2017

Psychophysiological measures of mental effort and emotion within user research

Chase Rubin Meusel
Iowa State University

Psychophysiological measures of mental effort and emotion within user research

by

Chase Rubin Meusel

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Human Computer Interaction

Program of Study Committee:
Stephen B. Gilbert, Major Professor
Michael Dorneich
Richard Stone
Elizabeth Stegemöller
Mack Shelley

Iowa State University

Ames, Iowa

2017

DEDICATION

To Sarah and our children, Aubrey and Bennett. Never was a man so blessed, lucky, and loved.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS

I would like to thank my advisor, Stephen Gilbert. Thank you for the years of encouragement, opportunity, and teaching. You fulfilled every duty of the academic advisor, but more impressively you mentored me in ways I couldn’t appreciate until after your lessons were over. Your support of my family, career, personal, and academic interests can’t be overstated. Thank you for the late-night saves. Thank you for the early morning revisions. Thank you for the insider tips at conferences. Thank you for the hours of growth simply talking in the car. One paragraph isn’t nearly enough room to write a proper acknowledgement, but there is a beauty in brevity. Thank you.

Thanks to my committee members Michael Dorneich, Rick Stone, Elizabeth Stegemöller, and Mack Shelley for their guidance, feedback, and support throughout the course of this work.

I also want to offer my appreciation to John Deere for funding multiple research projects with specific thanks to Brian Gilmore for supporting innovative research and Greg Luecke for letting me hitch my wagon to your combine (simulator). Thank you to my colleagues at Microsoft (Melinda, Mike, Umer, & Joe) who encouraged me to take on big projects and then let me own them.

Over the years at the Virtual Reality Applications Center, I have been extremely fortunate to meet and work with remarkable individuals. A few of these people played pivotal roles in my graduate career and professional development.

Andrea, you encouraged my interest in UX, took me under your wing, and taught me everything you could in what was the most intense and satisfying three months of graduate school. Thank you for building the framework of my career.

Leif, thank you for your unwavering kindness, support, and friendship. You shared your knowledge (and time, and whiteboard) with me to answer questions neither of us needed to work on for our theses, but we wanted to work on for ourselves.

Norene, you were never too busy to hear my questions or help me work through a problem, although we likely annoyed Amanda, our pod was the best pod.

Will, thanks for your interest and subversive teaching. I spent far too much of your time pitching ideas, but your optimism toward new research and knack for statistically sound experimental design are so rare I was obligated to run ideas past you. Thanks for the help, lessons, and 11th hour save.

To my friends who became family by putting up with mine. Thank you for loving our children and compromising your social lives for ours. Bhaskar, Andrew, Taylor, Amanda, Colin, Emma, and Will, thank you.

While I have a terrible memory for anything social, I would like to list a few people who also have positively impacted my life during graduate school: Meisha, Patrick, Lori, Caglar, Curtis, RGW, Whitley, Laura, Dan, Anna, Jamiahus, Mostafa, Tim, Desmond (you owe me two pizzas), Jordan, Chase, Carter, Kaitlyn, Karen, Daphne, and of course, Don.

Mom, you are why I am. You supplied a dangerous mixture of curiosity and freedom to nudge me forward, which (happily) continues to work to this day. You showed us the value of an education and literally did everything you could to make sure we could have one too. Your example continues to teach me new lessons. Thank you, Mom, I love you so much.

To my two children, both born during grad school, I love you oh so much. Aubrey, you arrived during my second semester and Bennett three years later. You not only helped me discover the value of 30 quiet minutes, but you literally forced me to build healthy routines and habits. Our lives are so much richer with you both in them.

Finally, to my wife, Sarah. You encouraged, supported, and loved me through every late night, dirty diaper, (rare) date, and deadline. We made it through our graduate school stage of life with more fun, love, and friends than I ever would have expected. It was difficult, but it was worth it and I could not have done it without you. I love you beautiful, thank you.

ABSTRACT

Psychophysiological measures have potential to aid the discipline of user research, but are currently under-utilized. Currently, across both academia and industry there is a need to increase the quality and quantity of feedback garnered from individuals during user tasks. Psychophysiological measures are beneficial in that they can collect data objectively, unobtrusively, and in real-time. The work put forth in this dissertation focuses on two separate contexts in which psychophysiological measures are used to increase the overall quality of user research data. The first context is described in Chapters 2 and 3, in which electrodermal activity (EDA) within a high fidelity combine simulator is used as a measure of mental effort. Due to both the natural complexity of operating a combine harvester and the relative lack of understanding of combine operators today, using psychophysiological measures within this environment serves to better understand the user without compromising the experience. The second context is described in Chapters 4 and 5, in which consumer level hardware is used to measure the emotional states of workplace employees. The hardware captured electrodermal activity and heart rate data from participants while they also submitted their emotional states as training data. These data were used to build a general emotion detection model which was then tested in real-time over the course of four weeks. Additionally, emotion reporting is explored through the lens of personality and models were built and evaluated to determine what, if any influence personality plays in emotional self-report. Both mental effort within the combine simulator and emotion detection using everyday technology seek to improve the overall understanding of the user and support the use of psychophysiological measures within user research.

CHAPTER 1

INTRODUCTION

This work represents two separate research efforts over four papers to bring psychophysiological measures into user research studies conducted within the field of human computer interaction. Specifically, mental effort within an agricultural combine simulator and emotion within everyday workplace behaviors. Examples of how psychophysiological measures, primarily electrodermal activity, were used in each context are included in each chapter.

Motivation

Measuring an individual's feelings toward a particular product, task, or experience is central to user research in both academia and industry. Typically, user feedback is gleaned with self-report measures or expert evaluation. While both techniques gather data, both fail to meet ideal scenario conditions where data can be captured in an objective manner, unobtrusively in real time. While there are plenty of scenarios where self-report or expert evaluation works well, there are other scenarios where the combination of objective, unobtrusive, and real-time measurements can be achieved using psychophysiological methods.

The measurement of human performance however can be measured unobtrusively and in real time, but stops there without introducing either of the previous limited feedback methods. Performance measures are currently a preferred manner to evaluate the state of a product, method, or process. This means recording items such as time to task completion, success rates, errors, and efficiency to name a few. In addition to performance measures, user research holds a fundamental curiosity to find out what the individual's thoughts, feelings, or points of concern are with the topic of evaluation. It can be as direct as evaluating how well the average user is able to send an email for a new mobile email application or as nuanced as discerning whether or not an individual is in the appropriate mood to receive bad news via that same app by means of real-time emotion detection.

Of the many psychophysiological measures available, electrodermal activity and heart rate stand out as two separate, but related measures commonly used to capture both sides of the autonomic nervous system. These measures act as proxies for how aroused, or relaxed, the body is in a physiological sense. These internal levels are inferred by monitoring the change in one’s autonomic nervous system activity which is comprised of the sympathetic and parasympathetic nervous system. Both the sympathetic and parasympathetic nervous system activity can be monitored via electrodermal activity (EDA) and heart rate, respectively. The current methodology for attaining these measures employs the use of small, external sensors which are commonly worn on the wrist. EDA is specifically measuring the minute changes in skin moisture, measured in conductance across two points on the body. As minute changes in skin moisture occur, conductance increases as moisture increases. This measure is incredibly sensitive and will register change well before noticeable moisture changes occur on the external portion of the skin. Moisture levels in the skin change as a reflection of the activity state of the sympathetic nervous system, which is the basis for arousal measures, (Boucsein, 2012).

For this work, EDA is the measure of physiological arousal that can be used to infer mental effort when used within a controlled experimental environment. With respect to emotion detection when EDA is coupled with heart rate, both halves of the autonomic nervous system are captured which allows both axes of affect, valence and arousal, to be measured and emotions to be placed accordingly. EDA is able to be used as both a measure of mental effort and emotion as the method of measuring EDA is different in each case. Mental effort looks at the overall activity of EDA over larger time epochs, while emotion is looking for specific feature changes within EDA responses that align with a specific emotion, at a moment in time.

These sensors fulfill the criteria of objective, unobtrusive, and real time and create an opportunity to leverage their use within user research scenarios. By pushing user feedback methodologies forward, the additional user research data gathered will lead to improved insights, higher quality products, and more fulfilling experiences.

This work focuses on two domains, but a wide variety of others could also be explored. The first area is the use of EDA to measure mental effort within agricultural operator research. The second area considers the use of EDA as a signal to measure emotion in real time.

Psychophysiological Measures Within Agricultural Operator Research

Gaining a sense of how difficult an individual task is can be daunting, especially within a complex human machine system such as a combine. Additionally, to measure how much effort an individual is attributing to one individual task is even more difficult. Currently, cognitive load is primarily measured either by way of self-report, more physically restrictive and complicated physiological measures such as EEG, or by analyzing performance metrics gathered during task completion.   While both methods offer results, neither are ideal for assessing cognitive load in real time or unobtrusively. By measuring psychophysiological measures during user research, an additional measure of cognitive load can be taken without compromising the individual's experience.

Measuring operator cognitive load within agricultural combines has become increasingly important over the past few years as more technology has been added without fully understanding the implications it has on the modern operator. A combination of both an increased number of features operators are expected to use effectively in addition to tightening markets requires today's operator to be an expert with software and systems they only spend two months out of the year in.

To understand today's agricultural combine operator and better serve their needs, new methods of measuring their cognitive load and understanding them as an operator have been employed within the combine simulator research platform. First, a measure of operator expertise was implemented to better understand the operator's knowledge of the combine itself. Second, electrodermal activity sensors were used within a variety of studies to gain a measure of overall arousal during various farming tasks, which gives insight as to the operator's mental effort when used within this experimental scenario. By better understanding who the operators are and tracking their individual mental effort, final product performance results and recommendations can be made with a greater degree of certainty.

Psychophysiological Measures Within Emotion Detection Research

Evaluating an individual's emotional state has previously been done by using a combination of self-report, expert observation, and more restrictive psychophysiological feedback measures (e.g., electroencephalogram, electromyography, facial analysis). Additionally, these emotional measurements are generally recorded in response to some type of intentional emotional stimulus. This prior work has successfully shown emotion detection in theory, but has not applied those techniques to the field. The next logical step then of this technology would be to measure the individual's emotional state in real time, without their knowledge (but with their consent), so that their technology can better serve their immediate needs and help them be the most productive version of themselves.

For example, imagine you are running late for your annual evaluation meeting. You are already two minutes late as you walk to your manager's office when you receive an email that states your car payment is late followed by a text from a friend which includes three worried face emojis. Now upon reading this email and text message you are both distracted from your imminent evaluation and have additional stress added to your mental load. Imagine now that your digital assistant is monitoring your emotional state and knows you are going to a meeting that is a high priority and that you are running late. Upon reading this information, your digital assistant could then defer delivery of the late payment email and unimportant text message until after your evaluation, helping defer that stress as well. This type of subtle, passive experience is a small but impactful use of affective computing which could be used to slightly improve a multitude of aspects throughout an individual's day. Helping improve your life by keeping track of your emotional state is one part of it, but taking action on that information steps into the next era of affective computing and more generally, adaptive systems. Additionally, as the population moves toward being entirely technology native, how long will it be until the population is entirely technology native with affective computing systems and competent artificial intelligence (AI)? While this work does not go into specific AI scenarios, building better affective computing systems is a piece of the overall landscape with respect to AI and personalized, adaptive systems.

To begin to incorporate truly affective computing experiences, though, these types of measures will need to become available with consumer hardware, be passively measured in the background, and have a supporting infrastructure that enables this information to make meaningful changes. The work presented here aims to take the existing measures from previous emotion detection work and couple them with consumer level hardware and modern machine learning solutions to begin building a consumer level, unobtrusive experience.

Limitations

Three major limitations are discussed here: susceptibility to outside stimuli, interpretation of results, and individual differences within data.

The major limitation of all psychophysiological measures is in the susceptibility of outside stimuli. Without setting specific experimental conditions or collecting data in a controlled environment outside factors can influence psychophysiological data that was unintended. In other words, the opportunity to introduce confounding variables is higher with psychophysiological measures than other measures. This is because the human nervous system responds to a wide variety of inputs that can be reflected in changes in heart rate variability and to a lesser extent, electrodermal activity. Because electrodermal activity is solely innervated by the sympathetic nervous system, there are fewer stimuli to influence that measure. Regardless, by having a participant perform tasks of varying degrees of difficulty, EDA may be targeted as a measure for mental effort but if the participant is also asked to physically run around a room while completing this task the EDA data will be extremely noisy within a high perspiration environment and control for the perspiration created due to physical activity would be extremely difficult given current technology. By focusing on a single experimental construct manipulation, psychophysiological measures are stronger when used without multiple influences. Future uses should be able to compensate for this added noise as better hardware becomes available to measure the different types of noise separately from the true physiological signals.

Psychophysiological measures can also suffer from high levels of individual differences. It is not uncommon to see two individuals respond in the same overall direction to a stimulus, yet have their measures values be an order of magnitude different. By standardizing data within an individual, additional comparisons across participants can then be made with greater confidence. Related to individual differences is also hardware quality. When low-cost hardware is used the likelihood of poor readings increases and greater individual differences may be recorded. This has become less of a concern though as high quality, research grade hardware has become largely available to general researchers. The last barrier in this is seemingly cost and those also, continue to drop as wearable technology continues to proliferate through the marketplace.

The final limitation is in the interpretation of results. When outside parties read research, which includes the use of psychophysiological data, a common mistake is made by assigning an inaccurate construct to the measure in use. For example, when reading a study which states electrodermal activity levels increased in condition B, a common remark would be to state that condition B resulted in higher stress among participants. While that could be true, accurate stress evaluation requires careful experimental condition manipulation and more measures than a just electrodermal activity to evaluate, ideally cortisol level measures. Although EDA can be used as a measure of stress if that is what is being manipulated, it ideally would only be used as an additional stress measure. Carefully understanding the experimental conditions and methodology allows psychophysiological measures to be interpreted accurately and provide a greater benefit. A good rule of thumb is to understand what construct was manipulated and then determine if the chosen measure can accurately describe that construct, as opposed to simply looking at the results.

A strong counter to the first two limitations exists in the form of increasing sample size. The impact of both susceptibility to outside stimuli and high individual differences is reduced by collecting more data. While often not a practical solution within a general research context, increasing sample size does help normalize differences in participant data and also can wash out outside stimuli effects.

Psychophysiological Constructs Within User Research

As the focus of this work is ultimately to gain better insights into user behaviors, thoughts, and feelings the major constructs used will be outlined here. Current constructs being measured are mental effort, emotion, and stress. Within this work, mental effort and emotion are considered the focus.

Mental effort is an extremely useful construct to measure with psychophysiological measures. The Air Force Research Lab found increased EDA activity during take-off and landing events, which are the highest cognitive loading events for a pilot to experience (G. F. Wilson, 2002). Similarly, EDA levels were shown to change with task difficulty in different driving environments, including two simulator scenario (Engström, Johansson, & Östlund, 2005). Within a desktop setting Ikehara & Crosby (2005) found increased EDA activity correlated to task difficulty, which was rated using a seven-point Likert scale. Nourbakhsh, Wang, Chen, & Calvo (2012) showed differences in EDA values when task difficulty was manipulated in two separate areas, text, and arithmetic based tasks. More recently Lyu et al. (2015) showed a novel measure, stress-induced vascular response index (sVRI) also showed differences between task difficulty in addition to traditional psychophysiological measures. Applying the psychophysiological measure of mental effort to the domain of agricultural operator research here is a novel use.

While mental effort is sometimes reported alongside stress, they are separate constructs that deserve their own measures and independent experimental procedures. While fundamentally different, they have been shown to correlate and some studies will report psychophysiological changes with both mental effort and stress (Haapalainen, Kim, Forlizzi, & Dey, 2010) or attempt to discriminate between the two (Setz et al., 2010).

Emotion may be the most difficult construct to measure by psychophysiological means as it is a multidimensional, complex phenomena (Rosalind W. Picard, 1997). Emotions are not often measured within user research, although they offer value as they influence individual decisions, social behavior, and learning aptitude (Hascher, 2010; McNamara, Jackson, & Graesser, 2009; Stowell & Nelson, 2007). Fridlund & Izard (1983) show the earliest attempt at emotional measurement using facial muscle sensors (EMG), but since then a variety of studies have utilized a wide variety of psychophysiological methods including blood volume pulse, EDA, HR, EEG, EMG, facial analysis, and even respiration (Ghiselin, Ekman, & Gruber, 1974; Katsis, Katertsidis, Ganiatras, & Fotiadis, 2008; K. H. Kim, Bang, & Kim, 2004a; Rosalind W. Picard, Vyzas, & Healey, 2001). Recent measures of emotion have focused on providing feedback to the user and the intrinsic value of understanding your immediate emotional state. Projects like “SmartHeliosity” displayed ambient lighting based on mood (Stefani, Mahale, Pross, & Bues, 2011) and “BioCrystal” displayed emotion via a desk mounted light (Roseway, Lutchyn, Johns, Mynatt, & Czerwinski, 2015). This effort to both measure emotion and make participants self-aware will continue to improve as the hardware which performs these measures continues to become more popular and less expensive.

Although not a focus of this thesis, a popular use for electrodermal activity and heart rate variability within user research has been to evaluate the participant’s overall arousal relative to varying task difficulty. This is often reported as stress. The construct of stress, as mentioned above, is rather complex, but in these cases only task difficulty was manipulated and in addition to psychophysiological measures, self-report stress measures were also used (Matthews et al., 1999; Zijlistra, 1993). Ward & Marsden (2003) utilized EDA and HR while exposing participants to a well-designed website vs. a poorly designed website. The poorly designed website showed higher EDA and HR levels. Similarly, Trimmel, Meixner-Pendleton, & Haring (2003) found increased EDA levels as they artificially increased web page load times. Lin, Omata, Hu, & Imamiya (2005) showed that as individual performance measures decreased, electrodermal activity increased as expected. Lin et al. also argued that psychophysiological measures can be used as a complementary measure to traditional usability measures.

With respect to the existing limitations of psychophysiological measures, the constructs mental effort, emotion, and stress continue to be useful within the world of user research. Operationalizing psychophysiological measures as proxies for these constructs then should continue to see use and increase in the coming years. EDA and HR measures will continue to be utilized as their availability, quality, and price becomes increasingly attractive for both personal and research uses.

The work included in this thesis proactively seeks to understand and address the limitations of the methods previously mentioned within their respective experimental protocols.

Research Questions

Each chapter has a high-level research question which will be outlined here and answered in Chapter 6.

Chapter 2: How well does electrodermal activity reflect mental effort in an agricultural equipment simulator?

Chapter 3: How much fidelity is required to represent the desired cue within the simulator?

Chapter 4: Can emotions be measured in real-time using everyday technology?

Chapter 5: Does personality predict emotion when observed with an extended data collection process?

Dissertation Organization

This work is made up of multiple publications which all together form the research of this dissertation. Chapters 2 and 3 focus on operator performance and cognitive load within the combine simulator. Over four years, the combine simulator projects at Iowa State University, sponsored by John Deere, have been a collaborative effort between Dr. Greg Luecke leading the simulator design and technical development with his student Don Kieu and Dr. Stephen Gilbert leading the experimental design and study efforts with his student, this author, Chase Meusel. All authors were involved in the experimental design for both publications. This author's unique contributions to this research include the experimental protocol and data collection, statistical analysis, and sole author of these chapters and the primary author of their corresponding publications.

Chapters 4 and 5 focus on emotion detection in general workplace scenarios using everyday technology. This work was sponsored by and took place at Microsoft. All authors from the Microsoft research team were involved in the experimental design and execution. Dr. Gilbert contributed to publication revisions and feedback after the conclusion of the experiment. Iowa State University Ph.D. student Will Stone assisted with the statistical analysis and writing of the results section for Chapter 5. This author performed the statistical analysis for Chapter 4 and was the primary author of these chapters and the primary author of their corresponding publications.

Chapter 6 offers conclusions from both areas with respect to the limitations of each. Lastly, future research directions are offered to pursue advancements in psychophysiological measures within user research.

References

Boucsein, W. (2012). *Electrodermal Activity* (2nd Ed). Wuppertal: Springer.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Research Part F: Traffic Psychology*, *8*(2), 97–120. http://doi.org/10.1016/j.trf.2005.04.012

Fridlund, A. J., & Izard, C. E. (1983). Electromyographic Studies of Facial Expressions of Emotions and Patterns of Emotions. In J. T. Cacioppo & R. E. Petty (Eds.), *Social Psychophysiology: A Sourcebook* (pp. 243–286). Guilford Press.

Ghiselin, M., Ekman, P., & Gruber, H. E. (1974). Darwin and Facial Expression: A Century of Research in Review. *Systematic Zoology*, *23*(4), 562. http://doi.org/10.2307/2412481

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-Physiological Measures for Assessing Cognitive Load. *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 301–310. http://doi.org/10.1145/1864349.1864395

Hascher, T. (2010). Learning and emotion: Perspectives for theory and research. *European Educational Research Journal*, *9*(1), 13–28. http://doi.org/10.2304/eerj.2010.9.1.13

Ikehara, C. S., & Crosby, M. E. (2005). Assessing Cognitive Load with Physiological Sensors. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, *0*(C), 1–9. http://doi.org/10.1109/HICSS.2005.103

Katsis, C. D., Katertsidis, N., Ganiatras, G., & Fotiadis, D. I. (2008). Toward emotion recognition in car racing drivers: a biosignal processing approach. *IEEE Trans. Systems, Man and Cybernetics - Part A: Systems and Humans*, *38*(3), 502–512.

Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, *42*(3), 419–427. http://doi.org/10.1007/BF02344719

Lin, T., Omata, M., Hu, W., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes? In *Proceedings of OZCHI 2005*.

Lyu, Y., Luo, X., Zhou, J., Yu, C., Miao, C., Wang, T., … Kameyama, K. (2015). Measuring Photoplethysmogram-Based Stress-Induced Vascular Response Index to Assess Cognitive Load and Stress. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (pp. 857–866). http://doi.org/10.1145/2702123.2702399

Matthews, G., Joyner, L., Gilliand, K., Campbell, S., Falconer, S., & Huggins, J. (1999). Validation of a comprehensive stress state questionnaire- Towards a state "big three"? *Personality Psychology*, *7*, 335–350.

McNamara, D., Jackson, G. T., & Graesser, A. (2009). Intelligent Tutoring and Games (ITaG).

Nourbakhsh, N., Wang, Y., Chen, F., & Calvo, R. A. (2012). Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks The University of Sydney. In *Australian Computer- Human Interaction Conference* (pp. 420–423).

Picard, R. W. (1997). *Affective Computing*. Cambridge: MIT Press.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191. http://doi.org/10.1109/34.954607

Roseway, A., Lutchyn, Y., Johns, P., Mynatt, E., & Czerwinski, M. (2015). BioCrystal : An Ambient Tool for Emotion and Communication. *International Journal of Mobile Human Computer Interaction*, *7*(3), 20–41.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *14*(2), 410–7. http://doi.org/10.1109/TITB.2009.2036164

Stefani, O., Mahale, M., Pross, A., & Bues, M. (2011). SmartHeliosity: Emotional Ergonomics through Coloured Light. In *Lecture Notes in Computer Science* (Vol. 5624, pp. 226–235). http://doi.org/10.1007/978-3-642-21716-6_24

Stowell, J. R., & Nelson, J. M. (2007). Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion. *Teaching of Psychology*, *34*, 253–258. http://doi.org/10.1080/00986280701700391

Trimmel, M., Meixner-Pendleton, M., & Haring, S. (2003). Stress Response Caused by System Response Time when Searching for Information on the Internet. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *45*(4), 615–621. http://doi.org/10.1518/hfes.45.4.615.27084

Ward, R. ., & Marsden, P. . (2003). Physiological responses to different WEB page designs. *International Journal of Human-Computer Studies*, *59*(1–2), 199–212. http://doi.org/10.1016/S1071-5819(03)00019-3

Wilson, G. F. (2002). An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation Psychology*, *12*(1), 3–18. http://doi.org/10.1207/S15327108IJAP1201_2

Zijlistra, F. R. H. (1993). *Efficiency in Work Behavior (Doctoral Dissertation)*. TU Delft.

CHAPTER 2

EVALUATING NOVEL HARVEST TECHNOLOGY WITHIN A HIGH-FIDELITY COMBINE SIMULATOR

This chapter was submitted to *Computers and Electronics in Agriculture*.

Author list:

Chase Meusel, Don Kieu, Stephen B. Gilbert, Greg Luecke, Brian Gilmore, Tim Hunt

Chase Meusel’s role in this research included contributions to the experimental design, participant recruitment, data collection, data coding and analysis, and primary authorship on the chapter/paper. Chase designed a novel UX observation technique to monitor iPad activity without interfering with user activity. Lastly, Chase used novel data fusion methods to report stronger findings by combining physiological, behavioral, performance, self-report, and interview data.

Abstract

Farming today is more complex than it has ever been. Operators are increasingly reliant on technology to aid and improve harvest performance. New harvest technology is under development that will advise harvest operators on the proper adjustment of machine harvest settings, as well as automatically adjust these machine settings without operator intervention, improving the harvest performance of the machine, and reducing the cognitive load of the operator. In this work a high-fidelity, interactive harvest combine simulator is used to understand how harvest operators currently use existing harvest technology, and to

evaluate the performance improvements provided by new prototype machine control algorithms and human control interface designs. The interactive harvest simulator is used to assess an intermediate advising step for machine controls adjustment compared with a path using fully autonomous machine adjustment. Testing novel harvest technologies using the virtual environment of the combine simulator introduces a specific set of constraints and challenges that are not found in most other vehicle simulation applications, including the need for accurate physical and visual crop flow representations and a requirement for realistic machine responses to a wide variety of operator input commands. Using a high-fidelity combine simulator for testing allows unique harvest scenarios to be repeated by experienced operators in a controlled virtual environment.

This study evaluates operator acceptance, performance, and feedback for two novel pieces of harvest technology, Advisor and Director. Advisor is an operator-in-the-loop system providing feedback on proper machine control adjustments during normal harvest operations. Director is designed to continuously monitor the overall harvest health and autonomously adjust the combine harvest settings. In this study, operators harvested the same virtual field twice, first using Advisor, and a second time using Director. Operators overwhelmingly perceived both the Advisor and Director systems as optimizing the harvest performance of the combine and recommended both Advisor and Director. The results presented in this work show that both systems improved the perceived harvest performance, although the Advisor was not as highly rated. Participants recommended the automated nature of Director, and both operator feedback and physiological measures indicates that this harvest technology reduced the cognitive load of the operator. This work demonstrates two main points. First, the interactive combine simulator can be used for evaluating novel harvest technology in the lab. Second, that operators can quickly acclimate to automation within the combine and were able to harvest in a more productive manner when using higher levels of automation.

Introduction

Harvest operators today face an increasing number of distractions and demands on their mental resources. Combine operators not only manage the physical crop harvesting process, they also must plan logistics for grain transport, analyze weather reports, communicate with outside operators, and take phone calls from a variety of sources. A potential solution for reducing the workload of the operator is to automate those aspects of tasks which demand high cognitive resources, such as the ongoing vigilance of driving and the complex input tasks required for machine adjustments. This approach has been shown to be effective in other comparable scenarios (Endsley and Kaber, 1999; Metzger and Parasuraman, 2001; Parasuraman et al., 2009). The tasks which demand the most of operators should be evaluated for potential automation benefits, such as the control and adjustments of the combine processing systems, including the fans, sieves, and implement arrangements. Automating the most important harvest controls can help reduce the overall cognitive load experienced by operators, as well as improve the performance from less experienced operators who might otherwise see low performance results. In this work, a new technology application is evaluated using two steps on the path to harvest automation, the first providing guidance for manual machine adjustments during harvesting and then the second fully automating the sensing and harvest adjustments required to improve the performance.

Two related technologies were evaluated in this study, Advisor and Director. Both technologies were developed by the research team and did not represent finished quality found in final production software interactions or robustness. Advisor technology offers expert level guidance to operators in real time via combine adjustment feedback and suggested actions. Performance gains have been demonstrated in other studies, where the assisted operator shows higher performance than fully manual or fully automated solutions in similar scenarios, (Endsley and Kaber, 1999; Endsley and Kiris, 1995). In this implementation, Advisor requires operators to input their observed harvest issues, accounts for the current system state of the combine overall, and delivers a recommended list of corrective changes in prioritized order. Because the Advisor must rely on the operator to identify and report issues, an implicit assumption is that the operators have enough basic knowledge of harvesting to initiate the system and report observed issues. After recommendations are made, the operator can either accept the current recommendation, view the next recommendation, or cancel the entire process. This affords the operator the opportunity to allow the adjustment to be made as suggested by Advisor, select an alternative action, or to cancel the process and make a manual change which may have been influenced by the earlier suggestions. The final step of the Advisor process then queries the operator to note whether the issue has been resolved or if a new issue is present. This answer can either end the engagement or begin anew with the new or modified issue.

Director is the next level of automation, where the system actively monitors the overall combine system state in real time and acts to improve harvest quality. After an initial setup to identify the harvesting preferences of the operator (e.g. Do you want a faster harvest with a lower quality sample or a slower harvest with a higher quality sample?) the system will make changes without interrupting the operator to improve the harvest process overall. Due to the ability of the Director to initiate change without involvement of the operator, operators with lower harvest knowledge stand to gain more benefit from this system as it has the capability to observe and autonomously make changes on issues that may have otherwise gone unnoticed. The system does notify the operator when a change is underway, but it does not have to wait for approval with every adjustment.

Both Advisor and Director represent incremental steps in available technology toward a fully automated harvesting system. These automation steps were designed to provide operator assistance without sacrificing quality. When comparing Advisor and Director to the established SAE Automated Driving Levels (SAE, 2014), Advisor falls within level 2 of partial automation, which requires multiple systems to be automated but ultimately requires the operator to still perform the remaining tasks to successfully operator the machine. Director then takes the next step and falls closer to level 3 of conditional automation where the operator hands over control of all aspects of the dynamic driving but needs to be present for intervention. With these automated driving levels to consider, the value of a guidance-based system, Advisor, can be adequately compared with the more automated system, Director. To understand the full value each of these systems provides, the current state of combine adjustment must be understood.

When a problem occurs during normal harvesting operations, current practice calls for the operator to use acquired knowledge to adjust the combine settings. When the operator does not know the correct solution, the process ends in one of three situations. The operator may 1) seek additional help, 2) ignore the potential issue, or 3) miss the harvest cue altogether. Seeking help requires time and will likely slow progress within the field because of the efforts required to contact an outside expert (e.g., “I have to call Dad.”), consult outside knowledge such as the harvest slide rule (Deere, 2013), review the troubleshooting guide (IH, 2009), or refer to the owner’s manual. If the operator simply ignores issues or misses harvest cues outright, the harvest process will result in lost grain loss and the operator is indirectly indicating low harvest knowledge. Both Advisor and Director can improve these known issues by providing a faster resource for outside information in Advisor and performing changes that would otherwise go untended with Director.

Several factors make it particularly difficult to test this highly specialized technology. First, it requires several factors—the right season, uniform crops in the field, an expensive harvest combine machine, and a human operator. The North American harvest season occurs only once per year, and most operators will not encounter these specific requirements outside of that window, so the technology is only sporadically needed. Testing the algorithms requires multiple runs through the field with a variety of crop conditions. Even the most uniform field and crops have unknown variations, and once a field is harvested, there is not a duplicate with which to compare results. Running an actual combine is expensive, and may be plagued with maintenance issues during the testing. Even obtaining the operator may be problematic, because the demand is high when the crop is ready to harvest. Occupying the time of an experienced operator may have a high cost in terms of lost harvest opportunities.

The limited time window of operation and infrequent use of this type of technology makes designing for this specific audience difficult and testing it prior to implementation nearly impossible. However, implementing the prototype harvest technology within the high-fidelity VR combine simulator gives the operator the opportunity to acclimate to the new automation system, provides a baseline for performance, and offers feedback for technology they have yet to encounter in the field, all without the pressure of monetary loss when using their own crops and equipment. Specific harvest scenarios can be built within the virtual environment; therefore, operators can make all normal adjustments that would occur in a real combine as both the operator and the technology are evaluated. Moreover, a simple reset of the simulation presents each operator with an identical field and set of crop conditions during the test.

Harvest scenarios include relevant exterior graphical cues (e.g. crop height and color), interior instrument cues (e.g. loss monitor, moisture monitor), and expected auditory cues. An emphasis is placed on observing operator feedback including verbal, performance, and physiological. All operators indicated preference to a system which helps them identify potential issues and the less experienced operators strongly prefer the system which helps them perform at a level closer to an expert. The software, hardware, and external components utilized to perform this study are outlined in the methods section.

Background

Previous studies have shown that virtual environments and simulators are effective at training new technologies and developing new products, especially within domains which have highly specific or constrained use cases. Simulators and virtual environment training transfer work has found success in areas such as repairing the Hubble space telescope (Loftin and Kenney, 1995), fire-fighting aboard a naval ship (Tate et al., 1997), or even performing highly specialized medical procedures (Calatayud et al., 2010; Kruglikova et al., 2010; Triantafyllou et al., 2014).

Automotive Simulators

The largest area of simulator research centers around the automotive industry, and this research has been using driving simulators since the 1960s (Weir, 2010). Driving simulators allow researchers to simulate actual driving conditions and situations within a controlled and safe environment (Lee et al., 1998). Simulator use in research studies also allows for greater flexibility and cost savings when compared to performing studies with a physical vehicle (Mann et al., 2014). Simulators today continue to see reductions in cost, improvements in computational performance, and increased use as a research tool in human factors research, interface design, and operator research (Bella, 2008; Birrell and Young, 2011; Jamson et al., 2014; Weinberg and Harsham, 2009).

Automotive simulators have been used to investigate a broad selection of topics within automotive research. Topics range from more basic driver performance work (Mclane and Wierwille, 1975) to very specific examples of investigating the impact of brake pedal stiffness in racing applications (de Groot et al., 2011). This work is specifically interested in applications of operator performance, workload measures, and automation applications. Automotive simulator research indicates that people are poor at dividing attention (Lee et al., 2005) even when only engaging on phone conversations (Horrey and Wickens, 2006). More closely related to this work, though, are topics of vehicle assistance, helping drivers achieve better performance, in this case, fuel economy (Hibberd et al., 2015; Jamson et al., 2014).

