctd -0.714 -0.067
DvcSpll-Chc -0.744 -0.031  0.609
DevcFll-Wrd -0.772  0.016  0.616  0.631
DevcPrt-Wrd -0.772  0.005  0.618  0.634  0.646
DvcAthghlgh -0.735 -0.079  0.616  0.627  0.634  0.637
DevcAtcrrct -0.752 -0.015  0.612  0.624  0.635  0.637  0.629
convergence code: 0
Model failed to converge with max|grad| = 0.00314442 (tol = 0.001, component 1)
```

Figure 115 – R^2 for Generalized Poisson Mixed Model (Transfer Test)



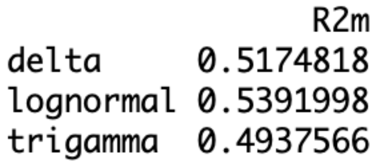
```
                R2m       R2c
delta     0.5968873 0.8560671
lognormal 0.5994886 0.8597980
trigamma  0.5941456 0.8521349
```

Figure 116 -Quasi-Poisson Model for Transfer Test Scores



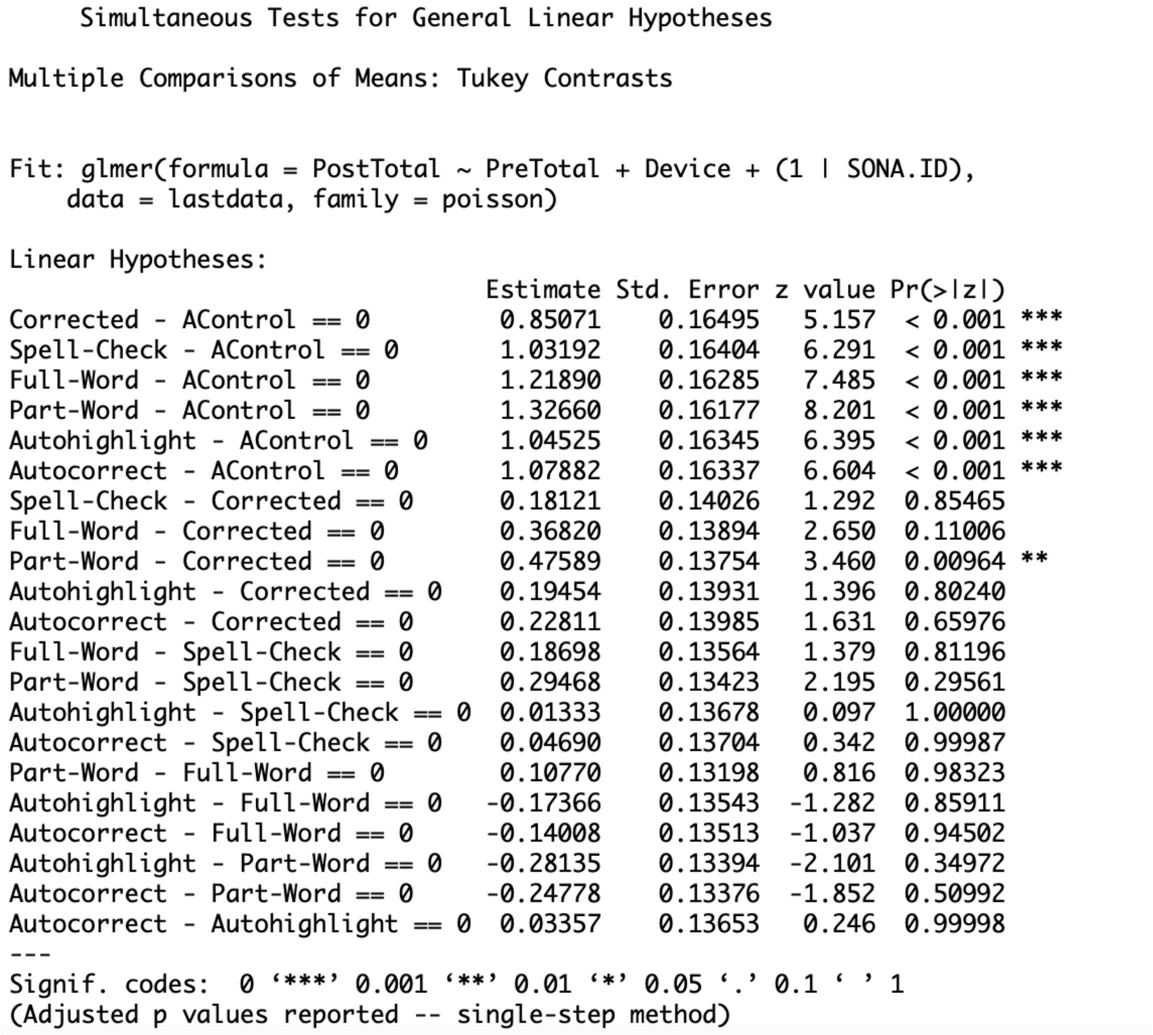
```
Call:
glm(formula = TransferTest ~ PreTotal + Device, family = quasipoisson,
    data = lastdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5310  -1.1819   0.1579   0.9332   4.0472

