Designing for Deep Engagement

by

David Bradford Ramsay

M.A., Massachusetts Institute of Technology (2016)
B.S., B.A. Case Western Reserve University (2010)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Authored by_____

David Bradford Ramsay
Program in Media Arts and Sciences
August 18, 2023

Certified by_____

Joseph A. Paradiso
Alexander W Dreyfoos Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by_____

Joseph A. Paradiso
Academic Head
Program in Media Arts and Sciences

Designing for Deep Engagement

by

David Bradford Ramsay

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 18, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Flow state represents the quality of meaningful experience– an effortless, depth of attention that is often undermined in our interrupt-driven, modern society. In this thesis, I present four novel interventions to promote states of deep engagement.

Evaluating whether one of these interventions has a meaningful impact on flow state is difficult to do. The bulk of my work, then, focuses on the methodological challenges of flow state research. Herein I tackle three weaknesses in our ability to make strong, generalizable predictions about the causal link between environmental stimuli and flow states: (1) I discuss advancing how we represent the environment (specifically for aural stimuli) using phenomenological principles; (2) I advance the state-of-the-art in how we represent and measure flow bio-behaviorally (with the goal of integrating physiology into our judgements); and (3) I evaluate methodological weaknesses in current experimental flow work. To do this, I present experimental work on models of auditory attention, new wearables and survey instruments for flow estimation, and an experiment that compares flow as measured in lab and at home across varying task structures.

This thesis contributes a suite of state-of-the-art psychophysiological and behavioral hardware tools designed to inform inference about flow in-the-wild; it also contributes two unique, open-source, naturalistic datasets collected with them. Combined with time-aware, probabilistic representations of cognition, this work sets the stage for a precise and explicit bio-behavioral definition of flow states that will improve our ability to understand its relationship to our environment. In so doing, it points to an improved approach for social psychology more generally.

Thesis Advisor:
Joseph A. Paradiso
Alexander W Dreyfoos Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Designing for Deep Engagement

by

David Bradford Ramsay

The following people served as readers for this thesis:

Thesis Reader

Rosalind W. Picard
Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Thesis Reader

Gloria Mark
Professor of Informatics
University of California, Irvine

Thesis Reader

Jan-Willem van de Meent
Associate Professor of Computer Science
University of Amsterdam

Contents

List of Figures

List of Tables

Acknowledgements

This thesis represents many amazing years and many important relationships in my life. I'm filled with gratitude when I look back on my time at the Media Lab.

I want to start by acknowledging the people that formed my support system outside of the lab that kept me grounded, had my back, and moved me forward. My roommates and close friends who I would see every week of this experience– Bernd Huber, Nathan Monroe, Sam Hendel, Dustin Gallegos, Nathan Dionne, and Dylan Robinson. My sister Tracy and my Mom, who have always been there to answer the phone and bring levity to my life. My Dad and biggest supporter, who religiously tuned into my twitch experiment every day. My wonderful girlfriend Kara, who actually agreed to part of my twitch experiment. You're all the reason I'm here and I love you all.

Second, I'd like to acknowledge my MIT family. Within my group and at the lab, I have made so many wonderful friendships. Within Resenv, Brian, Gershon, Mark, Spencer, Nan, Jie, Artem– thank you for welcoming me into the group and inspiring me and teaching me. Caroline, Ariel, Irmandy, Evan, Asaph, DDH, Fangzheng, Ali, Cedric, Sam, Nathan, Devora, Carlos, Elena, Amna– you made Resenv feel like a family and I'll miss you.

I started in our group with Juliana Cherston, with whom I've now shared more years in the same place than I've ever shared with anyone. You've been a great lab partner, friend, and support to me throughout these years. Also in those first years, I worked on 'Mindsprout' in Bill Aulet's Entreprenuership class. Though it didn't bring financial success, it did bring lasting friendship– Chetan, Will, and Kristy, you defined my first years at MIT and I'm grateful for our time together.

Over the years, I've collaborated closely with two incredible people– Ishwarya Ananthabhotla and Patrick Chwalek. I feel incredibly lucky to have spent so much time working alongside two of the most talented and hardest working people I know. They are great engineers, but even better friends.

Ishwarya– thank you for your structure, support, and friendship over all of these years. You're so talented and hard working, and you continue to be the voice of reason whom I trust and rely on.

Patrick– we've traveled all over and seen each other through a lot (and we made a pair of glasses). You'll always be like a brother to me and I'll miss coming into lab to do work and accidentally talking to you for 2 hours.

Outside of Resenv, the lab is full of people that have made my time so joyful– people like Matt Groh and Noah Jones, who co-taught the Replication Crisis class with me and are great friends. Rob Lewis, Zivvy Epstein, Dan Calacci, Javi Aguera, Fluid Interfaces (esp. Angela, Vald, Judith, Abhi, Camilo, Joanne, and Guillermo), my incoming Master's class (esp. Natasha, Saquib, Bianca, and Cameron) and the others that made the lab what it was before me– Nadya Peek, Ben Bloomberg, Rebecca Kleinberger, Scott Greenwald. I also

really appreciate the faculty that have gotten to know me and cheered me along the way, especially Tod Machover, Pattie Maes, and Ethan Zuckerman.

I'd like to thank Neha for mentoring me during my stint with DCI; Jeff Nichols/Gabe Shine and Kevin Kilgour/Matt Sharifi, my supervisors at Google, and the amazing folks I met in my summers there (Noshin, Ranjay, Ivo, Paul, Josep, Minttu). I'd like to thank the EC House Team in all its forms, especially Sandy Alexandre, Alexandria Clyburn, Carrie Wicks, Rob Miller, and the Medinas. I'd like to thank EC residents and the 3W pizens for years of joy and growth and excitement– EC is a sacred space and it was an honor to be a GRT for so many years. I'd also like to thank my little Terrell for all the fun memories.

To Dan, my original mentor at Bose, who has kept up with me and become a part of the lab himself– I've learned so much from you over the years, and I'm grateful that you have remained a constant force and a friend throughout my time at the Media Lab.

To Gloria, thank you for your guidance in the final stretch of this PhD. Your work has always been an inspiration to me. I was thrilled to have your insight, and even more thrilled when I realized how kind and wonderful you are as a mentor. Thank you.

JW. the first time I met you, you spent three hours talking with me and teaching me when we were scheduled for half an hour. You are not only one of the sharpest and most gifted educators I've ever met, you are one of the kindest. You've walked me through SVI and career decisions with equal rigor, and I've felt more confident in my path knowing that you were guiding me. Thank you for everything.

Bunnie. I've learned so much from you. More than anyone I know, I admire your ethical compass and your innovative way of making a difference in the world. You're a true hacker, and rightly held in high regard by everyone I know. I hope to be a lot like you. I learned a lot about engineering from you, but even more about living.

Roz. You have been an incredible guide and mentor to me throughout my time at the lab, and I'm so grateful for the role you've played in my life. I admire so much about how you live your life, how you take care of your students, and how you make a difference through your research. Thank you for supporting me along the way, I hope I can follow the example you've set as a professor in the way I mentor, teach, and research.

Joe. It's hard to know what to say. You took a chance on me nine years ago and it completely changed my life. This page is full of the people and experiences that have shaped me into who I am, touched my life, brought me joy, and sharpened me into who I am now. All of that I owe to you and I am so grateful.

I had no idea what I was getting into when I showed up to Resenv nine years ago. I can't believe how lucky I've been to have you as my advisor, mentor, and friend. It is obvious to everyone at the lab how you selflessly take care of your students and how you enable them to pursue their passions; I've felt totally supported since my very first day. Thank you for everything. It's been the best gift I've ever received to be part of your group, which is full of people that mirror their advisor– hard working, rigorous, and brilliant; qualities only outdone by their kindness, empathy, and humility.

Chapter 1

Introduction

This is a thesis about how our tools and environments alter our sense of meaning and fulfillment.

1.1 My Journey

I joined the Responsive Environments group in 2014 with a belief and a goal. I believed that our lives are powerfully and silently guided by subtle forces we never consciously consider; my goal was to harness the power of small changes in our environments and our world to nudge us to be fulfilled, joyful, and engaged.

1.1.1 A Belief Destroyed

My beliefs were shaped by many prevalent cultural and academic forces. Daniel Kahneman's Thinking Fast and Slow (2011) was taking the world by storm. It featured several chapters on social priming– our ability to drive behavior using subtle, unconscious influence. He wrote about this experimental work:

> When I describe priming studies to audiences, the reaction is often disbelief... disbelief is not an option. The results are not made up, nor are they statistical flukes. You have no choice but to accept that the major conclusions of these studies are true. More important, you must accept that they are true about you. [...] You do not believe that these results apply to you because they correspond to nothing in your subjective experience. But your subjective experience consists largely of the story that your System 2 tells itself about what is going on. Priming phenomena arise in System 1, and you have no conscious access to them.

Against the backdrop of this trend in experimental psychology, Richard Thaler– a behavioral economist who went on to follow Kahneman's footsteps with a 2017 Nobel Prize win– was rising to prominence with his 'Nudge Theory.' Thaler's work also suggested that tiny, trivial modifications in how you present a choice to someone would have a major impact on their decision making.

I bought heavily into the science behind these findings. Only over many years have I come to understand the flaws and limits of this worldview; the ways psychologists and behavioral economics have systemically and mistakenly misused statistics to 'prove' false things about human nature, judgement, happiness, and decision-making.

Kahneman eventually recanted. In an open letter to Yale's John Bargh in 2011, he warned about 'a trainwreck looming for the field', and would later admit "I put too much faith in underpowered studies" [287]. In turn, psychology has fallen into a replication crisis– the first major meta-analysis showed successful replications for only 36% of the findings in the top four major journals on the topic [88].

'Nudging' and behavioral economics has faced similar criticism. Vernon Smith had this to say in his Nobel Prize lecture in 2002:

> Behavioral economists have made a cottage industry of showing that the Standard Social Science Model [SSSM] assumptions seem to apply almost nowhere

> to real decisions. This is because their research program has been a deliberate search in the tails of distributions for "Identifying the ways in which behavior differs from the standard model..." (Mullainathan and Thaler 2001), a search that can only succeed. [400]

And he's right. While some nudges have made a difference, the differences are almost universally slight– the largest study to date on the topic (23 million people, 126 RCTs) shows a 1.4% difference in decisions, contradicting the 8.7% suggested in the research literature on the topic [110]. In that same paper, DellaVigna et al. showed that academic intuition on this topic is mistaken– the median academic predicted 4x the effect seen in reality, with a sizable minority reporting estimates an order of magnitude larger. Recent meta-analyses argue that the research literature, taken as a whole, has yet to provide any evidence at all for nudges once we control for publication bias [275].

I was one of those researchers with faulty intuition. I based many of my early studies on a belief that these interventions would have a real and meaningful effect that I could easily measure with 20 participants. I watch colleagues continue to ground their work on faulty principles from social science, and unwittingly perpetuate the underpowered studies and statistical malpractice that led us here. Unfortunately, while nudges of this sort may matter at the level of policy where hundreds of thousands of decisions are at stake, the odds that they matter for any one individual in their daily experience is vanishingly small.

So where does this leave those of us drawn to research and innovation in this huge area?

1.1.2 A New Goal

My goal remains to design tools and experiences that structure meaningful lives. There is no doubt in my mind that the design of our technology has had powerful, lasting effects on our cognition and our social fabric; that phones and browsers shape how we think and act. We're all aware of this– everyone I talk to has rules or habits or strategies to mediate how they interact with social media or phones– though our daily awareness may dwindle as we habituate to the patterns in our lives.

I remain committed to designing closed-loop, intelligent interventions that enhance how we live in a powerful way. The question becomes which kinds of interventions actually matter and which ones don't? What is the role of published social psychology in discerning the truth of the matter? How can we discern the truth using experimental techniques ourselves? While I retain my initial goal, my intentions have shifted towards our understanding of methodology in a few important ways.

Many people I know are aware of the replication crisis. They may understand some of the basic issues– underpowered research, p-hacking, file-drawers, etc. The more sophisticated among them understand some of the basics of meta-analysis and can list off a handful of the major replication failures. My impression, though, is that few understand the real weight of what is happening, and my first goal is to elucidate what's at stake. Yes, the misapplication of statistics is part of the problem; but as I've become an expert on this topic, I've come to realize the existential challenge facing experimental psychology as a whole.

Psychology as an experimental discipline is relatively young; it's also *much more difficult* to model, predict, or describe the complexities of human thought and behavior than any hard science. Frequently we try to understand fundamentally unobservable aspects of mental life– personality, beliefs, emotions. To do this properly is two steps beyond artificial intelligence; marrying the ability to reason with deep emotional awareness and theory of mind. Yet the brightest minds in statistical learning are focused on simpler-to-measure problems, and psychology continues to counteract a pervasive misunderstanding of statistics (80% of people teaching the statistics continue to hold incorrect beliefs about fundamental concepts [172]).

Parsing current experimental work thus requires a forensic mindset. Having done that myself, I'd like to present in this thesis my new perspective on *what* is true (Chapter 2) and *how* we get to truth in light of the crisis (Chapter 3). I will discuss the underlying tension between the Frequentist and Bayesian paradigms and its implication for our philosophy of science. This epistemological question abuts questions of artificial intelligence, where many great thinkers are critiquing and analyzing the limits of what's possible with approaches to statistical or differentiable learning that minimize inductive bias. The scientific endeavor and the goals of modeling intelligence are largely the same, and I will comment on the

parallels and shared learning between the discourse on these topics.

Having established our mistaken beliefs and laid the groundwork for a general approach to methodological rethinking, I will finish Chapter 3 by describing the implications for psychology and motivate my own research to build and quantify interventions to promote flow states. The remaining chapters will describe my work to advance the state of flow research, using it as a case study of broader methodological strategies:

Chapter 4, 5, and 6 focus on issues of ontology. In Chapter 4 I'll describe the new techniques I've developed to improve how we represent the external world and environment, focusing on auditory stimuli; in Chapter 5 I'll describe the new techniques I've developed to improve how we represent the mental states with an emphasis on flow states. Finally, in Chapter 6, I'll describe my experimental work that tests these ideas of representation and methodology.

Chapters 7 and 8 focus on naturalistic intervention design. In Chapter 7 I'll describe the interventions I've prototyped to induce a real impact on people's daily lives. In Chapter 8 I'll describe my self-experimentation to understand the impact of one of these interventions, and that study's broader implications.

Finally, in Chapter 9, I'll conclude with a brief summary of the contributions contained within this work.

Chapter 2

The Echoes of Poor Replicability

> I'm all for rigor, but I prefer other people do it. I see its importance— it's fun for some people— but I don't have the patience for it. If you looked at all my past experiments, they were always rhetorical devices. I gathered data to show how my point would be made. I used data as a point of persuasion, and I never really worried about, 'Will this replicate or will this not?'
>
> *Daryl J. Bem*
> *Engber, 2017*

Imagine that it's 2011. You're the editor of a well respected social psychology journal– the *Journal of Personality and Social Psychology (JSPS)*– and you have a big problem. You receive a paper from this professor Daryl Bem, from Cornell University. He has submitted a paper summarizing two years of work, his magnum opus. It follows all of the right methods. It's rigorously done. It looks just like all the other papers you accepted. And you really, really don't want to publish it.

That's because this paper has proved that we are capable of seeing the future– referred to as 'retro-causation' or 'psi'. It includes many different experiments that demonstrate this is a real phenomena. Daryl Bem really believes it's true.

In one such experiment, participants are presented with two curtains on a computer screen and asked to click on one, revealing what is behind it. Behind one curtain is nothing; behind the other is a picture. Over three thousand trials, he tested whether the participants would choose to see a picture of a daily object or a pornographic picture. If that picture was of benign daily objects, people chose the image 49.8% of the time; no different than chance. But if the image was a pornographic image of someone very attractive, 53% of the time people were able to choose the pornographic image (and 57% of the time for 'stimulus seekers'). That is more than random chance would predict, a statistically significant result [53]. Bem hypothesized that the motivation of seeing an attractive, erotic image was enough to encourage people to peer into the future. Accumulating the results across a series of seven studies like this, he showed that normal people are doing this with 74 billion to 1 odds.

What do you do with this paper? Well, to their credit, the journal editors held fast to objective statistical criteria and published it; however, alongside of it they wrote a small editorial:

> After a rigorous review process, involving a large set of extremely thorough reviews by distinguished experts in social cognition, we are publishing the following article by Daryl J. Bem... To some of our readers it may be both surprising and disconcerting that we have decided to publish Bem's article....We openly admit that the reported findings conflict with our own beliefs about causality and that we find them extremely puzzling. Yet, as editors we were guided by the conviction that this paper— as strange as the findings may be— should be evaluated just as any other manuscript on the basis of rigorous peer review...

The next day, it was a New York Times front page story: "One of psychology's most respected journals has agreed to publish a paper presenting what its author describes as strong evidence for extrasensory perception, the ability to sense future events. The decision is already mortifying scientists, generating a mixture of amusement and scorn."

Of course, these results have been handily disproven by meta-statistical work. And though the original work remains unretracted (as of 2023 there is still an active debate with JSPS

editors to get it removed [379])– and even though Daryl Bem has doubled down, publishing even more studies in favor of psi [52]– this paper was a postive catalyst for the paradigm shifting conversation now ongoing in social psychology. Its publication forced a broader reflection on our statistical standards– if this paper was just as rigorous as the others, which of the others are true?

Consequently, large scale replication efforts began. Kahneman penned his famous open letter about the 'trainwreck looming' for the field [287]. Devastating results started to pour in; it is now estimated that only 20% - 45% of social psychology papers will replicate [376] and prediction markets do a pretty good job of guessing which ones [77]. Moreover, incorrect results are cited 16x more a year than correct ones [387], and even retraction does not fix this [78] (With the most generous overestimate of the impact of retraction, the average non-replicated result will continue to be cited at 4 times the rate of an average replicable study).

Journal	% Replicated
Journal of Experimental Psychology: Learning, Memory, and Cognition	48
Journal of Personality and Social Psychology *(Social Psych Papers Only)*	23
Psychological Sciences *(Social Psych Papers Only)*	29
Psychological Sciences *(Cognitive Psych Papers Only)*	53
Overall	36

Figure 2-1: A reproduction of Table 1. from the Open Science Collaboration's 'Estimating the reproducibility of psychological science' [88]. Notice that Social Psychology papers specifically fare far worse than others. Based on 100 replication attempts carried out by 270 different authors.

Taken as a whole, these results have a couple earth-shattering implications: (1) within many of the top journals, 'significant' social psychology results are informative in a negative direction, worse than a coin flip– i.e. if you have no prior about a claim, you should assume it is less likely to be true than if you hadn't seen it published,[1] (2) if you've encountered the result outside of a journal, the bias in coverage makes the probability of accuracy much, much worse,[2] and (3) you may be better off asking your friends to make bets or trusting your intuition than peer-reviewed social psychology.[3]

Thus, the locus of truth for this field has shifted away from academic journals. Instead, replication-focused blogs run by statisticians are the most trustworthy source of information. Thankfully, it is possible to rescue objective insight from the accumulated literature with a healthy set of priors and a forensic attitude.