Non-automotive simulators

Automobiles and agricultural vehicles have certain commonalities but differ in terms of many purposes and functions. Thus, the development and use of simulators for agricultural vehicles has both similarities and differences from that of automobiles. For examples, tractors and automobiles are similar in requiring minor steering imperfections. However, the nature of the steering disturbances differ due to the differing forward speeds and driving surfaces of an automobile versus a tractor. Also, it is frequently the case that tractor operators are using a guidance system, which increases accuracy of straight-line driving. Karimi et al. (2008) developed and validated a simulation model of parallel swathing (driving in parallel paths to cover a field) in a tractor-driving simulator. The model accounted for tractor self-deviation and guidance system error. The study's field experiments were in close agreement with the simulator experiments regarding frequency composition of lateral deviations, thus showing the model's value in simulator fidelity.

While agricultural vehicle simulators differ from automotive simulators, there are also a variety of agricultural vehicles, which results in differing designs and research questions. For example, in the design of a tractor-air seeder driving simulator, the first step was to conduct a function-oriented task analysis to identify the required functions, tasks, and subtasks (Mann et al., 2014). A main finding was that operators allocated a substantial portion of their time to manually operating the air seeder. Additionally, operators had to monitor the air seeder that was mounted behind the tractor. To simulate this characteristic of a tractor-air seeder, researchers put two computer monitors behind the cab (one to rear-left and one to the rear-right of the operator's seat). Thirty-two images created a panoramic view using 32 images, forming the field boundary for the tractor-air seeder. This simulator is being used for two research issues: to determine an appropriate automation design for agricultural vehicles and to understand the impact of display design on an operator's situation awareness in a semi-autonomous agricultural vehicle (Mann et al., 2014).

A primary use for agricultural simulators is in their use as a tool to evaluate operator performance and novel agricultural technologies (Bashiri and Mann, 2014; Duncan and Turner, 1991; Mann et al., 2014). Research in this area though is still relatively scarce which leaves opportunity to evaluate innovative technologies prior to their release within particular markets. Industry has at times introduced new automation features without sufficient testing, resulting in user problems and complaints. Such evaluation prior to implementation is increasingly crucial as the vehicles become more complex, more sensors are integrated into the system, and more functions are automated. While automation seems to be an obvious positive for the operator, that is not necessarily the case. Automation may, for example, result in information overload for the operator. Predicting the impact on the operator of proposed automation is an important part of the design and development process. A simulator is ideally suited to conduct user studies, from which researchers can predict the impact of automation (Mann et al., 2014).

Overall, automotive simulators are the most popular medium for simulator based research platforms, but other there are examples of non-automotive vehicles that have been modeled as simulators and used for research purposes. From construction vehicles (Son et al., 2001; Yoon and Manurung, 2010) to agricultural equipment (Karimi et al., 2008; Karimi and Mann, 2008; Mann et al., 2014) non-automotive simulators have a large opportunity to gain operator feedback in a meaningful way. This work utilizes a harvester combine simulator specifically and is an updated model from the initial platform built by Luecke (2012). Luecke’s combine simulator is unique as it is constructed of many production parts enabling an active CAN bus, displays and intelligence features to provide fully functional and responsive John Deere combine setup which can integrate existing and future technologies for operator use and evaluation. Data collected in this way can be objectively measured to evaluate the value, issues, and expectations of operators long before they enter the field in a real combine.

The previous works measured an operator's ability to use the technology. More specifically in Duncan and Turner, (1991) a rank ordering of operator preference was also obtained. This paper goes beyond that to present and methodology to assess of an intermediate step in automation technology is needed for customer acceptance.

Methods

Research objectives

The primary research objectives of this work were to determine whether the two separate prototype harvest technologies could 1) increase operator performance 2) be accepted by the operator and 3) reduce operator workload. For operators today, this ultimately means they make more money. The two technologies evaluated in this work were Advisor and Director. Advisor is an operator in the loop system and by extension has more behavioral data to assess. Director takes the operator out of the primary loop and focuses on assessing the operator feedback and choice of interventions. Each operator used both the Advisor and Director systems in sequence as Advisor will be available prior to Director.

Both Advisor and Director were evaluated via operator feedback to standard questions, operator behaviors in response to system actions, physiological measures, and qualitative comment analysis.

Hypothesis

The first system, Advisor, was expected to increase operator performance and offload work from the operator to Advisor. Similarly, Director was expected to both increase performance and reduce workload by making beneficial adjustments without the operator's input. The systems were expected to perform comparably at the performance level because each system was delivering the same recommended change. If one system was to outperform the other, the expectation was that Director would do so as it would not wait for operator confirmation to make changes or give the operator a direct opportunity to veto any actions. Director was expected to have the greater effect at reducing workload though as it did not require any input for each suggested change where Advisor still required manual confirmation for each suggestion.

Participants

28 operators were recruited to take place in this study with the requirement that they have at least two years' experience operating a combine in the past four years. Additionally, operators were recruited with a diverse set of primary crops including corn, beans, and wheat. Operators travelled to the Virtual Reality Applications Center at Iowa State University to participate in this study from Iowa, Montana, and Illinois.

Operators were recruited from a large pool of individuals who had indicated their willingness to participate in research for an agriculture equipment company. Operators were compensated $150 for their effort and had travel expenses reimbursed. All operators were over 18 years old.

Tasks

All operators completed the same harvest scenarios where the independent variable was the crop being harvested. The overall study consisted of two separate phases, one using Advisor and a second using Director. Prior to the simulator portion of the study, operators completed demographic questions, system knowledge questions, and prior experience questions.

Independent variables

Independent variables used in this work were visual feedback in the cab, such as yield, moisture, and harvested crop quality and between groups two separate crop types were harvested, corn and wheat. Each crop presented the same changes with respect to crop variables yield, moisture, and quality. The crop type visible was determined by the operator's personal harvest experience. Crop yield was displayed via changes in the visual graphics of the simulation and in the instrumentation on the combine hardware as seen in Figure 1. Crop moisture was displayed via changes in the visual graphics of the simulation and within the combine instrumentation. Crop quality was not explicitly identified by a single metric, but was presented with changing visual representations via a simulated grain tank window as seen in Figure 2.

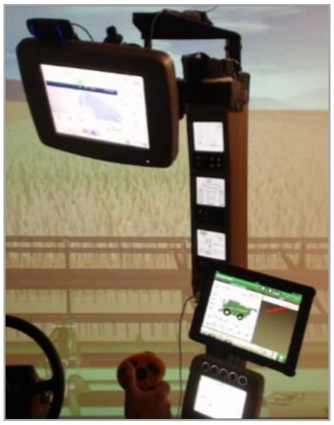


Figure 1. Harvest information displayed on four primary displays; from top down the navigation display, corner post display (3 small screens), the iPad with novel harvest technology, and the command arm display (bottom).

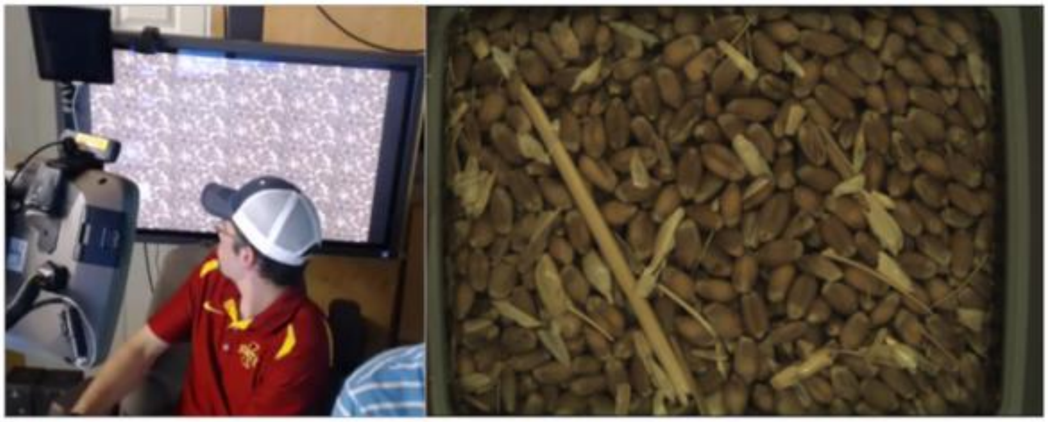


Figure 2. Operator inspecting grain tank window.

Dependent variables

Dependent variable measures included system performance metrics, operator feedback, cognitive load and operator ground truth comments. Performance metrics included items such as whether the operator reported the correct issue, how they chose to implement the suggested resolution from the system, whether they used the system, if they decided to turn the system off at any point, how many times they visually attended to the grain tank window, time spent in the field, and whether they slowed down during the use of Advisor or Director. Operator feedback items included the single ease of use question (SEQ) (REF), the system usability scale (REF), and a net promoter style question (REF). Cognitive load was measured continuously via electrodermal activity, (EDA) (REF). Lastly, comments made and feedback given throughout the duration of the study were noted for specific mentions of emergent themes such as estimated operator fatigue, estimated operation times, and operator trust in the system. A summary table of dependent variables can be seen in Table 1.

Table 1. Dependent variable summary table.

Dependent variable	**Metric(s)**	**Unit**	**Frequency of Collection**	**Data Type**
Performance	# slows, stops, errors		Per field condition	Ordinal
Cog load	EDA	Microsiemens	Continuous (32hz)	Continuous
Perception	SEQ	Scale 1-5	Per field condition	Ordinal
Satisfaction	Recommender	Scale 1-10	Per technology	Ordinal

Experimental design

Dependent on the operator’s experience, either the corn or wheat variation of the simulation was set to run. The hardware setup did not change with respect to crop. Crops were identical in size and field variation including transitions, e.g., harvesting from a tall crop to a normal height crop happened in the same place in each field. Both corn and wheat fields were structured the same way with respect to crop variation and changes in the information displays as well. The virtual field was comprised of seven, 30-inch-wide, half mile passes totaling 12 acres (Figure 3).

Each operator experienced all five field conditions including a normal pass, low moisture, high moisture, low yield, and high yield sections. The operator’s performance, cog load, and perception of changes were tracked for each of the four primary field conditions. Satisfaction was measured at the end of each field via recommender and intent to purchase questions.

Operators used Advisor in the first field, followed by Director in the second field. No counterbalancing was performed between fields as the release order of this technology was intended to be Advisor followed by Director.

Procedure

Operators initiated the research experience with an introductory survey on harvest knowledge and demographic information. After completion, they were brought into the combine simulator and prepared to begin harvesting the virtual field by explanation of the scenario and basic combine controls. No training was performed prior to the study. A researcher was present and sat next to the operator in the “buddy seat” for the duration of the study. The researcher was able to ask relevant questions during the study and answer any reasonable questions the operator may have had.

The first pass of each field was empty to give operators an opportunity to acclimate to the combine so they would be prepared for the first of four trials (or harvest events) in the remaining six passes. The first task was to complete the field using the Advisor system with only a description of what the system was intended to do. As gaining realistic operator feedback was an important goal, no specific instruction was given on how to use the novel harvest technology following best practices within UX testing guidelines (Krug, 2009). After task one was complete using Advisor, there was a short reset and then the second task of harvesting the field again using the Director system took place. Observations were noted during all harvest events to note if and how operators were engaging in technology use. In addition to observations as to the operator’s behavior and performance, operators also answered three questions after each interaction with the harvest technology when engaged during a harvest event (when field conditions changed). The three questions were 1) “Was the suggested solution appropriate?” 2) “Do you feel the combine is in an optimized state?” and 3) “How did you feel about the last adjustment overall? 1-5, 1 being poor, 5 being ideal.” Operators were also given an opportunity to list any issues or suggestions they wanted to share after the interaction had concluded. The target total time spent for completing both task one and task two was between 90 minutes to 120 minutes, in addition to questions before and after the simulation.

During both harvest tasks a second research team member was operating the “Wizard of Oz” station which controlled the information displayed within the combine simulator and the image presented in the grain tank window monitor behind the operator. By tracking the operator’s position in real time and monitoring the operator’s actions, the second research team member could update the information displayed within the simulator to reflect the current field conditions in real time. Information displays changed in this way included the yield monitor, moisture meter, and grain tank window image.

Once the harvest tasks were complete, an exit interview was conducted to cover a variety of experiences and finally a survey was completed to allow the operator to provide feedback on the harvest systems, simulator, and overall experience. The entire session lasted, on average, three hours.

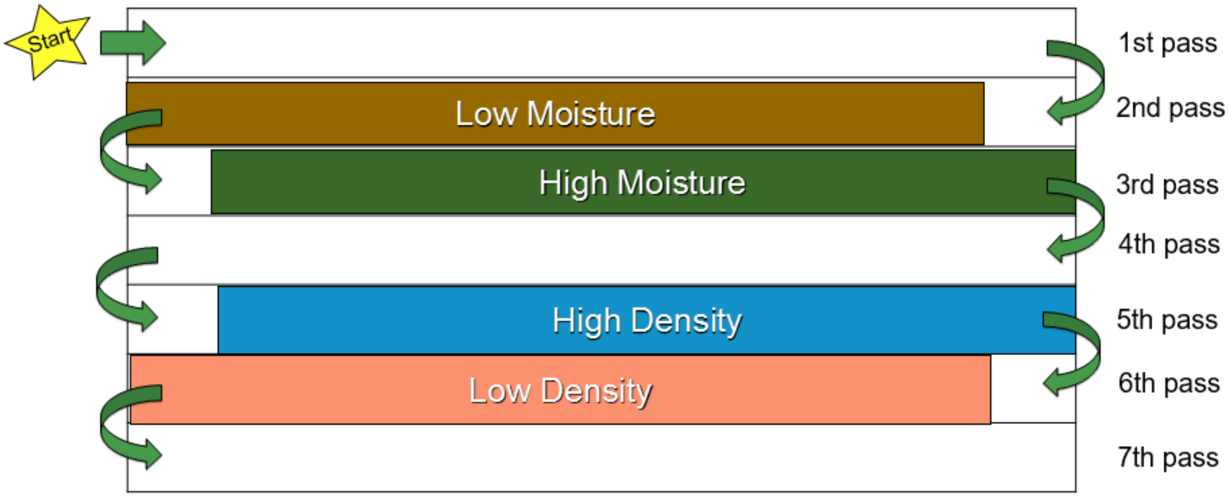


Figure 3. Top-down view of the 12-acre virtual field.

Limitations and assumptions

This study was not designed to measure literal crop harvest quality as the underlying simulation was not built to respond to all possible variable inputs. All changes displayed within the simulator were previously generated to appear as a realistic output for the corresponding section of field the combine was harvesting within. Additionally, this study was not designed to compare absolute ground speed as a measure of performance. This is due to the uniform nature of the field and the low number of constraints placed on the operator from a ground speed position. Future studies could potentially have a more sophisticated harvest model to account for all available inputs if harvest quality was an output of primary interest. While the actual harvest quality output is marginally useful, the freedom to run more than pre-scripted harvest events within the field would allow the study to commence without the second research team member adjusting in real time and allow a greater variety of field conditions to be evaluated.

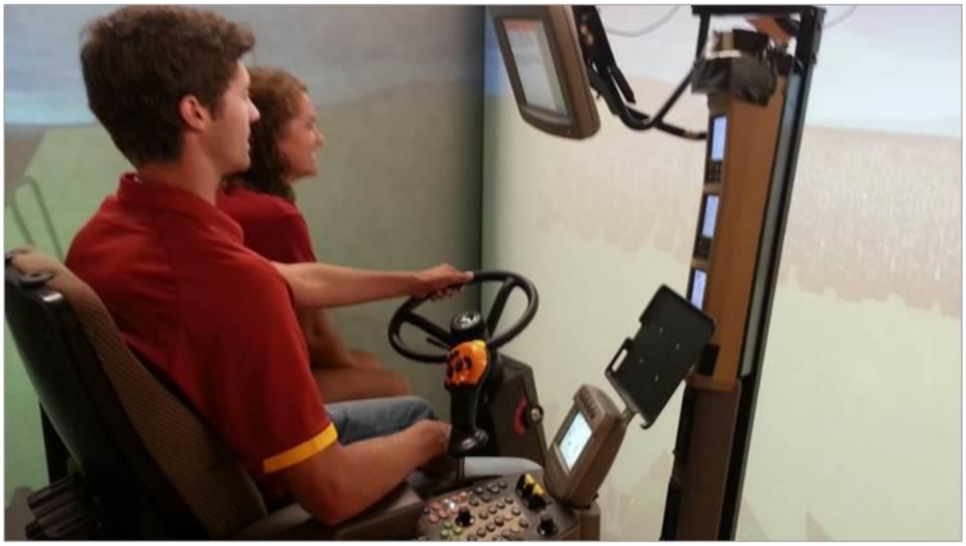


Figure 4. Operator driving the combine simulator with a research team member riding in the buddy seat.

Testing environment

The combine simulator featured a modified John Deere 9770 STS interior, with projected displays setup in front of and to the left of the operator to simulator immersive virtual farming, see Figure 4. The cab included a John Deere 2630 in-cab monitor running GreenStar 2 and an Apple iPad 3. The iPad was used to display the working prototypes of the novel combine technology software. The combine simulator software, Greenspace (Luecke, 2012) was run on Ubuntu 12, 64 bit with a 3 GHz dual core Intel processor, 8 GB of DDR3 ram, and an NVIDIA Quadro K600 graphics card. Two external stereo speakers were used to produce audio in addition to an 8” subwoofer. A Buttkicker bass shaker attached to the cab seat was also utilized to simulate the vibrations felt when operating a full size combine. Primary displays to simulator the virtual field were two short throw, rear projected, projectors at 1280x800, displayed on two 8’ x 6’ screens positioned in front of and to the left of the operator giving approximately 95° field of view on the front display and a full left peripheral view from the left display. The simulation was rendered monocularly using the

OpenSG graphics engine with displays handled by VRJuggler at an average frame rate of 31 frames per second. An additional 40" LCD television positioned immediately behind the operator and 4' off of the ground was used to simulator the grain tank window.

The second research team member operating the "Wizard of Oz" station utilized a Windows 7 computer to run the custom "wizard" application to monitor, record and manipulate combine simulator parameters in real time. Simultaneously the same team member was observing and recording the prototype interface evaluated on a second Apple iMac.

Results

28 operators completed the combine simulator study. Operators were primarily between 41-50 years old, 36%, with 20-30 and 31-40 each having 23% of the total. The majority, 55%, of operators have over 12 years of experience with almost all others have between 4-7 years' experience. Most operators, 64%, were either the owner of their farm or worked on a farm their family owns. Operators spent an average of 102 minutes in the simulator. Dependent variables measured within this study are noted and expanded upon within Table 1.

Performance

The two pieces of technology, Advisor and Director are both intended to improve operator performance, but Advisor was shown to require more input than Director. The amount of input was measured by number of interactions each piece of technology received from operators over the entire study. Two operator's data were removed due to video not being available for analysis. Advisor saw 13.7, 95% CI [8.832, 18.475] more interactions on average than Director over the course of the entire study, $t(25) = 5.8327$, $p < .0001$.

When investigating how operators used both Advisor and Director, no operators turned either system off. Performance within each system is reported as the % of operators who made manual adjustments instead of allowing the system to adjust, the % of operators who used the system as intended, (used as intended) and % of operators who did not make any errors, (no errors). Used as intended meant the operator successfully and intentionally used the system at least once by the final, of the four, trials available. Without errors represents % of operators who did not make a mistake through the entire field using that system. A mistake was noted when an operator would report the incorrect issue present, abandon the process prior to completion, or manually override the system suggestion. Fewer errors overall committed during the second half of the study, as Director had fewer errors than Advisor. All percentages are based on the total of 28 operators. The results can be seen in Table 2.

Table 2. Operator performance errors observed.

Technology	Manual adjustments	Used as intended	No errors
Advisor	7%	89%	46%
Director	0%	89%	79%

Operators reduced their ground speed fewer times when using Director when compared with Advisor, 95% CI [0.39, 1.39], $t(27) = 3.6728$, $p = .001045$. The difference also displays a medium effect size $r = .4456$.

No difference in number of times operators brought the combine to a stop using Director when compared with Advisor as the 95% CI includes zero, 95% CI [-0.13, 1.20], $t(27) = 1.6576$, $p = .109$. While not significant, a small effect size does exist, $r = .2148$.

Table 3 displays the number of operators who either slowed or stopped ground movement completely while harvesting.

Table 3. Operator ground speed changes observed.

Technology	# Operators who slowed	# Operators who stopped
Advisor	13	11
Director	1	6

Overall time spent in the field represents a measure of efficiency and potentially reduced operator fatigue. Operators spent less time in the field using Director when compared with Advisor, t(24) = 4.81, p < .0001. See Table 4 below for time spent by technology.

Table 4. Time spent (in seconds) in each section of the field in seconds, split by technology.

Technology	Total Time	Low Moisture	High Moisture	High Density	Low Density
Advisor	1305	374	347	314	271
Director	1081	273	281	285	242

Time spent in the field, (Figure 5) and time spent in each individual pass, (Figure 6) show larger standard deviations for Advisor when compared with Director. This can be interpreted as the process for Director was more in control as there was less variance in the time spent when using that technology.

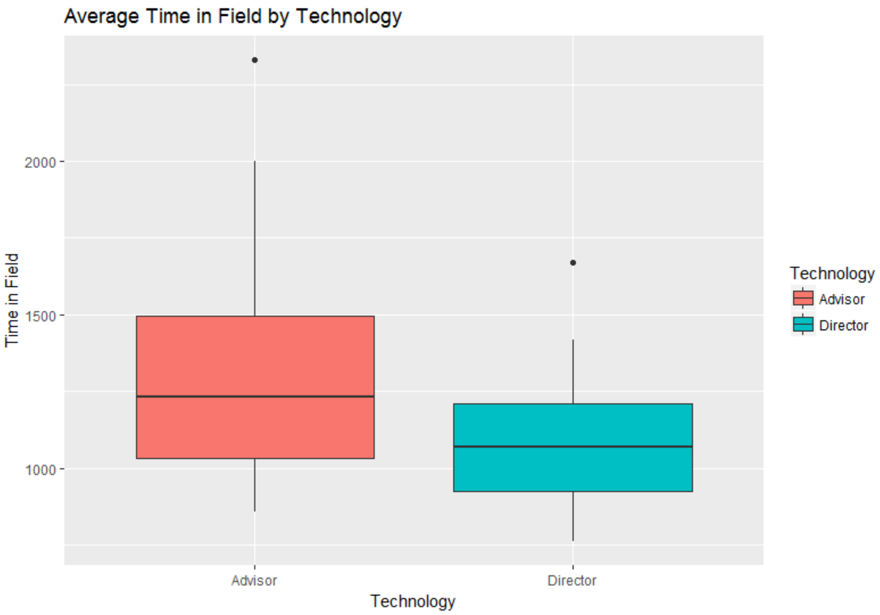


Figure 5. Boxplot of average time spent (in seconds) through entire field for Advisor and Director.

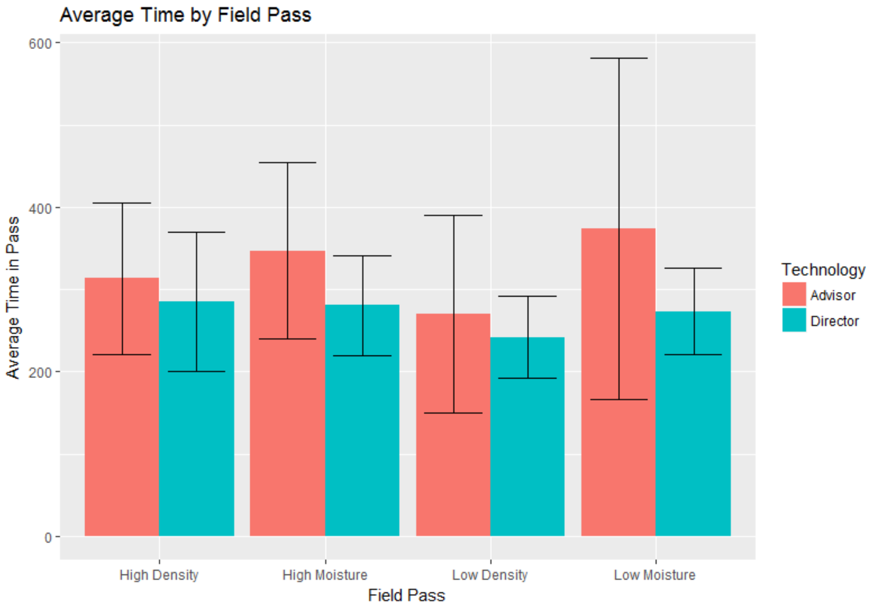


Figure 6. Average time (in seconds) all operators spent in each pass.

Grain tank window

24 of the 28 operators looked 6 times or fewer at the grain tank window through the entire study. The other four operators looked 27, 30, 33, and 36 times respectively. This results in a total of 190 individual grain tank window looks that occurred, 126 or 66% of

them were from four operators. Eight operators only looked at the grain tank window one time. The histogram of grain tank window looks by operator frequency can be seen in Figure 7.

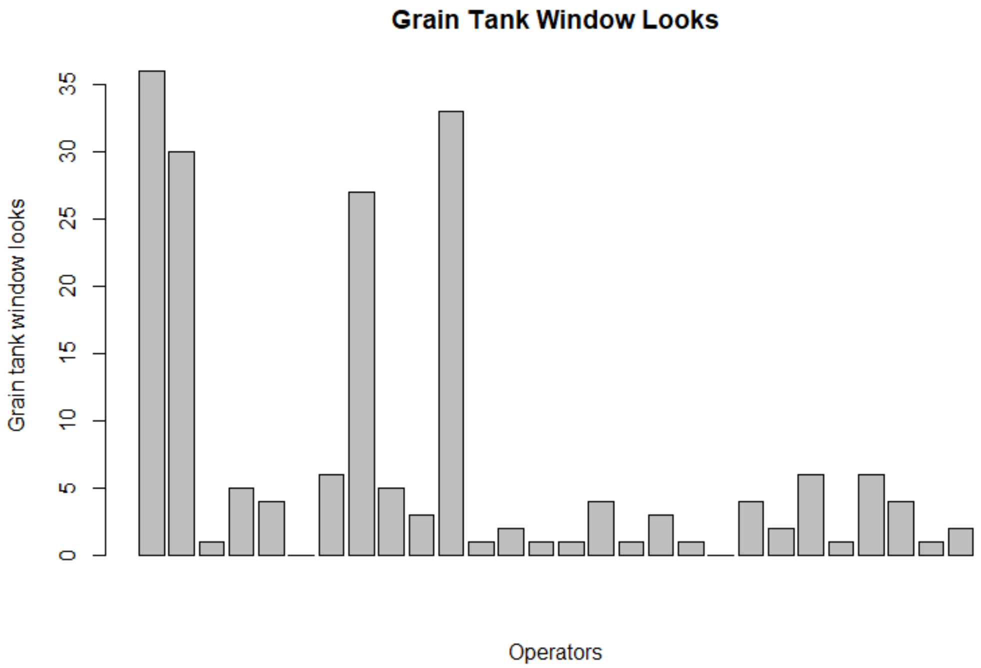


Figure 7. Number of grain tank window looks taken by operators, most operators do not look more than six times.

Operator harvest knowledge survey

Operators were surveyed on nine separate questions written to determine whether they would be able to correctly identify the correct adjustment needed to correct a harvest issue while harvesting.

Questions were created based on adjustments available within the combine simulator configuration, for conditions that commonly arise within farming large grain crops within the Midwestern United States. Original question answers were outlined based on research team knowledge, sponsor team knowledge, and agricultural extension office information

(Anderson, 2011; Fone, 2007; Mowitz, 2013; Wehrspann, 2004). Subject expert engineers from a large agricultural machinery company and three experienced combine operators were also consulted in the creation of these questions and their correct answers.

All 28 operators completed the operator harvest knowledge survey. The survey was comprised of nine questions taken during the general pre-survey questions. Of the nine questions, only eight were used for analysis as one of the questions was specific to a corn condition and 12 of the 28 operators were primarily experienced with wheat harvest. The eight questions were worth a total of 16 points, or two points per question. Of the eight questions used, operators scored an average of 11.21 (SD 3.24).

Groups were created by separating scores into low, medium, and high groups. Grouping was done by taking the average +/- the standard deviation and including those scores as the "medium", all scores above labeled as "high" and below labeled as "low." See Figure 8 for all scores.

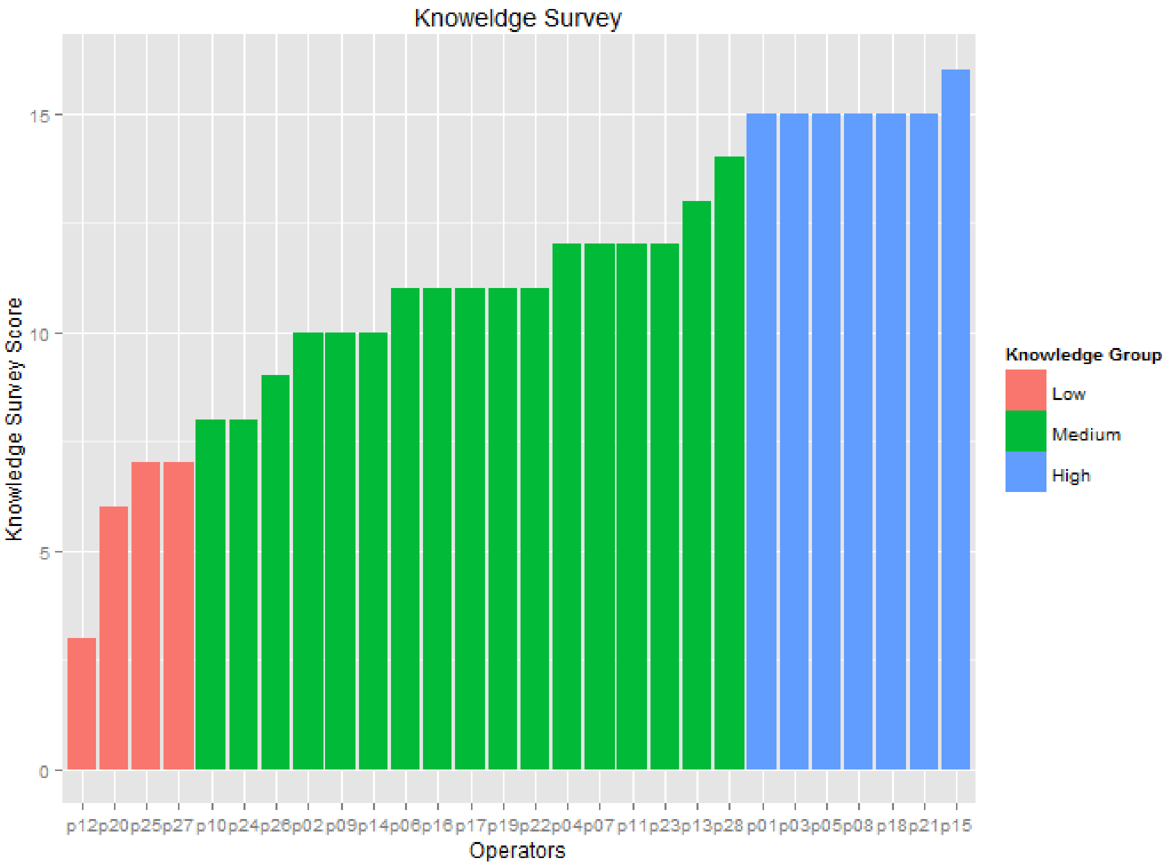


Figure 8. Operator knowledge scores, colored by group, max score of 16.

A one-way between subjects ANOVA was performed to determine whether knowledge scores between operator groups were different from each other. There was a difference on scores between knowledge groups for the three groups, F(2) = 53.71, p < .0001. Post hoc comparisons using the pairwise t-test with a Bonferroni adjustment indicated that the mean score for the Low group was different from the Medium group (p < .0001) and High group (p < .0001). Additionally, the Medium and High groups were also different from each other (p < .0001). These differences can be visually seen in Figure 9.

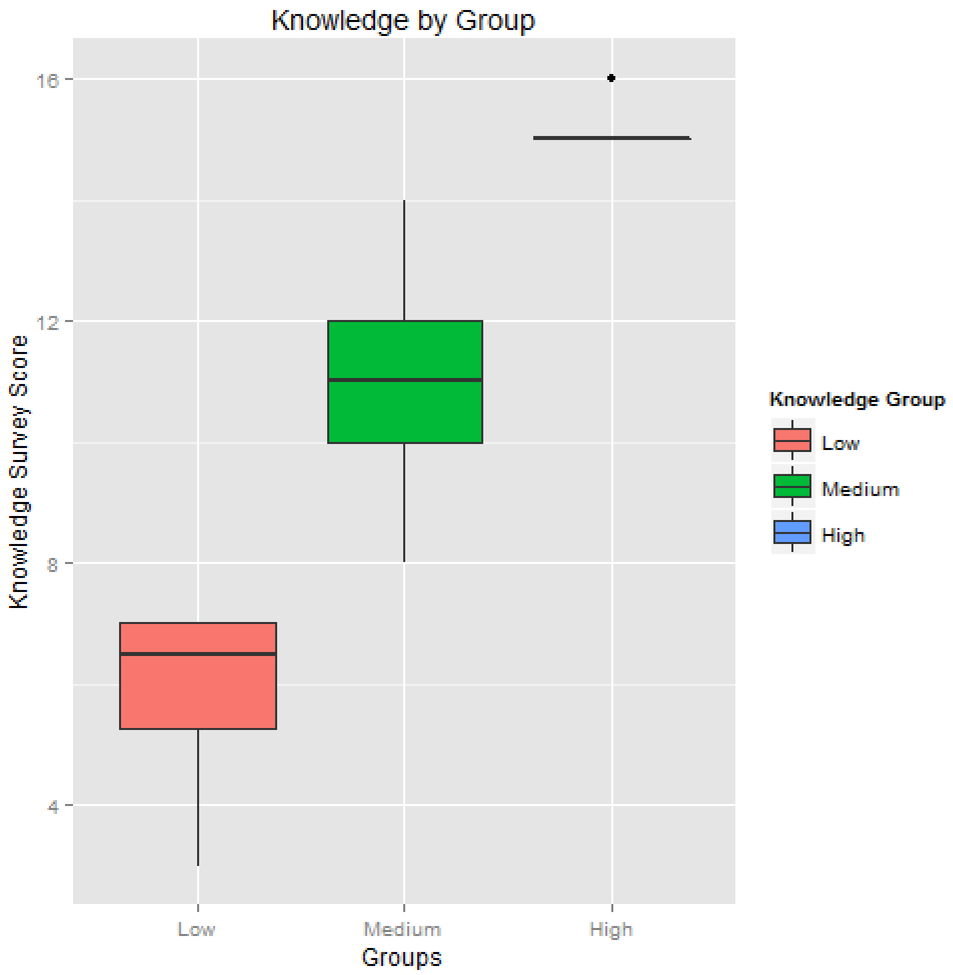


Figure 9. Knowledge scores split by group.