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.628214   0.153797  10.587  < 2e-16 ***
PreTotal            0.043090   0.005413   7.961 1.18e-12 ***
DeviceCorrected     0.657438   0.183531   3.582 0.000496 ***
DeviceSpell-Check   0.767725   0.180699   4.249 4.32e-05 ***
DeviceFull-Word     1.030760   0.174852   5.895 3.64e-08 ***
DevicePart-Word     1.086201   0.173569   6.258 6.50e-09 ***
DeviceAutohighlight 0.881812   0.177423   4.970 2.29e-06 ***
DeviceAutocorrect   0.890028   0.177808   5.006 1.97e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 2.55991)

    Null deviance: 663.79  on 125  degrees of freedom
Residual deviance: 339.82  on 118  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Figure 117 - R^2 for Quasi-Poisson Model (Transfer Test Scores)



```
                R2m
delta     0.5174818
lognormal 0.5391998
trigamma  0.4937566
```

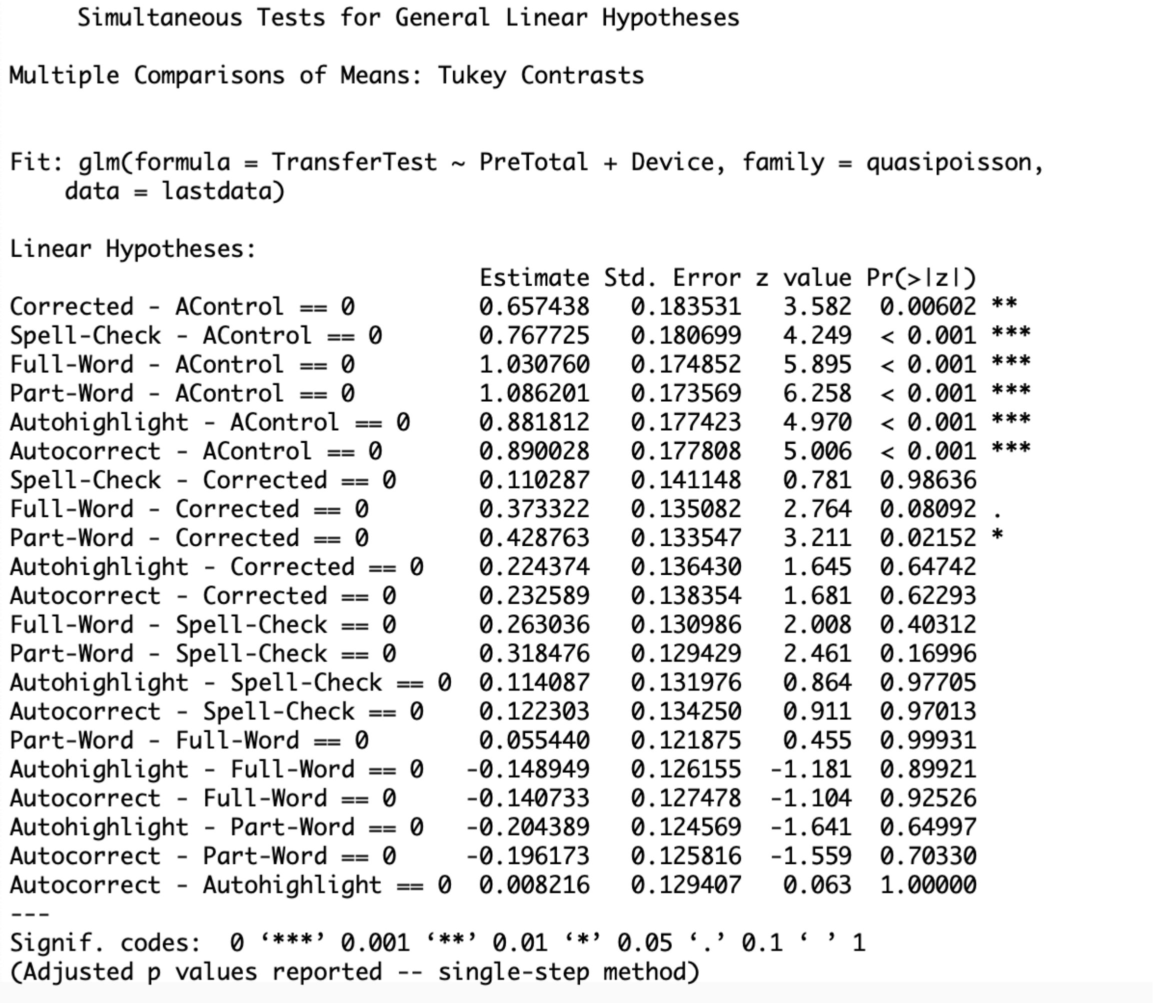
Figure 118 - Tukey Post-Hoc Test for Post-Test Scores



```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: glmer(formula = PostTotal ~ PreTotal + Device + (1 | SONA.ID),
    data = lastdata, family = poisson)

Linear Hypotheses:
                                 Estimate Std. Error z value Pr(>|z|)
Corrected - AControl == 0         0.85071    0.16495   5.157  < 0.001 ***
Spell-Check - AControl == 0       1.03192    0.16404   6.291  < 0.001 ***
Full-Word - AControl == 0         1.21890    0.16285   7.485  < 0.001 ***
Part-Word - AControl == 0         1.32660    0.16177   8.201  < 0.001 ***
Autohighlight - AControl == 0     1.04525    0.16345   6.395  < 0.001 ***
Autocorrect - AControl == 0       1.07882    0.16337   6.604  < 0.001 ***
Spell-Check - Corrected == 0      0.18121    0.14026   1.292  0.85465
Full-Word - Corrected == 0        0.36820    0.13894   2.650  0.11006
Part-Word - Corrected == 0        0.47589    0.13754   3.460  0.00964 **
Autohighlight - Corrected == 0    0.19454    0.13931   1.396  0.80240
Autocorrect - Corrected == 0      0.22811    0.13985   1.631  0.65976
Full-Word - Spell-Check == 0      0.18698    0.13564   1.379  0.81196
Part-Word - Spell-Check == 0      0.29468    0.13423   2.195  0.29561
Autohighlight - Spell-Check == 0  0.01333    0.13678   0.097  1.00000
Autocorrect - Spell-Check == 0    0.04690    0.13704   0.342  0.99987
Part-Word - Full-Word == 0        0.10770    0.13198   0.816  0.98323
Autohighlight - Full-Word == 0   -0.17366    0.13543  -1.282  0.85911
Autocorrect - Full-Word == 0     -0.14008    0.13513  -1.037  0.94502
Autohighlight - Part-Word == 0   -0.28135    0.13394  -2.101  0.34972
Autocorrect - Part-Word == 0     -0.24778    0.13376  -1.852  0.50992
Autocorrect - Autohighlight == 0  0.03357    0.13653   0.246  0.99998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```

Multiple comparisons were made between the estimated mean coefficients for each level of "Device" with a Tukey p-value adjustment. The 'glht' package in R was used to compute these contrasts (Hothorn, Bretz, & Westfall, 2008). All conditions performed better than the uncorrected control ("AControl"). Only the part-word editing device ("Part-Word") condition outperformed the corrected control ("Corrected").

Figure 119 -Tukey Post-Hoc Test for Transfer Test Scores (Quasi-Poisson)