2.0.1 The Overall Impact

Though I've introduced this crisis as it has unfolded in social psychology, the same story affects positive psychology and behavioral economics as well. In fact, the crisis is incredibly broad, affecting medical research, neuroscience, and AI. Besides the overall erosion of trust in science, each field has its own misconceptions to rectify.

For social psychology, the crisis is responsible for two pervasive misconceptions:

(1) **People are utterly lacking in agency, easily manipulated, and unaware of the forces that shape their lives.** In 'The Quick Fix: Why Fad Psychology Can't

[1] If we replicate an experiment exactly, and it was done at reasonable power (80%), we would expect a successful replication 80% of the time, and 20% failure shouldn't concern us; moreover, we should take all of the evidence together to estimate a real underlying effect size. In the cases we're discussing, however, the replications are pre-registered– thus averting systemic bias in reporting– and typically done at a huge multiple of the original sample size (i.e. 10x), giving much stronger evidentiary value. While the language of 'replication' and 'failed replication' perpetuates a flawed categorical statistical mindset, it is a convenient shorthand here for the state of affairs, as many of these failed replications are actually very strong evidence for negligible effect sizes.

[2] 20-45% of these studies replicate, and non-replicable results are cited 16x more on average. We'll use 35% and 16x as our touch points. That means roughly, if we assume citation count reflects news coverage, that the odds of a study you hear about failing to replicate is roughly 0.35*16 / (0.35*16 + 0.65*1) = 90%. In reality, popular press coverage is better represented by a Pareto distribution than simple 16:1 odds and the number is probably higher.

[3] Prediction markets have seemingly fared well as a tool to separate out replicable research, enough that DARPA and Berkley have funded betting market initiatives for behavioral science.

Cure Our Social Ills', Jesse Singal labels the worldview this bad science has perpetuated 'Primeworld' [395]. This idea has many corollaries– people deserve neither condemnation nor praise because they are simply the product of the subtlety of their environment (influencing our views of justice and self-worth). People are almost entirely shaped by their environment (diminishing the role of human nature– see, i.e., Steven Pinker's 'The Blank Slate' [338]). Perhaps most perniciously, as Kahneman famously said, "you must accept that [these studies] are true about you' despite the fact that 'you have no conscious access to them" [230]. Denying these facts is not an option; to do so means you are irrational and unwilling to confront the data (rhetoric that makes you either a 'believer in science' or not– a politicizing and factionizing identity).

These ideas diminish human nature and personal responsibility in a way that is incredibly disempowering. It also justifies overly-paternalistic legislation– people are *already* puppeteered by their environments so we might as well ensure that the controlling forces are pro-social.

(2) **The way to live a happy life, should you feel unfulfilled and disempowered, is to adopt one of many small habits that are grounded in a shallow philosophy with an under-powered positive psychology study behind it**. Mainstream academia has taken over the self-help industry, famous for peddling quick fixes and simple solutions. Faith in quick fixes aligns with point (1); it relies on the assumption that small changes are powerful enough to meaningfully shape our sense of fulfillment, and treats existing religious institutions as a collection of superficial nudges themselves (rather than a gestalt, memetic, adaptive social system in conversation with human nature).

2.0.2 YourAd: My Story of Reckoning

There are many, very important implications of the above-mentioned worldview– for politics and society, for design, and for our own sense of humanity. They also have a tremendous impact on science. I'd like to give an example of how these ideas have affected my work when I started my PhD, through a project I created called 'YourAD.'

Advertising

Advertisers are often scorned for their manipulative tactics and extractive techniques. The price of diamonds is famously the result of De Beers, an effective monopolist, artificially restricting supply in the late 1800s alongside their campaigns to solidify diamonds as a symbol of love and commitment. Cigarette sales– once lagging with women in the 1920s– soared after Edward Bernays (the father of Public Relations) handed them out to suffragettes at a large women's rights march and convinced newspapers to label them 'torches of freedom.'

Fast forward to the modern world. According to *Scientific American*, 'subliminal advertising'– famous as a massive fraud perpetuated by a marketing consultant who went on to make millions off of it when it was first 'discovered'– may actually be real [409]. Robert Cialdini– the world's top academic psychologists in marketing and persuasion– wrote in the Harvard Business Review that you can control your customers' decisions by changing the wallpaper behind your products[4] [277]. He attributes this to the 'subliminal mere exposure effect'– that subconscious exposure will make us view similar concepts favorably because of increased 'perceptually fluency' (i.e. we prefer things that feel familiar or are easy to understand). He speaks to this effect (with admittedly small effect sizes) at length in his NY Times best-selling book 'Pre-suasion' [85]:

> Additional research has found similarly sly effects for online banner ads—the sort we all assume we can ignore without impact while we read. Well-executed research has shown us mistaken in this regard. While reading an online article about education, repeated exposure to a banner ad for a new brand of camera made the readers significantly more favorable to the ad when they were shown it again later. Tellingly, this effect emerged even though they couldn't recall having ever seen the ad, which had been presented to them in five-second flashes near the story material. Further, the more often the ad had appeared while they were reading the article, the more they came to like it. This last finding

[4]Not in a reasonable sense like creating a cohesive brand identity; in the sense of 'clouds in your wallpaper unconsciously suggests comfort' and thus make a customer subconsciously more likely to weigh comfort heavily in their purchasing decisions.

> deserves elaboration because it runs counter to abundant evidence that most ads experience a wear-out effect after they have been encountered repeatedly, with observers tiring of them or losing trust in advertisers who seem to think that their message is so weak that they need to send it over and over. Why didn't these banner ads, which were presented as many as twenty times within just five pages of text, suffer any wear-out? The readers never "processed the ads consciously, so there was no recognized information to be identified as tedious or untrustworthy.
>
> These results pose a fascinating possibility for online advertisers: Recognition/recall, a widely used index of success for all other forms of ads, might greatly underestimate the effectiveness of banner ads. In the new studies, frequently interjected banners were positively rated and were uncommonly resistant to standard wear-out effects, yet they were neither recognized nor recalled. Indeed, it looks to be this third result (lack of direct notice) that makes banner ads so effective in the first two strong and stubborn ways. After many decades of using recognition/recall as a prime indicator of an ad's value, who in the advertising community would have thought that the absence of memory for a commercial message could be a plus?
>
> Within the outcomes of the wallpaper and the banner ad studies is a larger lesson regarding the communication process: seemingly dismissible information presented in the background captures a valuable kind of attention that allows for potent, almost entirely uncounted instances of influence.

Cialdini's work and similar research has started to bleed out into the mainstream conception of advertising. Shoshanna Zuboff– a Harvard psychologist– suggests in her influential book 'Age of Surveillance Capitalism' that the only way to compete in the modern economy is by manipulating people with more and more sophistication [464]. She believes that Google and Facebook advertising has gotten so good at eliciting desire in the people that they target that they have moved beyond traditional advertising and are instead trading 'in human futures'; they can 'coax, tune, and herd our behavior' and we are both deeply unaware of

Figure 2-2: Example of auto-generated ads: impersonal sugar ads (left), pair ads in which 'promoted' things are paired with surprising happy photos and negative with the opposite (center), and personalized ads (right) in a few of the IAB standard sizes.

the power of their manipulation and deeply controlled by it.

The implications for our society, our legislation, our economy, our lives, and our self-conception could not be greater.

YourAd

At the time, I believed in these ideas, so against this backdrop I created a project called YourAd [350]. As I wrote then:

> Advertisers have optimized the periphery of our attention to drive complex purchasing behavior, typically using persuasive or rhetorical techniques to promote decisions that are agnostic to our best interest... [and o]ur ability to recognize overt manipulation has pushed modern advertising towards subtle, integrated strategies.

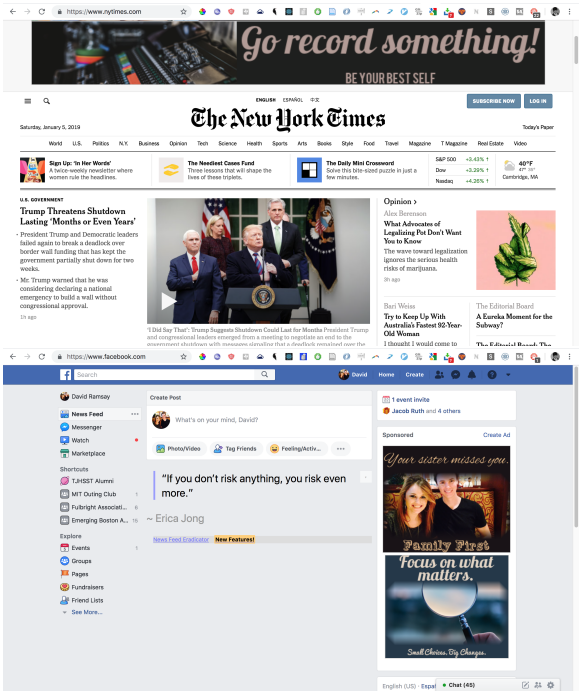


Figure 2-3: Example ads inserted into NYTimes and Facebook using YourAd.

> ...Instead of serving the ambition of companies with large marketing budgets, what if these techniques were used to reinforce the behaviors and attachments we choose for ourselves? YourAd is an open-source browser extension and design tool that allows users to supplant their internet ads with custom replacements– designed by and for themselves. YourAd incorporates industry best practices into a platform to facilitate experimentation with user-aligned advertisement ecosystems, probe the limits of their influence, and optimize their design in support of an end user's personal aspiration.

YourAd is a browser extension that allows people to create personalized ads targeted at themselves using personal messages and images, or pair concepts with random and surprising positive and negative images (see Figure 2-2). It then replaces the banner ads you normally see with them (Figure 2-3). I built this tool, and created several ads targeted at myself– ads to suggest I eat healthier, focus on work, spend more time playing music, and keep in touch with my family more regularly. They featured pictures of my loved ones; I then used this tool for several months. I found within a couple days, I started completely ignoring the ads.

Nothing about my behaviors changed; they had no impact on my life. I reverted to 'banner blindness' almost immediately, despite the novelty of the experience and my self-interest in their efficacy. I started to question the worldview promoted by Cialdini, Zuboff, and others like them.

Reality Strikes

A closer, skeptical reading of Cialdini's work betrays many methodological weaknesses.[5] While no one has attempted to replicate the 'subliminal mere exposure effect' it shows many hallmarks of poor methodological standards; in fact, Cialdini sometimes presents work as confirmatory even without hitting the $p < 0.05$ threshold.[6] The main study depends on a small, high variance dataset, with 12% of the outlier data removed. The evidence contained within this body of work simply isn't convincing against the backdrop of consistent failed replications for priming and subliminal advertising.

The Truth About Advertising

Cialdini and Zuboff both believe we live in Primeworld; Zuboff wants legislation to save us. These two individuals alone have had a dramatic influence on public discourse and marketing budgets (Zuboff's book Surveillance Capitalism is available in 17 languages; Cialdini's work is available in 41). They promote a dangerous narrative– that 'You are Now Remotely Controlled' (as the New York Times reports, citing Zuboff) [464].

[5] In the original work on wallpaper and furniture preference, he studies undergraduates when faced with a fake scenario of buying a couch in a fake online store. They are greeted with a large/obvious image of clouds or pennies, forced to rank their choices on 'comfort' and 'value', and then asked to make a purchase decision. There is indeed a real statistical effect; but the effect is explained by a common sense phenomena– when you ask someone to simulate a purchasing decision and put them in an obvious discount store, they will simulate a decision like a person who has decided to shop at a discount store. This study simply measures trust in obvious brand communication. Cialdini is puzzled by the lack of mediation of expertise (he expects couch-experts, or people with strong opinions about couches, to be immune to the prime); this is the obvious explanation. It also completely discounts the possibility that anyone was subconsciously manipulated by the wallpaper.

[6] See i.e. [184], where we find this conclusion: *'In line with [our hypothesis] H1, fear led social proof appeals to be more persuasive than the control (F(1, 305) = 3.84,* ***p = .051****, d = .22)'*

The reality is that, as Sinan Aral puts it in Harvard Business Review, 'the effectiveness of digital ads is wildly oversold' [34]. Large-scale analysis on Facebook and EBay suggests overestimates of around 4,000% [56, 179]; targeted ad effectiveness is overestimated by around 1,000% [143]. Frequently, causal influence (the ad driven purchasing behavior) is conflated with association; conversion rates simply suggest those that made a purchase saw a relevant ad at some point. This is especially pernicious for websites like Google, where the ads are shown as someone types in the name of that product category (frequently with purchase intent). Most tellingly– for all the hand-wringing about the power of targeted, digital ads– as of 2022 trends show advertisers returning to traditional ad spend [310].

As most companies have recognized, for an ad to work it needs to be consciously processed (hence the move to embedded ads that look like content and influencer product placement). For me, YourAd was a reality check; it made me look critically at the trust I put in the leading academics on influence and persuasion. I started reading the literature critically, and my views changed dramatically.

The most powerful scientific voices on this topic are getting it obviously wrong. Digital advertising funds the free access to huge, expensive infrastructure and life-changing services like Google; neutering this business model will only exclude those who can't easily afford this technology, in the name of protecting them from a chimera.

The Future of Ads

In a world where we don't accept a priori that people are easy to manipulate, ads are much less pernicious. In fact, in this world, they provide a service– they create an efficient, democratic, market-based system to allow innovative new products to get a foothold; they educate consumers about new things that could improve their lives (and on Google, that education can come on-demand when you express interest). Businesses with marketing budgets have either made good products or convinced a small handful of people with faith in the idea to give up cold, hard cash.

In this world, advertisers are not capable of fundamentally reaching into a consumer and

changing their preference using a thirty-second commercial or an unnoticed banner ad; they aren't capable of creating needs in people at all. They are forced to cater to our real needs as dictated by our nature. Fashion and jewelry– arbitrary items whose symbolic value is co-created culturally– are a feature of every human society. It is in-line with our own interests to have advertisers who efficiently shape and perpetuate the symbolic meanings behind our fashions. There is nothing wrong with this on its surface.[7]

Conclusion

This is one small example of how unreplicable, overexaggerated 'science' has impacted me– my own mistaken intuition and my own investment in a project that simply didn't have a meaningful impact. It also points to some of the real, powerful, society-impacting effects of replication crisis research in one, tiny domain. The broader implications are even more powerful.

2.1 The Crisis in Social Psychology

To put it bluntly, any modern research which suggests unnoticeable interventions cause measurable change in behavior is almost certainly false. This work can be broken down into several categories.

2.1.1 Social Priming

The first category is 'social priming', which posits an even more implausible step in the causal chain– something in our environment unconsciously 'primes' our mental state so we're biased towards actions, thoughts, or beliefs with weak associations to the stimuli. John Bargh is probably the most famous researcher on social priming; Kahneman penned

[7] Obviously ads can be predatory. They can misrepresent their product, overpromise, underdeliver, or prey on a consumer's insecurities or weaknesses; we need mechanisms and legal frameworks to keep advertising honest. It's certainly not all roses, but it's a lot closer to all roses than most people seem to believe.

his open letter addressed to him [461]. Some of his most famous results that have failed replication in high power trials include:

- *Warm beverages promote interpersonal warmth.* This is one of the most famous studies from Bargh's lab, featuring large effect sizes; it has failed to replicate in studies triple the size [81].

- *Hearing words associated with old age makes you walk slower and act older.* This finding sparked a ton of follow up work on priming– with nearly 4,000 citations; it is featured in many pop psychology books. It even replicated– when graduate researchers were in charge of surreptitiously measuring walk times using a stopwatch. The effect disappears with an automatic laser timer [121].[8]

Other famous examples of social priming that have failed replication include things like:

- *French background music makes you buy French wine.* This finding is a favorite of the marketing pop-psychology field; the original paper was based on only 82 purchases (so roughly 20 bottles each, across the four conditions of music type x wine type). This was replicated with five times the participants showing no effect [212].

- *Exposure to hostile words make you interpret ambiguous events as hostile.* This 1979 study by Srull and Wyer failed to replicate [286]; it's often cited as justification that small changes in verbiage matter, as in Cialdini's book where he describes a hospital deciding to ban the use of so-called 'aggressive phrases' like 'bullet points', 'attacking problems', and 'hitting targets' [85].

- *Feelings of cleanliness affect people's moral judgements.* Washing your hands after a disgusting movie scene or doing a word scramble with 'cleanliness words' cleans your conscience, thus it's easier for you to then accept immoral acts like falsifying a

[8]This finding has two important implications that are worth emphasizing: (1) tools for objective measurement improve the methodological quality of our work, and (2) people absolutely *do* respond to clearly articulated goals and expectations, sometimes in subtle ways. Priming as we have defined it includes a subconscious cue; our goals and beliefs manifest in our lives powerfully and automatically.

resume or keeping a wallet. This failed replication featured a public dispute between researchers [294].

- *The sight of money makes you more accepting of social norms.* Money priming makes up a whole subfield of priming research; the sight of money is claimed to produce all kinds of effects. In 2006, Vohs published a very influential paper in Science on money priming. Unfortunately Vohs work– alongside other work on money priming– has not replicated in large-scale pre-registered studies; and the original findings show tell-tale signs of p-hacking in the experimental design [79].

- *Female-named Hurricanes are more deadly.* This was supposedly because people have gender-based expectations, and thus don't flee from the gentler sounding female hurricanes. It turns out, this isn't true [398].

- *Imagining life as a college professor makes you better at recalling facts.* This is the so-called 'professor bias' made famous by Ap Dijksterhuis, which has failed in large scale replication [324].

- *Seeing the American Flag makes you more likely to vote Republican* [146].

- *Drawing smaller (vs. larger) distances on graph paper makes you feel closer to your family and reduces estimates of food calories* [329].

- *Seeing the color red or the word 'red' makes you perform worse on tests, moderated by self-control over emotional responses* [55].

- *Higher physical perspectives leads to more helping and sharing resources* [371].

- *Bitter foods that trigger physical disgust heighten our sense of moral wrongness* [167].

2.1.2 Physiology on Feelings

Another subclass of these issues appears with unconscious physiological drivers of emotions and behavior. In these experiments, the hypotheses suggest that engaging our musculature

in certain ways will unconsciously lead to cognitive or behavioral effects. The most famous three failed replications are:

Power Posing.

Harvard's Amy Cuddy popularized her research suggesting large hormonal effects (cortisol and testosterone) from striking a powerful pose. The idea that the posture itself matters– and not the act of self-coaching when striking the pose– has failed to replicate [353]. Dana Carney– the original first author– has said "I do not believe that 'power pose effects' are real...the evidence against the existence of power poses is undeniable" [334]. This is one of the most replicated studies in psychology and the results are quite definitive.

The Facial Feedback Effect.

To study whether recruiting smiling muscles will make you happy without the psychological implications of asking you to smile, researchers study this effect by asking participants to bite a pencil [416]. It turns out the happiness you might feel from forcing a smile is psychological; recruitment of the smile muscles themselves has yet to have a provable effect [442].

The Hungry Judges Effect.

This famous study showed judges make harsher parole decisions just before lunch by an astonishing margin (from 65% in the morning to 0% before lunch) [104]. This study has been cited 1,400 times and made a splash in the popular press; Stanford's Robert Sapolsky popularized this effect in his book 'Behave'. On NPR he reinforced how irrational people are when they question this data: "you get that judge two seconds after they made that decision, you sit him down at that point and say "Hey, so why did you make that decision?" And they're gonna quote, I don't know, Immanuel Kant or Alan Dershowitz at you. They're going to post-hoc come up with an explanation that has all the pseudo-trappings of free will

and volition, and in reality it's just rationalization. It's totally biological" [1]. In reality, it turns out the docket was ordered by case severity [177]. Meta-statistician Daniel Lakens points out the absurdity of the original finding: "if hunger had an effect on our mental resources of this magnitude," he writes, "our society would fall into minor chaos every day at 11:45" [250].