There was no difference found between operators of corn and wheat with respect to performance on the operator harvest knowledge survey, see Figure 10.

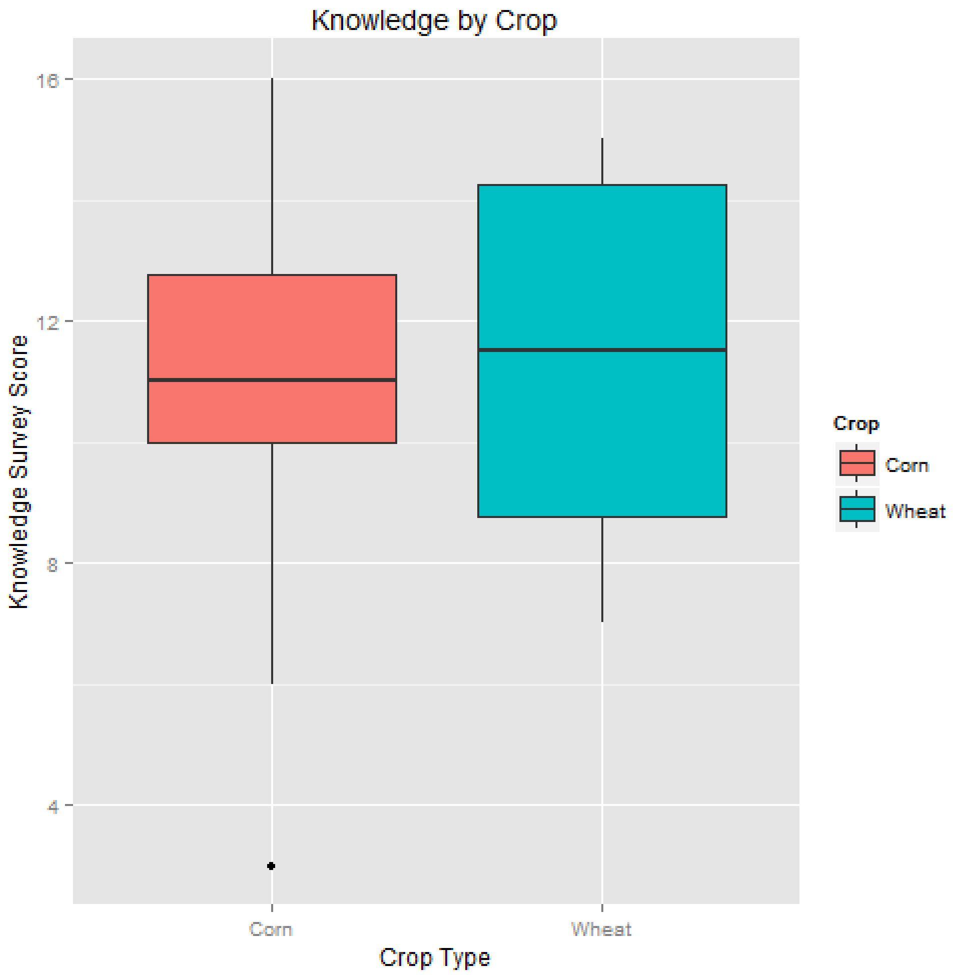


Figure 10. Knowledge scores split by crop.

Operator feedback

Table 5. Operator feedback on technology use in the simulator.

Technology	Interactions	Appropriate Solution?	Optimized combine?	Adjustment rating?
Advisor	75	89% Yes	89% Yes	4.2 (SD 1.0)
Director	96	NA	100% Yes	4.57 (SD 0.74)

Operators felt that Advisor offered an appropriate solution to the issue they reported in 89% of the issues. Advisor and Director saw operators report feeling that the combine was optimized in 89% and 100% of the scenarios, respectively. There was no statistical difference in mean adjustment ratings as seen in the last column of Table 5. A breakdown of which cues were acted upon by scenario can be seen in Table 6, the low-density scenario has the lowest observed interaction rate of 36% where the other three cues were >93% action rate. A low moisture scenario was not presented to corn operators, hence there is only a potential interaction for each of the wheat operators (n = 12).

Table 6. Advisor Interaction observations by scenario.

	Low Moisture	High Moisture	High Density	Low Density	Total
Interactions observed	12	27	26	10	75
Potential Interactions	12	28	28	28	96
Action rate	100%	96%	93%	36%	78%

Electrodermal activity results

Electrodermal activity (EDA) was measured continuously throughout the course of the study. Three specific results were tested in relation to change in operator mental effort. The first tested whether mean EDA values were different between Advisor and Director use, no statistical difference was found. The second tested if a correlation between SEQ scores and EDA data existed. Within Advisor, there was a moderate, negative correlation between SEQ and EDA r(20) = -.523, p = .0012. There was no significant correlation within Director. Third, high knowledge operators reported lower EDA levels during Advisor than low knowledge operators t(13) = 1.8386, p = .08971, 95% CI [-0.375, 4.565], r = .360. No difference found within Director.

Ground truth findings

While not specifically sought out, operators also made many comments which shed insight into their thoughts and feelings as to the real-world performance and repercussions of using these new technologies. Eight operators (29%) mentioned that Advisor was similar to having someone give you a second opinion on what change to make. 13 operators (46%) mentioned they would be more vigilant, be able to work more hours, and likely have less fatigue while using Director; a number of these individuals cited their experience with GPS steering as a baseline for reducing fatigue in the combine. Lastly, 14 operators (50%) commented on their feelings of trust toward Director, of those all but one noted they could come to trust the system over time.

System usability scale

System usability scale (SUS) scores are analyzed within this work as a scores arrived upon by using the suggested SUS scoring methodology as noted in the literature, (Brooke, 1986). The average SUS score for the entire study was 76.43 (SD 12.72) which places this technology above the industry average of 68 (Brooke, 2013). There was no difference found between operator’s SUS scores when compared between crop and wheat operators. Additionally, there was no significant effect found when comparing SUS scores among knowledge groups, F(2) = 2.27, p = .124.

Net Promoter Score

Only 27 of the 28 operators were used to calculate net promoters. One operator did not report net promoter scores. Net promoter is scored by determined by splitting operators into three buckets by their reported scores, detractors (0-6), passives (7-8), and promoters (9-10). The percent promoters, less the percent detractors gives the net promoter score. A perfect score, if all operators are promoters, would be 1. All scores can be seen in Table 7.

Table 7. Net promoter and purchase intent scores.

Technology	Recommender Score	Promoters	Detractors
Advisor Recommend	.33	15	6
Advisor Purchase	.11	10	7
Director Recommend	.78	21	0
Director Purchase	.59	18	2

Operators recommended Director over Advisor when comparing Net Promoter scores, t(26) = 3.98 value, p = .0005. Similarly, operators indicated they were more likely to purchase Director than Advisor when comparing Net Promoter scores, t(26) = 4.01 value, p = .0005.

Discussion

The 28 operators represented a diverse group of individuals who were all fairly experienced and many were the owner of their operation.

Performance

Advisor required more input than Director and that was evident in both observing operator behavior and when considering the needs of each system. Advisor required the user go through multiple steps with every issue encountered where Director required minimal user input. This large reduction in the number of interactions while harvesting allowed operators to engage more with other in cab requirements and should cause less general fatigue. Reduced load and therefore reduced fatigue will allow operators to be more vigilant for longer periods of time while operating the combine.

Additionally, operators did not turn off either system throughout the entire study. While this could partly be because operators were participating in a study to evaluate combine technology, operators ultimately still gave favorable reviews of both Advisor and Director. Similarly, only two operators (7%) elected to make a manual adjustment during the Advisor process which also indicates operators were willing to tolerate the novel technologies the majority of the time and it is possible they would have intervened more in a real combine or specifically when harvesting their own crop when their financial gains were in question.

Fewer errors were committed in the second half of the study when using Director relative to the first when using Advisor. It could be concluded that operators learned how to use the technology over the duration of their experience. Another consideration lies in the fact that the system operators were evaluating was not a final product, but a prototype experience built to expose operators to the concepts of automated harvest products.

The major performance results show that operators could harvest in a more productive manner when using Director. Both scenarios had the same speed limitations set in place by observing the loss monitor, but Director saw fewer reductions in ground speed. This is likely because Director was operating in real time and gave operators less opportunity to observe issues with harvest quality or had fixed any observed issues prior to slowing ground speed. As operators slowed less when using Director, they ultimately spent less time in the same field when using Director, the breakdown of time spent can be seen in Table 4.

Grain tank window

The measure of grain tank window looks by operator gives insight into the question "Do operators use the grain tank window during harvest? If so, how much?" While 18 of the 28 operators checked the grain tank window at least twice (once per field), eight operators only checked it at the time of explanation from the research team member and the remaining two of did not check it at all. The other interesting split though comes from the natural split in the data between 6 looks and 27 looks. 24 of the 28 operators looked 6 or fewer times, while the other four operators looked an average of 31.5 times. It appears the four operators who looked 31.5 times do use their grain tank window frequently, when all others use it less than once per pass.

Operator harvest knowledge survey

The harvest knowledge survey revealed a means to separate operators out based on their knowledge scores into ranked practice expertise scores, or scores which indicate their ability to make appropriate adjustments to the combine and gain a desired effect. A limitation of this work is the relatively small sample size, as more studies are completed, additional data will be gathered. A more complete discussion of the harvest knowledge survey can be seen in Meusel et al., (2016).

Operator feedback

Operators had 96 opportunities to act on each of the four scenarios over the course of all interactions with Advisor. Corn operators were only given three scenarios, hence the reduced potential interactions for the low moisture scenario as seen in Table 6. For the duration of the Advisor field, 67 of the 75 scenario interactions observed indicated operators had been given an “appropriate solution.” The same number of interactions were also reported to have ultimately placed the combine in an “optimized state.” Finally, Advisor received an average of 4.2 (SD 1.0) out of 5 when operators were asked “How do you feel about the last adjustment overall? 1-5, 1 being poor, 5 being ideal.” In comparison, Director received 100% positive feedback when operators were asked if they felt the combine was in an “optimized state” and received an average rating of 4.57 (SD 0.74) when asked the last adjustment question. Operators seemed to favor Director when discussing afterward and recognized the potential value of a system which would act without approval when operating correctly.

Electrodermal activity results

While Advisor and Director did not show overall differences in EDA as expected, EDA did change as expected when compared with SEQ values and between low and high knowledge operator groups. Higher SEQ values indicate the operator rated the specific encounter more favorably, indicating a lower level of imposed mental effort. As EDA should increase with higher mental effort, the inverse correlation found supports this. Similarly, because high knowledge operators have a deeper understanding of the combine as a system they should then exhibit lower EDA levels when exposed to novel harvest interactions as done within the study. This supports that experts with a greater number of mental schemas outperform novices who do not have advanced knowledge within their domain (Larkin et al., 1980; Simon and Newell, 1971). Both the inverse correlation between SEQ & EDA data, and the difference between high and low knowledge operators only occur within Advisor and not Director. This also makes sense considering the amount of knowledge required to successfully use each. Advisor requires the operator to have an existing knowledge of the combine use successfully, thus the gap between high and low knowledge operators is wider. As Director operators independently of the operator, the gap is much smaller as expected, hence no differences within EDA to report.

Ground truth findings

While operators seemed to prefer Director, positive comments were made about both systems, especially when compared to not having either instead of comparing with each other. Advisor was compared to having an expert or knowledgeable friend give you a second opinion or advice while out combining without having to stop to ask.

A strong case was made for Director when operators would make comparison between Director and using a GPS guided steering and tracking system. Operators who were familiar with GPS controlled steering inputs (which was the majority of operators) made the general comment that enabling GPS steering was able to free up cognitive resources from the operator so they could concentrate more fully on other measures and alerts that would have likely gone under-observed or neglected all together. The parallel was that Director could potentially free up additional cognitive resources for additional monitoring or even accomplish other tasks while completing the target harvest task.

Lastly, the operators trust within this system was also discussed with roughly half of all operators. Again, past successful interactions with in cab technology such as GPS steering applications have given operators confidence that future technologies will also work and have helped improve their likelihood to adopt and rate of adoption.

System usability scale

The average SUS score for the entire study was 76.43 (SD 12.72) which places this technology between “good” and “excellent” on the adjective ratings scale (Bangor et al., 2009). SUS has been used in a wide variety of technology domains such as Web, Cell Phones, GUIs, Hardware, and others. Applying the SUS measure to technology within the combine makes sense and is a measure that will be taken in future combine simulator studies going forward.

In addition to calculating the SUS, an ANOVA model was tested and found no differences between the three knowledge groups with respect to SUS scores. SUS scores broken into knowledge groups can be seen in Figure 11.

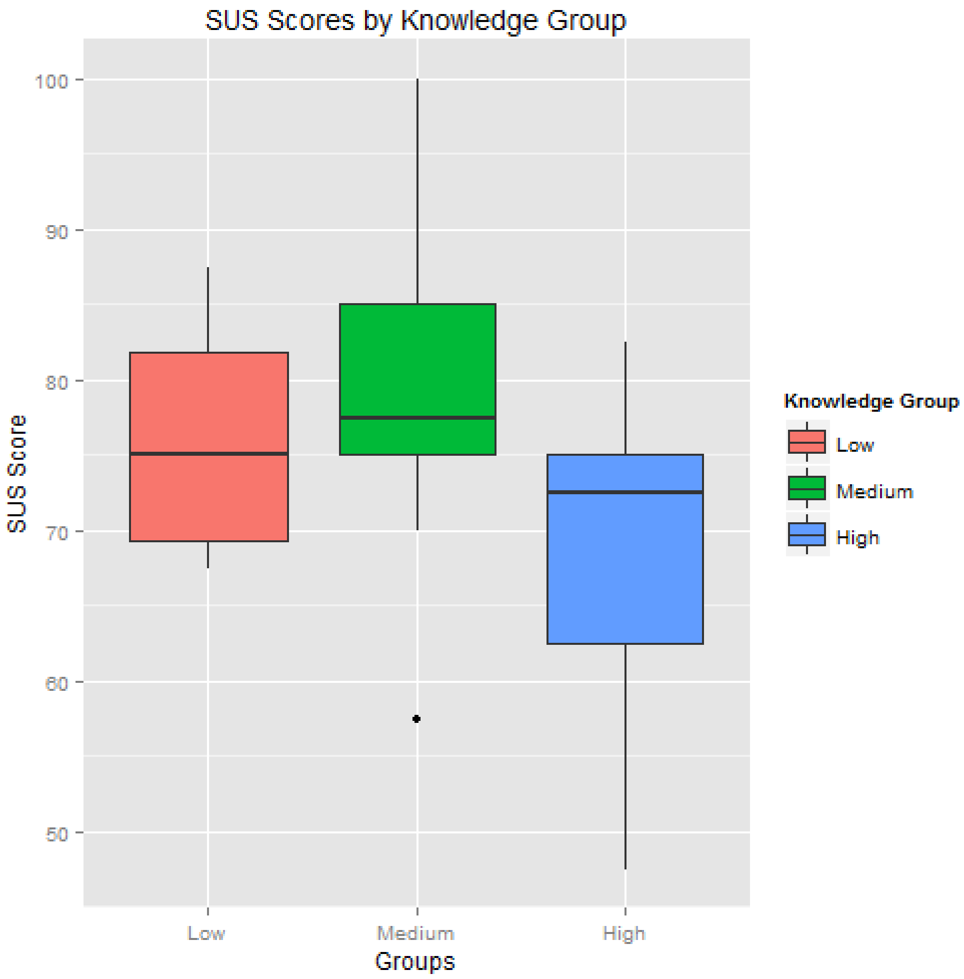


Figure 11. SUS scores by knowledge group, no significant differences.

Net promoter

Of the 27 operators who did report Net Promoter scores, they more highly recommended Director over Advisor and additionally reported they were more likely to purchase Director over Advisor. This is not surprising given the experimental setup as Advisor did require more input and attention than Director. Also, IC had the potential to perform poorly if given incorrect operator instruction either by missing a cue within a scenario or identifying the incorrect cue altogether. Director by default would perform optimally if never adjusted and therefore saw improved performance over Advisor at any time the scenario went unnoticed as it reduced the potential operator error to 0%. Overall, operator preference for Director is sensible and supported. The larger question is where does Advisor fit if all operators prefer Director? Expert operators made comments that Advisor has less value to them as they do not need help adjusting, but they do see value in full automation with Director.

Conclusion

Both combine technologies, Advisor and Director, were well received by operators and given positive recommender and purchase scores. Exit interview comments and positive scores indicate that operators are open to the idea of semi-autonomous and fully autonomous combine technology aids operating in real time while they harvest. A common thread of comparison with GPS based steering technology leads operators from all brand experience to be positive and welcoming of additional technological implementation. Assisting technologies such as Advisor and Director are welcome for both their reduction of operator workload and general fatigue reduction.

Ultimately though, operators in this study preferred and showed improved performance measures when using Director relative to Advisor. Operators performed fewer mistakes, opted to interrupt the system less, and spent less time in the field when using Director. These improvements to performance and efficiency cannot be understated in a domain where efficiency and quality of harvest are directly tied to the financial outcome of the operator in the combine. While there is no baseline data to identify the statistical improvement for Advisor, operators did review Advisor favorably as well and in particular novice operators appreciated the feedback offered to them without having to call someone else for help while harvesting.

Advisor and Director were not viewed as direct competitors by the operators, but as complimentary services on a spectrum of automation. Similar to the current discussion surrounding automated driving in commercial road vehicles (SAE, 2014), Advisor and Director can be seen as subsequent levels of automation following the initial step of GPS guided steering and other single system automation tools within the combine. It is important to note that although operators approved of these systems, many operators expressed their preference to take control for emergency and unusual scenarios.

Overall, the largest implication here is that when technology works as intended, humans seem to prefer the system which takes full responsibility and returns comparable or better results when compared with their own performance. If there are other tasks available for humans to spend their mental resources on, offloading other tasks becomes increasingly attractive. To offer a counterpoint though, some operators did mention that operating the combine manually was akin to "going fishing" and the brief time they are able to operate the combine during the year is somewhat therapeutic. In this scenario, the operator is not seeking automation, simply better tools and controls to successfully complete their task.

References

Anderson, D. (2011). Don’t Leave Yield Behind in the Field.

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114–123. http://doi.org/66.39.39.113

Bashiri, B., & Mann, D. D. (2014). Automation and the situation awareness of drivers in agricultural semi-autonomous vehicles. *Biosystems Engineering*, *124*, 8–15. http://doi.org/10.1016/j.biosystemseng.2014.06.002

Bella, F. (2008). Driving simulator for speed research on two-lane rural roads. *Accident Analysis and Prevention*, *40*(3), 1078–1087. http://doi.org/10.1016/j.aap.2007.10.015

Birrell, S. A., & Young, M. S. (2011). The impact of smart driving aids on driving performance and driver distraction. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*(6), 484–493. http://doi.org/10.1016/j.trf.2011.08.004

Brooke, J. (1986). SUS - A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*, 194.

Brooke, J. (2013). SUS : A Retrospective. *Journal of Usability Studies*, *8*(2), 29–40.

Calatayud, D., Arora, S., Aggarwal, R., Kruglikova, I., Schulze, S., Funch-Jensen, P., & Grantcharov, T. (2010). Warm-up in a virtual reality environment improves performance in the operating room. *Annals of Surgery*, *251*, 1181–1185. http://doi.org/10.1097/SLA.0b013e3181deb630

de Groot, S., Mulder, M., & Wieringa, P. A. (2011). Car Racing in a Simulator: Validation and Assessment of Brake Pedal Stiffness. *Presence: Teleoperators and Virtual Environments*. http://doi.org/10.1162/pres_a_00033

Deere, J. (2013). Combine Adjustment Guide S-Series.

Duncan, J. R., & Turner, R. J. (1991). Operator Performance with simulated Harvesting combine Automatic Controls. *The American Society of Agricultural Engineers*.

Endsley, M. R., & Kaber, D. B. (1999). *Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics* (Vol. 42). http://doi.org/10.1080/001401399185595

Endsley, M. R., & Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(2), 381–394. http://doi.org/10.1518/001872095779064555

Fone, N. (2007). Combine Settings: Sieves.

Hibberd, D. L., Jamson, H., & Jamson, S. L. (2015). The design of an in-vehicle assistance system to support eco-driving. *Transportation Research Part C: Emerging Technologies*, *58*, 732–748. http://doi.org/10.1016/j.trc.2015.04.013

Horrey, W. J., & Wickens, C. D. (2006). Examining the Impact of Cell Phone Conversations on Driving Using Meta-Analytic Techniques. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(1), 196–205. http://doi.org/10.1518/001872006776412135

IH, C. (2009). Case IH Combine Productivity Guide.

Jamson, A. H., Hibberd, D. L., & Merat, N. (2014). Interface design considerations for an in-vehicle eco-driving assistance system. *Transportation Research Part C: Emerging Technologies*. http://doi.org/10.1016/j.trc.2014.12.008

Karimi, D., & Mann, D. D. (2008). Role of Visual Cues from the Environment in Driving an Agricultural Vehicle. *The Ergonomics Open Journal*, *1*, 54–61. http://doi.org/10.2174/1875934300801010054

Karimi, D., Mann, D. D., & Ehsani, R. (2008). Modeling of straight-line driving with a guidance aid for a tractor-driving simulator. *Applied Engineering in Agriculture*, *24*(4), 403–408.

Krug, S. (2009). *Rocket Surgery Made Easy*. Berkeley, CA: New Riders.

Kruglikova, I., Grantcharov, T. P., Drewes, A. M., & Funch-Jensen, P. (2010). The impact of constructive feedback on training in gastrointestinal endoscopy using high-fidelity Virtual-Reality simulation: a randomised controlled trial. *Gut*, *59*, 181–185. http://doi.org/10.1136/gut.2009.191825

Larkin, J., McDermott, J., Simon, D. P., & Simon, herbert a. (1980). Expert and novice performance in solving physics problems. *Science (New York, N.Y.)*, *208*(4450), 1335–1342. http://doi.org/10.1126/science.208.4450.1335

Lee, W.-S., Kim, J.-H., & Cho, J.-H. (1998). A driving simulator as a virtual reality tool. *Proceedings IEEE International Conference on Robotics and Automation*, *May*, 71–76. http://doi.org/10.1109/ROBOT.1998.676264

Lee, Y., Lee, J. D., & Boyle, L. N. (2005). Change detection performance under divided attention with dynamic driving scenarios. In *Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design* (pp. 195–201).

Loftin, R. B., & Kenney, P. J. (1995). Training the hubble space telescope flight team. *IEEE Computer Graphics and Applications*, *15*, 31–37. http://doi.org/10.1109/38.403825

Luecke, G. R. (2012). GREENSPACE : Virtual Reality Interface for Combine Operator Training. *Presence: Teleoperators and Virtual Environments*, *21*(3), 245–254. http://doi.org/10.1162/PRES_a_00110

Mann, D., Bashiri, B., Rakhra, A., & Karimi, D. (2014). Development of a tractor driving simulator to research ergonomics of agricultural machines. *International Conference of Agricultural Engineering*, (1988), 6–10.

Mclane, R. C., & Wierwille, W. W. (1975). The Influence of Motion and Audio Cues on Driver Performance in an Automobile Simulator. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *17*, 488–501. http://doi.org/10.1177/001872087501700508

Metzger, U., & Parasuraman, R. (2001). The role of the air traffic controller in future air traffic management: an empirical study of active control versus passive monitoring. *Human Factors*, *43*(4), 519–528. http://doi.org/10.1518/001872001775870421

Meusel, C., Grimm, C., Gilbert, S., Luecke, G. (2016). An Agricultural Harvest Knowledge Survey to Distinguish Types of Expertise. *60th Annu. Meet. Hum. Factors Ergon. Soc.*

Mowitz, D. (2013). 7 Combine Tweaks to Boost Speed.

Parasuraman, R., Cosenzo, K. A., & De Visser, E. (2009). Adaptive automation for human supervision ocf multiple uninhabited vehicles: Effects on change detection, situation awareness, and mental workload. *Military Psychology*, *21*(2), 270–297. http://doi.org/10.1080/08995600902768800

SAE. (2014). Automated Driving: Levels Of Driving Automation Are Defined In New SAE International Standard J3016.

Simon, H. A., & Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*, *26*(2), 145–159. http://doi.org/10.1037/h0030806

Son, K. (2001). A driving simulator of construction vehicles - Son. *International Journal of the Korean Society of Precision Engineering*.

Tate, D. L., Sibert, L., & King, T. (1997). Using virtual environments to train firefighters. *IEEE Computer Graphics and Applications*, *17*, 23–29. http://doi.org/10.1109/38.626965

Triantafyllou, K., Lazaridis, L. D., & Dimitriadis, G. D. (2014). Virtual reality simulators for gastrointestinal endoscopy training. *World Journal of Gastrointestinal Endoscopy*, *6*, 6–12. http://doi.org/10.4253/wjge.v6.i1.6

Wehrspann, J. (2004). Combine Tips From the Pros.

Weinberg, G., & Harsham, B. (2009). Developing a Low-Cost Driving Simulator for the Evaluation of In-Vehicle Technologies. *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '09*, (AutomotiveUI), 51–54. http://doi.org/10.1145/1620509.1620519

Weir, D. H. (2010). Application of a driving simulator to the development of in-vehicle human-machine-interfaces. *IATSS Research*, *34*(1), 16–21. http://doi.org/10.1016/j.iatssr.2010.06.005

Yoon, J., & Manurung, A. (2010). Development of an intuitive user interface for a hydraulic backhoe. *Automation in Construction*, *19*(6), 779–790. http://doi.org/10.1016/j.autcon.2010.04.002

CHAPTER 3

EVALUATING SIMULATOR CUE VALIDITY WITHIN A HIGH-FIDELITY COMBINE SIMULATOR

This chapter will be submitted to *Computers and Electronics in Agriculture*. Author list:

Chase Meusel, Chase Grimm, Jordan Starkey, Stephen B. Gilbert, Brian Gilmore, Greg Luecke, Don Kieu

Chase Meusel's role in this research included contributions to the experimental design, participant recruitment, data collection, data coding and analysis, and primary authorship on the paper. Chase Grimm helped write the initial draft of both the background and methods section in addition to performing some statistical analysis. Jordan Starkey ran user studies and performed some qualitative analysis.

Abstract

Research, development testing of large and operator training of harvest equipment has become increasingly expensive and complex. Simulators can offset these costs, but it is critical to evaluate the fidelity of the simulator experience to ensure that it will lead to operator behaviors that match those in the real world. This research describes the validation process of new visual cues for the header within a combine harvester simulator. Results demonstrate that the visual cues within the combine simulator successfully communicated their intended message overall. Additionally, operator knowledge was positively correlated

with the correct actions taken based on the cues, indicating the cues were relatively realistic. This research demonstrates a robust approach for ensuring that simulator fidelity is sufficient for training and product evaluation within a simulator.

Introduction

Harvesting is an increasingly complex task. Today's harvest operators oversee an increasing number of controls and systems in their combines while also adding more tasks to the act of harvesting itself. The increased scope of today's harvest operation places increasing physical and mental workload on the operator. In addition to operating a combine, a farmer must manage grain transport, analyze weather reports, assess crop conditions, adapt to equipment issues, and constantly communicate with a variety of sources. As tasks have been added to the process of harvesting crops, the corresponding complexity has not been appropriately reduced for the operator.

The North American harvest season occurs once per year, and most farmers will not operate any harvest specific technology outside of that window. One way to provide training to allow operators to keep up their operational competency is by utilizing combine simulators. Simulator practice allows operators to utilize technology they otherwise would not use throughout the year and reduce the amount of reacclimating time needed during harvest season. Also, over the past decade, simulators have served as a valuable resource to research the feasibility of new products and the ergonomics of agricultural design (Luecke, 2012; Mann, Bashiri, Rakhra, & Karimi, 2014). Simulators allow year-round product testing of what can be considered a seasonal technology. One of the largest benefits of using a simulator is being able to create repeatable and reproducible research studies. However, in order to run an effective study, the graphics and visual cues in the simulator must meet operator expectations and accurately depict real life scenarios (Karimi & Mann, 2008).

This paper presents a cue validation study, which was run in order to ensure that 1) graphics in the virtual field displayed in a combine simulator met the expectations of farmers, and 2) the visual cues were distinguishable and accurately depicted. Whenever an external stimulus is required to trigger a human reaction within a system, the fidelity of that stimulus must be considered. This is particularly true of stimuli that are present within virtual environments, such as the one used within a combine simulator. The information that is presented must be of a high enough fidelity that it successfully allows the participant to take action, independent of how "real" it may be appear (Stoffregen, Bardy, Smart, & Pagulayan, 2003). Also, in the service of reducing operator cognitive load, these cues should clearly communicate their intended effect without imposing artificially higher cognitive load (i.e. bad cues are difficult to understand).

A possible resolution to lowering operator workload in general is to automate tasks which involve high physical and mental workload. This solution has proven effective in previous cases such as GPS-enabled guidance systems and automated harvest systems (Meusel, 2014). Automating specific harvest tasks not only can assist in reducing overall cognitive load experienced, but it also adds the potential for better system economy by lowering operation time, reducing equipment wear, increasing system reliability, and easier system maintenance (Duncan & Turner, 1991).

This research supports a larger effort, which is to improve the quality of the combine simulator platform for use in product development and testing (Kieu et al., 2017; Luecke, 2012). This cue validation study was partly motivated by a desire to evaluate operator use of harvester reel control. Some operators have been observed to adjust the reel frequently, and an automation system could relieve the operator’s cognitive load. For operators in the simulator to make appropriate judgments of the reel position and speed, a higher fidelity virtual environment was required in the current combine simulator used at the Virtual Reality Applications Center at Iowa State University. While the combine simulator graphics had previously provided sufficient fidelity to assess technologies related to ground speed, fan, sieve, and other combine settings, to the authors’ knowledge, the reel and the detailed process of grain passing through it and onto the belt had not been carefully simulated. Kieu et al. (submitted) describe the technology used to simulate grain in the reel and on the belt, which were advancements on the simulator platform first described by Luecke (2012). In this paper, the visual cues offered by that technology have been strenuously validated. It was crucial that the improvements made to the visuals in the combine simulator offered operators cues that were high enough fidelity that they could make realistic reel adjustment decisions. This research contributes a process for evaluating cue fidelity for simulators, as well as evidence that traditional agricultural simulator graphics are not sufficient for header and reel-based tasks.

Background

Running studies with the combine simulator requires an understanding of simulator testing within industrial applications and testing agricultural technologies. Simulators benefit from existing research which supports the use of virtual environments as training platforms in a few ways. Simulators within virtual environments are noted as successful with project-based learning, constructivism, exploratory learning, situated learning, and computer-mediated collaboration, which are all valuable learning attributes (Bransford, Brown, & Cocking, 1999). Dalgarno & Lee (2010) review the affordances of 3D virtual environments, which include items such as realistic display of environments, smooth display of changes, consistency of object behavior, control of environmental attributes and behavior, and others which directly apply to the combine simulator used in this work.

The automotive industry has been using driving simulators since the 1960s (Weir, 2010). Driving simulators allow researchers to simulate actual driving conditions and situations within a controlled and safe environment (Lee, Kim, & Cho, 1998). Flexibility and cost savings are also benefits of using a simulator in studies rather than the actual vehicle (Mann et al., 2014). Simulators have made it possible to observe operators with optimal stimulus and response control in any desired environment (Rizzo, Jermeland, & Severson, 2002). Increases in computational power and storage paired with decreasing costs has been a boon to driving simulators, which are currently used as a research tool in human factor studies and vehicle development (Karimi & Mann, 2008). There are other non-automotive domain virtual environment examples of training which have had positive results communicating virtual cues to users and then transferring that training to the real world, such as repairing the Hubble space telescope (Loftin & Kenney, 1995), fighting fires on board a naval ship (Tate, Sibert, & King, 1997), and medical students training to perform various medical procedures (Calatayud et al., 2010; Kruglikova, Grantcharov, Drewes, & Funch-Jensen, 2010; Larsen et al., 2009).

Simulators have been developed for vehicles other than automobiles, such as construction vehicles (K. Son, Goo, Yoo, Lee, & Lee, 2001) and agricultural equipment (Mann et al., 2014). However, while the automotive industry has been using driving simulators for many years, a relatively small number of such simulators exist for agricultural vehicles, resulting in a void in the literature (Karimi & Mann, 2008; Karimi, Mann, & Ehsani, 2008; Wilkerson, Asbury, Prather, & Lown, 1993)   Additionally, because there are so few agricultural simulators in existence, it can be more cost-effective to test feature development on actual equipment; some studies are well suited for simulator development by their design, but are simply executed on small scale tractor platforms (Pranav, Pandey, & Tewari, 2010; Pranav, Tewari, Pandey, & Jha, 2012).

Automobiles and agricultural vehicles have certain commonalities but differ in terms of many purposes and functions. Thus, the development and use of simulators for agricultural vehicles has both similarities and differences from that of automobiles. Where automotive simulators may spend additional resources modeling traffic and AI driving behavior, agricultural simulation is generally built for the individual operator and focuses on precision farming features. As tractor operators generally use a guidance system, which increases accuracy of straight-line driving, agricultural simulators require precision GPS control systems. An example of this type of work can be seen with Karimi & Mann, (2008) who developed and validated a simulation model of parallel swathing (driving in parallel paths to cover a field) in a tractor-driving simulator. The model accounted for tractor self-deviation and guidance system error. The study’s field experiments were in close agreement with the simulator experiments regarding frequency composition of lateral deviations, thus showing the model’s value in simulator fidelity. In a separate work, Karimi et al. (Karimi & Mann, 2009), used the same tractor simulator to work on developing the correct steering feel for the tractor. This allowed the researchers to explore what types of haptic feedback would provide the correct fidelity cue, as the steering is completely electronic and not connected to any physical wheels while in the simulator.

The variety of agricultural vehicles results in differing designs and research questions for simulators. For example, in the design of a tractor-air seeder driving simulator, the first step was to conduct a function-oriented task analysis to identify the required functions, tasks, and subtasks (Mann et al., 2014). A main finding was that operators allocated a substantial portion of their time to controlling the air seeder. Additionally, operators had to monitor the air seeder that was mounted behind the tractor. To simulate this characteristic of a tractor-air seeder, researchers put two computer monitors behind the cab (one to rear-left and one to the rear-right of the operator's seat). Thirty-two images created a panoramic view using 32 images, forming the field boundary for the tractor-air seeder. This simulator is being used for two research issues: to determine an appropriate automation design for agricultural vehicles and to understand the impact of display design on an operator's situation awareness in a semi-autonomous agricultural vehicle (Bashiri & Mann, 2014).