```
	 Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: glm(formula = TransferTest ~ PreTotal + Device, family = quasipoisson,
    data = lastdata)

Linear Hypotheses:
                                  Estimate Std. Error z value Pr(>|z|)
Corrected - AControl == 0         0.657438   0.183531   3.582  0.00602 **
Spell-Check - AControl == 0       0.767725   0.180699   4.249  < 0.001 ***
Full-Word - AControl == 0         1.030760   0.174852   5.895  < 0.001 ***
Part-Word - AControl == 0         1.086201   0.173569   6.258  < 0.001 ***
Autohighlight - AControl == 0     0.881812   0.177423   4.970  < 0.001 ***
Autocorrect - AControl == 0       0.890028   0.177808   5.006  < 0.001 ***
Spell-Check - Corrected == 0      0.110287   0.141148   0.781  0.98636
Full-Word - Corrected == 0        0.373322   0.135082   2.764  0.08092 .
Part-Word - Corrected == 0        0.428763   0.133547   3.211  0.02152 *
Autohighlight - Corrected == 0    0.224374   0.136430   1.645  0.64742
Autocorrect - Corrected == 0      0.232589   0.138354   1.681  0.62293
Full-Word - Spell-Check == 0      0.263036   0.130986   2.008  0.40312
Part-Word - Spell-Check == 0      0.318476   0.129429   2.461  0.16996
Autohighlight - Spell-Check == 0  0.114087   0.131976   0.864  0.97705
Autocorrect - Spell-Check == 0    0.122303   0.134250   0.911  0.97013
Part-Word - Full-Word == 0        0.055440   0.121875   0.455  0.99931
Autohighlight - Full-Word == 0   -0.148949   0.126155  -1.181  0.89921
Autocorrect - Full-Word == 0     -0.140733   0.127478  -1.104  0.92526
Autohighlight - Part-Word == 0   -0.204389   0.124569  -1.641  0.64997
Autocorrect - Part-Word == 0     -0.196173   0.125816  -1.559  0.70330
Autocorrect - Autohighlight == 0  0.008216   0.129407   0.063  1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Adjusted p values reported -- single-step method)
```