2.1.3 Implicit Bias

These issues are also pronounced amongst highly politicized literature on race or gender and implicit bias. Once again, these studies rest upon the premise that subconscious associations cause measurable differences in beliefs and behavior (or the inverse). Kelsey Piper reports "the conversation about diversity in tech is getting hijacked by bad research" [339]; Jesse Singal convincingly argues a similar point about the role of implicit bias research in racism [394]. Here are three famous examples:

Stereotype Threat

Stereotype threat is the idea that, by eliciting a stereotype before a task, performance drops (i.e. with the stereotype 'women are bad at math', calling a math test 'a math test' will cause female performance to drop while calling it a 'problem-solving task' will not). It was introduced in 1995 by Stanford's current Dean of Psychology, Claude Steele [406].

Stereotype threat was put in the replication spotlight as part of an NPR Radiolab episode featuring Steele and his collaborators [2]. As part of that episode, his collaborators admit they no longer believe in the concept and describe pervasive, inadvertent p-hacking. Statistical analysis of the original paper shows very strong evidence of this; psychologist Russel Warne writes [445]:

> The average sample size of African Americans in Steele and Aronson's (1995) four studies was 37.75. This means that the statistical power was too small to

> detect any but the strongest effects of stereotype threat. Assuming an effect of d = .50 (Cohen's, 1988, "medium" effect size and a reasonable default at the time for new research topics), the power to detect stereotype threat in Steele and Aronson's (1995) four studies ranged from .300 to .459. The joint probability for detecting stereotype threat in all four studies was just .014. Conversely, the probability of not detecting the stereotype threat in four out of four studies was .173 (Warne, in press, p. 279).
>
> In layman's terms, this means that all four studies were each (individually) more likely to fail to detect a stereotype threat than to detect the phenomenon. Moreover, the probability of all four studies failing to demonstrate stereotype threat was over ten times more likely (17.3%) than identifying the phenomenon in all four studies (1.4%). Based on these probabilities, the most likely result of a collection of four small-sample studies on stereotype threat was a mix of some studies supporting and not supporting the existence of the phenomenon. Steele and Aronson (1995) either got extremely lucky . . . or they engaged in questionable research practices that inflated the strength of the evidence for stereotype threat.

This corrobrates the overwhelming conclusions of meta-analysis on the phenomena. Rutger's psychologist Lee Jussim summarizes the state of stereotype threat in "Is Stereotype Threat Overcooked, Overstated, and Oversold?" [229]:

> Flore Wicherts (2015) performed the only meta-analysis of which I am aware that has subjected stereotype threat findings to a whole family of skeptical tests, such as p-curves, funnel and forest plots, and tests for excess of significant results. The results are not pretty and show that the effects primarily appear in the underpowered, small-scale studies, and either disappear or reverse altogether in the highly powered large-scale studies. Uli Schimmack has also shown that stereotype threat studies are likely to have considerable difficulty replicating.
>
> Furthermore, as Heterodox Academy's Amy Wax (2009) has so aptly pointed

> out, stereotype threat effects among African Americans have been mostly obtained in very select and unrepresentative samples. A slew of boundary conditions have been proposed, which is another way of saying, "it might only apply to select people under select circumstances."

Stereotype threat never meaningfully explained real achievement gaps; its claimed effect sizes have always been relatively minor [406]. Even in context, the best statistical minds of the replication crisis agree that there is no trustworthy evidence that the effect is real [162, 375].

Implicit Gender Bias in Recruiting

'Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians' [178] is a seminal paper in the world of bias– it made rounds in many major news outlets, with a culminating appearance in Malcolm Gladwell's best seller 'Blink'. It reports that adding a large visual barrier to obscure the audition from existing orchestra members could "explain 30 to 55 percent of the increase in the proportion female among new hires... from 1970 to 1996." The study suggests that implicit bias was perhaps the dominant reason women weren't getting hired; the authors attribute much of the success of the real-world success of women to removing gender cues. Upon deeper inspection, the claims from this paper are overstated; the Wall-Street Journal describes the results after they became controversial:

> the raw tabulations showed women doing worse behind the screens.... [t]he result was a tangle of ambiguous, contradictory trends. The screens seemed to help women in preliminary audition rounds but men in semifinal rounds. None of the findings were strong enough to draw broad conclusions one way or the other... After warning that their findings were not statistically significant, they declared them to be "economically significant." What does that mean in this context? "That doesn't mean anything at all," writes Columbia University data scientist Andrew Gelman.

The massive effect size reported by this study have been corroborated by a few other studies around blinded resumes– studies which Kelsey Piper reports are "literally incredible." She writes, "If these statistics were right, a relatively simple change to job applications could close the whole gender gap in tech" [339].

The real picture is much more complicated than these influential analyses imply; disparities between the sexes have many other possible explanations. 58% of women report experiencing sexual harassment in the workplace [397]. Women frequently deal with overt sexism/misogyny alongside a host of systemic, institutional challenges surrounding femininity and motherhood. Gender bias "operating at the unrecognized, unconscious level" [365] is only one possible factor in that conversation.

And despite early work to the contrary, the data does not suggest it is a dominant one. As Regner et. al. summarizes in a recent Nature Human Behavior paper on the topic, "whether gender bias contributes to women's under-representation in scientific fields is still controversial... [e]xperimental studies have examined the possible role of gender bias in contributing to women's under-representation, but have revealed conflicting patterns" [354]. Large scale studies on the topic across domains have had mixed results and occasional reversals (i.e. [203], [424], [339]). The totality of evidence suggests that– should unconscious attitudes manifest into real forms of discrimination– their real effect sizes, if we are able to measure them, will be very small.

The Implicit Association Test

> [I]n the early 1990s, social psychologists were enthralled by work in cognitive psychology that demonstrated unconscious or uncontrollable processes (Greenwald Banaji, 1995). Implicit measures were based on this work, and it seemed reasonable to assume that they might provide a window into the unconscious (Banaji Greenwald, 2013). However, ...[t]here is nothing implicit about being a Republican or Democrat, gay or straight, or having low self-esteem. Conflating implicit processes in the measurement of attitudes with implicit personality constructs has created a lot of confusion. It is time to end this confusion. The IAT is an implicit measure of attitudes with varying validity. It is not a window into people's unconscious feelings, cognitions, or attitudes.
>
> Ulrich Schimmack
> The Implicit Association Test: A Method in Search of a Construct (2021)

The Implicit Association Test is another concept featured heavily in Gladwell's 'Blink'. In this test, differences in response times when pairing words like 'good' and 'bad' with African American or European faces is taken as evidence of implicit racial bias.

The IAT has poor construct validity and reliability. It predicts effectively nothing about whether you exhibit biased behavior in your life; the largest meta-analysis, carried out by its proponents, suggest it can only explain 1% of the implicit bias it's trying to measure [151].[9] The authors of the IAT themselves– after public debate with a team led by UPenn's Philip Tetlock– concede it is only "good for predicting individual behavior in the aggregate, and the correlations are small" [266]. In their publication on the topic they write: "IAT measures have two properties that render them problematic to use to classify persons as

[9] They write: "we found little evidence that changes in implicit measures translated into changes in explicit measures and behavior, and we observed limitations in the evidence base for implicit malleability and change."

likely to engage in discrimination. Those two properties are modest test–retest reliability (for the IAT, typically between r = .5 and r = .6; cf., Nosek et al., 2007) and small to moderate predictive validity effect sizes. Therefore, attempts to diagnostically use such measures for individuals risk undesirably high rates of erroneous classifications" [182].

Despite these concessions, the IAT creators argue that its small correlational predictive validity for population level trends are still meaningful (it correlates with 4% of the variance in discrimatory behavior at this level). This interpretation of the IAT retains criticism. Ulrich Schimmack write in "The Implicit Association Test: A Method in Search of a Construct" [378]:

> Most of this valid variance stems from a distinction between individuals with opposing attitudes, whereas reaction times contribute less than 10% of variance in the prediction of explicit attitude measures. In all domains, explicit measures are more valid than the IAT.

In other words, the very small amount of variance explained by the IAT is also almost completely explained by simultaneous, simple measures of overt, conscious attitudes. Jesse Singal summarizes other critical views, from psychologists like James Jaccard and Hart Blanton, in his work on the measure [394]:

> ...after almost 20 years and millions of dollars' worth of IAT research, the test has a markedly unimpressive track record relative to the attention and acclaim it has garnered. Leading IAT researchers haven't produced interventions that can reduce racism or blunt its impact. They haven't told a clear, credible story of how implicit bias, as measured by the IAT, affects the real world. They have flip-flopped on important, baseline questions about what their test is or isn't measuring. And because the IAT and the study of implicit bias have become so tightly coupled, the test's weaknesses have caused collateral damage to public and academic understanding of the broader concept itself. As Mitchell and Tetlock argue in their book chapter, it is "difficult to find a psychological

construct that is so popular yet so misunderstood and lacking in theoretical and practical payoff" as implicit bias. They make a strong case that this is in large part due to problems with the IAT.

Despite the well publicized criticism, the IAT remains the centerpiece of implicit bias literature, perpetuating a flawed understanding. Harvard's 'Project Implicit' website– home to this work– has informed over a million visitors that their IAT scores suggest they potentially harbor unconscious racial bias against African Americans [43] with the disclaimer that "[researchers] make no claim for the validity of these suggested interpretations" [378]. Patrick Forscher– co-lead author on the last major meta-analysis of the IAT with the lead researcher at Harvard's Project Implicit– summarizes our understanding of implicit bias thusly: "Based on the evidence that is currently available, I'd say that we cannot claim that implicit bias is a useful target of intervention" [394].

In the wake of race riots in Ferguson, The Justice Department reported "emails circulated by police supervisors and court staff that stereotype racial minorities as criminals" [436]. As recently as 2008, 4% to 6% of Americans reported that they would be unwilling to vote for any African American candidate as president [331]– certainly an underestimate given social desirability biases. Overt racism still exists in American culture. Instead, the debate about racism has shifted focus to the unproven impact of unconscious attitudes on real behavior.

2.1.4 Summary

People make automatic judgments and engage in automatic, habitual behavior without conscious thought. This is a different claim than the claims we see above; that our automatic judgements and behaviors are shaped unconsciously, that to do so is trivial, and that we are fundamentally incapable of successful introspection about this process (and in fact, our earnest self-assessments, stated beliefs, and honest intentions may all oppose our implicit biases). There is no evidence for these pervasive and damaging ideas.

2.2 Behavioral Economics

Behavioral Economics has also come under criticism, though it has fared better than social psychology. Noble-prize winning economist Vernon Smith characterized it as "a deliberate search in the tails of distributions for "Identifying the ways in which behavior differs from the standard model..." (Mullainathan Thaler 2001, Vol. II, p. 1094), a search that can only succeed" [400]. Andreas Ortmann argues for the notion that economics models cannot be fundamentally improved by importing insights from psychology because psychology has no consistent and universalizable theory– it can only attach to standard economic models in an ad-hoc way (and the validity of these ad-hoc cognitive illusions and biases are hotly contested given the replication crisis, with many disappearing or reversing in more realistic decision making scenarios). Ortmann calls out specific examples [321]:

> [During replication] only about one out of ten participants violate the conjunction principle (rather than more than eight out of ten in Kahneman and Tversky's version of the problem.) The authors point out that their set-up mirrors real-life more closely than the scenario studies that Kahneman Tversky preferred here and elsewhere. Likewise the reality of endowment effects, sunk costs, income targeting, overconfidence, loss aversion, and self-control have all been contested, although you would not know if you read self-congratulatory accounts of the state of the literature like Thaler (2016). Notably, the last three concepts were identified by Thaler in his 2016 AEA presidential address as "three of the most important concepts of behavioral economics" (p. 1578)

Gerd Gigerenzer of the Max Planck Institute is one of the most vocal critics of the fundamental assumption behind most behavioral economics research– that humans are biased and irrational. He argues that our decision-making heuristics are quite rational and useful, especially in the context of uncertain environments. Gigerenzer and Kahneman/Tversky faced off in a somewhat heated, public debate in Psychological Review throughout the 90s [170, 231, 171]. While many of the specific initial results hold up empirically, the debate

centers around theory and contextualization– some of the original 'biases' fail to appear when tests are slightly re-framed (i.e. with raw numbers instead of percentages, or when people reason about a group of people instead of an individual). In general, Gigerenzer seems to desire a more rigorous underlying psychological theory of heuristics and biases to describe decision-making under uncertainty, and finds some clear weak points in the blanket descriptions put forth by prospect theory; Kahneman and Tversky, on the other hand, argue that the field and its data are too nascent to support a rigorous level of modeling and analysis.

2.2.1 Framing Effects

Framing effects share a lot with priming– subtle, changes in presentation that make a measurable difference– however in these cases, the outcome is a conscious decision and the 'frame' structures what a decision-maker considers and how they reason about their choice.

Framing effects are real [28]. Some have failed replication (i.e. framing effects to alter honest behavior in a high power, pre-registered study that showed no effect (N=1,200) [118]) or require further contextualization (i.e. in the political domain, a meta-analysis of 138 studies suggests framing effects are real but limited, with medium effect sizes reported on attitudes and emotions and very little effect on behavior– and these results significantly weaken if a competing frame is introduced [28]).

It is still difficult to identify which framing effects work under what circumstances; there is no unified theory. The debate continues over whether these effects should be described as fundamentally irrational (i.e. inconsistent with respect to stable preferences) cognitive biases or reasonable heuristics based on the implied information (language pragmatics) [171].

2.2.2 Nudges

Nudges are one of the areas that shares the most similarity with social psychology. As I mentioned in chapter 1, the largest meta-analysis on 'nudge-style' interventions– a 2020

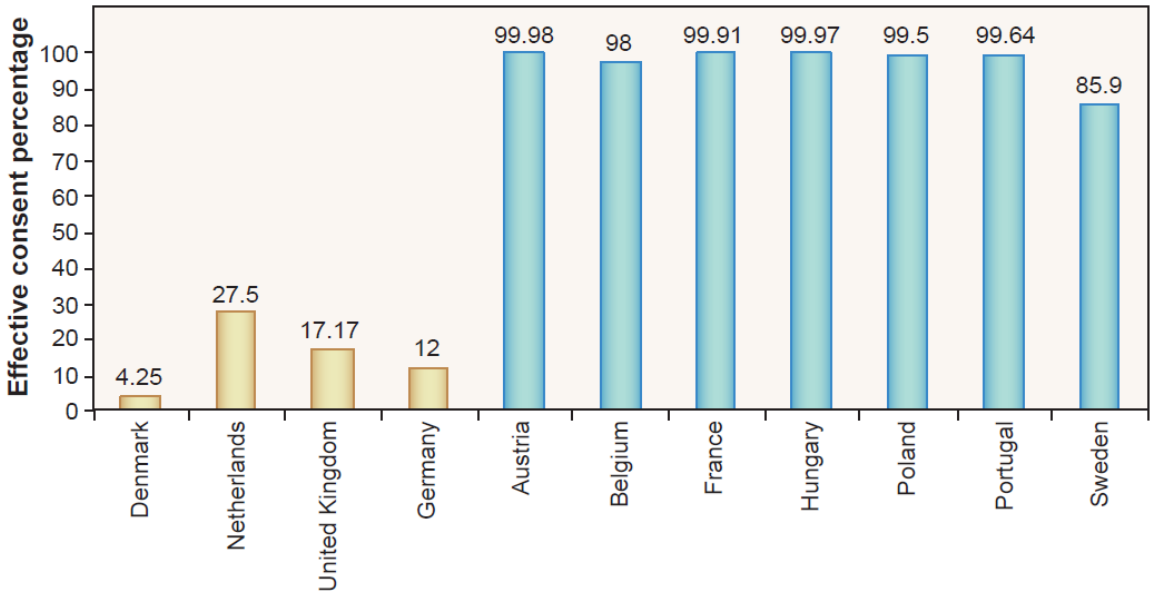
Berkeley study that looked at 126 randomly controlled trials performed by the US federal agencies (the OES and BIT-NA) on tens of millions of citizens— showed that nudges reported to have an effect size of 8.7% in academic contexts result in a real but marginal 1.4% impact on behavior in the real world [110]. While a 1.4% change in behavior might have an important impact at scale, it clearly plays a very, very minor role in individual, real-world decision making. The effects are real, but their impact is very small.

Ortmann [321] summarizes the literature as follows:

> Hummel Maedche ... provide a quantitative review of 100 empirical nudging studies in which 317 effect sizes from different research areas are reported.... They find relatively small effects for environmental, health, wealth, and privacy interventions (typically in the 0.2 – 0.3 range), even smaller effects for energy-saving interventions (around 0.1), and essentially imperceptible effects for policy-making (the major application area of behavioral insights units). Kristal Whillans (2020) report the outcomes of five naturalistic well-powered nudging experiments (involving almost 70,000 employees of a large organization) involving attempts to move people out of single-occupancy vehicle commutes. They report that the treatment effects observed in four of five of their experiments were statistically equivalent to zero.

The failures of large-scale nudge interventions goes on for multiple pages. Magda Osman has published a good critique suggesting that the focus on success– instead of documenting the failures and why– is holding back theory in this field [322].

That's not to say all nudges fail. One of the few nudges that seems to have succeeded well is Thaler's Save Tomorrow Campaign (pre-commitment to automatically save a higher percentage of your paycheck as you get raises). But to quote another skeptic of the field, John Cochrane– "The nudge for saving experience is good and solid. But the skeptical reader, who does not sing in the choir, wonders: you've been at it three decades, and this is all you got?" [87]



Effective consent rates, by country. Explicit consent (opt-in, gold) and presumed consent (opt-out, blue).

Figure 2-4: The famous figure from Johnson and Goldstein's 2003 Science paper 'Do Defaults Save Lives?' [227] demonstrating the alleged default effect.

2.2.3 Other Famous Effects

There are too many aspects of Behavioral Economics to thoroughly review, but a few select topics are included here:

The Default Effect

One of the most popular examples of the 'power of defaults' in behavioral economics comes from Johnson and Goldstein's 2003 Science paper 'Do Defaults Save Lives?' [227]. Within it is the following chart, which purports to show the difference in organ donation rates in countries that are opt-in vs opt-out:

This graph gets used by prominent behavioral economists to suggest that for complex decisions, we get overwhelmed and will simply choose the default. Dan Ariely has called this "one of his favorite graphs in all of social science" [35].

This argument that 'we rely on defaults in the face of complexity' is included by Nobel Prize winning economist Richard Thaler in his 2008 book Nudge, and his 2015 book 'Misbehaving' [427]. The graph was interpreted in 2013 NY Times: "Roughly 98 percent of people take part in organ donor programs in European countries where you have to check a box to opt out. Only 10 percent or 20 percent take part in neighboring countries where you have to check a box to opt in" [65].