Luecke (2012) developed a virtual reality interface for a combine simulator. The initial objective was to develop an accurate operator control interface which would enable training of the many combine functions available in a modern combine. The simulator that was developed allowed the operator to sit in a real combine seat with the actual operator controls and displays. Previous work, (Meusel, 2014; Meusel et al., submitted) used the technology developed from Luecke's (2012) work to develop and test the Advisor and Director automated harvest systems. Advisor was an in-cab technology which offers internal combine adjustment recommendations to the operator while Director automatically made adjustments to the system.

Research of automated technologies in agriculture is important and can help reduce the workload required to carry out tasks as well as increase performance (Scarlett, 2001). One study has compared the fuel usage rate as measured automatically by the CAN bus with physically measured fuel rates to assist in cost analysis for farmers (Marx et al., 2015). Another study has implemented particle filters in robots navigating through a corn field, furthering advancements in precision agriculture (Hiremath, van der Heijden, van Evert, Stein, & Ter Braak, 2014). A different study that looked at particles analyzed the particles' motion in a variable-amplitude screen box, improving cleaning and screening of agricultural materials (Ma, Li, & Xu, 2015). Another study developed a wireless system to automatically identify every mechanized operation on the farm to help improve data collection efficiency and optimize field logistics (Calcante & Mazzetto, 2014).

From wayfinding literature in traditional virtual environment studies, it is known that rich visual scenes are not required for successful virtual experiences to happen (Kelly, Sjolund, & Sturz, 2013; Ruddle & Lessels, 2006; Sjolund, 2014). These studies support the testing of cues on a platform that is not considered high fidelity visually but still allows operators to provide meaningful feedback while in the virtual environment. Thus, measures of graphical fidelity such as photorealism, that have been applied to computer graphics in movies and video games, are not as relevant to this research. The more critical components of fidelity in this context can be measured in one of two ways: experiential fidelity or action fidelity. Experiential fidelity is based on subjective user feedback about their presence within the virtual environment, while action fidelity is the user's ability to perform using the information available within the virtual environment (Stoffregen et al., 2003). This study focuses on action fidelity, whether operators can perform appropriately given the cues available. Additionally, because graphical fidelity is not the highest priority, other types of cues can be supported as ways to provide additional depth to the operators, such as rumble haptic feedback or detailed instrument feedback on the system diagnostic displays.

Methods

Overview

The objective of the cue validation study was to evaluate cues in a virtual farm field designed for the task of making realistic control adjustments to the header reel (height, fore/aft, speed). To prepare for this study, three expert agricultural engineers were initially solicited for open-ended feedback on the visual cues, and adjustments were made based on their feedback. During this phase, expert engineers gave critical feedback to address a missing cue within the simulator, which was the lack of activity in the header. While viewing the original visual cues, an expert engineer from John Deere stated, “This won't necessarily trigger me to make an adjustment because I don't see an issue with the material flow.” To drive the desired behavior of having operators adjust the reel, the expert suggestion was then given to allow operators to “see some crops sitting on the cutterbar [to indicate] you don't have reel engagement.” As this suggestion was echoed across experts, material flow and cues within the header itself were added. See below for details.

Current harvest operators were then recruited for this combine simulator study. Before the simulation portion of the study, participants answered demographics questions, general harvest questions, and questions regarding their personal experience operating a combine. Operators were then tasked with adjusting the combine within a variety of changing field conditions while also reporting which condition they were experiencing at the time of harvest. The researchers sought to gather participants' feedback on a potential reel automation technology. Researchers measured any adjustments made by the participant during the study that would indicate that the operator was overriding the automated system. Researchers also gathered participants' electrodermal activity (EDA) as a measure of cognitive load (Haapalainen, Kim, Forlizzi, & Dey, 2010; J. Son & Park, 2011) or stress (Katsis, Katertsidis, Ganiatras, & Fotiadis, 2008; Setz et al., 2010). Once participants completed their tasks in the simulator, they took a harvest knowledge survey.

The harvest knowledge survey was a key component, as it can be used as an indicator of how well an operator may perform in the combine by way of complex adjustments to the combine's harvest settings in a real field. Having a baseline of understanding about the operator gives an idea of how likely an operator might be to not only recognize the cue being shown, but also to take the correct action on that cue. Without understanding the operator's knowledge of harvest adjustments, the process to evaluate cues becomes entirely about graphical fidelity and less about the operator's ability to also take the right corrective action.

Figure 12 illustrates the possible outcomes when a cue is presented to an operator. The shaded regions illustrate the several ways that the desired outcome can go awry. The key point for this analysis is that if the cue is not perceived or the cue is perceived but then not interpreted correctly, it's not clear whether the cue or the operator is to blame. If the cue is not well designed, the error rests with the cue. If the operator is inexperienced, not recognizing the cue could stem from with the operator's lack of knowledge. To resolve this ambiguity, we used two approaches: 1) recruiting operators who had at least two years' experience harvesting in the past four years, and 2) using the harvest knowledge survey to establish a knowledge expectation for each operator.

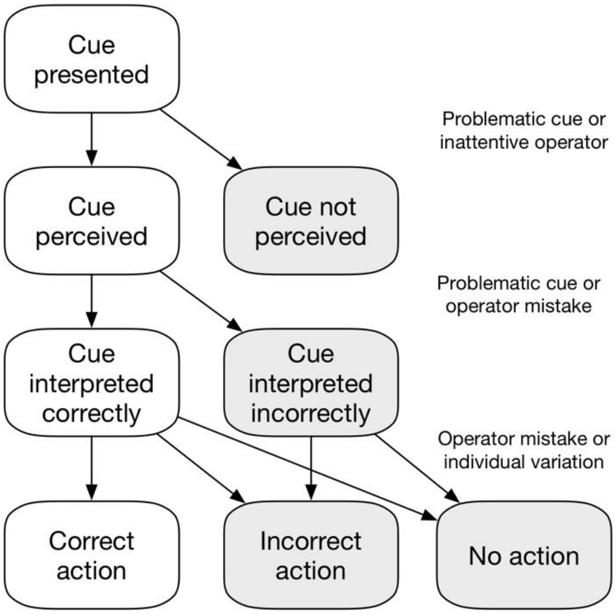

Cue presented
Problematic cue or inattentive operator
Cue perceived
Cue not perceived
Problematic cue or operator mistake
Cue interpreted correctly
Cue interpreted incorrectly
Operator mistake or individual variation
Correct action
Incorrect action
No action

Figure 12. The possible outcomes when an operator encounters a harvest cue.

Simulator Cues

This study was focused on the operator's ability to 1) perceive the cue presented, 2) successfully identify the cue, and 3) take the correct action on that cue. This type of fidelity measurement is considered a type of action fidelity and relies on the operator's ability to successfully perform based on the information the cue presents them (Stoffregen et al., 2003).

This study initially used signal detection theory (Abdi, 2007; Macmililan, 2002) as a framework to validate the varying virtual crop cues presented to the operators. This method is traditionally used to categorize signal vs. noise to identify the presence of a signal. Ultimately in this study, the signal was the cue presented to operators, and correct responses were counted for analysis. False alarms were not penalized differently than true misses.

Generally, visual cues within the combine simulator were considered high-enough fidelity to be actionable, even though not photorealistic. Previous studies (Luecke, 2012;

Meusel, 2014) demonstrated that the graphics present in the simulator were sufficient for operators to engage with farmer activities. The question in this study was whether the new cues, developed with the same graphics fidelity as previous cues, were sufficient in their look and behavior to convey to operators what was happening in the header. See Figure 13 for examples of some of the crop imagery used within the field cues.

In addition to the cues in the field, cues were added to the header itself based on feedback from the expert agricultural engineers. To the authors' knowledge, no harvest simulator has previously modelled the process of the plant passing through the reel, forming fragments on the belt, and periodically aggregating into clumps as it moves along the belt. Technical details of the cue graphics and the unique complexity of the grain particle simulation, including a stochastic clumping model, is discussed in Kieu et al. (2017).

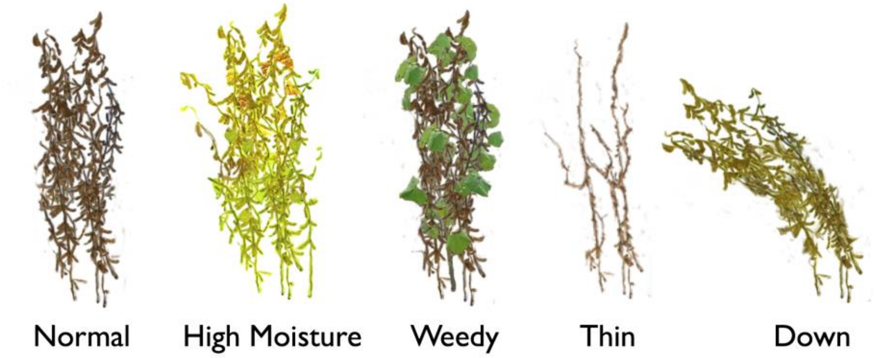


Figure 13. Examples of soybean plant models used within the simulator field to convey some of the cues.

Participants

The population sample consisted of 14 farmers with at least 2 years' experience operating a combine in the past 4 years. Operators were recruited from a large pool of individuals who had indicated their willingness to participate in research for an agriculture equipment company. Participants received $150 compensation for participating, which required approximately 1.5 hours, in addition to travel cost reimbursement. All participants were at least 18 years old; the majority were 31-40.

Independent Variables

The independent variable (IV) for this study was the cue presented to the operator. Levels of this IV included crop density, crop moisture, crop quality, and crop condition. All participants harvested the exact same field, with the same crop scenarios and visual cues placed at the same coordinates throughout the field. The moisture and density of the crop were both delineated in the simulator by changing the visual representation of the crop. The quality of the crop was identifiable via the information available on simulator controllers (e.g., the yield monitor and moisture meter), visual images in the virtual field, and visual representation on the draper head. A complete layout of the field, with cues, can be found in Figure 14. The complete list of cue types can be found in Table 8.

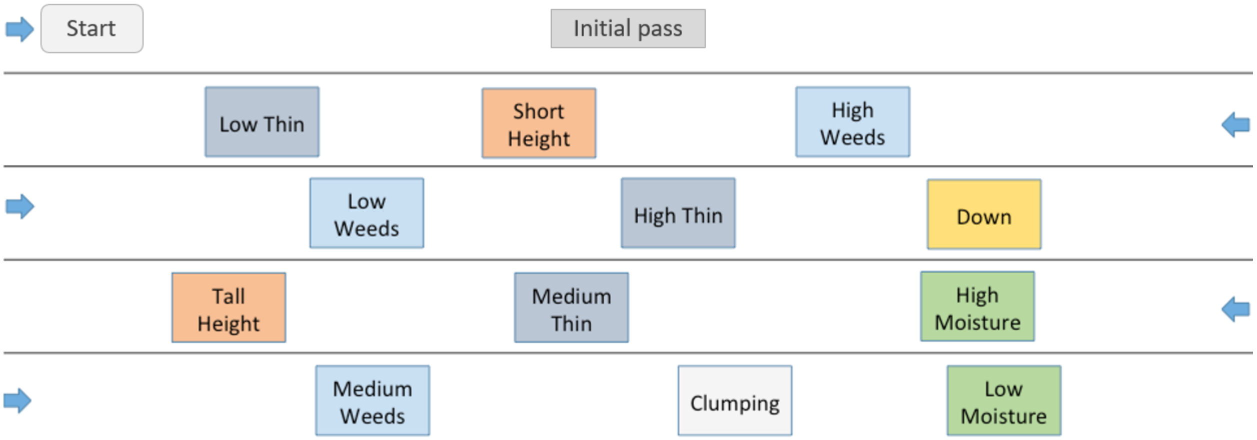


Figure 14. Virtual field layout with various cues, 1/2 mile passes.

Table 8. The types of cues presented.

Cue	**Expected Solution**	**Source of solution**
Weeds (high)	Reel down and back	Huitink, 2000; Operator’s Manual
Short height	Reel up/down	Operator’s Manual
Thin crop (low)	Reel down	Iowa State University Extension
Weeds (low)	Reel down and back	Huitink, 2000; Operator’s Manual
Thin crop (high)	Reel down	Iowa State University Extension
Down and tangled	Reel down and forward	Operator's Manual
Moisture level (high)	Slow combine	Iowa State University Extension; Operator’s Manual
Thin crop (medium)	Reel down	Iowa State University Extension
Tall crop	Reel up	Operator’s Manual
Weeds (medium)	Reel down and back	Huitink, 2000; Operator’s Manual
Clumping on the belt	Reel up	Huitlink, 2000
Moisture level (low)	Slow combine	Iowa State University Extension; Operator’s Manual

Dependent Variables

Every adjustment the participants made was recorded and used to gather the results from this study. Participants were also given the opportunity to comment on each cue they noted or acted on. Participants were told to harvest the field as they would an actual field, and were encouraged to make the same types of adjustments in the simulator as they do in a real harvest scenario. The variables measured are below in Table 9.

.

Table 9. Dependent variable list.

Dependent Variable	**Metric**	**Unit**	**Frequency of Collection**	**Data Type**
Performance	Correct cue	Count	Per cue condition	Ordinal
Performance	Correct action	Count	Per cue condition	Ordinal
Performance	Interface Interactions	Count	Per cue condition	Ordinal
Cognitive Load	EDA	Microsiemen s	Continuous (32hz)	Continuous

Experimental Design & Procedure

Prior to collecting feedback from operators, cues needed to be defined. To build the initial set of cues, agriculture extension documents and harvest publications were reviewed (Butzen, 2013; Huitink, 2000; Minnihan, Hanna, Isaac, & Couser, 2003). That led to a list of 44 cues including stalks dropping, uneven feed intake into the separator, tough green stems, and many more. That list was then reviewed and cut back based on what was first most likely to occur in the field, and which cues most frequently led to reel adjustments. That list of cues was then shared with three farmers and a round of initial feedback was given via one-on-one interviews. These farmers provided feedback on which cues operators utilized the most and what conditions would be the most important to test. Then, three expert agricultural engineers were consulted to provide the first round of feedback on visual cues from video recorded within the harvest simulator to be used for testing. Engineering experts consisted of both academic engineering professionals and agricultural engineering professionals. After the expert feedback was collected, changes based on that feedback were implemented, and a final version of the cues was built for the feedback of the recruited farmers.

Participants first were given the opportunity to review the IRB consent form and sign when they were ready to proceed. Participants were fitted with the EDA sensor to begin gather baseline data and then began the pre-survey and upon completion were seated in the combine simulator. A researcher was then seated beside the participant, similar to a passenger riding along in the combine. Each participant harvested a virtual soybean field in the simulator. The field consisted of 5 half-mile rows, with the first row being all normal crop (see Figure 14). Participants were informed to ignore tank unloading, as there was a virtual grain tank, and it was assumed to be infinite in capacity. Participants were instructed to use the first pass to familiarize themselves with the simulator controls and adjust to the graphics of the simulator. Participants were asked to harvest the entire field pass by pass and make any adjustments they would normally make in a real harvest scenario. Total time in the simulator lasted on average 54 minutes and entire study lasted approximately 1.5 hours per participant.

Participants were prompted to verbalize their actions and identify what crop condition they believed they were in. Participants were also asked brief questions about what visual changes could be made to improve the crop visuals.

After participants completed the simulator portion of the study, they were asked to complete a survey. The Harvest Knowledge Survey (Meusel, Grimm, Gilbert, & Luecke, 2016) was designed as a mechanism to evaluate operator's understanding of the combine at a system level. The results have been shown to represent how well an operator is able to make the correct adjustment to the combine given a set of conditions. What this questionnaire does not disambiguate, however, is whether operators understand why the changes they make will or will not work. Yet even as a tool for simply evaluating general operator understanding, the Harvest Knowledge Survey has been shown to successfully separate out operators into meaningful groups.

Cues

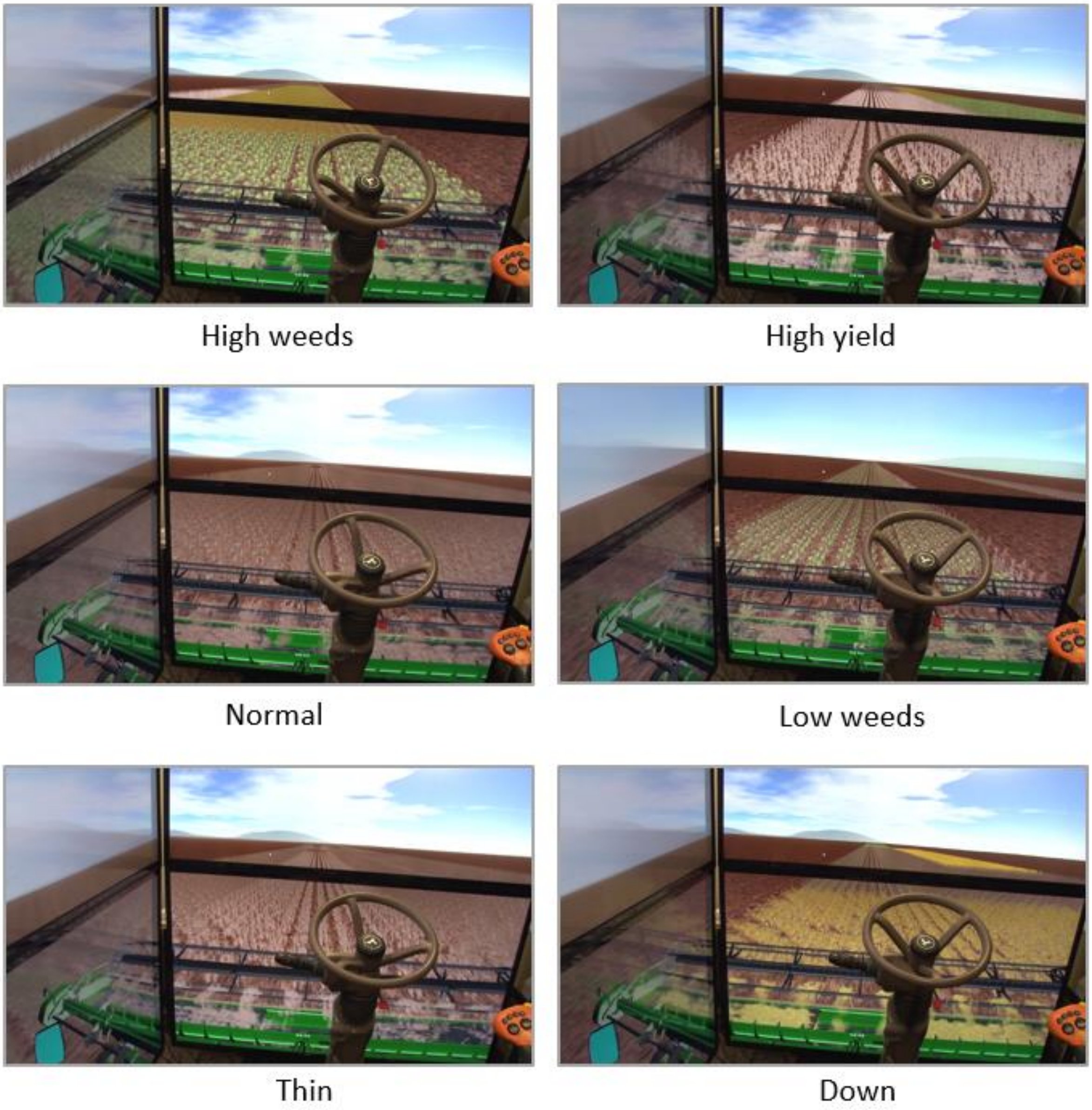
There were twelve cues shown to the operators throughout the course of the study (see Table 8), in addition to the control condition, which was shown everywhere else. These cues were translated into visuals within the combine simulator. In addition to visual components, many cues had secondary information that was changed simultaneously, e.g., when going into a low yield section, the yield monitor would also indicate a decrease. To see the complete field layout with all twelve cues presented, see Figure 14 and for examples of the visuals for each cue, see Figure 15.

Predictions

Cues that were primarily based on color change were expected to perform better than those which relied on subtler visual differences, such as green, high moisture crop being easier to identify than clumping occurring within the head. Additionally, participants who scored higher on the knowledge quiz were also expected to perform more highly on both the identification of cues and the number of correct actions taken on the cue. Lastly, participants who made more changes in the combine simulator were expected to experience higher workload as measured by higher phasic EDA.

Limitations and Assumptions

This study was not designed to measure literal crop harvest quality, as the underlying simulation was not built to respond to all possible variable inputs, such as threshing speed, sieve openings, chaffer settings, or fanspeed. While changes in the instrumentation were dynamic according to the type of crop the combine was in, these were only useful for the operators as input knowledge; no changes were made to the harvest output metrics such as yield and grain quality.



High weeds

High yield

Normal

Low weeds

Thin

Down

Figure 15. Photos from the simulator showing different grain cues, not all cues pictured but representative sample of size, color, and shape are shown.

Operators were assumed to have a working knowledge of the combine prior to participating in the study based on their recruitment criterion: that participants have at least two full seasons' harvest experience within the past four years. This criterion was used to avoid the operator experience being one of learning and exploration instead of the intended use and feedback to combine simulator cues within the field.

Hardware and Software

The combine simulator featured a modified John Deere S680 interior, with multi-TV displays setup in front of and to the left of the operator to simulator immersive virtual farming (Figure 16). The cab included a John Deere 2630 in-cab monitor running GreenStar 2. The combine simulator software, Greenspace (Luecke, 2012) was run on a PC running Ubuntu 12 with a 64-bit 3 GHz dual core Intel processor, 8 GB of DDR3 ram, and an NVIDIA Quadro K600 graphics card. Two external stereo speakers were used to produce audio in addition to an 8” subwoofer. A Buttkicker bass shaker attached to the cab seat that vibrated in correlation with the engine speed was also utilized to simulate the vibrations felt when operating a real combine. Primary displays to simulate the virtual field were four Vizio 70” televisions positioned in front of and to the left of the operator giving approximately 95° field of view on the front display and a full left peripheral view from the left display. The simulation was rendered monocularly using the OpenSG graphics engine with displays handled by VRJuggler at an average frame rate of 31 frames per second.

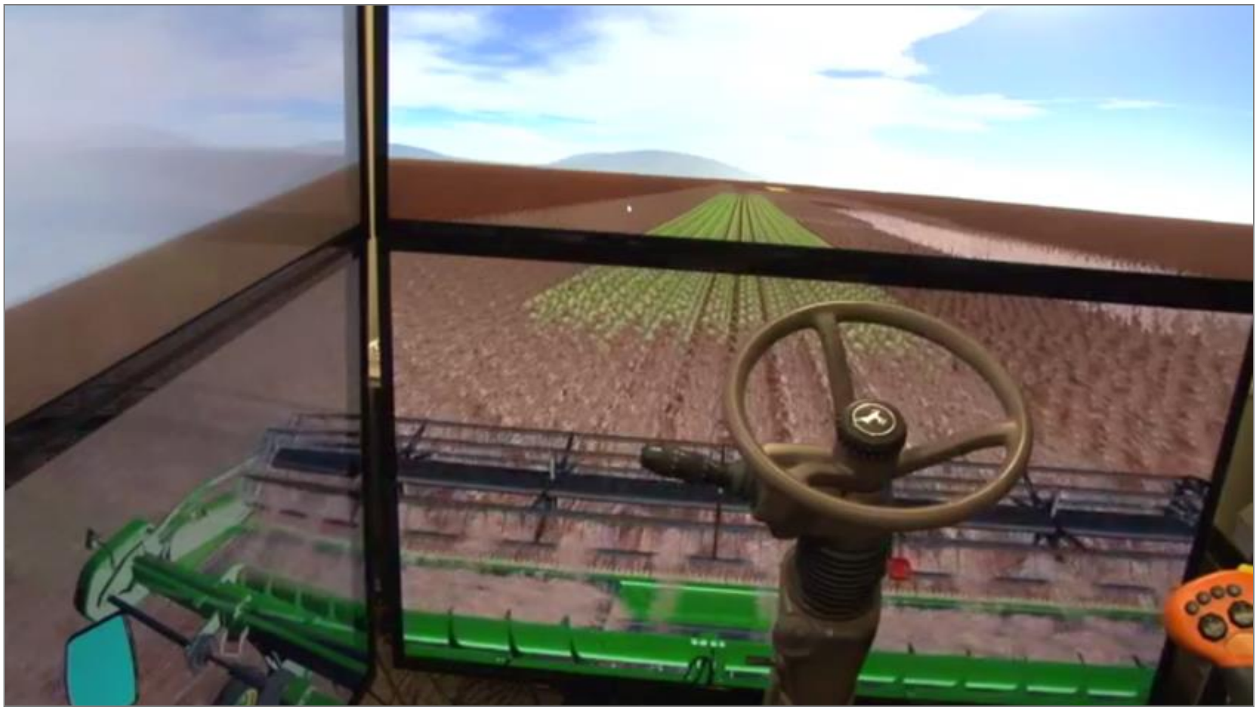


Figure 16. The combine simulator as configured for the cue study with a variety of field cues visible: the crop ahead shows a section of weeds in the beans and the header shows normal beans moving through the reel and along the belt.

Results

Resulting data from the cue validation study come in four primary forms: the results from 1) the knowledge survey, 2) the operator-system interactions, 3) the operators' performance in terms of cue identifications, their resulting actions, and ground speed, and 4) EDA correlations with the above variables.

Knowledge

These results aid in addressing the question of how the participants differed in their knowledge about the combine, which can be useful in grouping the results to the other research questions. Operator knowledge can be seen in Figure 17.

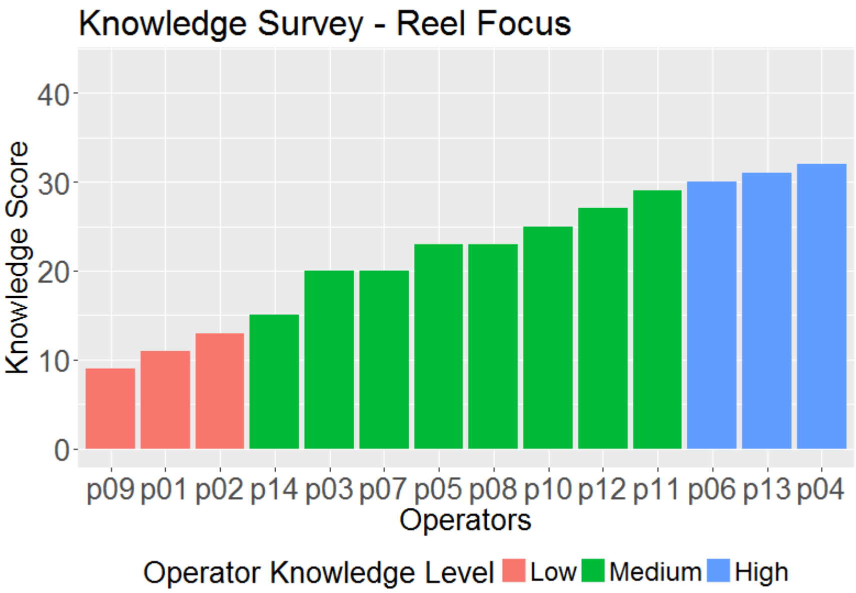


Figure 17. Operator knowledge score, colored by group, maximum possible score of 43.

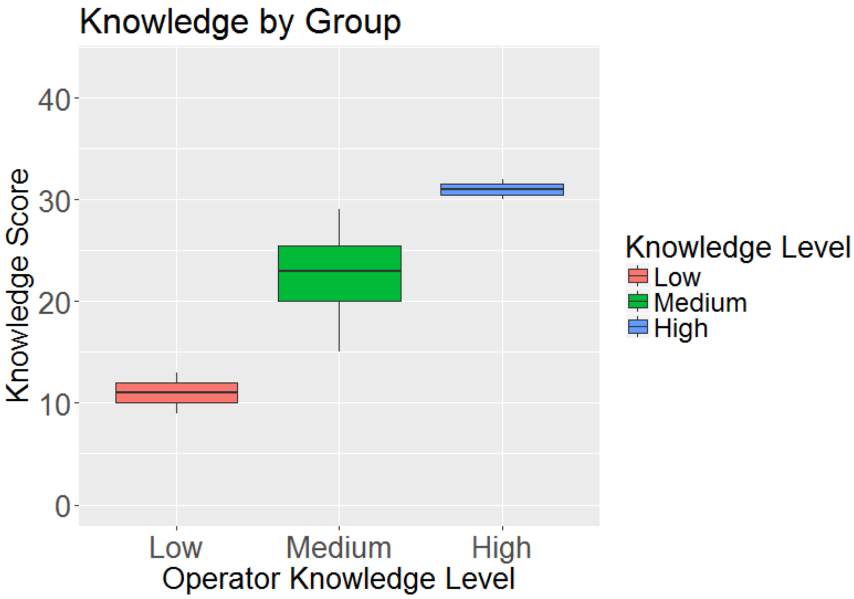


Figure 18. Knowledge scores split by group.

A one-way between subjects ANOVA was performed to determine whether knowledge scores between operator groups were different from each other. Groups were created by taking the mean knowledge score of the group, and then including +/- the standard deviation for medium knowledge. Scores above that threshold were put into the high knowledge group and below the low knowledge group. The process for dividing groups in more depth in a previous publication (Meusel et al., 2016).

Knowledge groups consisted of low knowledge (*n* = 3), medium knowledge (*n* = 8), and high knowledge (*n* = 3). There were no outliers, as assessed by boxplot seen in Figure 18; data were normally distributed for each group, as assessed by Shapiro-Wilk test (*p* > .05); and there was homogeneity of variances, as assessed by Levene's test of homogeneity of variances (*p* = .212). Knowledge quiz scores were statistically significantly different between different physical activity groups, *F*(2,11) = 22.76, *p* < .0005. Knowledge scores increased from the low (M = 11.00, SD = 2.00) to the medium (M = 22.75, SD = 4.43) and high (M = 31.00, SD = 1.00) knowledge groups, in that order. Tukey-Kramer post hoc analysis revealed that the mean increase from low to medium knowledge (11.75, 95% CI [5.07, 18.43]) was statistically significant (*p* < .005), the mean increase from low to high knowledge (20.00, 95% CI [11.95, 28.05]) was statistically significant (*p* < .001), and the mean increase from medium to high knowledge (8.25, 95% CI [1.57, 14.92]) was statistically significant (p < 0.05).

Operator Interactions with the System

A measure of activity within the combine is operationalized by interactions. One interaction indicates that the participant adjusted the reel height, the reel fore/aft position, or the reel speed one time. The interactions for each participant can be seen in Figure 19.

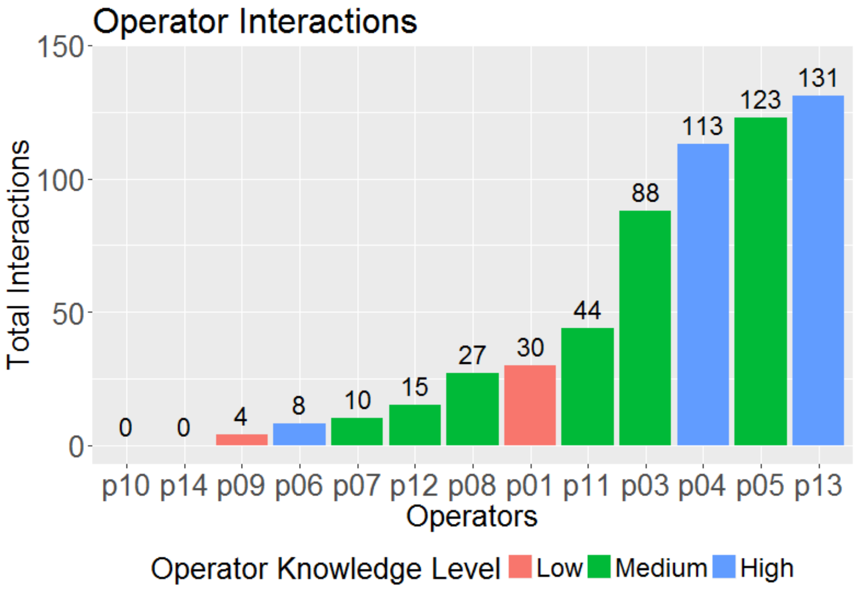


Figure 19. Operator interactions, colored by knowledge level (p14 & p10 are medium knowledge with 0 interactions; p02 had no interaction data due to a technical issue).

A Spearman's rank-order correlation was run to assess the relationship between knowledge and number of interactions. Preliminary analysis showed the relationship to be monotonic, as assessed by visual inspection of a scatterplot. There was a positive correlation between knowledge level and interactions, $r_s(11) = .422$, however the correlation was not significant, $p = .151$.

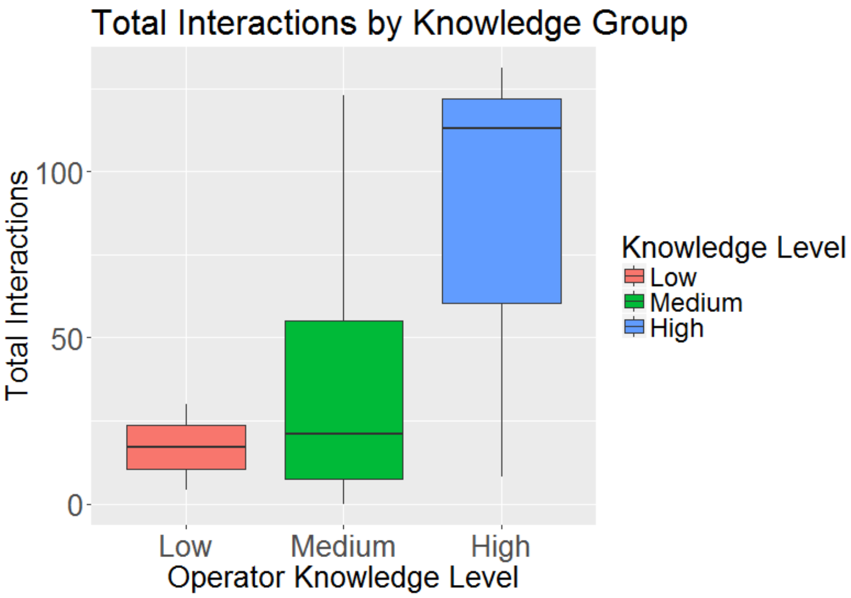


Figure 20. Box plot of total interactions separated by knowledge level

A Kruskal-Wallis H test was conducted to determine if there were differences in interactions between groups that differed in their knowledge level: the "low" (n = 2), "medium" (n = 8) and "high" (n = 3) knowledge level groups. Distributions of interactions were not similar for all groups, as assessed by visual inspection of a boxplot (Figure 20). Interactions from the Kruskal-Wallis H test increased from low (mean rank = 5.50), to medium (mean rank = 6.50), to high (mean rank 9.33) knowledge level groups, but the differences were not significant, $\chi^2(2) = 1.509$, $p = .470$.