Unfortunately, this graph obscures the truth. In the opt-out case, countries presume consent– in many cases they aren't ever given an explicit choice and must self-initiate a process to opt-out. Furthermore, the process for organ donation is complicated, so this presumed consent is frequently ignored in favor of family wishes. A systematic review of the actual changes in rates of organ donation from presumed-consent (where no choice is presented) suggests small actual increases in the range of 2-6 donors per million [362]. Most sources suggest little-to-no real effect from switching the default [368].

While the default effect is probably real for low-stakes decisions where people have weak preferences (i.e. where the decision cost is greater than the perceived benefit of making a marginally better choice), there is no evidence that it drives large changes in behavior; the canonical example of its impact– organ donation rates– is not at all definitive.

The Backfire Effect

A widely cited 2010 paper coined the term 'backfire effect' to describe a perplexing, counterintuitive cognitive bias– when people are given a correction to a news article, it actually increases their misperceptions. [317] This effect is more than simple source credibility (where new information from an untrustworthy source would be ignored); it actually pushes people further towards their pre-existing beliefs.

A more recent 2018 paper entitled 'The Elusive Backfire Effect' used the same methodology to test the effect on 10,100 subjects and found no evidence for it across 52 issues "despite testing precisely the kinds of polarizing issues where backfire should be expected" [457]. They included the exact condition of the prior article with 10x the population (n=977).

They conclude that presenting factual corrective information does not increase the strength of political misperceptions– "by and large" they write, "citizens heed factual information, even when such information challenges their ideological commitments." They note that ideology does shape the extent of receptivity to corrective information.

Honesty Priming

While fraud is not common in the academy, when it occurs it is very damaging to trust in science. The most recent famous case for social psychology involves flagrant evidence of fraud in papers on honesty primes by both Harvard's Francesca Gino and Duke's Dan Ariely [407]. This research is widely disseminated, having been detailed in Ariely's Netflix special and several New York Times bestselling books– and its irony makes for a great headline.

Ariely's work gained the spotlight in 2021 when researchers proved that fraudulent data underpinned his 2012 paper which showed that signing an honesty pledge at the start of a form made people more honest in reporting car odometer readings to their insurance company. Interestingly, the renewed scrutiny came after he himself published a follow-up paper in 2020 suggesting the original results didn't replicate with higher power experimentation [249]. Honesty priming does not seem to replicate [330].

Unconscious Thought Theory

Introduced by Ap Dijksterhuis in 2006, Unconscious Thought Theory (UTT) posits that we will make better decisions if we let our unconscious mind do the work. Dijksterhuis argues for the weighting principle– that unconscious thought naturally weighs attributes by their true relative importance, whereas conscious thought relies heavily on schemas and rules. The theory posits unconscious decision-making perform better than conscious decision-making as decisions increase in complexity.

A typical Dijksterhuis experiment involve showing participants several complex options where one is objectively the best but difficult to recognize. Participants either (a) aren't

allowed any time to think about choices, (b) given some time while distracted with a secondary task, or (c) allowed to carefully weigh the options. UTT predicts the second group– given time to unconsciously process the information– will perform the best, which is what Dijksterhui reports [117].

Replications of these findings fail to support the theory. From 'Four empirical tests of Unconscious Thought Theory' conducted with 480 participants [211]:

> Unconscious Thought Theory yields four specific predictions. First, an exact replication of Dijksterhuis et al. (2006a) study should indicate that unconscious decisions are superior to conscious decisions. Second, decisions should improve with duration of conscious thought. Third, unconscious decisions should be superior to conscious decisions, even if unconscious decisions are deliberated while having access to information. Fourth, unconscious decisions should be based on a weighting strategy. We report results of four studies, featuring 480 participants, that yield no evidence in favor of these predictions.

These failed replications appear elsewhere, across many labs and several variations [18, 446]. The original theory still has proponents, however. (see i.e Lassiter et al. [254], Payne et al. [332], Bargh et al. [44])

2.2.4 Summary

Behavioral Economics has also come under fire for its findings, some of which have failed to replicate. For the most part, the science behind 'nudging'– subtly controlling behavior or choice with differences in choice architecture– have very minor real-world effects. Framing effects do seem to impact conscious decision-making, but the debate over whether these are 'irrational biases' as oppose to 'intelligent interpretations of communicated information' is still contested.

It seems to me that Gerd Gigerenzer's interpretation of these effects is correct; a lot of useful information is conveyed in the subtlety of our language, and most of the evidence for

so-called 'irrationality' in when information is presented differently is actually evidence of our intelligence and awareness. In the real world it does not seem that these effects can be meaningfully exploited to unconsciously control behavior.

2.3 Positive Psychology

> So, a lot of positive psychology over the last ten, twenty years–I say this with respect to my friends in the field, I've been to conferences, I've given talks– and a lot of this is flimflam. A lot of it is oversold hokum... It is quick fixes based on shoddily done studies designed to get people to publish best-selling books and get TED Talks and so on. But, it's not grounded in science and, maybe worse than that, it has a limited and sort of parched philosophy behind it. Often a sort of simple-minded hedonism.
>
> *a famous Yale psychologist*
> *podcast interview, 2020*

'Positive Psychology' is the project of Martin Seligman (the president of the American Psychological Association) alongside Ed Diener and Mihaly Csikszentmihalyi (who popularized the concept of 'flow states') [156]. Its stated goal is to shift the focus of psychology research away from mental disorder and towards optimal experience– though humanistic psychologists of the seventies, led by Maslow, Rogers, and Bugental, already championed a decades-long tradition in this spirit.

In differentiating the two fields, Alan Waterman has pointed to a philosophical divide (positive psychologists have favored Aristotle, J.S. Mill, and modern eudaimonists over the existentialists and phenomenologists embraced by their humanistic predecessors) as well as nomothetic (general truths across populations) vs. idiographic (studying the individual) perspective [450]. Papers discussing the contentious relationship between the two 'sibling rival' fields regularly appear in major journals [359]. But the main focus has been epistemic.

To brand 'Positive Psychology', Seligman has described a turn from qualitative humanistic

techniques toward rigorous, quantitative ones. In 2000 he wrote "humanistic psychology did not attract much of a cumulative empirical base", and blames it for "10 shelves on crystal healing, aromatherapy, and reaching the inner child for every shelf of books that tries to uphold some scholarly standard" [57]. This derisive comment has led to tension and criticism from many others in the broader field, especially as its proponents face meaningful criticism of their own empirical practice.

In 'Positive Psychology Goes to War' [393] Seligman is accused of weak empiricism himself– incorporated poorly justified research into his talks and promoting his Penn Resilience Program (PRP). PRP is a resilience intervention with weak empirical support (one meta-analysis in the Journal of Adolescence concluded that a 'roll-out of PRP cannot be recommended'), which became a touchpoint when it was adopted to fight PTSD in the Army. At that point, it had never been tested for efficacy either for trauma or with soldiers. In the article, Harvard psychologist Richard McNally recounts trying to steer the Army away from a large roll-out in an early meeting; Nick Brown echoes McNally's sentiments [393]:

> The idea that techniques that have demonstrated, at best, marginal effects in reducing depressive symptoms in school-age children *could also* prevent the onset of a condition that is associated with some of the most extreme situations with which humans can be confronted is a remarkable one that does not seem to be backed up by empirical evidence.

In the single largest mental health intervention in history, the Army rolled out the Penn Resilience Program (rebranded the CSF) to the tune of $30 million. The roll-out was later extolled in a wholly-devoted issue of the APA's main journal, American Psychologist.

The program has since been criticized for its ethics and weak empirical backing in Time Magazine, Scientific American, Psychology Today, and by a team of psychologists in an article dubbed "The Dark Side of 'Comprehensive Soldier Fitness'" [132]. In it, the journal issue dedicated to celebrating the program has also been condemned— Psychologist Stephen Soldz writes:

> In addition to our deep concerns about Comprehensive Soldier Fitness, the American Psychological Association's unrestrained enthusiasm for the program is especially worrisome for what it says about the APA, the largest organization of psychologists in the country, indeed the world. As we have demonstrated, there are many complex issues regarding the CSF program's empirical foundations, its promotion as a massive research project absent informed consent, and the basis on which its psychologist developers justify the program. We would therefore expect a special issue of the American Psychologist, a journal edited by the APA's CEO Norman Anderson, to encourage an extended discussion of these matters.
>
> In contrast, guest editors Seligman and Matthews have assembled 13 articles that include no independent evaluation of the empirical claims underlying CSF. They contain no unbiased discussion of ethical issues raised by the program. ... Unfortunately, the APA's uncritical promotion of the CSF program reveals much about the current moral challenges facing the psychology profession itself.

Journalist Daniel Defraia quotes Columbia University's George Bonanno reacting to many over-exaggerated claims about the CSF program: "I was more or less floored. I've been studying resilience for 20 years, and I don't know of any empirical data that shows how to build resilience in anybody" [109].

After a decade of military use, there still does not appear to be strong empirical support for the program. No amount of resilience training seems to shield soldiers from the trauma of war.

2.3.1 Broader Criticism

The criticisms leveled at CSF extend to the entire field. Soldz contends 'Positive Psychology' is characterized by "its failure to sufficiently recognize the valuable functions played by 'negative' emotions like anger, sorrow, and fear; its slick marketing and disregard for harsh

and unforgiving societal realities like poverty; its failure to examine the depth and richness of human experience; and its growing tendency to promote claims without sufficient scientific support" [132].

Psychologist Richard Grant has suggested that the push towards 'quantitative' techniques incentivizes scientistic, pseudo-scientific 'rigor' [359]; Psychologist Barbara Held has argued that the very act of casting psychology as a positive/negative dichotomy is a terrible caricature of the complexity of human emotions that is reflected in the positive psychology literature [196].

Statistician Nick Brown has stated "positive psychology has claimed superiority... in terms of supposedly both using more rigorous science and avoiding popularizing nonsense... positive psychology did not live up to these claims... [it has] merely repeated whatever research problems that humanistic psychology might have had with its own romanticism, resulting in what are even more egregious problems" [155].

Daniel Horowitz, a historian who has outlined the rise of positive psychology in his book 'Happier?' characterizes it thusly: "Virtually every finding of positive psychology under consideration remains contested, by both insiders and outsiders... Major conclusions have been challenged, modified, or even abandoned. Even what happiness means has been up for grabs" [207].

2.3.2 The Critical Happiness Ratio

UNC psychologist Barbara Fredrickson is very influential in the positive psychology movement. In 2011 she introduced the 'critical happiness ratio' with a widely cited (>1,000) publication entitled 'Positive Affect and the Complex Dynamics of Human Flourishing' [153]. Fredrickson argued that if you experience more than 2.9013 positive emotions for every negative emotion, you would flourish. Otherwise, you would languish. She wrote of this discovery in a successful book on the topic, Positivity: "the 3-to-1 positivity ratio may well be a magic number in human psychology" [298].

Unfortunately, this ratio was born from the misapplication of complex fluid dynamics equations that few in the field understood. In fact, it was eight years before Nick Brown (who also revealed Cornell Food Scientist Brian Wansink's statistical errors) critically deconstructed the math behind this fantastic claim. He published his work in American Psychologist under the title 'The Complex Dynamics of Wishful Thinking: The Critical Positivity Ratio' [68]:

> We examine critically the claims made by Fredrickson and Losada (2005) concerning the construct known as the 'positivity ratio'. We find no theoretical or empirical justification for the use of differential equations drawn from fluid dynamics, a subfield of physics, to describe changes in human emotions over time; furthermore, we demonstrate that the purported application of these equations contains numerous fundamental conceptual and mathematical errors. The lack of relevance of these equations and their incorrect application lead us to conclude that Fredrickson and Losada's claim to have demonstrated the existence of a critical minimum positivity ratio of 2.9013 is entirely unfounded. More generally, we urge future researchers to exercise caution in the use of advanced mathematical tools such as nonlinear dynamics and in particular to verify that the elementary conditions for their valid application have been met.

The many years of acceptance for this idea were damaging to the field's reputation. Brown's co-author Sokal later noted "the main claim made by Fredrickson and Losada is so implausible on its face that some red flags ought to have been raised" [46].

As a response, Fredrickson removed the chapter from her book Positivity and issued a partial retraction, but turned to American Psychologist to defend herself. She responded to Brown's critique with 'Updated Thinking on Positivity Ratios'– a response that notably lacked involvement from her engineering co-author Losada [152]. In it, she argued that despite their flawed model, a 'tipping point ratio' of positive-to-negative emotions was still the correct conceptualization of human flourishing. Brown once again dispelled this idea with a retort entitled 'The Persistence of Wishful Thinking' [69].

This spat concluded in 2014; in 2018, Brown published 'Implications of Debunking the "Critical Positivity Ratio" for Humanistic Psychology' in the Journal of Humanistic Psychology, in which he takes aim at the broader field of positive psychology alongside several other critics [155]. Of this incident, he notes that the original paper gains citations as a much faster rate than the correction, and that co-author Losada continues to promote the retracted model (in parallel with other charlatanic behavior). To contextualize his encounter with Fredrickson, Brown had this to say:

> A number of unnamed adherents to positive psychology challenged us by claiming that Fredrickson's work on the critical positivity ratio was just an anomaly. As the avowed best-of-the-best researcher within positive psychology, they claimed her other work lives up to her top-notch reputation, so we turned our attention to some of that other work. In a recent high-profile paper in which she and her coauthors claimed that adhering to eudaimonic (i.e., altruistic-based) rather than hedonic (i.e., pleasure-based) happiness led to more favorable patterns of gene expression, we found that her conclusion again was more than problematic (Brown, MacDonald, Samanta, Friedman, Coyne, 2014), and when she and her coauthors published a follow-up paper on the same topic that doubled-down on her previous claim, that too we found to be severely flawed (Brown, MacDonald, Samanta, Friedman, Coyne, 2016); other authors have also pointed out very substantial problems with this work (Nickerson, 2017a; Walker, 2016). Similarly, in another recent high-profile paper we examined closely, Fredrickson and her coauthors claimed that practicing loving–kindness meditation led to better physiological outcomes (as indexed by vagal nerve tone), and yet again we found that conclusion untenable (Heathers, Brown, Coyne, Friedman, 2015).

2.3.3 The Happiness Pie

In 'Pursuing Happiness: The Architecture of Sustainable Change', Sonja Lyubomirsky presented a now popular concept in the positive psychology world– the 'happiness pie [271]:

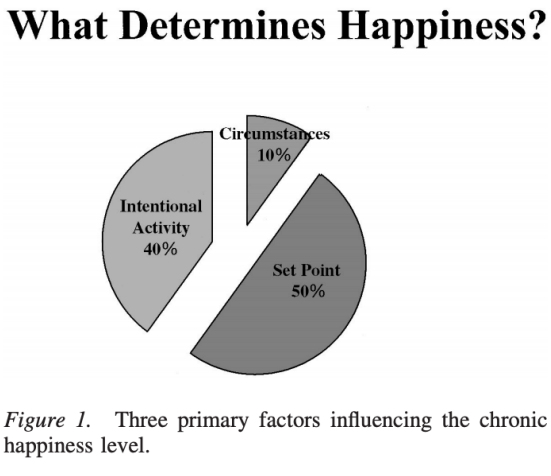


Figure 1. Three primary factors influencing the chronic happiness level.

Figure 2-5: Lyubomirsky's 'Happiness Pie' from [271].

> [E]xisting evidence suggests that genetics account for approximately 50% of the population variation (Braungart et al., 1992; Lykken Tellegen, 1996; Tellegen et al., 1988), and circumstances account for approximately 10% (Argyle, 1999; Diener et al., 1999). This leaves as much as 40% of the variance for intentional activity, supporting our proposal that volitional efforts offer a promising possible route to longitudinal increases in happiness. In other words, changing one's intentional activities may provide a happiness-boosting potential that is at least as large as, and probably much larger than, changing one's circumstances.

Jesse Singal describes the influence of this paper (cited over 3,000 times) on the field [393]:

> [the] concept went viral, leading to book contracts, speaking engagements, and other professional rewards for Lyubomirsky. Seligman transformed it into a "happiness formula" in his own work: H = S + C + V. That is, happiness, H, equals S (genetic set point) plus C (circumstances) plus V (things under the individual's voluntary control). In part on the basis of Lyubomirsky's finding, he argued that there was a great deal of potential for the average person to become significantly happier.

More than a decade later, Nick Brown took a critical look at this work in this 2019 paper

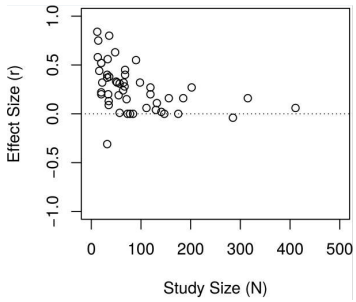


Figure 2-6: Figure 2 from 'Meta-analyses of positive psychology interventions: The effects are much smaller than previously reported' [454]. This is a re-analysis of the studies included in a previous meta-analysis conducted by Sin and Lyumbomirsky.

'Easy as (Happiness) Pie? A Critical Evaluation of a Popular Model of the Determinants of Well-Being' [67]:

> We conclude that there is little empirical evidence for the variance decomposition suggested by the "happiness pie," and that even if it were valid, it is not necessarily informative with respect to the question of whether individuals can truly exert substantial influence over their own chronic happiness level.

2.3.4 Positive Psychology Interventions

In a 2019 meta-analysis entitled 'Meta-analyses of positive psychology interventions: The effects are much smaller than previously reported', author Carmela White conducts a thorough re-evaluation of the field [454]. A small study bias is quite evident in previous meta-analysis work (Figure 2-6). White and her co-authors summarize their results as follows:

> the reanalyses and replications of Sin and Lyubomirsky [22] and Bolier et al. [23] indicate that there is a small effect of approximately r = .10 of PPIs on well-being. In contrast, the effect of PPIs on depression was nearly zero when based on the studies included in Sin and Lyubomirsky [22] and highly variable, and sensitive to outliers, when based on studies included in Bolier et al. [23]. Notably, Sin and Lyubomirsky [22] included nearly twice as many studies as Bolier et al. [23] in their meta-analysis of the effects of PPIs on depression.

This is revised down from effect sizes almost 3 times as large. Interventions of this type have barely demonstrated an impact on well-being and depression.

Furthermore, in a 2020 meta-analysis focused on gratitude interventions [94] the authors conclude: "the effects of gratitude interventions on symptoms of depression and anxiety are relatively modest. Therefore, we recommend individuals seeking to reduce symptoms of depression and anxiety engage in interventions with stronger evidence of efficacy for these symptoms."

Intervention studies based on positive psychological principles suffer from similar methodological problems that led to the replication crisis in the first place– under-powered work with high variance and small sample sizes, combined with publication bias and file drawer effects. There is currently little empirical support for interventions based on the principles championed in the positive psychology movement.

2.3.5 Mindfulness Interventions

Similar to positive psychology interventions, mindfulness interventions have started to come under fire for a lack of rigor (see i.e. 'Has the science of mindfulness lost its mind?' [144]). In a 2017 article in Perspectives of Psychological Science, fifteen psychologists penned an editorial entitled 'Mind the Hype: A Critical Evaluation and Prescriptive Agenda for Research on Mindfulness and Meditation' in which they argue that mindfulness is both a poorly-defined and difficult-to-measure concept whose "widespread use is premature" [438].

> ...there is a common misperception in public and government domains that compelling clinical evidence exists for the broad and strong efficacy of mindfulness as a therapeutic intervention (e.g., Coyne, 2016; Freeman Freeman, 2015). Results from some clinical studies conducted over the past 10 years have indicated that mindfulness-based cognitive therapy (MBCT) may be modestly helpful for some individuals with residual symptoms of depression (Eisendrath et al., 2008; Geschwind, Peeters, Huibers, van Os, Wichers, 2012; van Aalderen et al., 2012).

As a consequence of select results, published in high-profile journals, MBCT is now officially endorsed by the American Psychiatric Association for preventing relapse in remitted patients who have had three or more previous episodes of depression. Moreover, the U.K. National Institute for Health and Clinical Excellence now even recommends MBCT over other more conventional treatments (e.g., SSRIs) for preventing depressive relapse (Crane Kuyken, 2012). Mitigating such endorsements, a recent meta-analysis found that mindfulness-based stress reduction did not generally benefit patients susceptible to relapses of depression (C. Strauss, Cavanagh, Oliver, Pettman, 2014). Other meta-analysis have suggested general efficacy of MBIs for depressive and anxious symptoms (Hofmann, Sawyer, Witt, Oh, 2010), though head-to-head comparisons of mindfulness-based interventions (MBIs) to other evidence-based practices have resulted in mixed findings, some suggesting comparable outcomes, others suggesting MBIs might be superior in certain conditions, and others suggesting cognitive behavioral therapy is superior in certain conditions (e.g., Arch et al., 2013; Goldin et al., 2016; Manicavasgar, Parker, Perich, 2011). There is also mixed evidence comparing MBIs to interventions such as progressive muscle relaxation (e.g., Agee, Danoff-Burg, Grant, 2009; Jain et al., 2007). Direct comparisons of MBIs to empirically established treatments are limited.

In a recent review and meta-analysis commissioned by the U.S. Agency for Healthcare Research and Quality, MBIs (compared to active controls) were found to have a mixture of only moderate, low, or no efficacy, depending on the disorder being treated. Specifically, the efficacy of mindfulness was only moderate in reducing symptoms of anxiety, depression, and pain. Also efficacy was low in reducing stress and improving quality of life. There was no effect or insufficient evidence for attention, positive mood, substance abuse, eating habits, sleep, and weight control (Goyal et al., 2014). These and other limitations echoed those from a report issued just 7 years earlier (Ospina et al., 2007). The lack of improvement over these 7 years in the rigor of the methods used to validate MBIs is concerning; indeed if research does not extend beyond Stage

> 2A (comparison of MBI to wait-list control), it will be difficult, if not impossible, to ascertain whether MBIs are effective in the real world (cf. Dimidjian Segal, 2015). On balance, much more research will be needed before we know for what mental and physical disorders, in which individuals, MBIs are definitively helpful.

Few replication attempts for mindfulness based interventions have been attempted; some targeting on relapse prevention have failed [208] while some have succeeded [272].

2.3.6 The Paradox of Choice

Barry Schwartz wrote a book and delivered a famous TED talk on the 'Paradox of Choice'– more choices can paradoxically lead to decision fatigue, regret, and choice paralysis. You may never feel satisfied that you make the best choice when you have too many to consider. He states: "though modern Americans have more choice than any group of people ever has before, and thus, presumably, more freedom and autonomy, we don't seem to be benefiting from it psychologically" [384].

The study that is at the center of much of this narrative comes from the first study to investigate the concept, by Sheena Iyengar [219]. A tasting table for exotic jams is placed in a grocery store; in one condition, 24 jams are offered, and in the other only six. While more people came to examine the larger group, only 3% purchased a jam compared to 30% of customers confronted with the small assortment.

Another of her conditions, however, contradicts the 'paradox of choice' concept– she showed that people were more satisfied with their choice when choosing from 30 exotic chocolates instead of six, though they found the decision more frustrating.

Psychologist Ben Scheibehenne was intrigued by this, and conducted a meta-analysis of 50 experiments on the topic, with over 5,000 participants [374]. He found that "[t]he overall mean effect size across 63 conditions from 50 experiments in our meta-analysis was virtually

zero. On the basis of the data, no sufficient conditions could be identified that would lead to a reliable occurrence of choice overload."

He suggests more choices are definitely better when the choice is about quantity or when preferences are well-defined; there is no broadly applicable choice-overload effect. However, there were significant results in several of the experiments; they just fell equally on positive and negative sides of the average. For certain contexts, more choices may significantly improve or diminish our appraisal of our decisions.

The average effect size of zero in this meta-analysis led to a few high-profile articles challenging the concept in the Atlantic and the Financial Times; Barry Schwartz responded with a fair analysis of the original paper on PBS NewsHour [385]. Sometimes, more choices are better; sometimes they're worse; a lot of the time, they don't matter. We have yet to understand when and where these effects apply.

2.3.7 Hedonic Adaptation

One of the first, and most widely cited studies on hedonic adaptation– the idea that we have a setpoint for happiness, and regardless of the events in our life we end up at the same level of enjoyment– is a 1978 study called 'Lottery Winners and Accident Victims: Is Happiness Relative?' [64] This study shows relatively small differences in self-reported happiness between paraplegics and lottery winners in the first months after the event. This paper entered the popular consciousness when it served as a centerpiece of Dan Gilbert's famous TED talk on happiness.

This particular finding doesn't seem to be true as it's been presented. The lottery winners in the study didn't win 'life-changing money' (only 23% reported lifestyle changes). The life-satisfaction scores immediately after these life events showed parapalegics were, as you'd expect, significantly worse off. Moreover, Diener et al. write in 'Beyond the hedonic treadmill: Revising the adaptation theory of well-being' [115]: "Lucas (2005a) used two large, nationally representative panel studies to examine adaptation to the onset of disability. Participants in this study (who were followed for an average of seven years before and seven

years after onset) reported moderate to large drops in satisfaction and very little evidence of adaptation over time. For instance, those individuals who were certified as being 100% disabled reported life satisfaction scores that were 1.20 standard deviations lower than their nondisabled baseline levels."

Overall, however, a softer idea of hedonic adaptation is generally supported in the literature; it just requires several additional caveats. The above paper points out several key revisions that weaken the overall concept. Most people are happy most of the time, though there is high variability in setpoint across individuals due to temperament. Well-being setpoints might be heritable and distinct from overall life satisfaction setpoints. (Happiness is poorly defined as a unitary psychological concept, after all.) Individuals appear highly variable in how much and how quickly they adapt to new circumstances.

2.3.8 Agency and Well-Being

The experimental case for agency interventions is exceedingly weak. The main source of experimental work on agency comes from Harvard's Ellen Langer, who claims in her book 'Counterclockwise' [253] that being in charge of watering your plant at the nursing home makes you live longer (vs. simply having a plant that someone else cares for). Harvard's Dan Gilbert rehashes this study in his book 'Stumbling on Happiness' [174]:

> In one study, researchers gave elderly residents of a local nursing home a houseplant. They told half the residents that they were in control of the plant's care and feeding (high-control group), and they told the remaining residents that a staff person would take responsibility for the plant's well-being (low-control group). **Six months later, 30 percent of the residents in the low-control group had died, compared with only 15 percent of the residents in the high-control group.**

The shape of this anecdote should raise flags– control of watering a houseplant does not make you 2x more likely to die. In this example, participant age ranged from 65 to 90; and

this difference (6 deaths out of 45 people) doesn't control for age or health status prior to the intervention.

While self- and nurse-reported surveys indicate there may have been some real health differences– relative to baseline– between the two groups, this stark mortality statistic actually calls more plausible differences between the two groups (based on questionnaires) into question. If an additional six participants died in the control arm, it is likely that their health and age deteriorated more precipitously during the study for reasons other than the house plant intervention.

Gilbert uses a second, unrelated study [382] to reiterate his point about the importance of control, in which elderly adults could choose visitation times from college students (control), were simply told when to expect a visit (predictable), or happened to receive random visits (random).

Gilbert suggests again that, after this study, a "disproportionate number of residents who had been in the high-control group had died....[because they] were inadvertently robbed of control when the study ended." The paper actually reports no effect of control at all during the study ("the positive outcome of the predict and control groups is attributable to predictability alone.") and declares no link to mortality risk upon its conclusion (additionally, the risk they discuss was related to the lack of visitation, nothing to do the lack of control).

2.3.9 Growth Mindset Interventions

Carol Dweck is an eminent psychologist known for coining the idea of 'Growth Mindset'– the idea that your belief that you can grow your abilities and skills through effort (as opposed to a belief that you are fixed or limited by your genetic ability) is a key determinant of your success. For Dweck, this belief is the bedrock on which resilience and grit are built– challenges feel surmountable and enriching if you believe yourself to be growing as a result of your effort; without that belief, failure becomes a referendum on who you are and what you can accomplish.

The idea has become a popular one; interventions to teach growth mindset have been rolled out at large scale. In their wake there has been some controversy over their efficacy. A 2018 Meta-analysis by Brooke Macnamara of over 300 interventions (conducted on over 400,000 students) [396] concluded "overall effects were weak ... However, some results supported specific tenets of the theory, namely, that students with low socioeconomic status or who are academically at risk might benefit from mindset interventions." In two separate analyses, 64% and 86% of the interventions under review had no (or negative) impact on academic achievement; they also found evidence of publication bias, so they caution that the reported results are likely over-estimates.

Carol Dweck responded [129] by suggesting these mild effects were important and good value (given how cheap and quick mindset interventions are to implement), and with a 2018 replication (reported in Nature) [459] that showed 3% fewer high schoolers were off-track for graduation in the intervention group (N=12,542), with low-achievers and schools that more effectively support social norms showing the biggest effect. Much to Dweck's credit, this replication was pre-registered and randomized in 65 schools with a good control arm.

This result comes from a very minimal intervention– two, 20 minute sessions. The data largely agreed (empirically) with Macnamara's meta-analysis. At the same time, a similar large scale randomized controlled trial from England in 101 schools across the country (N=5,000) ran a much stronger intervention (eight consecutive weeks of 40 min to 2 hours a week, teacher training, and material for teachers to embed in their curricula). Similar age students showed negative or no difference from controls across the three main measurements– math, reading, and grammar.

In 'The One Variable that Makes Growth Mindset Interventions Work', Russel Warne points out– after reviewing a handful of the large scale trials of Growth Mindset Interventions– that the only ones that have successfully captured an effect are those run by Carol Dweck herself [444]. This is a known failure mode when scaling research interventions– having a team that specializes in the concept under study, communicates it well, believes in it, and has experience can dramatically improve the quality of what's delivered. It's another strike against the general idea of Growth Mindset Interventions (and instead a point for Carol

Dweck's specific Growth Mindset Interventions). As Macnamara stated in her article, these small effect sizes should serve as an upper bound.

The initial back and forth between Dweck and Macnamara was covered in Scientific American in their article 'Debate Arises over Teaching "Growth Mindsets" to Motivate Student' [113]. It falls on the side of Dweck, taking her reported effect sizes as accurate.

2.3.10 Summary

Positive psychology attempts to distance itself from a humanistic tradition of studying the well-lived life with its focus on 'empiricism.' Unfortunately, that has meant an epistemological shift in how work is conducted and evaluated; a shift that places too much faith in simple statistical tests and too much value on oversimplifications and contrivances that allow those tests to be used. By doing so, it abandoned the complex richness and deep intuition characteristic of the leading psychological minds of the preceding era– an epistemic approach much better suited to the task.

2.4 ...and more broadly

The preceding sections target the relevant parts of the replication crisis for my work; but it is worth briefly contextualizing how deep the crisis runs for the entirety of science.

2.4.1 Mainline Psychology

A look back on the history of experimental psychology shows that even the most basic, canonical findings are now facing skepticism (if not for p-hacking, than for the way they've been interpreted and simplified). Examples include:

- *The Milgram Experiments.* These famous experiments have been under renewed scrutiny for methodological issues. The results– while real– don't support Milgram's

'agentic state' interpretation (that we cede our moral agency to authority figures) [356].

- *The Stanford Prison Experiments* Reexamination of this famous experiments– purporting to show that we descend into abusive behavior given the right conditions– have shown that the experimenters coached the desired behavior out of the 'prison guards' [193].

- *The Asch Conformity Experiments* A famous experiment in which people are asked simple questions in a room full of confederates (who all intentionally answer incorrectly), this study shows 75% of people will conform at least some of the time. A 2015 study showed that the results of this study have been incorrectly overblown in 19 out of 20 major textbooks [183].

- *The Marshmallow Test* The results of this experiment hold– delayed gratification predicts future success– however their interpretation about why has been questioned. Does delayed gratification imply about executive functioning or more about childhood environment [451]? Even the critique has come under fire.

- *The Pygmalion ('My Fair Lady') Effect* This effect suggests that teachers' expectations account for real improvements in IQ. It turns out there are some serious methodological flaws in the original paper– and though expectancy effects are real, intelligence doesn't seem to be altered by a teacher's beliefs [431].

- *Robber's Cave* This famous experiment showed tribes of young boys cooperating when resources were plentiful, and attacking each other when resources were scarce. It turns out the whole experiment was contrived and took three tries and a lot of coaching to get the desired result [292].

- *The Bystander Effect* The story of Kitty Genovese's death in NYC lead to the creation of '911' to call the police. While she was murdered, almost no one witnessed it– the story was a fabrication by the NY Times editor Abe Rosenthal. Replications show the effect is real but small (r=0.2) [360].

- *Subliminal Advertising* The idea of subliminal advertising gained national attention in the wake of a 1957 study by James Vicary, in which he claimed imperceptible flashes of 'eat popcorn' and 'drink soda' during a movie resulted in sales increases of 18 and 58 percent. Vicary made $4.5 million as an advertising consultant in the wake of his fraud. Subliminal advertising still appears in the popular press as a real thing despite a complete lack of evidence for it.

- *Skinner Box Addiction* This study showed rats would forgo food and press a lever to death if it stimulated their pleasure centers; framing addiction as insurmountable once it is introduced to an environment [383]. This was famously challenged by the 'Rat Park' experiment, which showed it was the unnatural, isolating environment that caused rats to do this, not the drugs (rats in a pro-social rat park don't repeat the behavior) [157]. This study also suffered from a spotty replication record [335], but holds up conceptually (i.e. based on heroin use statistics in Vietnam veterans [189]).

2.4.2 AI

Generative Adversarial Networks (GANs) are a model architecture invented by Ian Goodfellow in 2014 that are responsible for numerous, impressive results. In 2018, a colleague of mine at Google published a paper comparing the results of all stated improvements to the original GAN paper, testing them over all sets of hyperparameters and on all standard test datasets. His results show that none of the published, claimed improvements on the original GAN architecture were actual improvements; they were overfit to the hyperparameters or test set used in the paper [270]. This result is corroborated by another paper benchmarking advances: "Deep metric learning papers from the past four years have consistently claimed advances in accuracy, often more than doubling the performance of decade-old methods... the actual improvements over time have been marginal at best" [311].

This raises some interesting questions for a field enabled by big data. When competitions are run for years, and multiple teams make small tweaks to the same fundamental architectures, the possibility for spurious/over-fit improvements starts to become a real problem.

Moreover, smaller research teams can't run huge tests like Google (comparing all architectures over all hyperparameters over all datasets). The types of assertions should probably be probabilistic, rather than the deterministic benchmarking that is standard practice today. This view is starting to gain traction; in 'Deep Reinforcement Learning at the Edge of the Statistical Precipice' [19] Agarwal et al. introduce the *rliable* tool to introduce uncertainty into metric reporting for reinforcement learning.

2.4.3 Medicine

> The case against science is straightforward: much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, science has taken a turn towards darkness.
>
> Lancet Editor Richard Horton
> 2015

The Lancet is one of the most respected journals in medical science; as we can see above, its editor has a pessimistic outlook on the quality of the science contained therein. He's not alone; Marcia Angell– previous editor of the New England Journal of Medicine– said this in 2009: "It is simply no longer possible to believe much of the clinical research that is published, or to rely on the judgment of trusted physicians or authoritative medical guidelines. I take no pleasure in this conclusion, which I reached slowly and reluctantly over my two decades as editor of The New England Journal of Medicine" [188]. That same year, Richard Smith, 25 year editor at BMJ, wrote: "Sadly we also know— from hundreds of systematic reviews of different subjects and from studies of the methodological and statistical standards of published papers— that most of what appears in peer-reviewed journals is scientifically weak" [399].

Medicine has also been ground-zero for the replication crisis, kicking off in 2005 with the famous publication of "Why Most Published Research Findings are False" by Stanford

Medical School's John Ioannidis. The low replication rate doesn't affect our quality of care because we have a rigorous approval process and a large pharmaceutical market to vet academic research. These companies– like Bayer and Amgen– report replication success of academic work between 11 and 24% [344, 49]. Major replications attempts show this problem is particularly stark in cancer research. [140]. This terrible track record for pre-clinical research costs an estimated $28 billion per year [154].

2.4.4 Neuroscience

fMRI is an imaging technique that looks at blood flow in the brain, with incredible power to tie high resolution images of brain activity to mental function. Unfortunately, many of the studies are also based on small sample sizes, and the statistical tools to run these studies can be opaque. In 2009, Dartmouth's Craig Bennett published a famous, tongue-in-cheek (but very real) study showing that a dead salmon had active voxels of brain activity– a statistically significant difference in their two (humorous) conditions– when they were shown different pictures of humans [54]. In an interview about his work, Bennett stated: "By complete, random chance, we found some voxels that were significant that just happened to be in the fish's brain, and if I were a ridiculous researcher I'd say, "A dead salmon perceiving humans can tell their emotional state."" [274] This 'Red Herring (Atlantic Salmon)' study has become a famous cautionary tale about statistical methods in fMRI research.

These issues have taken the spotlight again in 2020, with a big paper in Nature showing that the same fMRI data analyzed by 70 research teams– for the same pre-selected nine hypotheses– showed huge disagreement over which results were 'significant' [61]. Moreover, a team from Sungkyunkwan University in South Korea has shown that many successful fMRI 'replication studies' take advantage of the fact that– frequently– very large brain areas are functionally implicated, making them very 'forgiving'. In their paper entitled 'False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding' they recommend a pattern matching approach to test replicability [205]. These neuroscientific replication issues have been detailed by a large European team as well [60].

This might be the least of the issues for the 'social' branch of neuroscience, which attempt to tie behavioral or mental states to functional imaging. A 2020 meta-analysis has shown test-retest reliability of social neuroscientific studies land at about 0.4, "which is below the minimum required for good reliability (ICC = .6) and far below the recommended cutoffs for clinical application (ICC = .8) or individual-level interpretation (ICC = .9)" [135]. Other papers have found errors in common fMRI analysis software that lead to a false positive discovery rate of 70% with an engineered dataset that should give only 5% false positives [133], or pointed out that the correlations reported in many of these papers are above what is possible given well-characterized variability of the underlying personality/emotion data and imaging techniques [441].

2.4.5 Summary

The reproducibility crisis affects a variety of disciplines, from AI to medicine to neuroscience to psychology. A healthy statistical intuition about the system level incentives and reporting dynamics– i.e. basic meta-science– is crucial to parse the literature accurately. Tools to mitigate the crisis will become more important as the negative impacts become better publicized and better understood.