Cue Identification and Actions

Participants could identify 85% of the cues correctly over the entire course of the study (142/168). Participants were less likely to be able to take the correct action on the identified set of cues (53/142, 37%).

Each participant was asked to identify each patch they entered in the virtual field for all possible cues. Their verbal identification of the cue was recorded in addition to any adjustments they made to the reel position, reel speed, or ground speed. Their responses were then analyzed and can be seen in Figure 21.

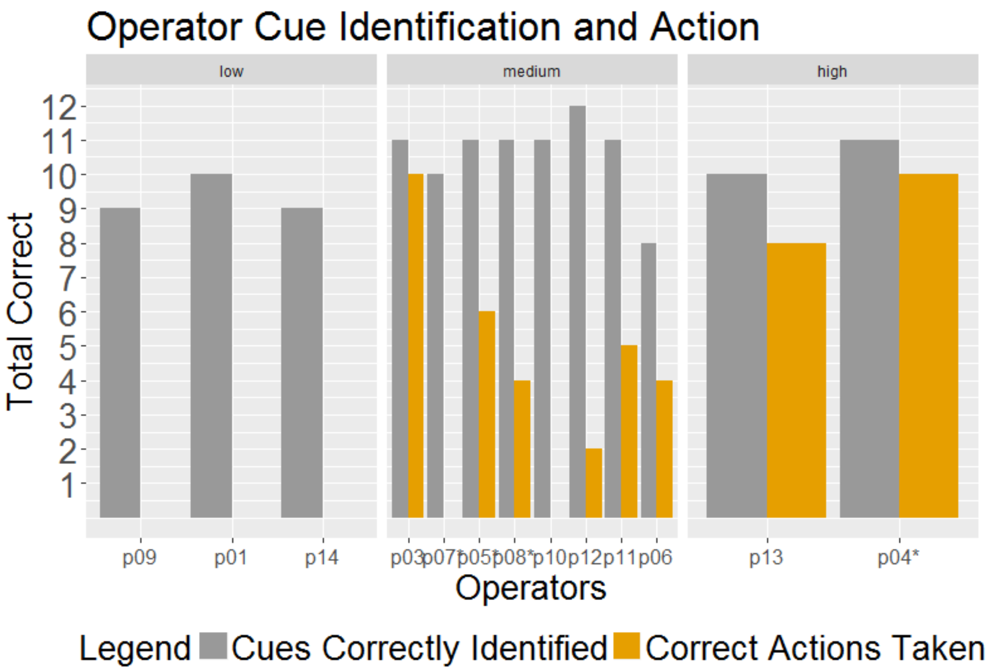


Figure 21. Total cues identified and actions taken, faceted by knowledge level, max score of 12. * indicates operator noted graphical limitations (p07, p05, p08, p04).

There is no significant difference between knowledge groups with respect to the total cues correctly identified. This was determined using a chi-square test of homogeneity to determine if one of the three knowledge groups differed in the number of total correct cues identified. No group was proportionally different from the others, *p* = .4043. These can be seen in Figure 22 and detailed by operator count in Table 10.

Table 10. Cue performance counts by operator, n = 14. Cue color and physical configuration noted; cues ordered how they were encountered in the field.

	Correct cue identified	Correct action taken	Color	Configuration
High weeds	12	6	Green weeds	Many weeds added in
Short height	10	6	Normal	Reduced height
Low yield, light	8	1	Normal	Slightly reduced volume
Low weeds	14	1	Green weeds	Minimal weeds added in
Low yield, extreme	14	4	Normal	Greatly reduced volume
Down crop	7	5	Normal	Crop bent over
High moisture	14	3	Green crop	Normal
Low yield, moderate	12	5	Normal	Normal
Tall height	13	7	Normal	Moderately reduced volume
Medium weed	13	3	Green weeds	Moderate weeds added in
Clumping	11	6	Normal	Crop clumps in header
Low moisture	14	6	Brown crop	Normal
Total	**142**	**53**		

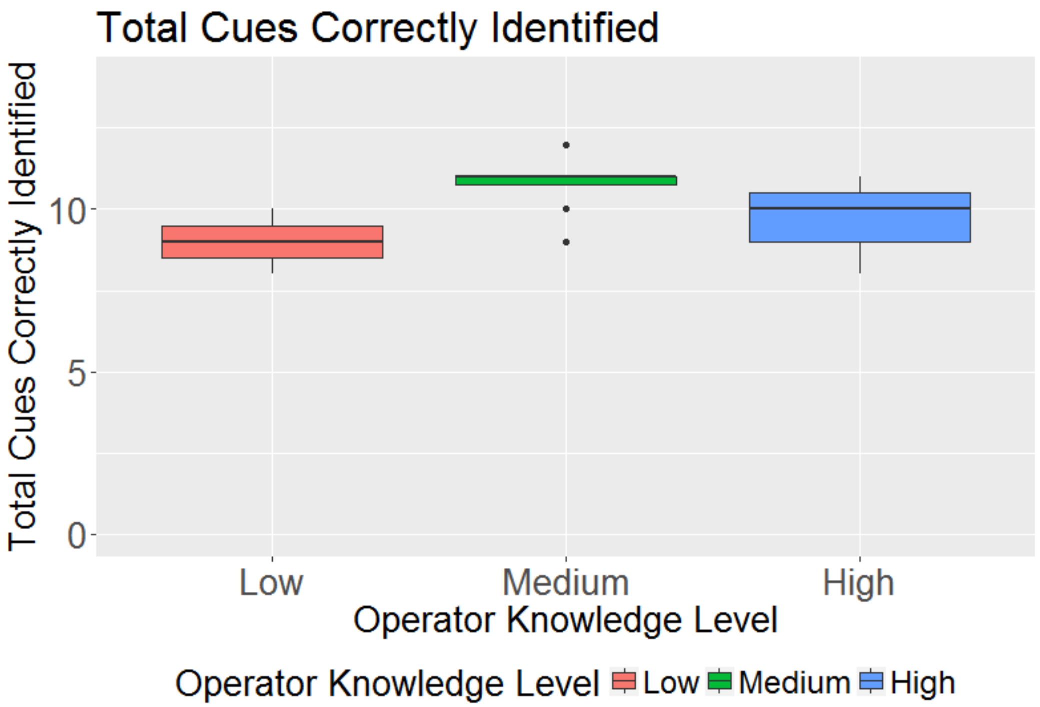


Figure 22. Boxplot of total cues correctly identified, separated by knowledge group; max score of 12.

When the number of correct actions taken as a proportion of the correct cues identified were compared across knowledge groups, there was a difference between the low knowledge group and medium knowledge group, *p* = .009, a difference between the low knowledge group and the high knowledge group, *p* < .005, and a difference between the medium knowledge group and the high knowledge groups, *p* = .0395. The data for correct actions, and why the proportions were different can be seen in Figure 23.

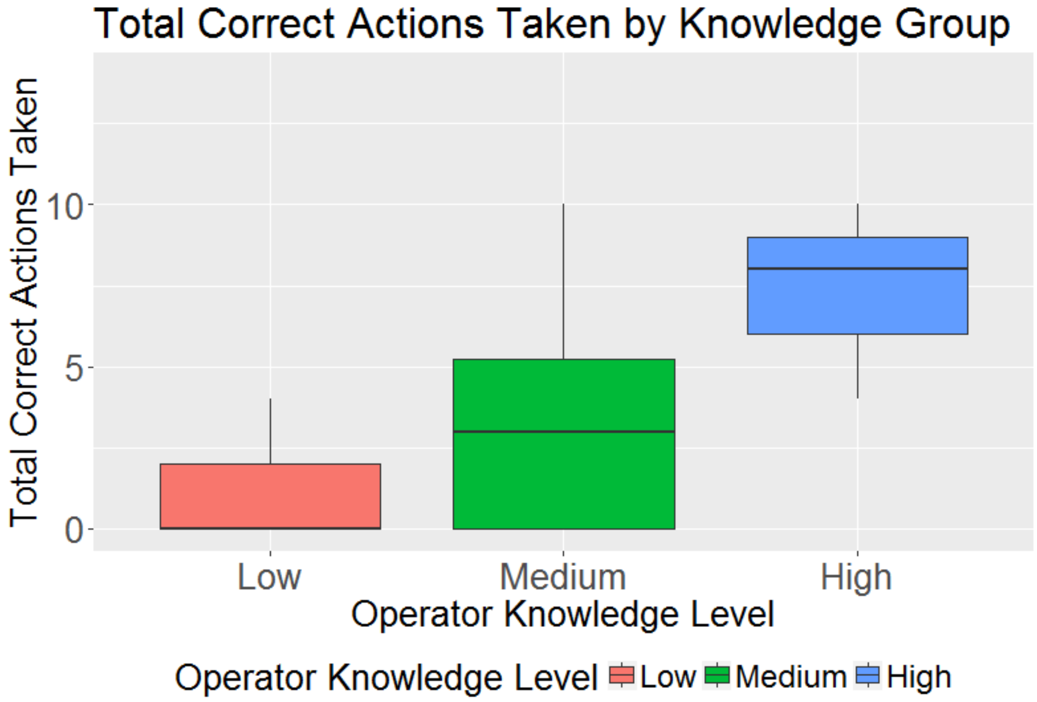


Figure 23. Boxplot of total correct actions taken, separated by knowledge group; max score of 12.

Additionally, correct actions taken on cues were also positively correlated with knowledge scores across the sample. The data were monotonic but failed Shapiro-Wilk test, *p* < .05. A Pearson's product-moment correlation was performed and results were similar, Pearson's results were kept due to similar results and that Pearson's correlation are relatively robust (Kang & Harris, 2012; Laerd Statistics, 2015). There was a large positive correlation between operator knowledge score and correct actions taken, *r*(11) = 0.596, *p* = .0315. The scatterplot of this relationship can be seen in Figure 24 below.

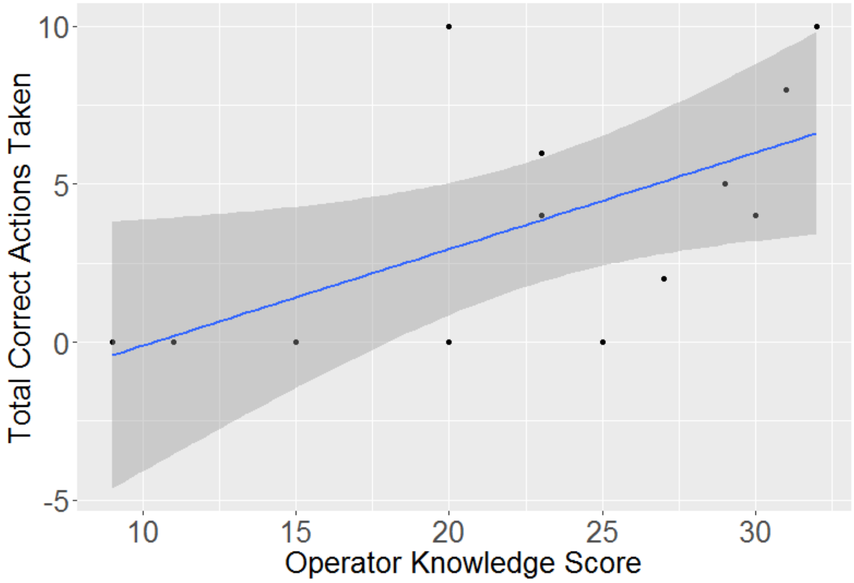


Figure 24. As knowledge score increases, correct actions taken increases.

Ground Speed

Ground speed did not vary by knowledge group, but the correlation between ground speed and knowledge score was tested and found to have a negative correlation. The data were monotonic but failed Shapiro-Wilk test, *p* < .05. A Pearson's product-moment correlation was performed and results were similar, Pearson's results were kept due to similar results and that Pearson's correlation are relatively robust (Kang & Harris, 2012; Laerd Statistics, 2015). There was a large negative correlation between operator knowledge score and average ground speed, *r*(11) = -0.558, *p* = 0.0476. The scatterplot of this relationship can be seen in Figure 25.

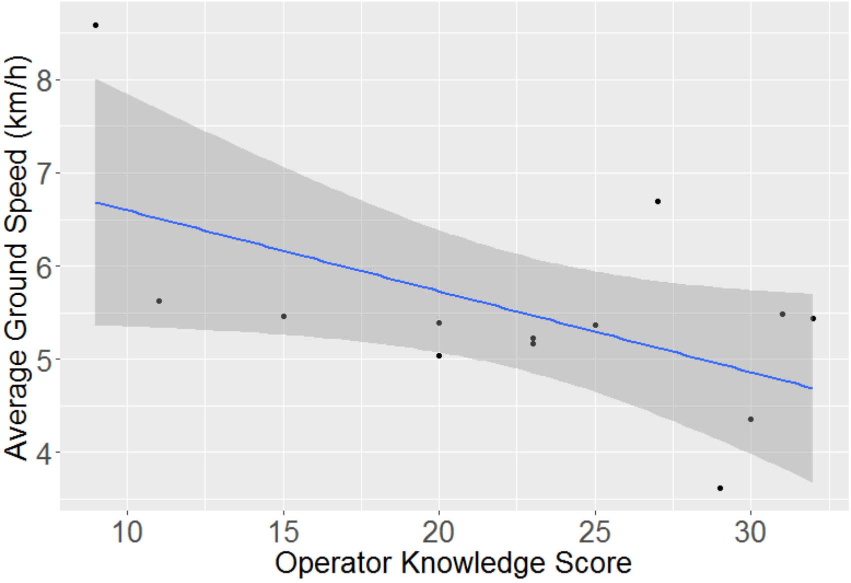


Figure 25. As knowledge scores increase, ground speed decreases.

Electrodermal Activity

Two different types of EDA were investigated within this data, tonic and phasic. Tonic EDA are the large, gradual changes which best represents knowledge as it is a factor which impacts everything. Phasic EDA is the amount of change in overall EDA and better represents the real-time activity of operator interactions as something that is changing moment-to-moment.

A Pearson's product-moment correlation was run to assess the relationship between operator's knowledge score and their tonic electrodermal activity. Preliminary analyses showed the relationship to be linear with both variables normally distributed, as assessed by Shapiro-Wilk's test (p > .05). There were outliers, but they were comparably large and in opposing directions. There was a moderate negative correlation between operator knowledge score and tonic electrodermal activity but it not was not significant, r(11) = -0.4324, p = .14.

A second Pearson's product-moment correlation was run to assess the relationship between the number of interactions an operator took and their phasic electrodermal activity.

Preliminary analyses showed the relationship to be linear, but data were not normally distributed. There were two outliers, but they were equally large in opposite directions at the same location. There was a large negative correlation between operator knowledge score and phasic electrodermal activity, *r*(11) = -.7538, *p* = .0026, 95% CI [-0.9235, -0.3566], with operator knowledge scores explaining 57% of the variation in phasic electrodermal activity. The scatterplot of this relationship can be seen in Figure 26**.**

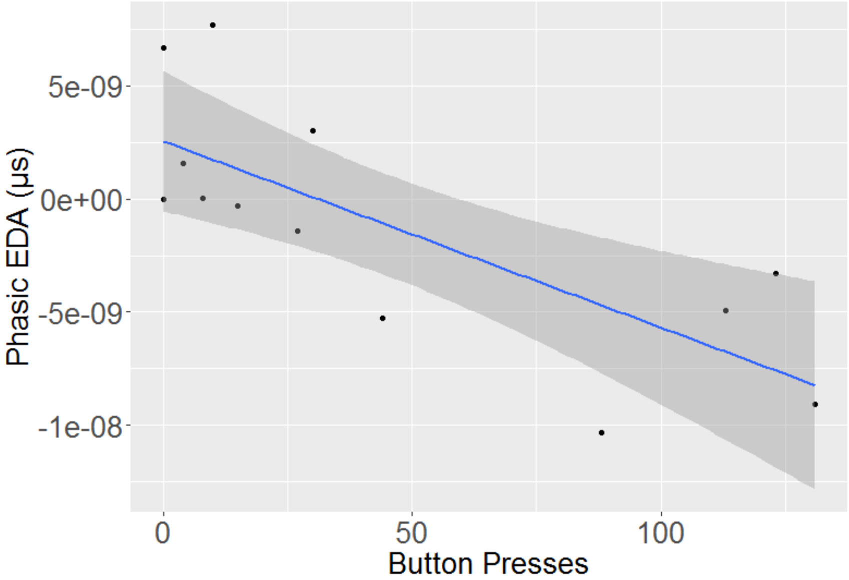


Figure 26. Scatterplot of phasic EDA and interactions (number of button presses).

A third Pearson's product-moment correlation was run to assess the relationship between the number of correct cues the operator identified within the field and their phasic electrodermal activity. There was a moderate positive correlation between total correct cues identified and phasic electrodermal activity, but the data were not significant, *r*(11) = 0.323, *p* = 0.259.

Additionally, another correlation was run to assess the relationship between the number of correct actions taken and phasic electrodermal activity. There was a small negative correlation between total correct actions taken and phasic electrodermal activity, but the data were not significant, *r*(11) = 0.166, *p* = 0.570.

Discussion

Knowledge

The operator knowledge survey is a method of evaluating functional knowledge the operator has about the operating of a combine that has been shown to successfully separate operators across a spectrum that is not completely represented by either age or experience (Meusel et al., 2016). This study was designed to evaluate harvest cues within the harvest simulator and explore whether were sufficient for operators to infer the appropriate field condition. Additionally, operators' actions taken on that cue were then evaluated and the harvest knowledge score was a primary variable for evaluating their performance. This approach stems from a continued interest in identifying differences between operator groups to better design applications for specific operator styles.

This study yielded a typical distribution of scores, encompassing three groups, low, medium, and high knowledge. Low, medium, and high knowledge groups were shown to be significantly different and were therefore utilized as a tool for comparative statistics via the other measures. It should also be noted, the knowledge quiz for this study was a modified version designed specifically to target reel interaction knowledge.

Operator Interactions with the System

Interactions within the combine simulator consisted of counting the number of adjustments to the reel controls for position and speed. A previous harvest simulator study showed that number of interactions with two different pieces of technology aligned with perceived difficulty of that technology (Meusel, 2014).

The findings from this study did not reveal differences in numbers of interactions between knowledge groups or as a function of knowledge itself. Previous harvest simulator work has indicated that lower knowledge operators tend to make fewer adjustments during the course harvesting the same field relative to higher knowledge participants. This trend may have been true within this data set, but due to the small sample size, there can be no conclusions drawn from this sample alone. Interactions should still be tracked, though, as a potential indicator of mental effort. The supporting evidence found that interactions can be used as an indicator of mental effort comes from the results that correlated EDA activity with operator interactions. More on this in the EDA section below.

Cue Identification and Actions

Successfully identifying the virtual cues within the harvest simulator was the driving factor behind this study as a piece of developmental work for the harvest simulator. Cue development was divided into two steps, the first of identifying the cue being presented and the second of taking the correct action on that cue. These results reflect aggregate cue performance. Most cues were recognizable, and those that were correctly identified fewer times (down & low yield) were modified based on feedback to improve implementation with future simulator studies.

Cues that were primarily based on color change were expected to perform better than those which relied on subtler visual differences, this was true of downed crop as it was only correctly identified by 7/14 operators compared to a cue like low weeds or high moisture, which were both identified correctly by all (14/14) operators.

Additionally, operators who scored higher on the knowledge quiz were also expected to perform more highly on both the identification of cues and the amount of correct actions taken on the cue. Operator knowledge did not correlate with how well operators would identify cues within this sample. However, operator knowledge did correlate with how well operators took correct actions and the large, positive correlation supports this prediction, $r(11) = 0.596$, $p = 0.0315$. The higher operators scored on the knowledge quiz, the more correct actions they took. The majority of operators who did not take the correct action were actually taking no action. This has been found in previous studies where lower operators were not necessarily taking incorrect actions, but failed to taken any action during sections of the field that required corrective action (Meusel, 2014; Meusel et al., submitted). This finding specifically supports the idea that the knowledge survey can be successfully used as a type of proxy measure for operator aptitude. By observing that higher knowledge scores also indicate higher correct actions taken, this study moves past the simpler process of only measuring whether cues could be successfully identified (as most successfully were) and into the area of cue understanding.

Additionally, operator knowledge groups did provide differences between proportion of correct actions taken as the low knowledge group did not perform any of the correct actions observed within the sample, while the medium and high knowledge groups performed an average of 3.875 and 9 correct actions per operator, respectively. Further data collection is needed to make a stronger claim.

These differences can be seen in Figure 21. This was expected, as taking the correct action is a task that requires a higher knowledge of operating the combine than simply identifying cues. This finding continues to support this research team's stance that lower knowledge operators have a greater potential benefit of automated harvest technology and should be the target users during that technology's development.

Ground Speed

During this study and in other harvest simulator studies, higher knowledge operators have anecdotally tended to drive more slowly and show more concern for the scenario within the simulator. The results from this study support that observation by showing a strong negative correlation between ground speed and knowledge. As knowledge scores increased, operator's average ground speed tended to decrease. This trend seemed to occur when higher knowledge operators encountered novel cues or stimuli within the field. They tended to slow down and make thoughtful corrections relative to their lower knowledge operator contemporaries. For reference, high knowledge operators averaged 1.1 km/h slower ground speed (5.46 km/h) than the low knowledge operators (6.56 km/h).

This finding will continue to be tracked in future studies as it seems to be consistent across operators and subject areas.

Electrodermal Activity

The first harvest simulator study (Meusel, 2014) successfully demonstrated EDA as a measure of mental effort within the cab. As this study was also concerned with operator workload, EDA was again measured during the study sessions as a proxy for cognitive load.

The key measure for EDA was interactions as a proxy for cognitive load. Operators who made more changes in the combine simulator were expected to experience higher workload as measured by higher phasic EDA. The number of interactions each participant took were noted and then compared with their phasic EDA, as phasic activity more accurately reflects activity sensitive to individual actions. As EDA was shown to have a large negative correlation (-.7538) with interactions, this measure will continue to be monitored and inspected as a potential proxy for mental effort in future studies.

Tonic EDA was evaluated with operator knowledge scores and provided inconclusive results. Tonic EDA represent the slow, more gradual changes over time which makes it the fitting measure to compare with knowledge scores. EDA was also compared with operator cue performance. Successful cue actions were shown to change relative to operator knowledge, but because knowledge and tonic EDA were not shown to have a relationship within this sample, the lack of findings between EDA and cue performance was not surprising.

Conclusion

First, this study successfully helped identify technical and perception issues with cues shown in the harvest simulator. These findings directly contributed to improvements in the simulator and the ability to test future combine technology.

The work here demonstrates that thoughtful, iterative development on a research platform can be successfully used to adapt the platform to a variety of research topics. Previous work (Kieu et al., 2017; Luecke, 2012) was required to build high enough fidelity experiences to consider testing in new areas. Additionally in this study, the simulator platform was used to outline a process of cue validation that others designing novel simulator cues could also follow.

This study also successfully followed up on findings previously highlighted within this research groups harvest simulator work. This work specifically offered an additional set of varied knowledge scores within an operator sample to contribute to the larger harvest knowledge survey data set. Harvest knowledge groups were shown to perform differently with respect to action taken on cues and ground speed. Both of those show higher knowledge operators performing more successfully than low knowledge operators while operating more slowly to make those correct actions.

Lastly, EDA was shown to highly correlate with interactions as a non-invasive measure of mental effort in real time. This allows greater flexibility with future study configurations with respect to gauging operator effort.

Operator feedback was taken both directly (comments on experience) and indirectly (performance & EDA measures) to provide high quality data for future harvest simulator development. The cues that were identified as having more issues both in performance and direct feedback could be addressed and improved prior to the next study utilizing the harvest simulator platform.

As the harvest simulator platform continues to be improved for a specific population, these small sample studies are particularly helpful as expertise within this domain can be difficult to source through traditional means. The harvest simulator platform continues to move forward as a tool to conduct operator centered research on a wide variety of topics that impact the lives of billions through the continued increased demand on the agricultural system worldwide.

Acknowledgements

Thank you to Norene Kelly for assisting in initial data collection and pre-study cue evaluation. Also, thank you to John Deere for the opportunity to conduct meaningful and interesting user research.

References

Abdi, H. (2007). Signal Detection Theory (SDT). In *Encyclopedia of Measurement and Statistics*. Sage.

Bashiri, B., & Mann, D. D. (2014). Automation and the situation awareness of drivers in agricultural semi-autonomous vehicles. *Biosystems Engineering*, *124*, 8–15. http://doi.org/10.1016/j.biosystemseng.2014.06.002

Bransford, J., Brown, A., & Cocking, R. (1999). *How people learn: Brain, mind, experience, and school.*

Butzen, S. (2013). Reducing Harvest Losses in Soybeans.

Calatayud, D., Arora, S., Aggarwal, R., Kruglikova, I., Schulze, S., Funch-Jensen, P., & Grantcharov, T. (2010). Warm-up in a virtual reality environment improves performance in the operating room. *Annals of Surgery*, *251*, 1181–1185. http://doi.org/10.1097/SLA.0b013e3181deb630

Calcante, A., & Mazzetto, F. (2014). Design, development and evaluation of a wireless system for the automatic identification of implements. http://doi.org/10.1016/j.compag.2013.12.010

Dalgarno, B., & Lee, M. J. W. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology*, *41*(1), 10–32. http://doi.org/10.1111/j.1467-8535.2009.01038.x

Duncan, J. R., & Turner, R. J. (1991). Operator Performance with simulated Harvesting combine Automatic Controls. *The American Society of Agricultural Engineers*.

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-Physiological Measures for Assessing Cognitive Load. *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, 301–310. http://doi.org/10.1145/1864349.1864395

Hiremath, S. a., van der Heijden, G. W. a M., van Evert, F. K., Stein, A., & Ter Braak, C. J. F. (2014). Laser range finder model for autonomous navigation of a robot in a maize field using a particle filter. *Computers and Electronics in Agriculture*, *100*, 41–50. http://doi.org/10.1016/j.compag.2013.10.005

Huitink, G. (2000). Harvesting Soybeans. In *Arkansas Soybean Handbook* (pp. 1–12).

Kang, Y., & Harris, J. R. (2012). Investigating the Impact of Non-Normality, Effect Size, and Sample Size on Two-Group Comparison Procedures: An Empirical Study. In *Presented at the Annual Meeting of the American Educational Research Association (AERA)* (p. 29).

Karimi, D., & Mann, D. (2008). Role of motion cues in straight-line driving of an agricultural vehicle. *Biosystems Engineering*, *101*(3), 283–292. http://doi.org/10.1016/j.biosystemseng.2008.09.006

Karimi, D., & Mann, D. (2009). Torque feedback on the steering wheel of agricultural vehicles. *Computers and Electronics in Agriculture*, *65*(1), 77–84. http://doi.org/10.1016/j.compag.2008.07.011

Karimi, D., Mann, D. D., & Ehsani, R. (2008). Modeling of straight-line driving with a guidance aid for a tractor-driving simulator. *Applied Engineering in Agriculture*, *24*(4), 403–408.

Katsis, C. D., Katertsidis, N., Ganiatras, G., & Fotiadis, D. I. (2008). Toward emotion recognition in car racing drivers: a biosignal processing approach. *IEEE Trans. Systems, Man and Cybernetics - Part A: Systems and Humans*, *38*(3), 502–512.

Kelly, J. W., Sjolund, L. A., & Sturz, B. R. (2013). Geometric cues, reference frames, and the equivalence of experienced-aligned and novel-aligned views in human spatial memory. *Cognition*, *126*(3), 459–474. http://doi.org/10.1016/j.cognition.2012.11.007

Kieu, D., Luecke, G., Gilbert, S., Hunt, T., Gilmore, B., Kelly, N., & Meuse, C. (2017). Hearing the Customer Voice Through a Combine Simulator: Innovative Techniques for Product Development. *IEEE Transactions on Human-Machine Systems*.

Kruglikova, I., Grantcharov, T. P., Drewes, A. M., & Funch-Jensen, P. (2010). The impact of constructive feedback on training in gastrointestinal endoscopy using high-fidelity Virtual-Reality simulation: a randomised controlled trial. *Gut*, *59*, 181–185. http://doi.org/10.1136/gut.2009.191825

Laerd Statistics. (2015). Pearson’s product-moment correlation using SPSS Statistics.

Larsen, C. R., Soerensen, J. L., Grantcharov, T. P., Dalsgaard, T., Schouenborg, L., Ottosen, C., … Ottesen, B. S. (2009). Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. *BMJ*. http://doi.org/10.1136/bmj.b1802

Lee, W.-S., Kim, J.-H., & Cho, J.-H. (1998). A driving simulator as a virtual reality tool. *Proceedings IEEE International Conference on Robotics and Automation*, *May*(May), 71–76. http://doi.org/10.1109/ROBOT.1998.676264

Loftin, R. B., & Kenney, P. J. (1995). Training the hubble space telescope flight team. *IEEE Computer Graphics and Applications*, *15*, 31–37. http://doi.org/10.1109/38.403825

Luecke, G. R. (2012). GREENSPACE : Virtual Reality Interface for Combine Operator Training. *Presence: Teleoperators and Virtual Environments*, *21*(3), 245–254. http://doi.org/10.1162/PRES_a_00110

Ma, Z., Li, Y., & Xu, L. (2015). Discrete-element method simulation of agricultural particles' motion in variable-amplitude screen box. *Computers and Electronics in Agriculture*, *118*, 92–99. http://doi.org/10.1016/j.compag.2015.08.030

Macmililan, N. A. (2002). Signal Detection Theory. In *Steven's Handbook of Experimental Psychology - Volume 4: Methodology in Experimental Psychology*.

Mann, D., Bashiri, B., Rakhra, A., & Karimi, D. (2014). Development of a tractor driving simulator to research ergonomics of agricultural machines. *International Conference of Agricultural Engineering*, (1988), 6–10.

Marx, S. E., Luck, J. D., Hoy, R. M., Pitla, S. K., Blankenship, E. E., & Darr, M. J. (2015). Validation of machine CAN bus J1939 fuel rate accuracy using Nebraska Tractor Test Laboratory fuel rate data. *Computers and Electronics in Agriculture*, *118*, 179–185. http://doi.org/10.1016/j.compag.2015.08.032

Meusel, C. (2014). *Exploring mental effort and nausea via electrodermal activity within scenario-based tasks*. Iowa State University.

Meusel, C., Grimm, C., Gilbert, S., & Luecke, G. (2016). An Agricultural Harvest Knowledge Survey to Distinguish Types of Expertise. *60th Annual Meeting of the Human Factors and Ergonomics Society*.

Meusel, C., Kieu, D., Gilbert, S., Luecke, G., Gilmore, B., Hunt, T., & Kelly, N. (n.d.). Evaluating Novel Harvest Technology within a High Fidelity Combine Simulator. *Computers and Electronics in Agriculture*.

Minnihan, J., Hanna, M., Isaac, N., & Couser, B. (2003). *Setting Combines for Harvesting Best Soybean Seed Quality and Maximum Yield*.

Pranav, P. K., Pandey, K. P., & Tewari, V. K. (2010). Digital wheel slipmeter for agricultural 2WD tractors. *Computers and Electronics in Agriculture*, *73*(2), 188–193. http://doi.org/10.1016/j.compag.2010.05.003

Pranav, P. K., Tewari, V. K., Pandey, K. P., & Jha, K. R. (2012). Automatic wheel slip control system in field operations for 2WD tractors. *Computers and Electronics in Agriculture*, *84*, 1–6. http://doi.org/10.1016/j.compag.2012.02.002

Rizzo, M., Jermeland, J., & Severson, J. (2002). Instrumented Vehicles and Driving Simulators. *Gerontechnology*, *1*(4). http://doi.org/10.4017/gt.2002.01.04.008.00

Ruddle, R. A., & Lessels, S. (2006). For efficient navigational search, humans require full physical movement, but not a rich visual scene. *Psychological Science*, *17*(6), 460–465.

Scarlett, A. J. (2001). Integrated control of agricultural tractors and implements: A review of potential opportunities relating to cultivation and crop establishment machinery. *Computers and Electronics in Agriculture*, *30*(1–3), 167–191. http://doi.org/10.1016/S0168-1699(00)00163-0

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine : A Publication of the IEEE Engineering in Medicine and Biology Society*, *14*(2), 410–7. http://doi.org/10.1109/TITB.2009.2036164

Sjolund, L. A. (2014). *Cue integration and competition during navigation*. Iowa State University.

Son, J., & Park, M. (2011). Estimating cognitive load complexity using performance and physiological data in a driving simulator. *AutomotiveUI, 2011, Adjunct Proceedings*.

Son, K., Goo, S. H., Yoo, W. S., Lee, M. C., & Lee, J. M. (2001). A driving simulator of construction vehicles - Son. *International Journal of the Korean Society of Precision Engineering*.

Stoffregen, T. A., Bardy, B. G., Smart, L. J., & Pagulayan, R. J. (2003). On the Nature and Evaluation of Fidelity in Virtual Environments. In L. J. Hettinger & M. W. Haas (Eds.), *Virtual and Adaptive Environments: Applications, Implications, and Human Performance Issues*. Lawrence Erlbaum Associates Publishers.

Tate, D. L., Sibert, L., & King, T. (1997). Using virtual environments to train firefighters. *IEEE Computer Graphics and Applications*, *17*, 23–29. http://doi.org/10.1109/38.626965

Weir, D. H. (2010). Application of a driving simulator to the development of in-vehicle human-machine-interfaces. *IATSS Research*, *34*(1), 16–21. http://doi.org/10.1016/j.iatssr.2010.06.005

Wilkerson, J. B., Asbury, B. C., Prather, T. G., & Lown, J. B. (1993). Development of a tractor driving simulator. *American Society of Agricultural Engineers*.

CHAPTER 4

REAL TIME EMOTION DETECTION USING EVERYDAY TECHNOLOGY

This chapter will be submitted to *IEEE Transactions on Affective Computing*.

Author list:

Chase Meusel, Mike Bortnick, Jin Jun, Joe Munko, Umer Farooq, Stephen B. Gilbert

Chase Meusel's role in this research included the experimental design, participant recruitment, data collection, data coding and analysis, and primary authorship on the paper. Chase also designed the feature extraction, analysis process, and machine learning models. Jin Jun built the final, deployed version of the feature extraction process and implemented it to work with Azure machine learning services. The design and development of the data collection application was done by Mike Bortnick & Jin Jun.

Abstract

This work outlines the process and evidence for measuring an individual's workplace emotional state in real time using an electrodermal activity (EDA) sensor within daily meeting scenarios over two phases. In the first phase, 89 participants wore an EDA sensor, the Microsoft Band, and reported their current emotional state. Data collected were used to train and validate an emotion detection model using machine learning. In the second phase, real time model validation, 16 participants from the first phase wore the EDA sensor and responded to a prediction of their emotional state on a phone-based mobile application. Participants reported the emotion detection model was correct (perceived accuracy) 76.43%

of the time. Specific emotions reported by participants matched with the emotion detection model's emotional state (true accuracy) over 52.52% of the submissions. By implementing this process with additional sensors and improved sensor quality, real time emotion detection should increase in accuracy and feasibility.

Introduction

Emotion is known to be a complex, important, and largely underutilized part of today's technological ecosystem (R. W. Picard, 2010; Rosalind W. Picard, 1997).   Not only do emotions represent how we're feeling in an immediate sense, they also influence a number of factors from social interactions to learning aptitude (Graesser, 2009; Hascher, 2010; Stowell & Nelson, 2007). Emotional intelligence (EQ) has been shown to be as important as one's intelligence quotient (IQ) for both professional and personal health based scenarios (Goleman, 2006; Schutte, Malouff, Thorsteinsson, Bhullar, & Rooke, 2007). For individuals who may not have naturally high EQ, assistance in recognizing their own and potentially others emotional state can be exceedingly useful. Some populations with known emotional intelligence deficiencies such as those on the autism spectrum (El Kaliouby, Teeters, & Picard, 2006; Golan, Baron-Cohen, & Hill, 2006) or those who have prosopagnosia (face blindness) could benefit from a system that monitors emotional states. Additionally, people are generally poor at reflecting on their own emotional state in meaningful ways without assistance from activity tracking tools (McDuff, Karlson, Kapoor, Roseway, & Czerwinski, 2012). This research supports the increase in personal activity and emotion tracking software currently available.

To improve affective computing experiences, emotional measurement should be improved, as it is a key mechanism in the communication of emotion from user to system (Rosalind W. Picard, 1997). Additionally, because emotion is an essential piece of the rational thought process, understanding one's own emotional state can help improve general thinking for any individual with a low EQ (Damasio, 2005; LeDoux, 1996). Improving emotional skills training is an area of increased importance in efforts to highlight areas where technology can have the largest impact (Slovák, Gilad-Bachrach, & Fitzpatrick, 2015).

This work aims to outline a system to evaluate emotional states in real time for daily use by taking advantage of existing hardware that is available at the consumer level. To describe recent emotion detection research, the authors propose three additional factors (6-8 below) that build on Picard, Vyzas, and Healey's original five factors of eliciting emotion (1-5) (Picard et al., 2001).

1. *Subject-elicited* vs. *event-elicited*: Subject eliciting the emotion at will vs. the subject expressing an emotion in response to an outside stimulus.
2. *Laboratory setting* vs. *real-world*: Data collected within a controlled setting vs. outside the lab within a naturalistic setting.
3. *Expression* vs. *feeling*: Externally observed emotions vs. internally felt and reported emotion.
4. *Open-recording* vs. *hidden-recording*: The subject is aware of data recording vs. unaware of being observed and recorded.
5. *Emotion-purpose* vs. *other-purpose*: The subject is aware that the purpose of their participation involves emotion vs. they are unaware the work involves emotions.

6. *Single participant* vs. *multiple participants:* Data collected from an individual participant vs. collected from multiple participants.
7. *Short-term* vs. *long-term data collection*: Data collected for a brief period of time vs. multiple sessions over an extended time period, multiple days at minimum.

   Research hardware vs. *consumer hardware*: Data collected using purpose built, high quality research hardware vs. data collected using consumer level, or low cost, hardware.

Picard et al. described the most natural setup for collecting emotion data as one that has participants experiencing emotion due to an outside stimulus (*event-elicited*), happening outside of the laboratory (*real-world*), felt as an internal feeling (*feeling*), monitored unknowingly (*hidden-recording*), for a reason unrelated to the collection of emotion detection (*other-purpose*). In addition the authors suggest that data should be collected from a large sample (*multiple participants*), over an extended period of time (*long-term data collection*), using everyday technology (*consumer hardware*).

This work uses *event-elicited* emotions from *real-world* scenarios which are reported as internal *feelings* from the participant. Data were collected as an *open-recording* as participants were aware they were submitting data for a study investigating an *emotion-purposed* study. *Multiple participants* were observed over an extended period, or *long-term data collection*, using *consumer hardware* as the device to record EDA and other physiological signals. Most studies within the field of emotion detection have used emotion-elicitation protocols (*subject-elicited*) and remained within the *laboratory setting*, where this work uses naturally occurring *event-elicited* emotions in *real-world* workplace meeting scenarios.

The motivation for this work lies in a desire to build a tool which can objectively measure an individual's emotional state in real time, using everyday technology. This tool ideally can be used as an objective measure of participant emotion toward a particular experience or product within or outside the laboratory setting. Additionally, this tool could move beyond the research scenario and help augment the emotional self-awareness and communication for individuals with known low EQ. Other systems currently achieve components of this goal, but none have yet met all of these criteria. This leads to the primary research question: can emotions be objectively measured in real time using everyday technology? If so, how accurate is this system?

Background and Related Work

Research on recognition, monitoring, and tracking of emotional states has continued to grow in recent years, but the vast majority of this work has been done in controlled laboratory conditions, using emotion elicitation protocols with research equipment. This current work attempts to address this gap by leaving the laboratory and using methods for measuring emotion that are available with everyday technology.

The first major effort to systematically identify an individual's emotional or affective state began with Ekman's work highlighting how to extract affective states from facial clues (Ekman & Friesen, 1975). Ekman's facial affect coding system (FACS) was built and tested as a computer vision solution targeted to discriminate between multiple emotional states, including neutral, anger, disgust, fear, joy, sadness, and surprise (Essa & Pentland, 1994; Littlewort, Bartlett, Fasel, Susskind, & Movellan, 2006). Ekman (Ekman, Levenson, & Friesen, 1983) then suggested emotion was difficult to observe physiologically (e.g., via

EDA, heart rate, pupillometry, etc.) due to the recording window, or epoch size, being too large and multiple emotions adding noise to the data. This critique was improved upon as others outlined how physiological measures could be used as methods for measuring emotional states (John T Cacioppo, Berntson, Larsen, Poehlmann, & Ito, 2000; Healey, 2000).

The earliest study to use physiological sensors as a method of measuring emotion is from Fridlund and Izard in 1983 (Fridlund & Izard, 1983). They used four separate facial electromyography (EMG) sensors to differentiate happiness, sadness, anger, and fear. Later, Picard et al. (Rosalind W. Picard et al., 2001) were able to classify emotions with 81% accuracy differentiating between eight different states using facial EMG, blood volume pulse (BVP), EDA, and respiration effort with a single participant over a six-week period. Similarly, Haag, Goronzy, Schaich, & Williams (2004) found success measuring arousal with 89% accuracy and valence with 96% accuracy, using an artificial neural network to classify physiological sensor data with a single participant. This trend of using machine learning classification models continues today. Kim, Bang, and Kim (Kim, Bang, & Kim, 2004b) found 61.8% accuracy using ECG, skin temperature, and EDA within a four state model and used a generalized model with 50 participants. Katsis, Katertsidis, Ganiatras, and Fotiadis (2008) observed 10 subjects, specifically race car drivers, in their vehicles and found 79% accuracy with four states and a variety of physiological sensors. Additionally, the studies mentioned here all used sensors that were either research grade or purpose built, as opposed to something that could be purchased at the consumer level.

Moving past the use of physiological sensors for emotion detection alone, the following studies outlined methodologies for building systems to analyze incoming data in real time using generalized models. Leon, Clarke, Callaghan, and Sepulveda (2007) validated an autoassociative neural network model using multiple physiological sensor inputs with 71.4% accuracy when tested with an individual who had not contributed to the training set data. Bailenson et al. (2008) designed a system for real time emotion classification using facial features and physiological sensors that was successfully able to differentiate between sadness and amusement with 98% and 94% accuracy, respectively. In addition to these studies using facial and physiological sensor inputs, others have done similar work using voice as the sole measure (S. Kim, Georgiou, Lee, & Narayanan, 2007; Vogt, André, & Bee, 2008).

In recent years, emotion detection work continues to improve real time analysis and additionally places an emphasis on emotion tracking, self-awareness, and communication. Cowie et al. (2000) introduced and validated “Feeltrace,” a tool to track perceived emotions in real time, yet still used a basic emotion elicitation protocol and required constant input from the participant via a mouse controlled emotion wheel. El Kaliouby, Teeters, and Picard (2006) proposed an “emotional prosthetic” which would allow users, notably those with autism spectrum disorder, to interpret the emotional state of those around them by using facial affect software. McDuff, Karlson, and Kapoor (2012) more recently built and tested an emotion tracking and self-awareness tool “AffectAura,” which recorded physiological, performance, and behavioral data over time to measure emotional states and subsequently provide a reflection tool.

Others have designed a physical representation of emotion for self-awareness such as Stefani, Mahale, Pross, and Bues’ work with “SmartHeliosity” (2011) which used facial emotion detection software to change ambient lighting dynamically based on the users emotional state.   Similarly, Roseway, Lutchyn, Johns, Mynatt, and Czerwinski (2015) designed and tested an emotional recognition, tracking, and communication tool, “BioCrystal,” which promoted emotional self-awareness, improved stress-management, and acted as a communication tool for others to observe the user’s emotional state. Other areas of emotion communication have explored how haptic feedback (Obrist, Subramanian, Gatti, Long, & Carter, 2015) and thermal feedback (Wilson, Davidson, & Brewster, 2015) can be used to augment or create emotional messages. Though not direct communication, it has also been shown that people perceive footfalls with an emotional interpretation which lined up with measured EDA and self-report values (Tajadura-Jiménez et al., 2015). Hollis, Konrad, and Whittaker (2015) reported that users who tracked behaviors of a self-selected, unwanted behavior reported more successful behavioral change and greater engagement when using an emotion-focused system of tracking when compared to those who used a fact focused-system.

The previous body of work displays a shift from modeling a single user to achieve emotion detection to a new emphasis on generalized models which promote emotional tracking, self-awareness, and communication with multiple users in real time with increasingly common and less obtrusive pieces of technology. The present research continues toward this goal by providing real time feedback of emotional states using everyday technology. What then separates this emotion detection work is that the model aims to evaluate the user’s emotional state in real time, using consumer level hardware, relies on naturally elicited emotion, and tracked participants over weeks instead of a single lab session.

Methods

Phase 1: Building the model

Design

As the motivation for this work is to use everyday technology, the Microsoft Band was selected as the sensor (Figure 27). The Microsoft Band fits all criteria of being a 1) common consumer product, 2) capable of measuring EDA, and 3) capable of transmitting data in real time. By using a consumer device, data were able to be collected unobtrusively as individuals were already acclimated to wearing the sensor for extended periods of time throughout daily activities. The EDA sampling rate for the band was increased to 4 Hz, the maximum value for this hardware. The sampling rate was increased to provide a usable signal for processing, feature extraction, and analysis based on previous empirical studies.

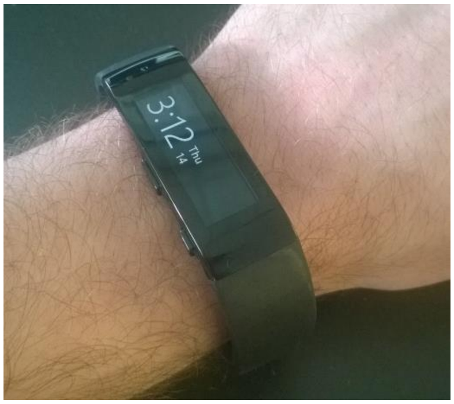


Figure 27. Microsoft Band, suggested orientation.

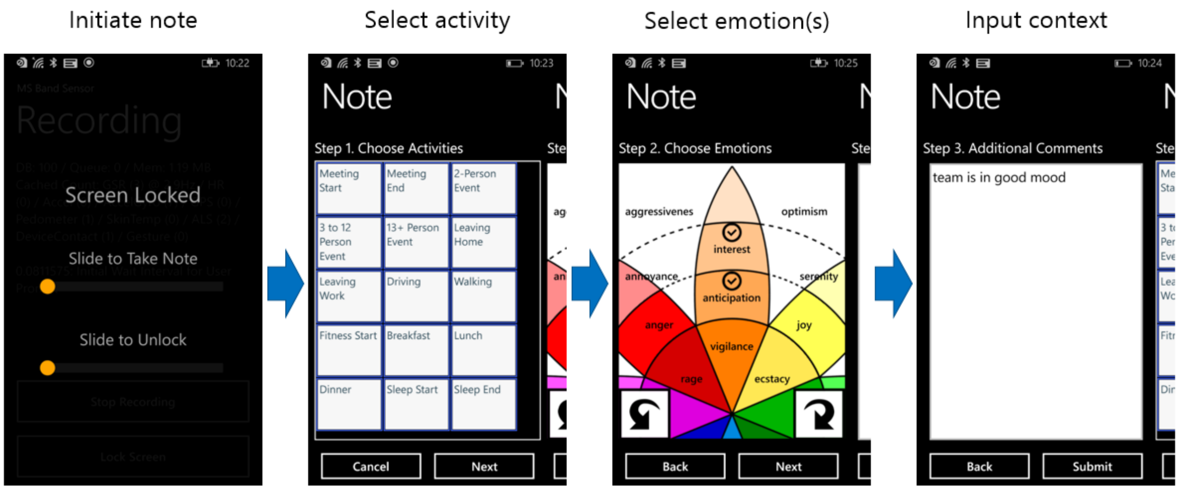
A motivating factor for this work was to collect data as part of natural scenarios outside of the laboratory. Participants were instructed to record only during meetings where they would be able to submit emotion data. Meetings were defined as being an interaction between the participant and at least one other individual. The suggested minimum meeting time length was 30 minutes. Meetings were selected as a natural scenario for interpersonal emotions to occur and to reduce noise via physical activity.

To record Band data, a custom Windows Phone application was created by a colleague to record and store both physiological and self-report data. The data collection application allowed the sensor data to be recorded passively while also providing a mechanism to provide emotional state feedback (Figure 28). The data collection application allowed participants to submit their data at will and would subtly prompt for feedback once every ten minutes. This phase focused on the EDA sensor data collected and the corresponding emotional state data submitted.

Participants

Participants involved in this work were recruited by a third party from within a large technology corporation. Potential participants were told they needed to have a Microsoft Band and Windows Phone and that they would be using these devices to record and submit their emotions. This study offered participants $25 for recording and submitting four separate meeting events over a two-week period, suggested as recording two meetings per week for two weeks. Additionally, the same incentive was offered for a second set of two weeks, and if both data submission periods were met (meaning eight separate meeting events recorded), the participant received a bonus $25. Participants were also offered the opportunity to opt-in for a second month of data collection.

118 participants were recruited, and of those, 89 were successfully setup for data collection. Those that were unable to join either had technical difficulties (e.g., phone unable to install application or band connection issues) or did not have the required hardware (i.e., used an iPhone or Android device).



DB: 100 / Queue: 0 / Mem: 1.19 MB

Pedometer (1) / SkinTemp (0) / ALS (2) /

DeviceContact (1) / Gesture (0)

0.0811575: Initial Wait Interval for User

team is in good mood

Figure 28. The note submission process participants followed.

Protocol

Upon arrival, participants completed a research project participation form informing of them of their rights and relevant risk information. Participants were instructed to begin the recording process when their meeting began and submit an initial "note" immediately. To continue recording data, the app was instructed to be left running in the foreground. The note taking process was the sequence of steps taken to submit data at a specific moment. The steps taken to submit a note were 1) initiating the note via "slide to take note" action, 2) selecting the activity going on (if applicable), 3) selecting which emotions were currently being felt, and 3) adding any additional comments for context (if applicable). This entire process is outlined in Figure 28.

The note submission process was used every time the participant would submit their emotion. The quickest note submission process required the participant to only input their emotion(s) and allowed the activity and additional comments to be skipped. Emotion data was collected using a visual representation of Plutchik's emotion wheel within the data collection app, seen in Figure 29, (Plutchik, 1984). The emotion wheel was subdivided into five "fuzzy" states as seen in Figure 29: joy, sadness, fear, anger, and neutral. Additionally, this wheel provided participants with both an emotional choice (the smaller text labels) and a sense of valence or intrinsic attractiveness or aversiveness (stronger valence towards the center of the radial axes) (Frijda, 1986).

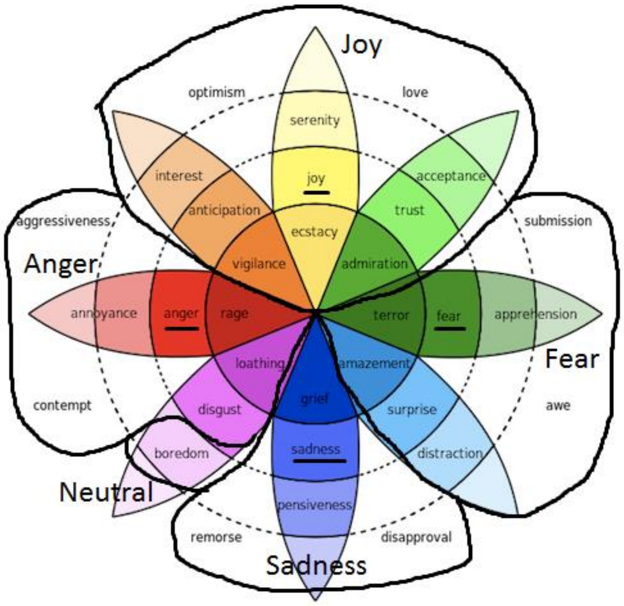


Figure 29. Plutchik's wheel of emotion with “fuzzy” groups.

Data collected was then analyzed on an individual participant basis to check for conformity to the study guidelines and for sensor data cleanliness. Of the 89 who submitted data, 32 gave data that were of excellent quality. Excellent quality was defined as data which had emotions regularly submitted only within the meeting scenario, and did not have missing EDA data. This was due to some participants recording sensor data without submitting any emotions or recording outside meeting scenarios. The 32 high-quality datasets were used to build a singular training set.

Sensor data collected were then run through multiple filters to smooth the signal, to identify skin conductance peaks preceding reported emotions, and lastly to extract meaningful feature sets both as outlined in Boucsein's Electrodermal Activity (Boucsein, 2012) and according to additional features selected by the research team. Example features utilized were prior emotion reported, time to prior emotion, peak amplitude, maximum incline, rise time, recovery time, area under the curve, and mean, standard deviation, maximum and minimum values for each. This process was used identify peak locations and extract meaningful features. Those features were attached to a single peak and were associated with emotions that users submitted via the emotion wheel. An example of features extracted can be seen in Figure 30.

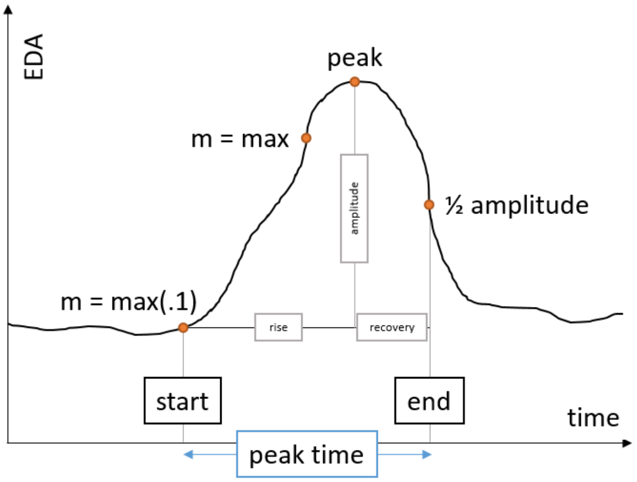


Figure 30. Example features extracted from sensor data.

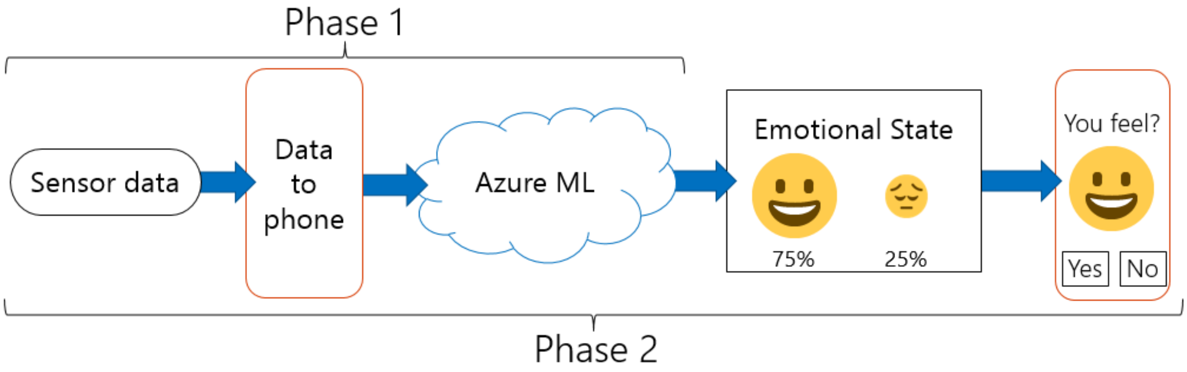


Figure 31. Data collection, analysis, and feedback process for phase 1 & phase 2. The participant is only shown the final step.

The training set built from submitted emotions and the features extracted were then used to train and validate a machine learning model using Azure Machine Learning. The supporting documentation was used to determine what learners and parameters to test and adjust ("Azure Machine Learning Documentation," 2016). A variety of both two-class and multi-class learners were tested, including multiclass decision jungles, one-vs-all multiclass learners, boosted decision trees, and support vector machines. Ultimately, the multiclass decision forest trained on measuring the five emotional states, or fuzzy emotions, from the emotion wheel subset was used as the primary learner for phase two model testing as having additional discrete states to classify as it allowed for the greatest variety of projected emotional states when compared with the two-class model. Accuracy values can be found under the results section of Phase 1. The overall process from participant data submission to validation within the machine learning model can be seen in Figure 31.

Phase 2: Testing the model

Design

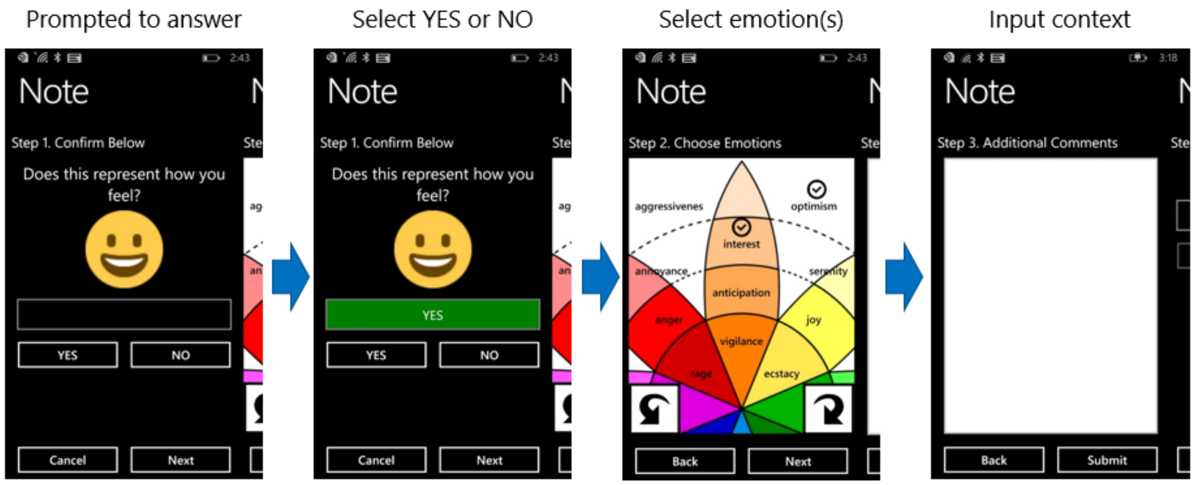
The goal of Phase 2 was to measure emotion in real time using the model trained from Phase 1. Phase 2 was a validation of the model working in real time with participants providing live feedback. Participants passively sent their sensor data in real time using the data collection app on their phone and provide feedback when prompted. These data were processed in real time and run through the trained emotion detection model to identify active emotional states.

When the emotion detection model had satisfied a set validation criterion, the participant was then prompted for feedback at a maximum of once every seven minutes.

Seven minutes was determined via pilot testing to find an interval which was frequent enough to collect useful data, but not too frequent as to frustrate the participant. Additionally, the participant was able to skip any entry prompt. At any given data submission point, the emotion detection model returned a confidence value related to each of the five fuzzy emotional states with a total value of 1 (e.g., joy .5 + sadness .25 + fear .25 = 1). The validation criterion was met when the emotion detection model placed majority confidence in a single fuzzy state (meaning at least >=50% confident in the primary state) and the second highest state was less than 60% of the primary state's value. This approach was used to prevent prompting for feedback when the emotion detection model was equally confident in two different states (e.g., 50/50 joy/sadness). Alternately, if no peaks had been detected from the incoming EDA data after 60 seconds, a prompt was sent with the estimated state of neutral. Feedback prompts were sent in one of two methods, the "emoji" method and the "blinded" method.

The first, "emoji" method would display a message stating "Does this represent how you feel?" and show the appropriate emotion for the participant to respond with yes or no, as seen in Figure 32 from one of five separate emojis (joy, anger, fear, sadness, and boredom. The default state had both yes and no unselected, users had to take action for either yes or no to be chosen. Then, participants were asked to input their emotional state via the emotion wheel. The "blinded" method prompted users to select their emotional state on the wheel, but the app did not show what the model's estimated emotion was. The "blinded" condition was setup as a within subjects independent variable to measure the bias from being shown an emoticon. The type of feedback offered alternated between the emoticon and blinded methods.

Data collected were run through the same analysis process as Phase 1. This phase focused on identifying the perceived accuracy of the model via participant feedback and the performance accuracy of the model via participant emotion data submitted.


Prompted to answer

Select YES or NO

Select emotion(s)

Input context

Step 1. Confirm Below

Does this represent how you feel?

Step 2. Choose Emotions

Step 3. Additional Comments

Figure 32. Emoticon feedback process, phase two. Participant answer to “yes or no” used to calculate perceived accuracy and the actual emotion they submitted was used to calculate measured accuracy.

Participants

To complete the live validation, a total of 16 participants from Phase 1 were enrolled to continue data collection within the meeting scenario. Participants enrolled in Phase 2 were invited from the pool of 32 individuals from Phase 1 who submitted high quality data. Those who did not participate in Phase 2 either directly declined or were unable to continue for a variety of reasons (moved, no longer used required hardware, did not respond, etc.).

Protocol

Phase 2 participants followed the same instructions as they had for Phase 1, with the primary difference being that instead of submitting their emotion at will, they would respond to a prompt asking them to report their emotion at a specific moment still only within meetings. Participants would receive prompts throughout the entire course of the meeting until the meeting ended or they closed the data collection application.

Participants were asked to respond to prompts as quickly as possible. Responses included selecting yes or no to the emoticon presented with the text, “Does this represent how you feel?” and also selecting any relevant emotion states being felt as done in Phase 1. If no emoticon was present, only the emotion states were selected. At the conclusion of the meeting time, participants would end their recording.

Results

The majority of the results described here are from Phase 2, as Phase 1 was primarily data collection, system building, and emotion detection model training. The first step was to validate whether or not the emotion detection model produced better than chance results using only the EDA sensor within a multiclass machine learning model. Once trained, the model was then enabled to allow live data to be evaluated and returned for participant feedback. Participant responses were then evaluated to identify the difference between perceived model accuracy and measured accuracy. Additionally, emoticon vs blinded responses were evaluated to determine whether a biasing effect was present.

Phase 1: Building the model

The multiclass emotion detection model used produced an overall accuracy of 51.5%. In this case the overall accuracy of the model is determined by the total number of correct state measurements from the emotion detection model when compared with the actual submitted responses of the 32 participants whose data was of excellent quality. Excellent quality data in this work required participants to record data during meeting scenarios and have a usable signal from the sensor (a loose Band would record EDA values of 0).

The final training set from those 32 participants was comprised of 1,721 data points (individual submissions) collected over four weeks. Each data point was comprised of up to 39 unique features. Phase 1 data collection continued, but was not used in Phase 2, the additional data collected will be used to build and test a new model. The first model to be built and tested used the five fuzzy emotional states. The training set was randomly split 80:20 into training and validation for model training. Multiple learner models were tested, a minimum of five runs per learner, with the multiclass decision forest having the highest reported overall accuracy of 51.5%.

Additionally, a positive and negative two-class model was trained from a subset of the five state, fuzzy data. The two-class boosted decision tree returned the highest results of 59.6% accuracy and 62.5% area under the curve when evaluating the receiver operating characteristic (ROC) curve. This model was not used for further testing within this work.

Phase 2: Testing the model

Emotional states were measured in real time with a perceived accuracy of 76.43% and a performance accuracy of 52.52%. Perceived accuracy was measured as the number of times participants replied yes to the emoticon presented to them with the text “Does this represent how you feel?” Performance accuracy was the number of times the participant-reported emotion matched the evaluated emotion from the emotion detection model. This difference between perceived accuracy and performance accuracy indicates that participants overestimated the accuracy of the emotion detection model by 23.91%. 16 participants were enrolled to record data in real time and provide feedback when prompted. Those participants submitted a total of 289 responses, of which 140 were emotion detection model prompted responses. These 140 responses were used to measure the perceived and actual accuracy. The other 149 responses were either self-initiated submissions or neutral prompts based off of a lack of physiological activity.

Figure 33 depicts an example of what the measured EDA response looks like from the live streaming view seen by the researchers. The data here reflect an individual's real time EDA value and the returned emotional state after being processed by the emotion detection model. The perceived accuracy comes from the participant's yes or no response. The performance accuracy is compared against the submitted emotions, here seen as "interest, optimism, and serenity."

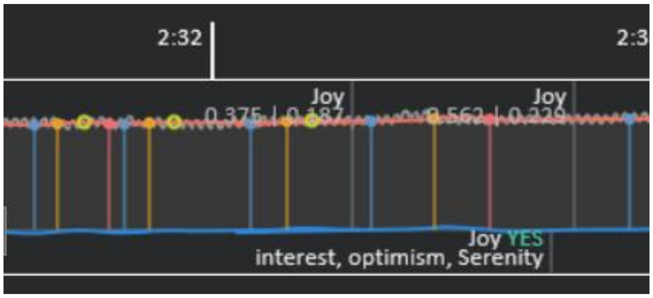


Figure 33. Real time data viewing interface. Here the participant is estimated to be feeling joy (above colored line), and when prompted with smile emoticon (for joy), participant responds YES (at bottom), and reports interest, optimism, and serenity.

There was no statistical difference in emotion response accuracy when comparing those which followed the emoticon vs. those that followed the blinded condition. The blinded condition yielded an accuracy of 56.25%, and the emoticon condition yielded an accuracy of 49.33%. The difference was not significant, $t(131) = 0.8104$, $p = .307$ and had a small effect size, $r = .129$.

Discussion

The goal of creating a model which performed at better than chance levels was successfully met. Data collected and validated shows that using an everyday piece of wearable technology can lead to better than chance measurement of an individual’s emotional state. This section discusses the aforementioned results including the constraints, assumptions, and limitations within. Revisions to this work, plans for future work, and opportunities for this emotion detection model are also covered.

Phase 1: Building the model

Overall, EDA was shown to be able to measure emotions in real time using everyday technology with an overall accuracy of 51.5%, but with constraints. Limited data, only 1,721 data points from 32 participants were used for the training and validation of the emotion detection model tested in Phase 2. The majority of the data collected from other participants were not used due to poor data quality, which was primarily caused by either loose fitting connections, high physical activity, or incorrect scenario recordings. This restriction meant that this work was not tested in a truly general environment, but the goal of moving outside the lab environment was met. Participants were also relied upon to give accurate self-report of their emergent emotional status instead of using an emotion elicitation protocol as seen in previous studies (Bailenson et al., 2008; Cowie et al., 2000; Haag et al., 2004; Rosalind W. Picard et al., 2001; Salimpoor, Benovoy, Longo, Cooperstock, & Zatorre, 2009).

Multiple machine learning trainers were tested to determine the best results. Some of the other models compared with the final multiclass decision forest were the multiclass decision jungle and one-vs-all multiclass. Two-class models were also tested, including decision forest, decision jungle, boosted decision tree, and a support vector machine. The two-class boosted decision tree returned the best results.

Two-class models were tested using the positive, negative emotion split and returned an accuracy of 59.6%. Additionally, one two-class model was trained solely on a single participant's data and returned an accuracy of 75%. Other single models could be trained to investigate whether this increased accuracy holds true for more individually trained, non-generalizable models as seen in previous work (Haag et al., 2004; Rosalind W. Picard et al., 2001).

Additional data collected during this work both from Phase 1 and Phase 2 will be added to a larger set and then used to retrain the emotion detection model for future testing. Increased high quality data should improve the model using the same multiclass decision forest model.

Phase 2: Testing the model

Participants overestimated the accuracy of the emotion detection model. The delta between the two measures was 24%, (76% perceived – 52% actual). This result can be seen as an advantage when considering self-perception theory, which indicates that individuals derive their emotional state based off of their own perception and behavior. A system which recognizes this could potentially use this suggestive power appropriately, (Chang, Resner, Koerner, Wang, & Ishii, 2001). It should also be considered that participants were susceptible to the Hawthorne Effect since they provided self-report data that they knew was contributing to a study, (Cowie et al., 2000). In an attempt to counter this effect, the participants were shown both the emoticon prompts and also blinded prompts to alleviate any bias they may have to provide correct answers.

Researchers anticipated that participants would be more likely to report emotions which aligned with the emoticons shown. There was no measured difference between the emoticon and blinded conditions with respect to accuracy. As the suspected bias was not displayed in the data, one of two possibilities seem most likely: 1) participants were providing answers they thought the researchers would want or 2) participants may have partially been feeling what the emoticon represented, as emotions are not always mutually exclusive states (Paul Ekman & Davidson, 1994). The first option is possible, but more likely, participants could agree with the displayed emoticon for some amount, but it may not have been the only emotion they were feeling. Regardless, more data is needed to determine if there is a meaningful difference between model accuracy for the emoticon vs blinded conditions.