Chapter 3

Methodological Implications

In the last chapter, we looked at many of the mistaken findings that have been promoted as a result of the replication crisis. In this chapter we'll talk about the methodology that led us here, the underlying philosophical issues they stem from, and their implications for psychology.

Many people believe that the replication crisis starts and ends with a simple misapplication of statistics. In actuality, it is the fine point of a *deeper* crisis– a much broader debate about statistical philosophy with implications for general scientific practice at multiple levels– as well as a *broader* one– it's one of many methodological challenges facing experimental psychology.

In this chapter, we briefly look at the surface level problems of replication before discussing the deeper issues of a Humean philosophy that attempts to side-step causal intuition and prior belief. This mistaken perspective underlies the broad misconception that Popperian falsifiability is the correct approach to theory building. Once we dispel with this notion, the scientific endeavor looks more like a process of formalizing how we reason at scale– a direct analog to the goals of artificial intelligence.

This shift in methodological perspective has important implications for the broad issues facing psychology, a discipline which has long served as the lynch-pin for debate in philos-

ophy of science.[1] Even if we solve the statistical problems,[2] there are still broad challenges of generalizability (*i.e. When and how does what we observe apply to other contexts, stimuli, and people? What moods, dispositions, and personalities is it true for?*) and ontology (*How do we describe what makes two people, situations, or stimuli more similar or more distinct?*) to contend with.

We can thus marry insight from a few disciplines to improve our scientific practice in psychology. We can treat scientific theories as data generating models, and treat ontological claims as scientific theories. We can also move towards statistical frameworks that allow for stronger modeling assumptions (required when data is scarce/contextual) and that incorporate causality and uncertainty as first class primitives (matching the underlying structure that we are modeling).

3.1 The Problem with P-values

In [376], Ulrich Schimmack taxonomizes many popular academic perspectives minimizing the replication crisis: that 'crisis' is an overly dramatic label, that these issues are not endemic, that replications themselves are flawed or fail to accurately reproduce original findings, or that results regress to the mean or shift over time. Schimmack writes:

[1] Karl Popper famously advocated for falsifiability because he didn't think Freudian psychoanalytic theory should be considered 'science' (the so-called Demarcation Problem). Kuhn credits Piaget with his inspiration for paradigm shifts (Piaget– a great psychologist– "would probably be rejected from most modern journals on methodological grounds of sample size, non-standard measurement, and lack of inter-rater reliability" [243].) Within psychology, there is debate between humanistic and positive psychology about the role of empiricism for claims about flourishing; a debate that echoes a long history of similar methodological debate in psychology. Kurt Lewin famously discussed the interactions of people and environments in the first half of the twentieth century, and described much of modern psychological practice as 'like trying to build a model of gravity by studying balls rolling down inclined planes and simply averaging the results' (paraphrased). In the 1970s, Walter Michel famously argued against personality (Big 5) as useful for prediction in 'Personality and Assessment' which led to a dramatic decrease in experimental personality work in favor of social psychology's focus on contextual factors. Meta-analysis suggests environments have fared worse as a predictive tool. Almost all of the great psychologists have had a clearly articulated and novel epistemic perspective; much of the richest innovation in a field this young and complex comes at this level of reflection.

[2] i.e., if we were to continue with categorical notions of 'statistical significance', that 'significant' results replicate roughly 80% of the time in line with suitably powered testing; reported effect sizes contain unbiased variance about the true underlying effect size; base rates of 'statistically significant' results in publication match the actual base rates of true hypotheses.

> In an opinion piece for the New York Times, [the] current president of the Association for Psychological Science commented on the OSC results and claimed that "the failure to replicate is not a cause for alarm; in fact, it is a normal part of how science works". On the surface, [she] makes a valid point. It is true that replication failures are a normal part of science. If psychologists would conduct studies with 80% power, one out of five studies would fail to replicate, even if everything is going well and all predictions are true. Second, replication failures are expected when researchers test risky hypotheses (e.g., effects of candidate genes on personality) that have a high probability of being false. In this case, a significant result may be a false-positive result and replication failures demonstrate that it was a false positive. Thus, honest reporting of replication failures plays an integral part in normal science, and the success rate of replication studies provides valuable information about the empirical support for a hypothesis. However, a success rate of 25% or less for social psychology is not a sign of normal science, especially when social psychology journals publish over 90% significant results (Motyl et al., 2017; Sterling, 1959; Sterling et al., 1995). This discrepancy suggests that the problem is not the low success rate in replication studies but the high success rate in psychology journals.

This failure to understand the fundamental statistical issues is pervasive. In his paper 'Statistical Rituals', Gerd Gigerenzer suggests– contrary to theorists who point to misaligned system-level academic incentives– that the primary failure mode is a fundamental misunderstanding of p-values. He also points to the 'null ritual'– and the mindless application of formulaic null hypothesis significance testing (NHST)– as the pervasive pedagogical misstep that perpetuates this crisis. [172]

The data bare him out– interpreting p values as 'the probability of success in replication' is a pervasive misunderstanding which exists "among 20% of the faculty teaching statistics in psychology, 39% of the professors and lecturers, and 66% of the students." In a meta-review of research on this topic, the data suggest between 56%-80% of statistics and methodology instructors in psychology departments hold at least one major misconception about null hy-

pothesis testing; that increasing to 74%-97% for other university level psychology educators [173].

Modern statistical practice for journal acceptance centers around $p<0.05$ as a categorical line between 'significance' and 'non-significance', but the $p<0.05$ threshold was never meant to be a categorical cutoff. Ronald Fisher himself– the father of modern statistics and creator of null hypothesis testing– stated about p-values: "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas" [148].

Fortunately, it seems there is momentum to end an era of categorical thinking in statistics. In a series of 2019 editorials in Nature and the American Statistician, the American Statistical Association released a special issue with over 800 researchers signed on to put an end to the categorical term 'statistically significant' [448, 449, 131]:

> In 2016, the American Statistical Association released a statement in The American Statistician warning against the misuse of statistical significance and P values. The issue also included many commentaries on the subject. This month, a special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on 'Statistical inference in the 21st century: a world beyond $P < 0.05$'. The editors introduce the collection with the caution "don't say 'statistically significant'". Another article with dozens of signatories also calls on authors and journal editors to disavow those terms.
>
> We agree, and call for the entire concept of statistical significance to be abandoned. [27]

Their analysis of of 791 articles in 5 major journals (mostly outside of psychology) showed that 49% included conclusions or text that betray a fundamental misinterpretation of this concept.

3.1.1 The Correct Way to Think

Null Hypothesis Significance Testing and corresponding p-values are useful. The problem comes from (1) the reduced emphasis on statistical measures that matter, (2) their categorical interpretation, and (3) the misuse of language around 'statistical significance.'

P-values depend on sample sizes just as they do on effect sizes; the conflation of language between 'statistically significant' (implying the effect is likely to be real and not an artifact of noise) and 'significant' (implying it is meaningful and important) works against the pursuit of good science. Results that are significant (as in they matter) should be judged instead by the effect size (d) or the normalized effect size (Cohen's d).

Gene Glass– the statistician who invented the meta-analysis- writes "statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude– not just 'Does a treatment affect people?' but 'How much does it affect them?'" Jacob Cohen– for whom Cohen's D was coined– states: "The primary product of a research inquiry is one or more measures of effect size, not P values."

Interpreting P-values

As we know from introductory statistics, the p-value is simply the probability that the results we see were generated by noise. With a large enough sample size and with enough variables controlled, trivial and contingent relationships can be 'proven true' even when they have no bearing on the real-world.[3]

The p-value is *not* the probability that a result is true, which depends on how likely a hypothesis is to be true before testing. The probability that a hypothesis is true– known as the positive predictive value (PPV)– is very frequently conflated with the p-value itself.

[3] Jacob Cohen makes this point– that statistically speaking, the difference of means sampled from two real sub-populations will never be perfectly zero– in his 1990 article 'Things I Have Learned (So Far)'. He states that "[t]he null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world... If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal about rejecting it?"

The PPV is defined by the equation $(1-\beta)*R/((1-\beta)*R+\alpha)$ where α is our threshold, β is our study power, and R is our odds ratio (or prior); any one of these four quantities is completely defined by the other three. We can apply this formula to get the number we really wish to use the p-value for– an assessment of how likely a result is to be true given we meet our α criteria, or $P($ real effect $= 1 \mid$ significance $= 1$ $)$.

3.1.2 Real World P-Hacking

Unfortunately, applications of the PPV are guesses, as we never actually know these quantities in the real world. Moreover, if we try to apply this formula directly, we must assume we have direct access to all evidence (or at least an unbiased sample of it). In the real world, our evidence is obscured due to p-hacking, file drawer effects, and publication bias. Our ideal evidentiary value $P($ real effect $= 1 \mid$ significance $= 1$ $)$ needs to be considered against the backdrop of $P($ significance$_{actual} = 1 \mid$ significance$_{observed} = 1$ $)$ in the real world, which is unknown but certainly very low due to biases in scientific reporting.

P-hacking takes many forms. The typical example is a researcher testing increasingly obscure or combinatorial hypotheses until one hits (i.e. 'subliminal advertising works, but only if people are primed to be thirsty first and only if the brand is not a major coca-cola brand' [440]). These can be easy to spot as the hypothesis was clearly not generated a priori.

More pernicious forms of p-hacking exist however; we might imagine one of our researchers pre-processing the data a few different ways, re-engineering features, and trying several different models to test the same hypothetical underlying relationship. They run slight variations of the same test, taking 10 or 100 or 1000 'forking paths' to examine one hypothesis and reporting significance if any one hits. In this case, the hypothesis might appear like a reasonable one; to spot this kind of behavior, it's important to assess the data analysis techniques against your own intuition and consider how standardized they are. Variations on standard techniques, dubious treatments of outliers, shifting procedures across similar studies, and/or a large independent variable space can be signs that this has happened;

pre-registration of the methodology guards against this.

Finally, a researcher might look at the data, see a relationship, and create a hypothesis that perfectly explains it. Here, there might actually be just one hypothesis under test, and it may seem like incredibly powerful evidence when it is not. For example– when predicting divorce– finding a correlation in personality traits or communication styles that is consistent with 5 couples out of 10 after you know which ones divorced is trivial; there are many sets of traits that will separate any two arbitrary groups of five couples. Predicting the specific relationship beforehand, however, is much more difficult and evidence of incredible insight. Conflating exploratory and confirmatory work by hypothesizing after the results are known (HARKing) has been a source of confusion in the popular discourse around researchers like John Gottman, who did exactly this experiment (the first FAQ on his website, which address his ability to predict divorce, states: "statements about the 94% accuracy rate of divorce prediction have become a source of confusion. What Dr. Gottman is able to say is that a particular couple is behaving like the couples that were in the group that got divorced in his 1992 study (Buehlman, K., Gottman, J.M., Katz, L.)" [3]). Three significant findings are completely unsurprising if we test 1000 theories— they are probably a result of noise. Three significant findings out of three hypotheses, on the other hand, is strong evidence that the theory contains deep, accurate insight.

These forms of p-hacking are harder to spot and easier to inadvertently fall prey if you are not disciplined as a researcher. As a research consumer, we must try and spot this kind of behavior in the papers we read. To do so requires an evaluation of the researcher's methods; are they deeply aware of how to use statistics or are they 'motivated researchers'? We actually have a lot of clues when evaluating an academic's work– do they seem to know the results before they collect data? Are their data analysis techniques consistent and in-line with common sense? Do they appear aware of these statistical issues? Researchers who pre-register, choose sample sizes based on a power analysis, revise their alpha levels, report null results and full p-values, and interrupt their results accurately and conservatively probably use a trustworthy method; researchers who only use N=20, over-interpret their findings in a flashy way, and consistently report strong significance for counter-intuitive effects should

raise suspicion. Even conscientious researchers doing high-power research are probably not powering their studies to 100%, the recommendation is 80%; that means even if they only study real effects, we'd expect one null result for every five hypotheses they test.

P-hacking at the level of the researcher is only one form of this 'biased observer' issue– we also have file-drawer effects and publication bias. We can imagine 10 or 100 researchers all testing roughly the same hypothesis; because journals only like to publish 'interesting' results, only the one with a significant result will get published (publication bias). This publication bias causes researchers to discard their null results instead of bothering to submit them (the file-drawer effect).

This kind of system-level bias occurs most frequently when a topic enjoys media success and becomes popular to fund; one of the most covered replication-failures of this nature is stereotype threat (as discussed in Chapter 2). Evidentiary value from papers on popular topics need to be severely discounted in light of this.

3.1.3 What to Do with Existing Research?

Looking Backward

This is a pretty damning critique of the state of scientific evidence; it's very hard to separate sound research from unsound work. Fortunately, it is possible to evaluate literature critically and identify useful experimental insights. As we recall from the introduction to this crisis, we can trust at least 25% of research, and people are generally quite good at predicting that subset. We simply need to bring a critical eye to the work we read; evaluate the power of the study, the quality of the methodology, and the plausibility of the conclusions. We need strong priors– informed by meta-statistical analysis and replication work– as we integrate new data into our beliefs. Individual papers can still be very informative if we put in the effort to critically evaluate them.

On a system level, there is actually a fair amount to be extracted from the literature. We can evaluate patterns in the data over journals, fields, individual researchers, and topics to

identify bias. Two of the most useful, standard tools for this are funnel plots and p-curves.

Funnel Plots We expect, if we plot sample size on one axis and effect size on the other, to get a triangle– more power in the study will get us closer to the real effect size, less powerful studies should have unbiased variance around the real effect size. Failure to match this pattern gives quantitative evidence of bias in the literature) [410]. See the example in Figure 3-1.

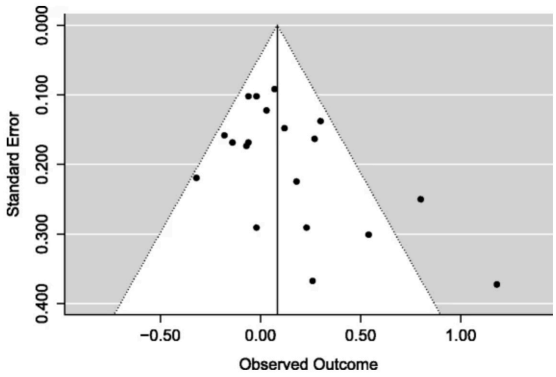


Figure 3-1: Example funnel plot for a meta-analysis of the Pygmalion effect (how teacher expectations affect student IQ) from [390].

P-Curves We expect, if a venue is unbiased, that the distribution of reported p-curves should follow certain trends (it is uniformly distributed when testing null effects, and logarithmic when testing real effects at high power). If we see a big spike just below p=0.05 or p=0.01 and few results below that, it represents clear evidence that people are rolling the die many times and only reporting results that just fall below those cutoffs. Any left-skew in p value distributions is inconsistent with either a null effect or a real one [392].

There are other tools that are proposed for meta-statistical analysis of bias:

(1) Ulrich Schimmack has created the Z-Curve [47], which estimates the replication rate of a researcher and ranks journals and researchers. It creates this replication rate by comparing the fraction of significant results reported (the observed discovery rate) with the expected true discovery rate (which relies on a post-hoc estimate of study power). We can estimate the average power of a body of work by looking at the distribution of p-values; we take significant results, and fit a curve to them to estimate the true power. Any discrepancy

between the expected and reported discovery rates is evidence of bias. In my experience looking at results from this test, it's a very noisy estimate of research quality and bias.

(2) There are also forensic tools that automate simple checks for implausible data like SPRITE or GRIM [194, 66]. These are surprisingly useful– while outright fraud represents a tiny fraction of research output, mistakes are not uncommon (for instance, 30% of Genomics papers are believed to contain excel auto-correct errors based on a study of 10,000 papers [260]).

Taken together, there is a rich community of meta-statisticians that run analyses, correct results, and publish accessible articles summarizing the quality of research. One of the best places to start to gain expertise in these disciplines is the blogosphere.

The Continued Problem with Replication

One major part of helping us to understand past work is the practice of replication. Replication will always be important to bring us closer to a true understanding of an effect size and measurement error. If we continue to carve out a place for NHST– which may be a reasonable thing to do in certain circumstances– we'd expect even high-power, unbiased research to result in Type I and Type II errors at relatively high rates.

Current replication attempts, unfortunately, perpetuate the flawed assumptions they seek to correct. They imply a categorical truth value on a topic by reinforcing the notion that 'successful' replications are once again 'statistically significant' while 'unsuccessful' replications are not. Of course we shouldn't be thinking this way at all; we should be looking at the accumulated literature on a topic and use all of it to better predict the real effect size (and, in the background, many meta-statisticians are doing just that).

Unfortunately, we still are retro-fitting proper statistical formulations into the entrenched categorical simplifications that underpin modern scientific practice. Because the evidence in prior literature is so distorted, 'replication' becomes shorthand for a pre-registered, very high-power study whose methodology we can trust; its 'success' or 'failure' becomes short-

hand about the trustworthiness of the original evidence when we re-apply the original flawed statistical categories.

Moving Forward

There are certain things we can do as individual researchers to avoid the statistical missteps of recent work.

As individual researchers, we have an obligation to use and promote sound research practices; commit to our data processing beforehand, pre-register our studies, open our datasets, rigorously divide exploratory and confirmatory work, interpret our results cautiously and correctly, and publish all of our results. It's important to go in with a true 'scientific' mindset– asking an earnest question of the data– rather than a rhetorical or motivated one.

Pre-registration and open data, while a step in the right direction, still leave plenty of researcher degrees of freedom on the table. Though they doesn't solve the problem, their practice brings methodological awareness into the spotlight and exerts subtle pressure to commit to both hypotheses and data processing techniques a priori. It also serves as a signal to the broader community that you're aware and care about these issues, increasing external trust in your work.

It's also important to teach statistical intuition to new researchers; to establish research programs that promote and reward high quality methodology and transparency. Much of the work we need to do involves pressuring journal editors and stakeholders who design system-level incentives to require these changes. NIH is leading the way in how they've addressed the estimated $10-50 billion [39] lost to irreproducible medical research; they now require open-data for the research they fund [248]. The signs are positive for medical research; hopefully other disciplines will fall in line. A general push for open publishing of all research–especially carving out a place for negative findings and explicit exploratory analyses– would go a long way to improve the state of affairs.

These, however, are only band-aids on a larger wound. The statistical rituals that got us to this point are born from a philosophy of statistics that undervalued the critical importance

of causality and intuition in science; the implications are much larger than a slight revision of how we present the results of our t-tests. Statistical misapplication is also only one in a set of broader methodological concerns for facing psychology.

3.2 A Broader Crisis: Research Methodology

Psychology has seen many great debates about the role of empiricism (see i.e. the debate between humanist and positive psychologists in the last chapter). Outside of statistical concerns, broader methodological concerns fall roughly into categories of generalizability and ontology:

- **What are we measuring?** Psychological phenomena are unobservable, sparking substantial debate at the level of basic concepts. When it comes to states of attention like flow, it's not uncommon to find papers that are critical of the core concepts themselves [204, 24]. Where a concept defies formal definition, multiple definitions compete and science does not build; we fail to accumulate a literature around a single topic.