Design implications

When implementing this type of emotion detection technology, the considerations for use within research and daily lives of the general consumer are very different.

As a research tool, this system's interface, data collection process, and information presentation does not need to consider the participant as heavily, especially if the participant is not involved in the data collection process. This tool will likely be used to gain additional objective measures without the participant's direct intervention. There may be scenarios in which the participant would need to take part in the data collection process, such as the work presented in this paper, but in this case only select pieces of the process would be shown and can be improved to shorten their interaction time and reduce the amount of mental effort required to record and submit data.

The larger design implications stand with using this type of emotion detection technology with the general population or even with specific populations looking to improve their emotional self-awareness and communication skills. The primary purposes of this technology would be recording, self-recognition, communication, and self-reflection of one's emotional state. Additional research in how this technology would specifically be used is needed, but many of the first steps have been taken. Previous research outlines how to maximize autism research (Carter & Hyde, 2015; R. W. Picard, 2010), what an "emotional prosthetic" may look like (El Kaliouby et al., 2006; Roseway et al., 2015), and even how emotions can be communicated in everyday interactions (Chang et al., 2001; Olivier & Wallace, 2009; G. Wilson et al., 2015).

Conclusions

This work shows promise that everyday technology can be used to measure individual emotional states in real time. While this work has limitations with a positive emotion bias data set due to a single scenario (meetings only) and a single sensor data type (EDA), it is a step toward a generalizable emotion detection model which can be utilized in a variety of different ways to bring about improved emotional self-awareness, communication, and affective computing scenarios in the future.

Limitations and Future Work

Some limitations of this work are: 1) the data is heavily bias toward positive emotions, 2) only the EDA data were used for emotion detection, 3) the sensor quality on the Microsoft Band is relatively low when compared with other research quality physiological sensors, and 5) data were collected only within meeting scenarios.

As seen in Figure 34, the fuzzy emotions pictured are presented by the size of their representation within each area (separated by color). The blue emotion wheel displays emotions available by their count on the emotion wheel, no participant data used. The red displays the proportion of times an emotion was submitted by a participant during the data collection. The gray displays the proportion of times an emotion was prompted for by the emotion detection model.

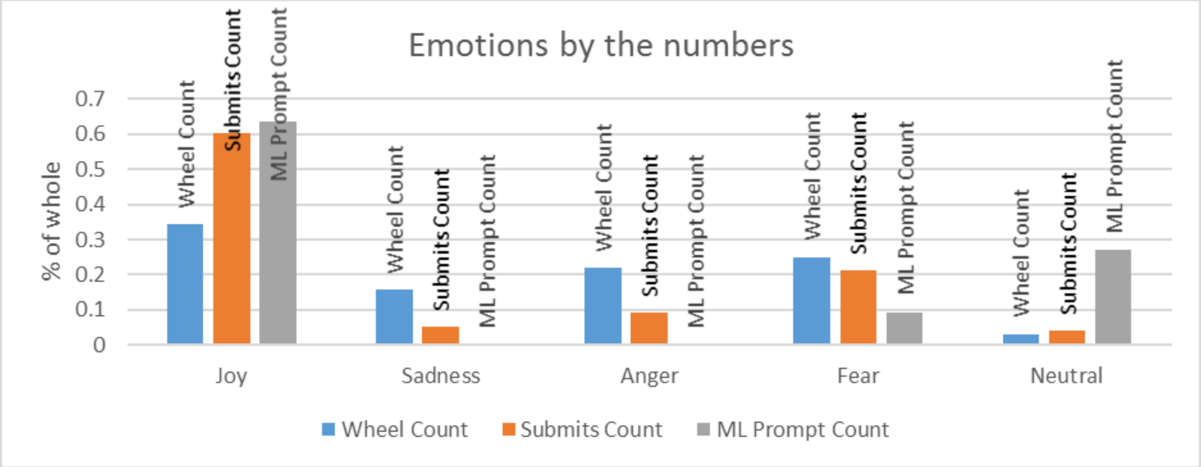


Figure 34. Emotion fuzzy states represented by their count after normalizing for each of the three areas.

Of the five fuzzy emotional states, (joy, sadness, anger, fear, neutral), sadness and anger were never selected as a validation point by the emotion detection model. This suggests the model was biased toward joy and leads to a suggestion of further data collection to strengthen the model’s confidence in the less often reported emotions such as sadness and anger. This is not surprising though as cultural norms in the workplace encourage positive attitudes and do not often foster sadness or fear. This suggests that data should be collected in additional scenarios outside meetings alone.

Emotion is known to be a two measure construct of both arousal and valence (Boucsein, 2012; Russell, 1980). EDA is generally used as a measure of arousal and heart rate measures (normally photoplethysmography, PPG) are generally used as a measure of valence (Greenwald, Cook, & Lang, 1989; Johnsen, Thayer, & Hugdahl, 1995; Rosalind W. Picard et al., 2001; Roseway et al., 2015; Winton, Putnam, & Krauss, 1984). This study also collected participant’s PPG data but was unable to use it at the time of testing. This work continued without the use of PPG data as those data were not available to use at the time of the study due to technical limitations. Levenson, Ekman, and Frisen reported after three repeated experiments EDA was shown to differentiate between positive and negative emotions (Levenson, Ekman, & Friesen, 1990). Boucsein also indicates that the traditional limits of EDA as a measure of arousal could also be due in part to the common emotion elicitation technique of emotionally evocative image stimuli, which this study did not utilize.

Lastly, this study utilized Plutchik’s emotion model as the visual medium for participants to input their emotions. Plutchik’s model was selected as it allowed multiple emotions to be selected at once and is used in other areas of emotion research. Retrospectively, Russell’s circumplex model of affect would have been simpler to interact with for the participant and been more consistent with the majority of emotion detection research today (Russell, 1980).

The immediate future work to improve this emotion detection model lies in retraining the model with collected PPG sensor data and also include new data collected after this analysis was performed. Additionally, demographic data will be utilized to investigate potential differences within the participants.

References

Azure Machine Learning Documentation. (2016). Retrieved from https://msdn.microsoft.com/en-us/library/azure/

Bailenson, J. N., Pontikakis, E. D., Mauss, I. B., Gross, J. J., Jabon, M. E., Hutcherson, C. a C., … John, O. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human Computer Studies*, *66*(5), 303–317. http://doi.org/10.1016/j.ijhcs.2007.10.011

Boucsein, W. (2012). *Electrodermal Activity* (2nd Ed). Wuppertal: Springer.

Cacioppo, J. T., Berntson, G., Larsen, J., Poehlmann, K. M., & Ito, T. (2000). The Psychophysiology of Emotion. In R. Lewis & J. M. Haviland-Jones (Eds.), *The Handbook of Emotion* (2nd ed., pp. 173–191). Guilford Press.

Carter, E. J., & Hyde, J. (2015). Designing Autism Research for Maximum Impact. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2801–2804).

Chang, A., Resner, B., Koerner, B., Wang, X., & Ishii, H. (2001). LumiTouch : An Emotional Communication Device. In *Proceedings of the ACM 2001 Conference on Human Factors in Computing Systems* (pp. 2–3). http://doi.org/10.1145/634067.634252

Cowie, R., Douglas-Cowie, E., Savvidou, S., Mcmahon, E., Sawey, M., & Schröder, M. (2000). “Feeltrace”: An instrument for recording perceived emotion in real time. *ISCA Workshop on Speech & Emotion*, 19–24. http://doi.org/citeulike-article-id:3721917

Damasio, A. R. (2005). *Descartes’ Error: Emotion, Reason, and the Human Brain*. Penguin.

Ekman, P., & Davidson, R. J. (1994). *The Nature of Emotion: Fundamental Questions*. Oxford University Press.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. *Journal of Personality*. ISHK. http://doi.org/10.1163/1574-9347_bnp_e804940

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science (New York, N.Y.)*, *221*(4616), 1208–1210. http://doi.org/10.1126/science.6612338

El Kaliouby, R., Teeters, A., & Picard, R. W. (2006). An exploratory social-emotional prosthetic for autism spectrum disorders. *Proceedings - BSN 2006: International Workshop on Wearable and Implantable Body Sensor Networks*, *2006*, 3–4. http://doi.org/10.1109/BSN.2006.34

Essa, I. a., & Pentland, a. (1994). A vision system for observing and extracting facial action\nparameters. *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, *1994*(247), 76–83. http://doi.org/10.1109/CVPR.1994.323813

Fridlund, A. J., & Izard, C. E. (1983). Electromyographic Studies of Facial Expressions of Emotions and Patterns of Emotions. In J. T. Cacioppo & R. E. Petty (Eds.), *Social Psychophysiology: A Sourcebook* (pp. 243–286). Guilford Press.

Frijda, N. (1986). *The Emotions*. Cambridge University Press.

Golan, O., Baron-Cohen, S., & Hill, J. (2006). The Cambridge Mindreading (CAM) Face-Voice Battery: Testing complex emotion recognition in adults with and without Asperger Syndrome. *Journal of Autism and Developmental Disorders*, *36*(2), 169–183. http://doi.org/10.1007/s10803-005-0057-y

Goleman, D. (2006). *Emotional Intelligence*. Bantam.

Graesser, A. (2009). Deep learning and emotion in serious games. In *Serious games: Mechanisms and effects* (pp. 81–100).

Greenwald, M. K., Cook, E. W., & Lang, P. J. (1989). Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of Psychophysiology*.

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004). Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System. *Affective Dialogue Systems*, *i*, 36–48. http://doi.org/10.1007/978-3-540-24842-2_4

Hascher, T. (2010). Learning and emotion: Perspectives for theory and research. *European Educational Research Journal*, *9*(1), 13–28. http://doi.org/10.2304/eerj.2010.9.1.13

Healey, J. A. (2000). *Wearable and automotive systems for affect recognition from physiology*. MIT.

Hollis, V., Konrad, A., & Whittaker, S. (2015). Change of Heart. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (pp. 2643–2652). New York, New York, USA: ACM Press. http://doi.org/10.1145/2702123.2702196

Johnsen, B. H., Thayer, J. F., & Hugdahl, K. (1995). Affective judgment of the Ekman faces - a dimensional approach . *Journal of Psychophysiology*, *9*(3), 193–202.

Katsis, C. D., Katertsidis, N., Ganiatras, G., & Fotiadis, D. I. (2008). Toward emotion recognition in car racing drivers: a biosignal processing approach. *IEEE Trans. Systems, Man and Cybernetics - Part A: Systems and Humans*, *38*(3), 502–512.

Kim, K. H., Bang, S. W., & Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, *42*(3), 419–427. http://doi.org/10.1007/BF02344719

Kim, S., Georgiou, P. G., Lee, S., & Narayanan, S. (2007). Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *2007 IEEE 9Th International Workshop on Multimedia Signal Processing, MMSP 2007 - Proceedings* (pp. 48–51). http://doi.org/10.1109/MMSP.2007.4412815

LeDoux, J. E. (1996). *The Emotional Brain*. Simon & Schuster.

Leon, E., Clarke, G., Callaghan, V., & Sepulveda, F. (2007). A user-independent real-time emotion recognition system for software agents in domestic environments. *Engineering Applications of Artificial Intelligence*, *20*(3), 337–345. http://doi.org/10.1016/j.engappai.2006.06.001

Levenson, R. W., Ekman, P., & Friesen, W. V. (1990). Voluntary facial action generates emotion-specific autonomic nervous system activity. *Psychophysiology*, *27*(4), 363–384. http://doi.org/10.1111/j.1469-8986.1990.tb02330.x

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, *24*(6), 615–625. http://doi.org/10.1016/j.imavis.2005.09.011

McDuff, D., Karlson, A., Kapoor, A., Roseway, A., & Czerwinski, M. (2012). AffectAura: An Intelligent System for Emotional Memory. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (p. 849). New York, New York, USA: ACM Press. http://doi.org/10.1145/2207676.2208525

Obrist, M., Subramanian, S., Gatti, E., Long, B., & Carter, T. (2015). Emotions Mediated Through Mid-Air Haptics. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (pp. 2053–2062). New York, New York, USA: ACM Press. http://doi.org/10.1145/2702123.2702361

Olivier, P., & Wallace, J. (2009). Digital technologies and the emotional family. *International Journal of Human Computer Studies*, *67*(2), 204–214. http://doi.org/10.1016/j.ijhcs.2008.09.009

Picard, R. W. (1997). *Affective Computing*. Cambridge: MIT Press.

Picard, R. W. (2010). Emotion Research by the People, for the People. *Emotion Review*, *2*(3), 250–254. http://doi.org/10.1177/1754073910364256

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191. http://doi.org/10.1109/34.954607

Plutchik, R. (1984). Emotions: A General Psychoevolutionary Theory. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion*. New York: Psychology Press.

Roseway, A., Lutchyn, Y., Johns, P., Mynatt, E., & Czerwinski, M. (2015). BioCrystal : An Ambient Tool for Emotion and Communication. *International Journal of Mobile Human Computer Interaction*, *7*(3), 20–41.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. http://doi.org/10.1037/h0077714

Salimpoor, V. N., Benovoy, M., Longo, G., Cooperstock, J. R., & Zatorre, R. J. (2009). The rewarding aspects of music listening are related to degree of emotional arousal. *PLoS ONE*, *4*(10). http://doi.org/10.1371/journal.pone.0007487

Schutte, N. S., Malouff, J. M., Thorsteinsson, E. B., Bhullar, N., & Rooke, S. E. (2007). A meta-analytic investigation of the relationship between emotional intelligence and health. *Personality and Individual Differences*, *42*(6), 921–933. http://doi.org/10.1016/j.paid.2006.09.003

Slovák, P., Gilad-Bachrach, R., & Fitzpatrick, G. (2015). Designing Social and Emotional Skills Training. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (pp. 2797–2800). New York, New York, USA: ACM Press. http://doi.org/10.1145/2702123.2702385

Stefani, O., Mahale, M., Pross, A., & Bues, M. (2011). SmartHeliosity: Emotional Ergonomics through Coloured Light. In *Lecture Notes in Computer Science* (Vol. 5624, pp. 226–235). http://doi.org/10.1007/978-3-642-21716-6_24

Stowell, J. R., & Nelson, J. M. (2007). Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion. *Teaching of Psychology*, *34*, 253–258. http://doi.org/10.1080/00986280701700391

Tajadura-Jiménez, A., Basia, M., Deroy, O., Fairhurst, M., Marquardt, N., & Bianchi-Berthouze, N. (2015). As Light as your Footsteps. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI ’15* (pp. 2943–2952). ACM Press. http://doi.org/10.1145/2702123.2702374

Vogt, T., André, E., & Bee, N. (2008). EmoVoice - A framework for online recognition of emotions from voice. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5078 LNCS*, 188–199. http://doi.org/10.1007/978-3-540-69369-7_21

Wilson, G., Davidson, G., & Brewster, S. (2015). In the Heat of the Moment: Subjective Interpretations of Thermal Feedback During Interaction. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (pp. 2063–2072). ACM Press.

Winton, W. M., Putnam, L. E., & Krauss, R. M. (1984). Facial and autonomic manifestations of the dimensional structure of emotion. *Journal of Experimental Social Psychology*, *20*(3), 195–216. http://doi.org/10.1016/0022-1031(84)90047-7

CHAPTER 5

EMOTION PATTERNS AND MEASURES RELATIVE TO PERSONALITY

This chapter will be submitted to *IEEE Transactions on Affective Computing*.

Author list is:

Chase Meusel, Will Stone, Stephen B. Gilbert, Mike Bortnick, Jin Jun, Joe Munko, Umer Farooq

Chase Meusel's role in this research included the experimental design, participant recruitment, data collection, data coding, and primary authorship on the paper, including all sections but results. The statistical analysis and results section were completed by Will Stone.

Abstract

This work outlines the exploration of self-reported emotion tracking responses relative to personality scores. Emotions were recorded via an emotion tracking mobile application in the participant's workplace in conjunction with their physiological data. Personality scores were recorded via the Big Five Inventory personality survey on the first day of data collection. Over the course of two months, 92 participants opted to wear a physiological sensor, the Microsoft Band, and record their emotional state. Data collected was first used to build an emotion detection model with machine learning (Meusel, et al., submitted). This effort raised questions of whether personality characteristics might be correlated with emotional state as estimated by the model, and whether the particular emotion

detection model might be skewed based the distribution of personalities in its population sample.   Personality scores indicated the sample was more conscientious, more agreeable, less neurotic, and less open with no difference in extraversion. Agreeableness was negatively correlated with Neutral and Sad emotion response rates (τ = -.21 [-.36, -.06] and -.18 [-.32, -.02], respectively). As anticipated, Neuroticism was positively correlated with Sad emotion response rate (τ = .32 [.18, .45]). These results provided additional support for the common finding that Neuroticism is correlated with negative emotion but does not provide support either direction for the common finding Extraversion correlates with positive emotions. These findings are useful as an exploration of how personality and emotion are related when data are collected in real-world scenarios over a two-month period and not from a single emotion capture event.

Introduction

Emotion is known to be a complex, important, and largely underutilized part of today's technological ecosystem (R. W. Picard, 2010; Rosalind W. Picard, 1997).   Not only do emotions represent how we're feeling in an immediate sense, they also influence a number of factors from social interactions to learning aptitude (Graesser, 2009; Hascher, 2010; Stowell & Nelson, 2007).

Personality has a long relationship with emotion through a variety of topics. Early work discussed how emotions existed relative to self. Magda Arnold's work (Arnold, 1960) outlined the view that emotion exists in a relationship between an individual's self-perception and the object of their emotion relative to their own appraisal of that object. Arnold also outlined the concept of appraisals which would influence multilevel appraisal theories which have become popular in emotion research (Ellsworth, P. C., & Scherer, Ellsworth, & Scherer, 2003; Moors, 2014; K.R. Scherer, 1999).

A large part of the relationship between emotion and personality comes from a desire by researchers to understand how these two influence the other and better understand humans in general (Fossum & Barrett, 2000; Georgi, Grant, Georgi, & Gebhardt, 2006; Wang, Shi, & Li, 2009). Various attitudes toward evaluation have been investigated depending on what emotional model is being used. Russell's circumplex model of affect is the most popular and simplest method of structuring emotion on a two-dimensional mapping of valence and arousal (Russell, 1980). Yet others suggest considering a third dimension for evaluation such as social desirability of a mood (Feldman Barrett, 1996), or variations on the approach-avoidance spectrum (Morgan & Heise, 1988). Similarly, personality itself is not a construct made of a single dimension, both description and evaluation have their own independent spectrums when looking at personality constructs or even each factor of the Big Five, (Peabody, 1967; Gerard Saucier, 1994). Additionally, description is the more important spectrum where evaluation requires subjective judgement of desirability and as that varies more widely, is inherently less useful.

Within this work, the Big Five personality traits are the primary method for identifying individual personality traits. The Big Five traits are Conscientiousness, Agreeableness, Neuroticism, Openness, and Extraversion (Goldberg, 1990; Robert R. McCrae & Costa, 1999). There is also previous evidence that emotion and personality share at least two accepted associations. The first, neuroticism is correlated with negative emotions and the second, extraversion is correlated with positive emotion (Costa & McCrae, 1980; R. R. McCrae & Costa, 1991; McNiel & Fleeson, 2006; Wang et al., 2009).

Within emotion and personality research, emotions or affect are typically reported as a single event at the same time as the personality information is gathered. Surveys such as the Postive and Negative Affect Scale (PANAS) or The Multiple Affect Adjective Checklist (MAACL) are traditionally used to gather this information (David Watson, Clark, & Tellegen, 1988; Zuckerman & Lubin, 1965). While these tools are useful for gathering affective measures at a single point in time, they do not offer the utility of a long-term emotion tracking solution. As emotion tracking tools become more prevalent, solutions which allow the capture of longitudinal data will become increasingly important in understanding the relationship that personality and emotion have with each other over time.

Picard et al., (2001) described the most natural setup for collecting emotion data as one that has participants experiencing emotion due to an outside stimulus (event-elicited), happening outside of the laboratory (real-world), felt as an internal feeling (feeling), monitored unknowingly (hidden-recording), for a reason unrelated to the collection of emotion detection (other-purpose). In addition, the authors suggest that data should be collected from a large sample (multiple participants), over an extended period (long-term data collection), using everyday technology (consumer hardware), in this case a simple application to record emotion.

This work uses event-elicited emotions from real-world scenarios which are reported as internal feelings from the participant. Data were collected as an open-recording as participants were aware they were submitting data for a study investigating an emotion-purposed study. Multiple participants were observed over an extended period, or long-term data collection, using consumer hardware as the device to record emotions.

This work explores the relationship between reported emotion and personality when participants track their emotions in real-life scenarios while interacting with other people without being prompted by a specific emotion elicitation protocol over a long-term period of data collection.

Background and Related Work

Research on personality and emotion tracking has continued to grow in recent years, but the clear majority of this work has been done in controlled laboratory conditions, using emotion elicitation protocols with research equipment. This current work attempts to address this gap by collecting data outside of the laboratory and allowing participants to experience and describe their own feelings.

The Big Five, A Brief Review

The Big Five personality factors began as a much larger, exhaustive list of words to be evaluated and organized by their language attributes. The initial effort took nearly 18,000 terms and aimed to evaluate them by using lexical analysis, or using the known dictionary explanation of those terms which include all terms relevant to personality, then uses that information to group the terms into clusters (Allport & Odbert, 1936; Baumgarten, 1933; Klages, 1926). A few years later, Cattell used this work as a starting point to create his own list of trait terms to analyze (Raymond B. Cattell, 1943, 1945a; Raymond B Cattell, 1945b). Cattell took a subset of the larger list, 4,500 terms and conducted a new analysis to further reduce that set down to 35 variables. This drastic reduction in terms was primarily an artifact of computational limits for the time (John, Angleitner, & Ostendorf, 1988). After a number of separate analyses, Cattell arrived at 12 personality factors from this work and those 12 factors would ultimately become part of the 16 personality factors (16PF) survey later on (R. B. Cattell, Eber, & Tatsuoka, 1970).

The next step toward the Big Five was in taking a further reduced set (22 terms) of Cattell's 35 and evaluating those based on self-ratings, peer rating, and expert psychology staff member ratings (Fiske, 1949). These ratings were ultimately very similar and were roughly shaped into what was becoming the Big Five set of factors. Tupes and Christal then reanalyzed data from different samples consisting of varying populations (military personnel to graduated college students). The five recurring factors emerged as the Big Five. The model began to gain popularity and a name, the "Big Five" by Goldberg, (1981). The Big Five was setup to be five diverse categories which could, from a very high level, organize the underlying personality factors within each group. The Big Five would, at minimum, provide a common framework for future personality research to consider and work from. Future work would display that the Big Five was the only set of personality criteria which would endure various sample sizes, types, and data collection methodologies (G Saucier, 1997).

The Big Five factors listed in the common CANOE orientation are:

I. Conscientiousness: (efficient/organized vs. easy-going/careless)

II. Agreeableness: (friendly/compassionate vs. analytical/detached)

III. Neuroticism: (sensitive/nervous vs. secure/confident)

IV. Openness to experience: (inventive/curious vs. consistent/cautious)

V. Extraversion: (outgoing/energetic vs. solitary/reserved)

The Big Five have since been used and validated many times over, including multiple sets of research teams, including Tupes and Christal (1961), Goldberg (1981; Lewis R.

Goldberg et al., 1990), and Costa and McCrae (1980; 1991; 1999). Costa and McCrae specifically worked to reduce the factors to three, (Neuroticism, Extraversion, & Openness, NEO) but their later work does include all five factors.

More recently, Anusic, Schimmack, Pinkus, & Lockwood, (2009) have suggested that to draw firm conclusions multitrait-multimethod personality data must be gathered. This may be because the Big Five are orthogonal dimensions, but it is expected that single rate data will have correlations between factors. Alternatively, they also state that because you can estimate variance based on the data collection method, in this case single reporter (self), some weaker conclusions can be drawn. Ultimately, they suggest single reported personality scores be recorded and utilized with the caveat that the sample personality distribution is not an unknown.

Measuring Emotion

The first major effort to systematically identify an individual's emotional or affective state in addition to self-report data began with Ekman's work highlighting how to extract affective states from facial clues (Ekman & Friesen, 1975). Ekman's facial affect coding system (FACS) was built and tested as a computer vision solution targeted to discriminate between multiple emotional states, including neutral, anger, disgust, fear, joy, sadness, and surprise (Essa & Pentland, 1994; Littlewort, Bartlett, Fasel, Susskind, & Movellan, 2006). Ekman (Ekman, Levenson, & Friesen, 1983) then suggested emotion was difficult to observe physiologically (e.g., via EDA, heart rate, pupillometry, etc.) due to the recording window, or epoch size, being too large and multiple emotions adding noise to the data. This critique was improved upon as others outlined how physiological measures could be used as methods for measuring emotional states while still primarily using existing self-report measures of emotion as ground truth (Cacioppo, Berntson, Larsen, Poehlmann, & Ito, 2000; Healey, 2000).

While measuring emotion via physiological means has become a viable option in limited scenarios, the most common way to capture emotion data still lies within self-report data. Self-report is still the easiest and most reliable way to measure emotion (Feldman Barrett, 1996). Self-report can be split into 3 areas: 1) social desirability of a mood, 2) hedonic tone, and 3) level of arousal. While not surprising, this reiterates there is still an opportunity to design a system to systematically detect emotion and potentially personality, without relying on self-report. While self-report has understood bias, that bias has not been shown to be consistent in such a manner that all the observed variance is accounted for. This inherent randomness in self-report data is the worst part of the best tool available for researchers today.

Big Five Findings

The existing evidence that emotion and are connected via neuroticism and extraversion to negative and positive emotions continues to see affirmative testing in the data. Neuroticism has been also called emotional instability, or a likelihood to experience psychological distress where extraversion has been sociability or the likelihood to act more socially (R. R. McCrae & Costa, 1991). A number of studies have shown this relationship's pattern to hold true across samples (Costa & McCrae, 1980; McNiel & Fleeson, 2006; Rusting & Larsen, 1997; Wang et al., 2009; D; Watson & A, 1992). This connection is so strong that Neuroticism is sometimes used interchangeably with negative emotion and Extraversion is taken to mean positive emotion.

To understand a normal Big Five distribution, a large sample using this framework should be referred to (Srivastava, John, Gosling, & Potter, 2003). Srivastav, et al., collected personality responses from 132,515 participants, ages 21-60. The distributions of percentage of maximum possible (POMP) scores were noted to compare with the sample from this works collected data set.

This research focuses on the self-reported emotions using mobile technology in conjunction with personality scores to explore the emotion-personality relationship in real-life scenarios over an extended data collection period.

Methods

Design

As the motivation for this work is to use everyday technology, a custom Windows Phone application was created to record and store both physiological and self-report data. In addition to the self-report emotion data, participants were also instructed to wear their Microsoft Band wearable device to collect physiological data. The physiological data was not analyzed for the purposes of this report. The data collection application allowed the sensor data to be recorded passively while also providing the interface for participants to input their emotional feedback (Figure 35). The Windows Phone emotion submission application allowed participants to input their data at will and then the application would prompt the participant to input feedback every ten minutes. By using a mobile application, data could be collected unobtrusively as individuals already carried their device with them into most scenarios.

Experimentally, the motivating factor for investigating personality and emotion together inside of this study was to explore the data collected as part of natural scenarios outside of the laboratory. Participants were instructed to record only during meetings where they would be able to submit emotion data. Meetings were defined as being an interaction between the participant and at least one other individual. The suggested minimum meeting time length was 30 minutes. Meetings were selected as a natural scenario for interpersonal emotions to occur and to reduce noise by not encouraging data collection at all times.

Participants

Participants involved in this work were contacted and recruited by a recruitment service, participants were exclusively made up of employees of a large technology corporation. Potential participants were told they needed to have a Microsoft Band and Windows Phone and that they would be using these devices to record and submit their emotions. This study offered participants $25 for recording and submitting four separate meeting events over a two-week period, suggested as recording two meetings per week for two weeks. Additionally, the same incentive was offered for a second set of two weeks, and if both data submission periods were met (meaning eight separate meeting events recorded), the participant received a bonus $25. Participants were also offered the opportunity to opt-in for a second month of data collection.

118 participants were recruited in total. Of those 118, 92 were successfully setup for data collection. Of the 92, 73 were male and 19 were female. Those that were unable to join

either had technical difficulties (e.g., phone would not support the data collection or had band related issues) or did not have the required hardware (i.e., used an iPhone or Android device).

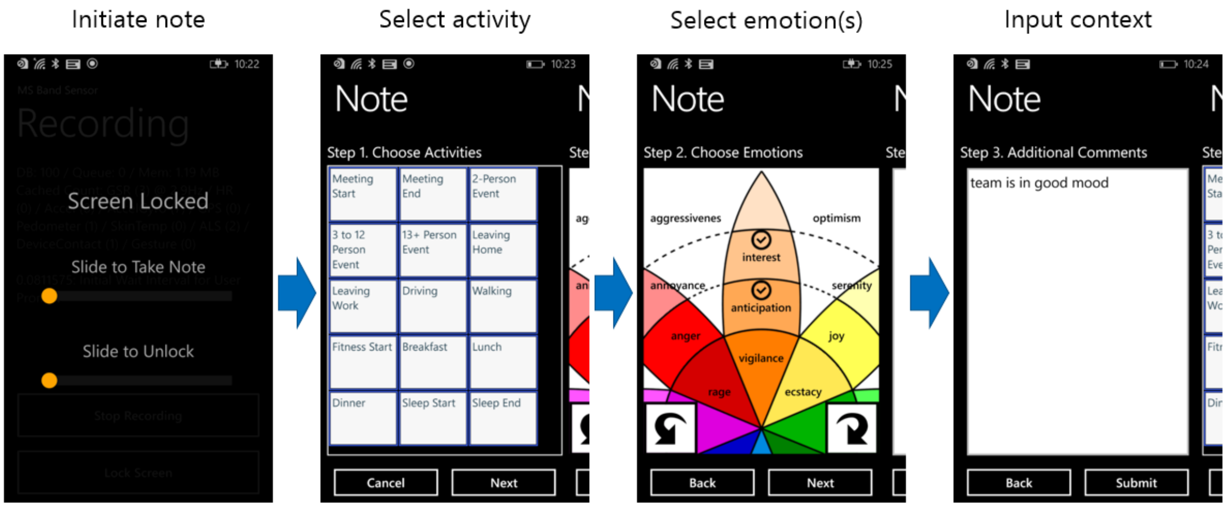


Figure 35. The note submission process participants followed.

Protocol

Upon arrival, participants completed a research project participation form informing them of their rights and what potential risk they would be taking on. Immediately following that, participants completed the 44 item Big Five Inventory personality survey. Participants were instructed to begin the recording process when their meeting began and submit an initial "note" immediately. To continue recording data, the app was instructed to be left running in the foreground. The note taking process was the sequence of steps taken to submit data at a specific moment. The steps taken to submit a note were 1) initiating the note via "slide to take note" action, 2) selecting the activity going on (if applicable), 3) selecting which emotions were currently being felt, and 3) adding any additional comments for context (if applicable). This entire process is outlined in Figure 35.

The note submission process was used every time the participant would submit their emotion. The quickest note submission process required the participant to only input their emotion(s) and allowed the activity and additional comments to be skipped. Emotion data was collected using a visual representation of Plutchik’s emotion wheel within the data collection app, seen in Figure 36 (Plutchik, 1984). The emotion wheel was subdivided into five "fuzzy" states: joy, sadness, fear, anger, and neutral. Additionally, this wheel provided participants with both an emotional choice (the smaller text labels) and a visual scale of valence or intrinsic attractiveness or aversiveness (stronger valence towards the center of the radial axes) (Frijda, 1986).

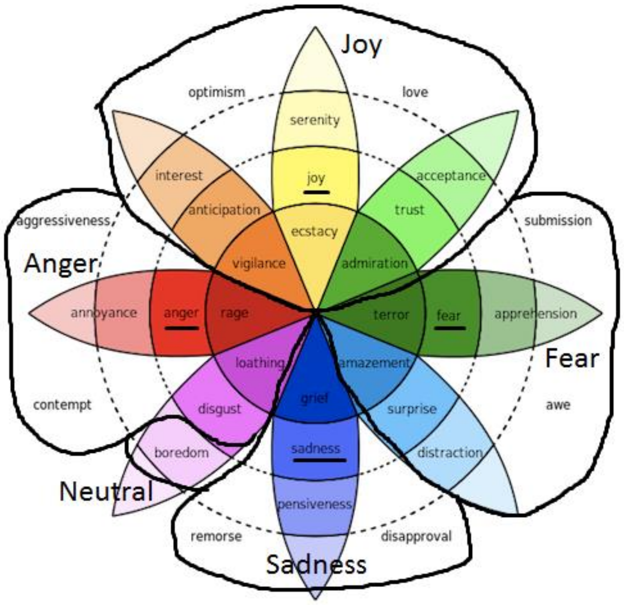


Figure 36. Plutchik's wheel of emotion with “fuzzy” groups.

Data collected was then analyzed on an individual participant basis to check for conformity to the study guidelines and for sensor data cleanliness.

Predictions

Although emotion data is captured over an extended period of time, it is expected that individuals with a higher Extraversion score will report higher positive emotions. Similarly, it is expected that individuals with a higher Neuroticism score will report higher negative emotions.

Results

Emotion responses were available for 92 participants, however, responses from some participants were excluded from analysis for two primary reasons. First, responses were excluded from 14 participants for not completing the personality measures. Then, a remaining 3 were excluded because they wore the band for less than 10 minutes. Results were then based on 73 participants, 23% of which were female. Average participation time was 6 hours, but the empirical distribution was positively skewed (*M* = 6, *Median* = 5.3, *SD* = 4, range: 0.4—18.6; Figure 37).

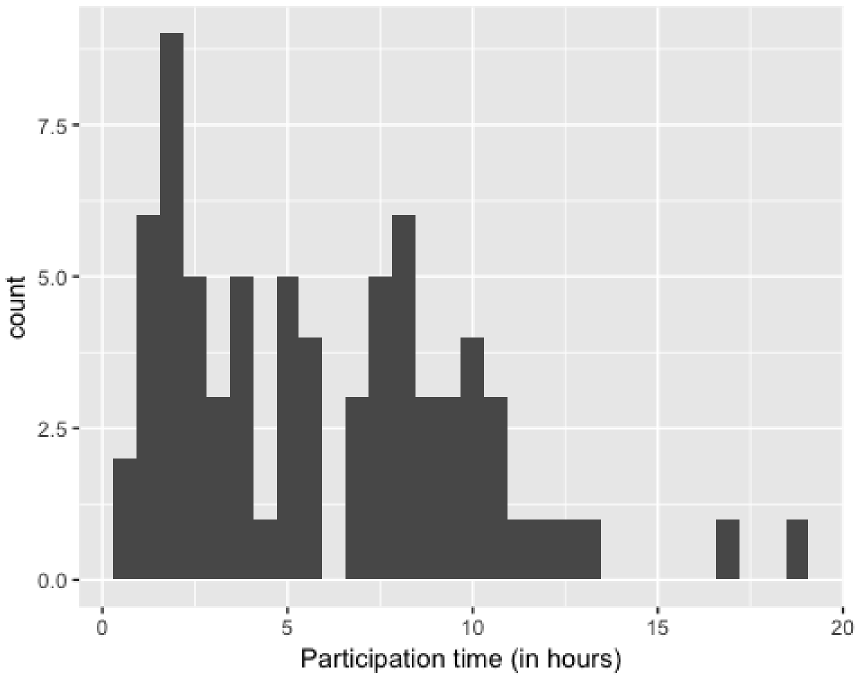


Figure 37. Empirical distribution of study participation time, n = 92.

Personality Scores

Following the recommendation of Srivastava, John, Gosling, and Potter (2003), factor scores for the Big Five Inventory (BFI) scale were computed as percentage of maximum possible. That is, scores for each factor were transformed to a 0 to 100 range by subtracting the factor score by 1 and multiplying by 25. The empirical distributions of factor scores were approximately normal and no outliers were identified (Figure 38).

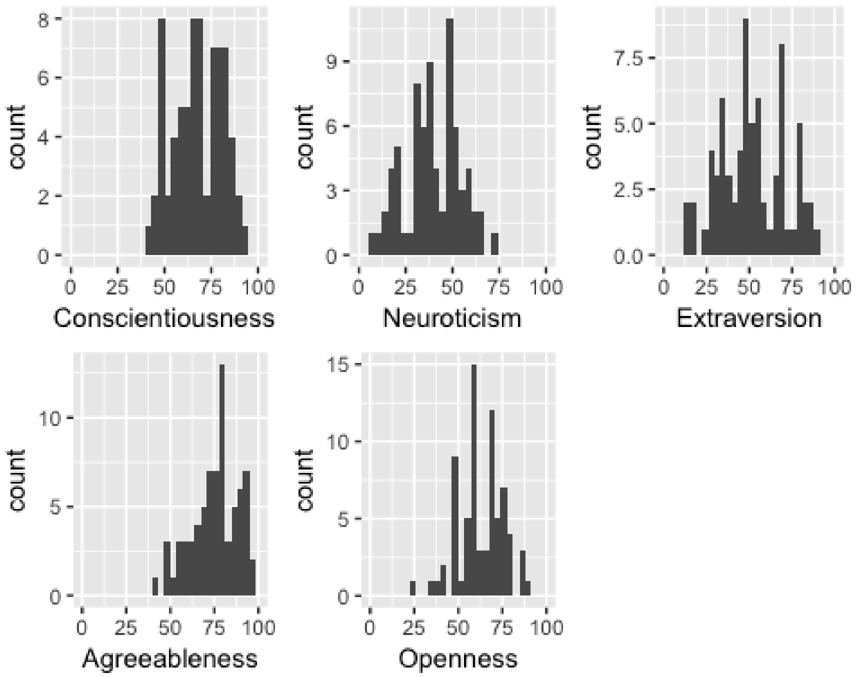


Figure 38. Empirical distributions of Big Five Inventory factor scores, transformed to percentage of maximum possible scoring, n = 92.

Given the study population, personality factor scores were predictably non-representative of a general population. In the present sample, means and standard deviations were: Conscientiousness *M* = 67.9, *SD* = 12.9; Agreeableness *M* = 75.2, *SD* = 13.1; Neuroticism *M* = 39.7, *SD* = 14.8; Openness *M* = 62.9, *SD* = 13.1; Extraversion *M* = 51.9, *SD* = 19.3. Compared to results reported by Srivastava et al., the present sample was more conscientious, more agreeable, less neurotic, and less open, but there was no difference in extraversion (all *t*(72.11), *p* = 0.008, 0, 0, 0, and 0.232, respectively). However, in the present sample, there was no evidence of a difference between genders for any factor except for Neuroticism; women were more neurotic than men (*Mdiff* = 10.76 [3.04, 18.47], *t(28.35)* = 2.85, *p* = 0.008).

Emotion Responses

Empirical distributions of the number of emotion responses approximated negative binomial distributions (Figure 39). Joy was the most frequent response and Sad was the least. The sample means and standard deviations were: Anger *M* = 8, *SD* = 9.9; Fear *M* = 11.1, *SD* = 13.3; Joy *M* = 30.3, *SD* = 26; Neutral *M* = 5.3, *SD* = 7.1; Sad *M* = 2.9, *SD* = 4. Given high variability in participation time, the rates of emotion response are more comparable than the counts. The means and standard deviations for rate of emotion response per hour were: Anger *M* = 1.4, *SD* = 1.5; Fear *M* = 2, *SD* = 2.3; Joy *M* = 5.6, *SD* = 4.5; Neutral *M* = 0.9, *SD* = 1; Sad *M* = 0.5, *SD* = 0.6.

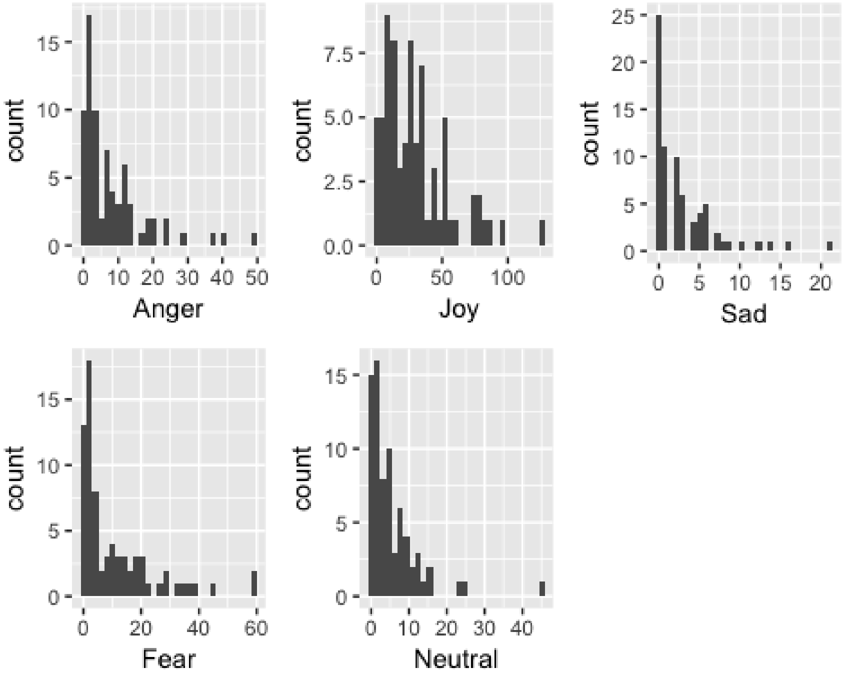


Figure 39. Empirical distributions of emotion response counts.

Correlations between variables of interest

See Table 11 for correlations (Kendall's tau) between variables of interest. Note that emotion response rates were used for the table, not emotion response counts. Women were more neurotic than men (τ = .26 [.11, .40]). As expected, due to single mode (individual) measurement of BFI, some of the personality factors were correlated. Agreeableness was negatively correlated with Neuroticism (τ = -.21 [-.36, -.06]) and Extraversion was positively correlated with Openness (τ = .21 [.05, .35]).

Table 11. Correlations (Kendall's tau) between variables of interest. Absolution values greater than .15 are statistically different from 0 (p<.05), noted with *.

	1	2	3	4	5	6	7	8	9	10	11
1 Female											
2 Participation time	.07										
3 Conscientiousness	.10	.02									
4 Agreeableness	.01	-.10	.11								
5 Neuroticism	.26*	.01	-.05	-.20*							
6 Openness	.08	.09	.09	.10	.08						
7 Extraversion	.03	.08	.09	.09	.00	.21					
8 Anger rate	.03	.05	.01	-.10	.04	-.08	.05				
9 Fear rate	.05	.00	.10	-.09	.13	.04	.03	.38*			
10 Joy rate	.04	-.06	-.05	-.06	.06	-.02	.04	.33*	.36*		
11 Neutral rate	-.02	.07	-.03	-.21*	.10	-.05	-.03	.41*	.28*	.35*	
12 Sad rate	.07	.09	.03	-.18*	.32*	.15*	.02	.18*	.28*	.29*	.31*

There was no evidence of differential emotion response rate by gender, however, some personality factors did predict different emotion response rates. Agreeableness was negatively correlated with Neutral and Sad emotion response rates (τ = -.21 [-.36, -.06] and -.18 [-.32, -.02], respectively). Neuroticism was positively correlated with Sad emotion response rate (τ = .32 [.18, .45]). Positive correlations between emotion response rates are comparably moderate in degree and likely indicate variability in overall responsiveness between participants.

Predicting emotion response by personality

To better understand how personality relates to emotion response, general linear models were fit to the response count for each emotion as a function of personality factors and gender. The outcome variables of interest are discrete counts, however; there is no evidence that the mean and variance for each emotion response count are equal. Therefore, negative binomial general linear models were fit to the data. Because of the variability in participation time, an offset component of the log of participation time was included to account for variable exposure. Models were fit using the R language and environment for statistical computing (R Core Team, 2017) and the glm.nb function from the "MASS" package (Venables & Ripley, 2002). Because the primary objective of the present analysis is to predict emotion response rate as a function of personality, an exhaustive model selection process was implemented. Models were fit based on every linear combination of predictors. The model of best fit was then chosen by determining the model with the lowest Akaike Information Criterion (AIC), (Hu, 2007). Given moderate correlation between some of the predictor variables (multicollinearity), statistical significance of predictor coefficients is not informative of the predictive value of an individual predictor. Note that the observed multicollinearity may limit inference of predictive information for any individual predictors.

Anger

The best model for prediction of Anger response rate (that is, the model with lowest AIC) included a single predictor, Agreeableness. As indicated by residual deviance, the model was a modest fit to the data ($\chi^2(71) = 82.97$, $p = 0.157$). An index-deviance plot did not indicate any extreme instances of poor model fit (see Figure 40 for index-deviance plots for all models). However, the standard error for the predictor coefficient was too large to infer the magnitude or sign of the coefficient.

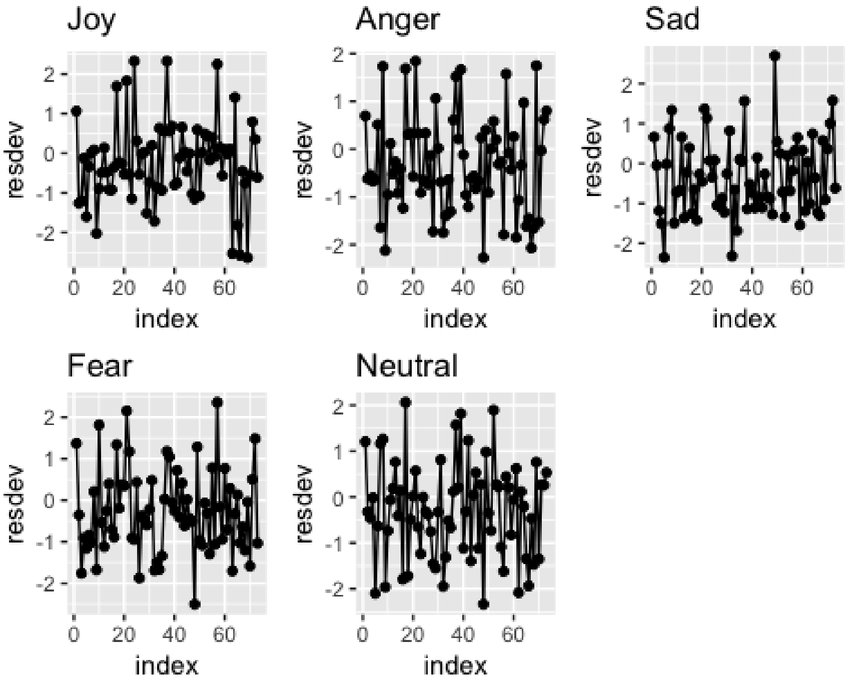


Figure 40. Index-deviance plots for models of best fit for each emotion response rate.

Fear

The best model for prediction of Fear response rate included Conscientiousness, Agreeableness, and Neuroticism as predictors. As indicated by residual deviance, the model was a modest fit to the data ($\chi^2(69) = 79.82$, $p = 0.18$) and an index-deviance plot did not indicate any extreme instances of poor model fit. However, the standard errors for the predictor coefficients were too large to infer the magnitude or sign of any of the coefficients.

Joy

The best model for prediction of Joy response rate included Neuroticism as the single predictor. As indicated by residual deviance, the model was a modest fit to the data ($\chi^2(71) = 81.55$, $p = 0.18$) and an index-deviance plot did not indicate any extreme instances of poor model fit. However, the standard error for the predictor coefficient was too large to infer the magnitude or sign of the coefficient.

Neutral

The best model for prediction of Neutral response rate included Agreeableness as the single predictor. As indicated by residual deviance, the model was a modest fit to the data ($\chi^2(71) = 83.3$, $p = 0.15$) and an index-deviance plot did not indicate any extreme instances of poor model fit. A positive unit difference in Agreeableness predicted a Neutral response rate 0.98 [0.96, 0.99] lower.

Sad

The best model for prediction of Sad response rate included Agreeableness and Neuroticism as predictors. As indicated by residual deviance, the model was a modest fit to the data ($\chi^2(70) = 75.33$, $p = 0.31$) and an index-deviance plot did not indicate any extreme instances of poor model fit. Given the same Agreeableness, a positive unit difference in Neuroticism predicted a Sad response rate 1.03 [1.01, 1.05] higher. The standard error for the predictor coefficient for Agreeableness was too large to infer the magnitude or sign of the coefficient.

Discussion

In this work, personality data was used to discern if reported emotions were truly different when collected in real-world scenarios over an extended collection period. While not all general population differences were found, some predicted findings did occur within this sample.

Personality scores

Given the study population, personality factor scores were predictably non-representative of a general population. This is due to the sample being employees at a large technology company which are not representative of the general population. According to the 2016 Diversity in High Tech report, relative to private industry, the high tech sector employed a larger share of whites (63.5% industry to 68.5% tech), Asian Americans (5.8% industry to 14% tech) and men (52% industry to 64% tech), and a smaller share of African Americans (14.4% industry to 7.4% tech), Hispanics (13.9% industry to 8% tech), and women (48% industry to 36% tech), (U.S. Equal Employment Opportunity Commission, 2016). In addition to these differences, the expected gender outcome of higher Neuroticism in women held, while no other factors did.

This sample was more Conscientious, more Agreeable, less Neurotic, and had lower Openness than the general population results but that is not surprising for a sample that is generally more highly educated and has been habituated to work within a positive, social environment. Also, as females are more highly associated with Neuroticism and this sample had a lower female representation, 21% (19/92 participants were female) than the general high tech industry, this result makes sense.

Emotion responses

Joy was the highest reported emotion both by count and in rate. This again, makes sense, given the sample population being employees who are generally in a positive work environment and are less likely to report negative emotions in social settings. Unfortunately, though, this means that the other emotions that were reported were underrepresented relative to joy. For the purposes of this exploratory analysis though, there was enough variance to identify models which have a reasonable fit.

Correlations between variables of interest

First, as expected, some of the BFI factors were correlated. This was expected due to the single-rater (self-report) data collection mechanism for personality data. The best way to attain orthogonal facts within the BFI is to collect personality data through multiple methods. That said, the negative correlation of Agreeableness with Neuroticism and positive correlation of Extraversion with Openness is not particularly surprising given the general valence of each of those factors.

Agreeableness was negatively correlated with Neutral and Sad emotion response rates while Neuroticism was positively correlated with Sad emotion response rates. Both factors make sense as Agreeableness is generally positive and Neuroticism is negative, reporting higher negative emotions (sad here).

Predicting emotion response by personality

Of the five fuzzy emotions, only Neutral and Sad displayed a meaningful difference based on personality factors. More agreeable individuals were predicted to report slightly fewer neutral responses, which on the surface makes sense as agreeable individuals should be less likely to be neutral and more likely to identify a positive agreement characteristic.

Also, given the same Agreeableness, higher Neuroticism should indicate a higher Sad response rate. This finding supports one of the two major points of the emotion-personality investigation where Neuroticism is associated with negative emotion.

Conclusions

Summary

While there was no support for the previously established link between Extraversion and positive emotion, there was evidence to support Neuroticism with negative emotion. In this study, it was an emotion within the Sad category of reported emotions.

We know there was indication of differential response by gender with respect to women reporting with Neuroticism than men. That aligns with previous research and the reduced effect aligns with the reduced female representation within the sample.

Additionally, the lower neutral response rates for high Agreeableness individuals and higher negative (sad) response rates for high Neuroticism individuals makes sense.

While the sample size for this exploratory analysis on personality and emotion was relatively low, the novelty of measuring emotion over time, with focus on a single emotion at once provides some additional value to these models. The researchers believe that this exploration of emotion and personality is enough to warrant further investigation into longitudinal data collection with not only emotion, but also a more robust and multiple-method channel of personality assessment.

Limitations

This study was restricted by the single-rater, self-report personality measures. To take full advantage of the existing personality literature, multi-method personality ratings should have been used. Additional limitations include the poor general population representation. The limited number of female participants reduces the certainty we have in any findings we may have wanted to present with females.

Future directions

Future emotion and personality work will not only attempt to gain personality data via multi-method channels, but also will collect physiological data to measure emotion in an objective manner. Physiological data offers a potential view past the known, yet highly variable, biases self-report presents. A controlled lab study with an emotion elicitation protocol may be appropriate to explore the physiological responses in addition to multi-method personality data that will be collected. The goal would then be to gain better estimates of personality using multiple methods so that personality factors aren't correlated. This would allow significant predictors to indicate real differences in personality and emotion responses.

Additionally, as this work captured emotional data over time, future studies will attempt to also gather personality data over time and control the amount of time spent recording emotions from a more representative sample.

References

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*, 171–220. http://doi.org/10.1037/h0093360

Anusic, I., Schimmack, U., Pinkus, R. T., & Lockwood, P. (2009). The nature and structure of correlations among Big Five ratings: the halo-alpha-beta model. *Journal of Personality and Social Psychology*, *97*(6), 1142–56. http://doi.org/10.1037/a0017159

Arnold, M. (1960). *Emotion and Personality*.

Baumgarten, F. (1933). Die Charaktereigenschaften. *Beitraege Zur Charakter Und Persoenlichkeitsforschung*.

Cacioppo, J. T., Berntson, G., Larsen, J., Poehlmann, K. M., & Ito, T. (2000). The Psychophysiology of Emotion. In R. Lewis & J. M. Haviland-Jones (Eds.), *The Handbook of Emotion* (2nd ed., pp. 173–191). Guilford Press.

Cattell, R. B. (1943). The description of personality: basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, *38*(4), 476–506. http://doi.org/10.1037/h0054116

Cattell, R. B. (1945a). The Description of Personality: Principles and Findings in a Factor Analysis. *The American Journal of Psychology*, *58*(1), 69–90. http://doi.org/10.1177/036354657800600202

Cattell, R. B. (1945b). The Principal Trait Clusters for Describing Personality. *Psychological Bulletin*, *42*(3), 129–161. http://doi.org/10.1037/h0060679

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the sixteen personality factor questionnaire (16 PF) in clinical, educational, industrial, and research psychology, for use with all forms of the test,*. Institute for Personality and Ability Testing.

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: happy and unhappy people. *Journal of Personality and Social Psychology*, *38*(4), 668–678. http://doi.org/10.1037/0022-3514.38.4.668

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. *Journal of Personality*. ISHK. http://doi.org/10.1163/1574-9347_bnp_e804940

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science (New York, N.Y.)*, *221*(4616), 1208–1210. http://doi.org/10.1126/science.6612338

Ellsworth, P. C., & Scherer, K. R. (2003), Ellsworth, P., & Scherer, K. (2003). Appraisal processes in emotion. *Handbook of Affective Sciences*. http://doi.org/2009-07773-029

Essa, I. a., & Pentland, a. (1994). A vision system for observing and extracting facial action\nparameters. *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on*, *1994*(247), 76–83. http://doi.org/10.1109/CVPR.1994.323813

Feldman Barrett, L. (1996). Hedonic tone, perceived arousal, and Item Desirability : Three components of self-reported mood. *Cognition and Emotion*, *10*(1), 47–68. http://doi.org/10.1080/026999396380385

Fiske, D. W. (1949). Consistency of the Factorial Structure of Personality Ratings From Different Sources. *Journal of Abnormal and Social Psychology*, *44*(3), 329–344. http://doi.org/10.1037/h0057198

Fossum, T. A., & Barrett, L. F. (2000). Distinguishing Evaluation From Description in the Personality-Emotion Relationship. *Personality and Social Psychology Bulletin*, *26*(6), 669–678. http://doi.org/10.1177/0146167200268003

Frijda, N. (1986). *The Emotions*. Cambridge University Press.

Georgi, R. Von, Grant, P., Georgi, S. Von, & Gebhardt, S. (2006). *Personality, emotion and the use of music in everyday life: Measurement, theory and neurophysiological aspects of a missing link - First studies with the IAAM -. Buch*.

Goldberg, L. R. (1981). Language and individual differences: the search for universals in personality lexicons. *Review of Personality and Social Psychology*, *2*(1), 142–165. http://doi.org/10.1037/0022-3514.59.6.1216

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. http://doi.org/10.1037/0022-3514.59.6.1216

Graesser, A. (2009). Deep learning and emotion in serious games. In *Serious games: Mechanisms and effects* (pp. 81–100).

Hascher, T. (2010). Learning and emotion: Perspectives for theory and research. *European Educational Research Journal*, *9*(1), 13–28. http://doi.org/10.2304/eerj.2010.9.1.13

Healey, J. A. (2000). *Wearable and automotive systems for affect recognition from physiology*. MIT.

Hu, S. (2007). Akaike Information Criterion. Raleigh, NC: Center for Research in Scientific Computation.

John, O. P., Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*.

Klages, L. (1926). The Science of Character (Translated 1932).

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, *24*(6), 615–625. http://doi.org/10.1016/j.imavis.2005.09.011

McCrae, R. R., & Costa, P. T. (1991). Adding Liebe und Arbeit: The full Five-Factor Model and Well-Being. *Personality and Social Psychology Bulletin*, *17*(2), 227–232. http://doi.org/10.1177/014616729101700217

McCrae, R. R., & Costa, P. T. (1999). No. 349 - The Five-Factor Theory of personality.pdf. *Handbook of Personality: Theory and Research*, 139–153.

McNiel, J. M., & Fleeson, W. (2006). The causal effects of extraversion on positive affect and neuroticism on negative affect: Manipulating state extraversion and state neuroticism in an experimental approach. *Journal of Research in Personality*, *40*(5), 529–550. http://doi.org/10.1016/j.jrp.2005.05.003

Moors, A. (2014). Flavors of appraisal theories of emotion. *Emotion Review*, *6*(4), 303–307. http://doi.org/10.1177/1754073914534477

Morgan, R. L., & Heise, D. (1988). Structure of emotions. *Social Psychology Quarterly*, *51*(1), 19–31. http://doi.org/10.1016/0191-8869(88)90141-9

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. *Journal of Personality and Social Psychology*, *7*(4p2), 1–18. http://doi.org/10.1037/h0025230

Picard, R. W. (1997). *Affective Computing*. Cambridge: MIT Press.

Picard, R. W. (2010). Emotion Research by the People, for the People. *Emotion Review*, *2*(3), 250–254. http://doi.org/10.1177/1754073910364256

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191. http://doi.org/10.1109/34.954607

Plutchik, R. (1984). Emotions: A General Psychoevolutionary Theory. In K. R. Scherer & P. Ekman (Eds.), *Approaches to Emotion*. New York: Psychology Press.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. http://doi.org/10.1037/h0077714

Rusting, C. L., & Larsen, R. J. (1997). Extraversion, neuroticism, and susceptibility to positive and negative affect: A test of two theoretical models. *Personality and Individual Differences*, *22*(5), 607–612. http://doi.org/10.1016/S0191-8869(96)00246-2

Saucier, G. (1994). Separating Description and Evaluation in the Structure of Personality Attributes. *Journal of Personality and Social Psychology*, *66*(1), 141–154. http://doi.org/10.1037/0022-3514.66.1.141

Saucier, G. (1997). Effects of variable selection on the factor structure of person descriptors. *Journal of Personality and Social Psychology*, *73*(6), 1296–1312. http://doi.org/10.1037/0022-3514.73.6.1296

Scherer, K. R. (1999). Appraisal theory. *Handbook of Cognition and Emotion*. http://doi.org/10.1002/0470013494.ch30

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, *84*(5), 1041–1053. http://doi.org/10.1037/0022-3514.84.5.1041

Stowell, J. R., & Nelson, J. M. (2007). Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion. *Teaching of Psychology*, *34*, 253–258. http://doi.org/10.1080/00986280701700391

Tupes, E. ., & Christal, R. . (1961). *Recurrnt Personality Factors Based on Trait Ratings*.

U.S. Equal Employment Opportunity Commission. (2016). *DIVERSITY IN HIGH TECH*.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer.

Wang, L., Shi, Z., & Li, H. (2009). Neuroticism, extraversion, emotion regulation, negative affect and positive affect: The mediating roles of reappraisal and suppression. *Social Behavior and Personality: An International Journal*, *37*(2), 193–194. http://doi.org/10.2224/sbp.2009.37.2.193

Watson, D., & A, C. L. (1992). On traits and temperament: General and specific factors of emotional experience and their relationship to the five-factor model. *Journal of Personality and Clinical Studies*, *60*(June 1992), 441–476.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. http://doi.org/10.1037/0022-3514.54.6.1063

Zuckerman, M., & Lubin, B. (1965). The Multiple Affect Adjective Checklist (MAACL). *Educational and Industrial Testing Service (EdITS)*.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

Conclusions

This work discussed how psychophysiological measures can be used within user research scenarios to improve the quality and quantity of feedback gained from people in various user research scenarios.

Chapter 2 and Chapter 3 discussed how electrodermal activity can be used within a high fidelity combine simulator to help understand operator workload through a variety of harvest scenarios. While the combine simulator is a unique platform for conducting user research, the challenge that it attempts to address is one that exists across a variety of research scenarios. That challenge is, at its core, how do we as researchers better understand a population that is not well understood, difficult to observe, and relatively small? Farmers in this case are a relatively small group of individuals who are not representative of the general population, they work in a time sensitive manner with high amounts of cognitive load, and there is not much documentation on operator behavior in general and even less when it comes to how they interact with new technologies. By incorporating psychophysiological measures, such as electrodermal activity, we as researchers are attempting to allow that person to give us more information than they normally would be able to without impeding their process or performance. A key component of psychophysiological measures within this work was the cost of gaining that insight to the participant. The measures here were chosen as they were relatively low-cost, did not require excessive prep or clean-up, and did not require the participants themselves to perform any additional work. Asking the participant to take on additional work is counter-intuitive to the mission of observing people in the most natural setting. Electrodermal activity then serves as a proxy measure for mental workload and by observing this measure in conjunction with the task difficulty and their performance, a clearer image of their overall workload emerges.

Chapters 4 and 5 dealt with a different, but related topic of user research, emotion. One aspect of user research is to have empathy for the user. It is our role to best understand those that utilize the products and services we build and communicate their feelings to others who may otherwise not be aware. Those feelings that we can observe, the external affective state of another person, is one way to get that information. Another is through evaluating emotion based on the behaviors of others (e.g., was there a high abandonment rate on a particular step of a complicated process?). The study which Chapters 4 and 5 are built on attempted to, again, use psychophysiology as a method of objective understanding without having to rely on those we are observing to report their own feelings to us. Not only is the process of producing an affective state not agreed upon (Does affect happen post cognition? Simultaneously? Before?) emotion itself covers all areas from the initial feelings to the expressions, outward or quiet internal expressions, people experience. By understanding what emotions someone is experiencing we can better understand what their struggles and delights are, which in turn will help us create a better experience for them. Chapter 4 attempted to cover this by building an emotion detection model which could be used in normal, day-to-day life, using a psychophysiological sensor which was commercially available. The idea of passively being able to track emotions either solely for the person tracking them self or for specific research purposes both have very compelling arguments. While the hardware available is not quite yet sensitive enough to measure and discriminate all the noise from the actual signal with respect to EDA, the concept of real time emotion detection is still promising and pending improved hardware, should be a service offered relatively soon. Chapter 5 explored a separate emotion issue, one of understanding how personality influences emotion. The research tying personality and emotion together is well founded and has been explored in a variety of scenarios. Yet, in an effort to capture emotion in the least intrusive, most natural way this effort attempted to understand how emotion was reported through the lens of personality. Even though more data need to be collected, the first results warrant additional investigation, as at least one of the typical emotion-personality connections was replicated (increased neuroticism associated with increased sadness). By recording emotion outside of the lab, with naturally occurring emotions and a large number of people across two months, this work attempted to capture emotion in a different scenario than many others. Both Chapters 4 and 5, then, ultimately attempt to better understand humans through the lens of emotion. Both the model of emotion via external sensors and considerations of personality help to better understand others.

Ultimately, increased understanding of people helps those that want to design products and services for them and can help those same people better understand themselves. While the mechanisms to this understanding can be increasingly complex and even subtle, there is an opportunity to better serve others while also improving our own practice. While psychophysiological measures such as electrodermal activity and heart rate may not be the single method to gain all insight, just as emotion is not the single factor to consider when communicating with others, both psychophysiology and emotion have their place and are tools on the researcher toolbelt that serve a specific purpose and have the potential to improve this discipline overall. Psychophysiology is currently most useful to measure participant states such as mental effort and emotion (in their respective experimental designs) in scenarios where traditional measures may not be suitable or possible. Also, though, psychophysiology is not meant to be a one-size-fits-all solution that can inform multiple measures simultaneously without regard for the scenario itself. Understanding when and where these tools are best applied is part of the craft we can all work at improving for ourselves and those we aspire to understand.

Research Questions

The answers to the guiding research questions are summarized here.

Chapter 2: *How well does electrodermal activity reflect mental effort in an agricultural equipment simulator?* EDA was successfully measured in the combine simulator and was observed to 1) decrease as satisfaction increased and 2) decrease with higher knowledge operators. EDA positively correlates with mental effort, therefore the observed increases in both satisfaction and operator knowledge and lower EDA levels supports EDA as a proxy for mental effort in the combine simulator.

Chapter 3: *How much fidelity is required to represent the desired cue within the simulator?* EDA was observed to 1) decrease as interactions increased and 2) decrease with the number of correct actions taken. Although fidelity can be measured in different ways, the EDA findings from this study served as a successful measure of cue fidelity in addition to the observed behaviors in the combine simulator.

Chapter 4: *Can emotions be measured in real-time using everyday technology?* EDA and heart rate were used to successfully measure emotion with 51.5% accuracy with the data training set. When tested with participants in real time, accuracy was reported by the participants to be 76.4%.

Chapter 5: *Does personality predict emotion when observed with an extended data collection process?* While there were no models produced that would reasonably predict anger, fear, or joy, one model produced decreased neutral responses given increased agreeableness scores, and one other that was produced increased sad responses given increased neuroticism scores when agreeableness was held constant. Thus, personality could be seen to partially predict emotion, but further research is needed to create a reliable predictive model.

Future Work

Based on the two primary areas of this work, the first future study to investigate would be the continued implementation of psychophysiological measures within user research scenarios. Electrodermal activity was primarily used in this work, but other measures such as heart rate variability (HRV), blink rate, pupillometry, or general eye tracking are becoming more accessible without forcing people to feel intruded upon or take on additional work during their normal process. The approach taken in the work discussed here could be modified and replicated for other measures in a variety of user research scenarios. Implementing these other measures will give additional insight into workload and other internal states that are otherwise difficult to communicate.

For the work concerning emotion, next steps are more specific. The follow-up study should focus on a more controlled environment to build an emotion detection model for. This work attempted to use low-cost sensors while also collecting data in the real-world.

Instead of adding unnecessary noise, these steps should be taken one at a time to improve the model accuracy. With respect to personality, the next study should attempt to normalize the amount of time spent recording emotions, address the discrepancy between emotions submitted, and collect personality information from more than just the single recorder (self-report) method. Additionally, ties between personality and physiological data should be investigated to understand what type of connection, if any, they have on emotion together. All improved emotion recognition for technology helps improve the general state of affective computing, which is something human computer interaction topics in general will have to continue to improve in with the goal of improving the relationship between us and the technology we use.

Closing

As I was sitting with my nephew and daughter watching a cartoon in the summer of 2016, my phone chimed the ICQ notification sound, "uh-oh!" to alert me that I had received a new email. To me the notification is a reminder of past years when I used ICQ as my primary messaging client to talk with friends while playing video games on the internet while also alerting me that an email had been delivered. To my nephew (just over 3 years old) it was a somewhat unsettling experience. He immediately turned to me and earnestly asked "What was that?!" To which I explained "That was just my phone, that sound means I have a new message. Like, ‘Uh-oh!’ someone wants to talk to you!" He looked at me, quizzically, unsatisfied, and said "Oh…ok." The entire exchange happened in less than 10 seconds, but spoke volumes. To me, the "uh-oh!" just means I have an email, but to him my phone was sending out a distress beacon, he had heard the literal “uh-oh!” This exchange showed me two separate concepts from the view of a 3-year-old who has grown up with technology in hand. 1) Emotion can vary in meaning from technology, and the variance can be learned, just as it is from people. In the same way that a sibling may sarcastically say "very funny" to mean not funny at all, I understand the "uh-oh!" to be less severe than literal meaning. 2) My nephew was willing, without hesitation or thought, to accept that my phone was emoting a genuine emotional statement of warning and surprise as opposed to just a sound with affective ambivalence. Overall, an interaction like this helps to illustrate that people readily anthropomorphize their technology and are also willing to accept emotional statements from it. It is this acceptance of emotion from technology that will help usher in the next steps of affective computing. We currently assign personality traits, genders, age, etc. to our digital assistants and it will not be long before we are thinking of more technology in that manner.