  This is not to say debate about the core topic is unwarranted; on the contrary, this debate is one of the most important steps in the scientific endeavor. The current process of reasoning, comparison, and consolidation relies on an ill-defined, social academic marketplace. If instead we were to treat ontological claims as theories, we might improve and formalize this process with an eye toward its predictive utility.

- **Laboratory Stimuli** A few years ago, my colleagues at the Media Lab created an immersive exhibit at the Museum of Fine Arts. Titled 'Hot Milk', they invited patrons to sit in a chair and experience a plethora of 'textbook' positive stimuli from laboratory psychology experiments– they would be given an mouth brace to engage their smile muscles, their head would be gently stroked, they would be shown images of clowns, soothing music would play, and they would be fed placebo pills to make them happy.

Needless to say, this experience was utterly unsettling for for anyone brave enough to allow themselves to try it out.

This art piece raises an important issue. Stimuli cannot be considered in isolation; their impact is part of a larger, gestalt whole. If we hope to make a robust, generalizable, scientific claim, we need to operate at the level of the larger, gestalt percept. Most concepts in psychology face this challenge– in many cases, the percepts and the experiences that contextualize them are not analytically decomposable.[4]

Moreover, laboratory stimuli include social features that are rarely considered. They may be perceived (we know a happy picture should make us happy) as much as they are felt (you actually feel happy when looking at them); a tension that can be difficult to tease out of survey results. They imply intentionality. (For example, a positive picture implies the experimenter is trying to make you feel happy; a negative one implies that they are trying to make you feel sad. How does this knowledge make you feel? Do you believe their stimuli is well designed and effective? Are you agreeable or resistant to it, and how does that depend on your personality, your opinion about the experimenter, or the perceived hypotheses?)

These considerations– how the subject perceives both the goal (i.e. to make you feel happy) and the intent (i.e. to learn about emotion) of a stimuli– are frequently more important than the typical naive physicalist description (it's a picture of a puppy). If we wish to build theories about how kinds of stimuli affect people, or draw broader conclusions about life from laboratory work, we first require an ontological framework to describe the stimuli accurately in phenomenological terms.

- **Laboratory Settings** For many of the reasons just described, we know that results from contrived, laboratory settings don't always hold in the real world. Laboratory settings are generally sterile, high pressure environments where people are very aware of being observed and more likely to be vigilant. They are social settings; the way an experimenter treats a subject can have a huge impact on their comfort, the demand

[4]Concerts are a great example of this. The experience can be utterly transformative, but if you try to alter any dimension of that experience to understand its importance– you remove the lights, or the sounds, or the music– the whole psychological experience falls apart. The whole is much more than the combination of the parts, and cannot be described in more fundamental terms.)

characteristics and desire to be a good subject, etc.

Moreover, people who have time and/or willingness to come in to a lab are typically a biased subset of people– WEIRD,[5] interested in the topic/result, or motivated by study incentives. This once again presents a problem for generalizability.

- **Representations** As we can see, none of the ontological primitives we work with– the stimuli we design, the environments that envelope them, the states and traits of the individuals we test, or their resulting cognitive experiences we care to measure– are directly observable (they are all phenomenological experiences and percepts). When we measure them, however, we frequently consider them with deterministic confidence– for instance, our representations of cognitive states usually come in the form of a single average from a survey; we frequently treat our treatments/stimuli with similar simplicity.

  This paradigm fails to represent the underlying stochastic nature of our concepts; better representations would treat their inherent uncertainty as a first-class primitive. With regards to cognition, many great thinkers have suggested bio-behavioral representations, where multiple physiological, behavioral, and self-report measures are used in concert to estimate the underlying state. This fusion across multiple modalities provides us with a mechanism to assess the trustworthiness and utility of each individual modality in our estimation.

- **Physiology in the Lab** The desire to create these better representations has driven significant research to understand the underlying psychophysiological relationships. The in-lab paradigm allows us to collect clean data and tightly control sources of noise. We can elucidate the types and orders of magnitude we might expect of the relationships that exist between cognition and measurable physiology.

  Physiology in the lab, however, also faces a challenge of generalizability. This work suffers from the same critiques above (the stimuli and environment may elicit different responses than naturalistic ones), in addition to two more: (1) real world systems and measurements are significantly noisier, so relationships in the lab may not be feasi-

[5] shorthand for Western, Educated, Industrialized, Rich and Democratic.

ble to capture under real-world conditions; (2) the relationship between physiological signals and cognition is roughly a many-to-few-to-many bottleneck of sympathetic and parasympathetic activation. Nose temperatures, for instance, have been shown to drop for fear, stress, joy and guilt; they remain untested for a wide variety of other emotional stimuli, some of which assuredly have a similar influence. This makes reasoning about cognition from physiology very challenging under realistic conditions (even after we attempt to control for the primary mechanisms driving nose temperature like metabolism, ambient temperature, and clothing choice).

- **Interaction Effects**. The story of personality psychology's demise appears in Ulrich Schimmack's 'Personality Science' [377], Brian Little's 'Me, Myself, and Us', [264] and John Gottman's 'The Mathematics of Marriage' [180]: personality psychology experienced a boon in the first half of the 20th century leading to the taxonomy of the Big-5 personality factors. These factors were used experimentally to predict behavior, showed little predictive value, and were famously criticized by Walter Michel in his influential 1970's work 'Personality and Assessment'. His book led to a refocusing on social psychology and environments as predictors (after all, environments must be the primary factor driving behavior if personality explains so little of the variance in behavior).

  Recent meta analysis of this body of work shows environments have actually fared worse than personality in prediction; neither personality nor environments have much explanatory value when considered in isolation [377]. As Kurt Lewin would argue nearly a century ago, this implies that behaviors must be considered within a transactionalist frame; they are shaped through the interaction of the person and their environment. We cannot consider just the person or just the environment alone.

  This insight has problematic implications for data collection (a high effort process for those of us deeply studying individuals in context)– as Andrew Gelman discusses, a simple interaction effect requires roughly sixteen times the amount of data for similar power compared with a main effect [163]. For much psychological work, the scale of data for appropriately powered research looks infeasible; without large scale and a

robust treatment of the entire constellation of person/environment interactions, we can't draw general conclusions.

These points suggest a few important, fundamental challenges to how we study psychological phenomena. Our core concepts must move towards phenomenological, gestalt representations that capture the uncertainty inherent to the process. This is tricky to formalize; it would be useful to have a way to compare and evaluate ontological claims. Bio-behavioral models that provide more insight into the psychological state under study are preferable.

Building a solid model of how and when the real-world looks like the laboratory setting is equally difficult; it may well be that the laboratory setting simply cannot be designed to generalize to normal living. Thus, in order for our data to apply in the real-world, we need to move our study to the real-world as well.

Tools for naturalistic study can alter naturalistic behavior; it is important to design them in a way that minimizes their psychological footprint. These naturalistic scenarios also ground our intuition about the value of physiological markers when we consider real-world noise and confounding.

Finally, the current paradigm– randomizing out individual differences to get at an average treatment effect– makes little sense from a transactionalist worldview; a view with reasonably strong empirical support. If a concept under study is dominated by interaction effects, data requirements grow exponentially and likely become untenable.

This suggests a return to currently out-of-favor idiographic approaches that were popular a generation ago– approaches we abandoned as 'unscientific' because they don't carry with them an implicit claim of generalization. Unfortunately, this implicit notion of generalization rarely holds even for our nomothetic studies; moreover, randomizing the interaction effects across treatment groups not only destroys our picture of the true underlying relations (i.e. our intervention could be very positive or negative for people depending on something about their personality, which is obscured in the average), we should also expect this powerful antecedent of our outcome to only balance across groups in expectation.

3.3 A Deeper Crisis: Humean Philosophy

The impacts of the replication crisis represent not just a broader methodological crisis in psychology, but also a deeper philosophical confrontation at the root of statistics, philosophy of science, and applied statistical learning theory (AI/ML). The origins of this crisis are frequently attributed to the 18th century Scottish philosopher David Hume.

Hume famously articulated that causal beliefs are really unjustifiable extrapolation of 'constant conjunction' with his 'problem of induction'– i.e. we believe all swans are white until we see a black one; a chicken believes the farmer is a benevolent source of food until the day he is not. For Hume, no amount of consistent correlation ever justifies a causal belief. This argument has an important corollary; that the strength of a belief should only be considered in direct proportion to the amount of evidence we observe. We should not solidify an underlying causal model in the face of constant conjunction; we should simply treat our statements of future prediction as expressions of observed frequencies. The certainty of our predictions is not derived from a presumed understanding of causality but from the regularity of observed phenomena.

His skepticism about our causal intuition had a deep influence on the statistical theory which serves as the basis for modern scientific research. Though the Bayesian counterpoints are gaining mainstream acceptance (in which evidence is considered in light of prior beliefs and intuitive causal reasoning is embraced) the sub-structure of science is still built on Humean foundations.

3.3.1 The Individual Researcher

The first place we see this difference playing out is in theory building and analysis we perform as researchers. Beyond a reinterpretation of p-values in light of prior belief, researchers like Judea Pearl have been advocating for a causal reappraisal of mainstream associative statistical practice (in the tradition of Hume and his adherents like Pearson).

Pearl effectively argues that causal assumptions are a necessary prerequisite for any valid analysis. For example, correlation can be spurious if we control for the wrong variables ('colliders'[6] as in Berkson's paradox) or *fail* to control for the wrong variables (confounders). In the extreme case, it's possible for Simpson's Paradox to occur, as it did famously in a 1994 BMJ kidney stone analysis (showing one treatment was more successful both for people with small kidney stones and people with large kidney stones– the only two groups– but was a worse treatment when considering the two groups together). In this real life example, kidney stone size affects treatment selection; thus, the overall results are misleading, and we should trust the results for each subgroup [59]. Pearl advocates for 'do-calculus' and the adoption of Bayesian Networks as primary tools to formalize our causal intuition and perform statistical analysis.

While it is common belief that randomized controlled trials side-step these problems, they too require causal assumptions– randomized controls are controls in expectation (not perfect controls) and thus require repeated runs to approach an unbiased estimate of the average treatment effect. Single trials are only balanced for a sufficiently large sample size, which we can only estimate given our priors about the underlying data generating process. If con-founders are well understood, it is better to precisely control for them rather than to rely on randomization. (The debate between pure randomization vs. a Bayesian attempt to minimize the effect of likely confounders goes back to Fisher himself) [108].

When it comes to drawing conclusions, even from a randomized study, we again run into questions that rely on tacit hueristics about the underlying causal structure which would be better made explicit– how representative is our sample? Did we introduce bias? How representative are our stimuli, our environments, the moods of our participants? How much do each of these matter? What kinds of variation and underlying data generating processes do we expect in our data, so we can analyze it appropriately? In all cases, scientific studies require explicit, a priori causal assumptions to contextualize the results and use them to build generalizable scientific theories.

[6] In Pearl's language a simple example of a 'collider' would be controlling for university admission when assessing the link between intelligence and athleticism; if a school selects for either criteria, controlling for it will introduce a negative correlation between these two otherwise unrelated concepts.

3.3.2 The Scientific Process

The issues Pearl identifies at the level of the individual study also apply to our philosophy of science. Karl Popper was a Humean; and though Kuhn famously attacked many of his notions, the general belief amongst scientists seems to be that Kuhn described the scientific enterprise as the flawed, sociological enterprise that it is, and Popper described the falsification standard it as it should be. This is not the case.

Richard McElreath, author of 'Statistical Rethinking' [289], writes "The greatest obstacle that I encounter among students and colleagues ... is a kind of folk Popperism, an informal philosophy of science common among scientists but not among philosophers of science. Science is not described by the falsification standard, as Popper recognized and argued. In fact, deductive falsification is impossible in nearly every scientific context."

He supports this argument with a story from gene science (in which one theory was proven by predicting the distribution of alleles over time very closely) in which no theories were falsifiable. Much like Pearl, he argues that statistical techniques follow from theory building, and theory building looks like a 'data generating model' or a 'process model'. Falsifiability can be seen as a subset of this broader abductive approach over hypothesis spaces; it works well when a particular theory predicts data that is unlikely under all other theories (I think of this as the 'Babe Ruth' criteria)– $P(D \mid H_1)$ is high, and the $P(D|H_{all\ others})$ is low, thus $P(H_1 \mid D)$ dominates the hypothesis space after inference. Of course, there are infinitely many hypotheses we could create, so we rely on priors over our hypotheses just as much as the data (i.e. Occam's Razor– simple hypotheses are better). As soon as we've seen that data, all viable future theories will incorporate them a priori. As a result, competing theories will consolidate, making similar kinds of predictions and becoming increasingly difficult to differentiate empirically.

This worldview puts forth a very different fundamental requirement for scientific theory– theories must be data generating, not falsifiable.

3.3.3 Lessons from AI

This model of science starts to look like a model of human reasoning from computational cognitive science (see i.e. Josh Tennenbaum's work): many AI researchers argue that intelligence *is* a process of causal model-building, a process which enables counterfactual reasoning about the world from a small set of observations. This view is not without its critics, however– a similar Humean/Bayesian debate exists in AI. As Frank Wood writes in 'Introduction to Probabilistic Programming':

> How do we engineer machines that reason? This is a question that has long vexed humankind; answering it would be incredibly valuable. There exist various hypotheses. One major division of hypothesis space delineates along lines of assertion: are random variables and probabilistic calculation more-or-less a requirement (Ghahramani, 2015; Tenenbaum et al., 2011), or the opposite (LeCun et al., 2015; Goodfellow et al., 2016)? The field ascribed to the former camp is roughly known as Bayesian or probabilistic machine learning; the latter as deep learning. The first requires inference as a fundamental tool; the latter optimization, usually gradient-based, for classification and regression.

The debate between intelligence-as-pattern-matching vs intelligence-as-model-building has direct analogs to how we think about structuring scientific theories. More importantly, practitioners in this field are at the cutting edge of modeling assumptions– quantitatively exploring the limits of our inductive biases, the structure of our data, and the optimal ways to represent it statistically. There is a lot to gain by applying their work to our philosophy of science in psychology.

Psychologists rely on Structural Equation Modeling (SEM) as their main tool for causal reasoning, which has been largely limited to assumptions of deterministic linear relationships until Judea Pearl (relatively recently) started extending the technique towards probabilistic and nonlinear ones. Pearl advocates for Bayesian Networks (BNs), a type of Directed Acyclic

Graph (DAG) that allows inference over a limited set of conditional probability distributions. Probabilistic programming languages (PPLs) allow more flexibility and expressivity than BNs (in the relationships between variables, control flow, and distribution choice) at the cost of computational complexity at inference-time. These languages treat random variables as first class primitives. They represent a large-scale effort to create general-purpose, expressive computing tools that allow practitioners to fully express data-generating theories in precise terms.

3.4 Opportunities for Improvement

The replication crisis is one example of the underlying Humean/Bayesian tension that we see in discussions of statistical practice, philosophy of science, and artificial intelligence. Combining the best points of these discussions leads us to consider our goal as scientists in terms of creating precise, data generating theories– deeply informed by domain knowledge– using state-of-the-art tools like PPLs, and to compare them with abduction.

We can apply this logic to the methodological challenges we discussed above. We can (1) treat our ontological claims as scientific theories (implying they should be data-generating, causal models that we can express in the language of PPLs). Our latent cognitive concepts will be represented by probability distributions, making the inherent uncertainty a first class feature of the representation. For these models to have a chance of converging to something that usefully characterizes real-world experience, we need (2) multiple streams of naturalistic, real-world bio-behavioral data to feed them, a requirement that demands convenient, always-on wearable monitoring that are socially acceptable enough to disappear into the background. Finally, we (3) should attempt to understand our work idiographically, paying special attention to the interpersonal variation. Focusing on longitudinal data– one person across activities, days, and contexts– is a powerful way to inform our models and draw personalized insights that are robust before we contend with generalization across groups.

In the following chapters, I will apply these principles to flow research. I will show (1)

how we might improve representations of environmental stimuli that can grab our attention and distract us based on phenomenological principles, (2) introduce a suite of novel tools to inform naturalistic, bio-behavioral models of flow itself, to allow us to move towards probabilistic, bio-behavioral representations, and (3) collect data that show the limits of current practice and can serve as a useful jumping off point for exploratory flow modeling work. Additionally, I will introduce several interventions I've designed in light of my new beliefs about what interventions can have an impact, with an eye towards idiographic, longitudinal data collection. I provide a dataset an open-source, longitudinal dataset based on one month of living with one of these interventions. These are philosophically motivated choices, in light of the broader methodological and statistical challenges adjacent to the replication crisis, as well as the state-of-the-art solutions proposed to address them.

Chapter 4

Models of External Stimuli

If you're in the jungle, intently listening to the soundscape enveloping you, and a jackhammer intrudes in the distance, you'll notice it– it will be the main thing that you recall when asked about the scene. Similarly, if we introduce a lion roaring into an urban soundscape, it will be the primary thing you notice and remember amongst dozens of sounds to choose from.

These sounds grab out attention because they stand out in context. The jackhammer is not memorable in the city scene; in nature, however, it powerfully holds out attention. The lion roar– acoustically similar to a motorcycle revving, and easily masked and obscured in our urban spectrogram– does not stand out acoustically. These cases trivially demonstrate that acoustic features alone– even contextualized ones– are insufficient to predict how we attend to and recall the sounds of our environment.

To make generalizable claims about how the world shapes and interacts with our experience, we must represent the world as we perceive it. This problem of representation has a rich philosophical and psychological history; as we discussed in the last chapter, our representations of stimuli and environments must move away from naive-realist descriptions of pitch and duration and towards gestalt, phenomenological ones.

This chapter introduces the work I've done with my colleague Ishwarya Ananthabhotla to

develop phenomenological representations of auditory stimuli. Our goal is to introduce, in a general way, phenomenological/gestalt features to the conversation around predicting human auditory attention and memory. We do this by (1) creating a dataset, (2) running some basic aural memory tasks using this dataset, and (3) exploring the role of these features in estimating the results. We end with a discussion of the implications of our work on the future gestalt modeling and some specific examples of the types of user interfaces it might enable. This work points to general principles around how psychology should represent our tools and environments.

4.1 Background

4.1.1 Auditory Perception

In the 1950's, Colin Cherry conducted a famous dichotic listening experiment in which he proved 'the cocktail party effect'. He showed that a background conversation– of which you have no conscious recollection– would draw your attention to it if the speaker uttered your name, or invoked danger ('fire!'), or mentioned something explicit.

This effect points to a fascinating insight about auditory processing– at the lowest levels of pre-conscious processing, we are highly attuned to the meaning of sounds (it's your name), not just their acoustic structure (the constituent phonemes). The latest neuroscience supports the idea that gestalt auditory pre-processing and attentive filtering precedes conscious perception. [401, 455] Distinct and measurable pre-conscious variations in neuronal event-related potentials (ERPs) occur both when a contextualized stimuli is acoustically novel (i.e. unexpectedly softer or lower pitched than others) as well as when it is conceptually novel (i.e. a machine sound in a group of animal sounds). [380]

The stored gestalt representation of the current sound context– necessary to explain these change-driven ERPs– can be thought of as the first stages of auditory memory [214]. This immediate store, known as 'echoic memory', starts decaying exponentially by 100ms after a sound onset [269]. Measurements of ERPs suggest immediate storage of rhythmic stimuli

on the order of 100 ms with a resolution as low as 5 ms [313]; other studies have shown this immediate store is complimented with an additional echoic mechanism that lasts several seconds [93, 23]. On these time-scales, our auditory system compresses its perceptual representation of textures based on time-averaged statistics [288]. These short-term effects sit atop a larger literature of 'perceptual learning' which emphasises the role of meaning and experience in shaping the most fundamental and visceral aspects of perception. [169]

These fascinating experiments motivate the importance of high-level conceptual and contextual features for our representations of aural percepts– a field currently dominated by simple time-frequency models.

4.1.2 Taxonomies of Acoustic Stimuli

Despite what scientists understand about perception, our attempts to taxonomize sounds-as-perceived are relatively simplistic.

In 1993, Gaver introduced an ecological model of auditory perception based on the physics of an object in combination with the class of its sound-producing interaction [160]. He suggests that everyday listening focuses on sound sources, while musical listening focuses on acoustic properties of a sound, and that the difference is experiential.

Research in which participants are forced to categorize a collection of sounds corroborates this distinction– listeners primarily group sounds by category of sound-source, sometimes group sounds by location/context, and only in certain conditions favor groupings by acoustic properties [278, 187]. Open-ended sound labeling tasks encourage more detailed descriptions along valence/arousal axes (i.e. for animal sounds) or using acoustic properties (i.e. for mechanical sounds) if sound-source distinctions are too insufficient for the categorization task [58].

It has also been suggested that non-verbal sounds from a living source are processed differently in the brain than other physical events [261]. Symbolic information tends to underly our characterization of sounds from humans and animals (i.e. yawning, clapping), while

acoustic information is relied on for other environmental sounds [175, 20, 340]. Along these lines, Dubois et al. [126] demonstrated that, for complex scenes, the perception of pleasant/unpleasantness was attributed to audible evidence of human activity instead of measurable acoustic features.

4.1.3 Taxonomies of Sound Objects

It is clear from the above research that acoustic features are insufficient to capture the complex cognition underlying sound phenomenology; we must start with something more, for example a characterization of a sound's interpreted cause (and the meaning of that source for the subject). In many cases however, a sound's cause can be ambiguous. In [41] Ballas introduced a measure of casual uncertainty (H_{cu}) based on a large set of elicited noun/verb descriptions for 41 everyday sounds: $H_{cui} = \sum_j^n p_{ij} log_2 p_{ij}$. (For sound i, p_{ij} is the proportion of labels for that sound that fall into category j as decided by experts reviewing the descriptions). He shows a complicated relationship between H_{cu} and the typicality of the sound, its familiarity, the average cognitive delay before an individual is able to produce a label, and the ecological frequency of the sound in his subjects' environment. H_{cu} was further explored in [259] using 96 kitchen sounds. Lemaitre et al. demonstrated that H_{cu} alters how we classify sounds: with low causal uncertainty, subjects cluster kitchen sounds by their source; otherwise they fall back to acoustic features.

4.1.4 Summary

Many factors influence the cognitive processes underlying human aural attention, processing, and storage. Research shows a complicated interdependence between attention, acoustic feature salience, source concept salience, emotion, and memory; furthermore, verbal, pictorial, and phonological-articulatory mnemonics can have a significant impact on auditory attention, processing, and recall.

4.2 Advancing the Paradigm

Current state-of-the-art models for auditory salience operate on acoustic features in a collapsed time-frequency space. Multiple sounds in a scene may overlap in this representation. With advances in machine learning, it is now possible to transform raw audio into accurate predictions of individual sound objects, and base predictions of attention, salience, and memory on sound objects and their meanings in addition to acoustic properties.

Of course, this new feature space comes with its own set of questions. The label of sound sources alone aren't necessarily useful; we need to understand the fundamental meaning of that sound source in context to make predictions; the emotions, expectations, and associations the source evokes. The work summarized below is our first attempt to explore a first-order, phenomenological feature space for audition.

4.3 HCU400: A New Annotated Dataset for Exploring Aural Phenomenology through Causal Uncertainty

To begin our work on on gestalt/phenomenological representation of sound perception, we needed a dataset of environmental sounds which:

- captured common environmental sounds from everyday life,
- labeled those sounds on dimensions of 'meaning' – what emotions and images they evoke,
- sample the full space of causal uncertainty.

This will allow us to test our ability to improve the representation of sounds, and the relationships we expect between acoustic and semantic features. We could not find an existing dataset with sounds of reasonable quality, so we created our own– the HCU 400. This open-source dataset includes 400 sounds– many obvious, and many ambiguous in origin– alongside a collection of 12,000 hand-annotations (30 per sound). We collected hand-labeled, high-level ratings of emotional features (valence and arousal) in line with previous work on affective sound measurement [62]; we also collected features that provide other insights into the mental processing of sound– familiarity and imageability [380, 74]. Finally, we introduce a per-sound estimate of H_{cu} based on the clustering of free-text labels in word-embedding space.

Our preliminary analysis shows reliable trends across users in their labeling, agreement between the automatic H_{cu} metric, deliberation time for a label, and our a priori design of ambiguous vs. non-ambiguous samples. This dataset is useful for explorations of sound representation and cognition. For details and to review the data, see [32].

The HCU400 dataset consists of 402 sound samples sourced from the Freesound archive (`https://freesound.org`) alongside 3 groups of features: sound sample annotations and

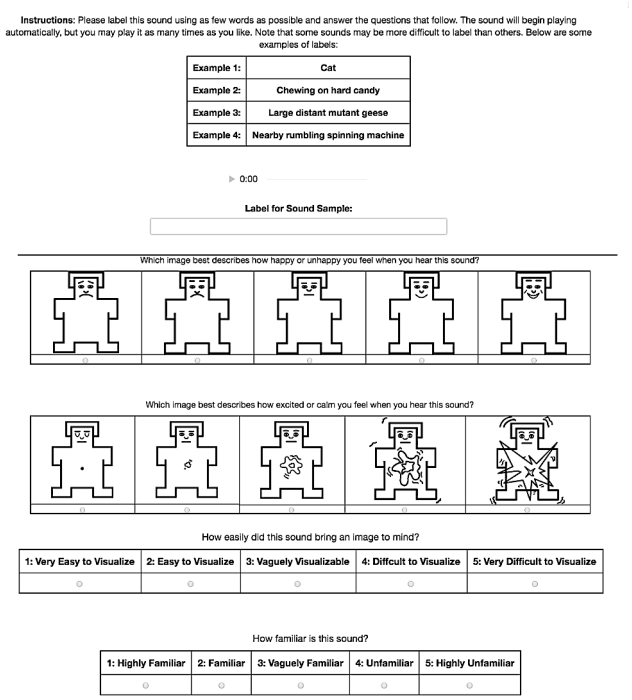
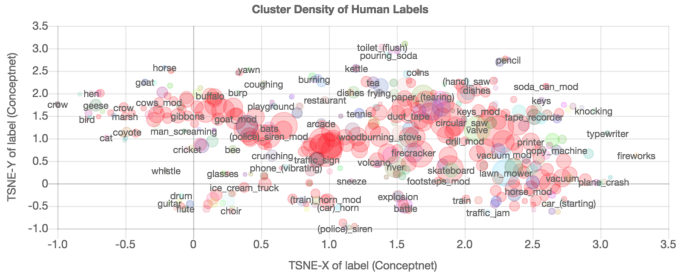


Figure 4-1: A screenshot of the interface shown to AMT workers.

associated metadata, audio features, and semantic features. It is available at `github.com/mitmedialab/HCU400`.

A major goal in our curation was to find audio samples that spanned the space from 'common and easy to identify' to 'common but difficult to identify' and finally to 'uncommon and difficult to identify'. We sought an even distribution of sounds in each broad category (approximately 130 sounds) using rudimentary blind self-tests. In sourcing sounds for the first two categories, we attempted to select samples that form common scenes one might encounter, such as *kitchen, restaurant, bar, home, office, factory, airport, street, cabin, jungle, river, beach, construction site, warzone, ship, farm,* and *human vocalization.* We avoided samples with explicit speech.

To source unfamiliar/ambiguous sounds, we include a handful of digitally synthesized samples in addition to artificially manipulated everyday sounds. Our manipulation pipeline applies a series of random effects and transforms to our existing samples from the former categories, from which we curated a subset of sufficiently unrecognizable results. Effects



Typing	Modified Chair Sliding
Cluster Radius = 6.3	Cluster Radius = 8.7
typing on a keyboard	*bowling ball*
Typing on keyboard	*electric tube*
typing	*PVC pipe building pressure and release*
typing	*Error message on computer*
Typing on a keyboard	*High Speed Frisbee*
someone typing	*beer mug sliding on bar*
Someone typing on keyboard	*driving a car*
keyboard	*toy car hitting wall*
typing	*filling up a tub*
...	...

Figure 4-2: Left: Average ConceptNet embedding where the radius represents our H_{cu} metric; red bubbles and the '_mod' suffix are used to indicate sounds that have been intentionally modified. Right: examples from our free-text capture, demonstrating the difference between labels of large radius and small radius clusters.

include reverberation, time reversal, echo, time stretch/shrink, pitch modulation, and amplitude modulation.

Annotated Features

We began by designing an Amazon Mechanical Turk (AMT) experiment as shown in Figure 4-1. Participants were presented with a sound chosen at random, and upon listening as many times as they desired, provided a free-text description alongside likert ratings of its familiarity, imageability, arousal, and valence (as depicted by the commonly used self-assessment manikins [62]). The interface additionally captured metadata such as the time taken by each participant to complete their responses, the number of times a given sound was played, and the number of words used in the free-text response. Roughly 12,000 data points were collected through the experiment, resulting in approximately 30 evaluations per sound after discarding outliers (individual workers whose overall rankings deviate strongly from the global mean/standard deviation).

Audio Features

Low level features were extracted using the Google VGGish audio classification network, which provides a 128-dimensional embedded representation of audio segments from a network trained to classify 600 types of sound events from YouTube [201]. This is a standard

feature extraction tool, and used in prominent datasets. A comprehensive set of standard features extracted using the OpenSMILE toolkit [142] is also included.

Semantic Features

A novel contribution of this work is the automation and extension of H_{cu} using word embeddings and knowledge graphs. Traditionally, these are used to geometrically capture semantic word relationships; here, we leverage the 'clustering radius' of the set of label embeddings as a metric for each sound's H_{cu}.

We employed three major approaches to embed each label: (1) averaging all constituent words that are nouns, verbs, adjectives, and adverbs– a common/successful average encoding technique [210]– (2) choosing only the first or last noun and verb, and (3) choosing a single 'head word' for each embedding based on a greedy search across a heavily stemmed version of all of the labels (using the aggressive Lancester Stemmer [83]). In cases where words are out-of-corpus, we auto-correct their spelling, and/or replace them with a synonym from WordNet where available [296]. Labels that fail to cluster are represented by the word with the smallest distance to an existing cluster for that sound (using WordNet path-length). This greedy search technique is used to automatically generate the group of labels used in the H_{cu} calculation. Both Word2Vec [295] and Conceptnet Numberbatch [402] were tested to embed individual words.

After embedding each label, we derived a 'cluster radius' score for the set of labels, using the mean and standard deviation of the distance of each label from the centroid as a baseline method. We also explore (k=3) nearest neighbor intra-cluster distances to reduce the impact of outliers and increase tolerance of oblong shapes. Finally, we calculate the sum of weighted distance from each label subgroup to the largest 'head word' cluster– a technique which emphasizes sounds with a single dominant label.

We also include a location-based embedding to capture information pertaining to the likelihood of concept co-location in a physical environment. In order to generate a co-location embedding, we implement a shallow-depth crawler that operates on ConceptNet's location

relationships ('Located-Near', 'Located-At', etc) to create a weighted intersection matrix of the set of unique nouns across all our labels as a pseudo-embedding. Again, we derive the centroid location and mean deviation from the centroid of the labels (represented by the first unique noun) for a given sound sample.

Given the number of techniques, we compare and include only the most representative pipelines in our dataset. All clustering approaches give a similar overall monotonic trend, but with variations in their derivative and noise. Analysis of cluster labels in conjunction with scores suggests that a distance-from-primary-cluster definition is most fitting. Most embedding types are similar, but we prefer ConceptNet embeddings over others because it is explicitly designed to capture meaningful semantic relationships.

Our clustering results from a Processed ConceptNet embedding are plotted in Figure 4.3. Intentionally modified sounds are plotted in red, and we see most sounds with divergent labeling fall into this category. Sounds that have not been modified are in other colors– here we see examples of completely unambiguous sounds, like human vocalizations, animal sounds, sirens, and instruments.

Baseline Analysis and Discussion

We find that the likert annotations are reliable amongst online workers, using a split ranking evaluation adapted from [38]. Each of the groups consisted of 50 % of the workers, and the mean ranking was computed after averaging N=5 splits. The resulting spearman rank coefficient value for each of the crowd-sourced features is given in Figure 4-3. This provides the basis for several intuitive trends in our data, as shown by Figure 4-4 – we find a near linear correlation between imageability and familiarity, and a significant correlation between arousal and valence. We also find a strong correlation between imageability, familiarity, time-based individual measures of uncertainty (such as such 'time to first letter' or 'num of times played'), and the label-based, aggregate measures of uncertainty (the cluster radii and H_{cu}).

We next see strong evidence of the value of word embeddings as a measure of causal un-

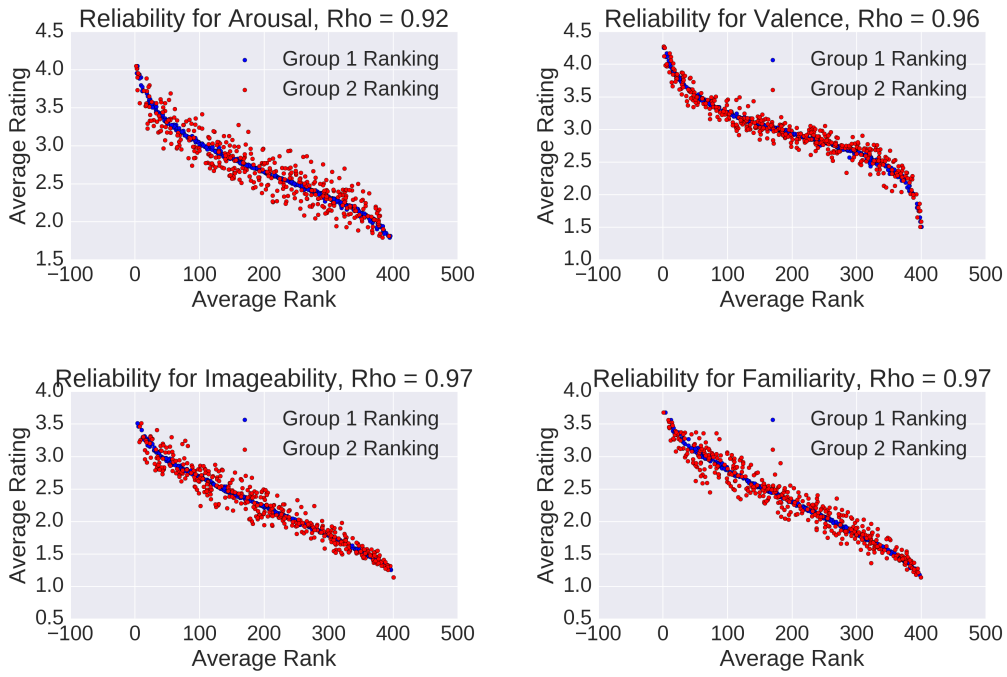


Figure 4-3: Split ranking correlation plots and Spearman rank coefficient values for the four Likert annotated features following the split-half procedure outlined in Bainbridge et al. [37]. We split the participants randomly into two groups several times and each time plot the score of the first group against its ranking by score in each sub-group.

certainty – the automated technique aligns well with the split of modified/ non-modified sounds (see Fig. 4.3) and a qualitative review of the data labels. Our measure also goes one step beyond H_{cu}, as the cluster centroid assigns representative content to the group of labels. Initial clustering of sounds by their embedded centroids reveals a relationship between clusters and emotion rankings when the source is unambiguous, which could be generalized to predict non-annotated sounds (i.e., sirens, horns, and traffic all cluster together and have very close positive arousal and negative valence rankings; similar kinds of trends hold for clusters of musical instruments and nature sounds).

Furthermore, we use this data to explore the causal relationship between average source uncertainty and individual assessment behavior. In Figure 4-5, we plot the distributions of pairs of features as a function of data points within the 15th (red) and greater than 85th (blue) percentile of a single cluster metric ('Processed CNET'). It confirms a strong relationship between the extremes of the metric and individual deliberation (plot e), as reported by

Figure 4-4: Correlation Matrix displaying the absolute value of the Pearson correlation coefficient between the mean values of annotated features, metadata, and four representative word embedding based clustering techniques.

Ballas [41]. We further find that more ambiguous sounds have less extreme emotion ratings (a); the data suggest this is not because of disagreement in causal attribution, but because individuals are less impacted when the source is less clear (b). This trend is not true of imageability and familiarity, however; as sounds become more ambiguous, individuals are more likely to diverge in their responses (d). Regardless, we find a strong downward trend

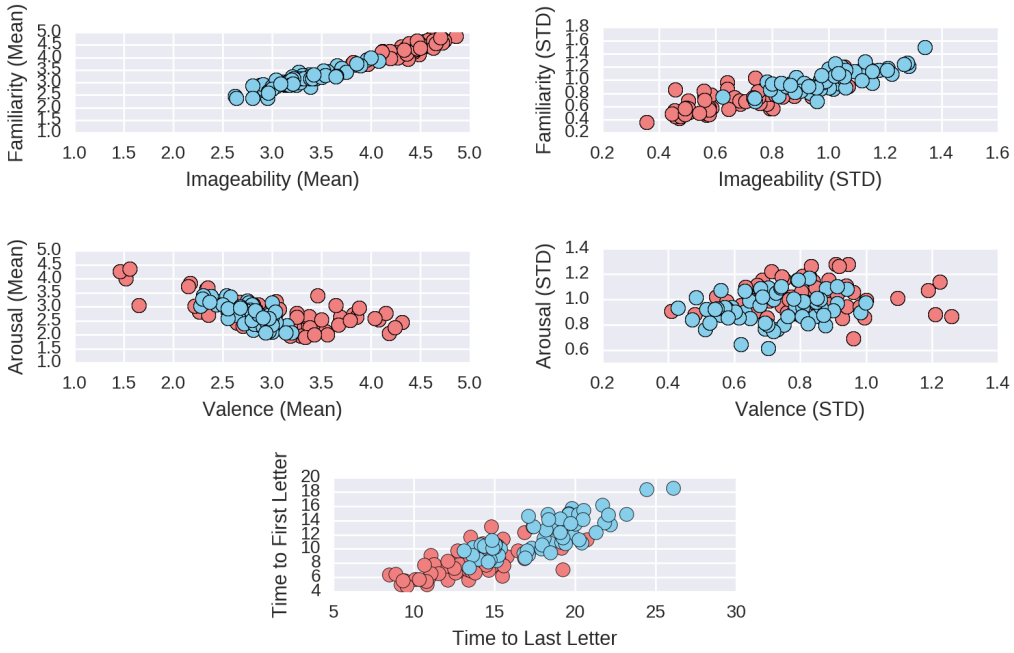


Figure 4-5: Feature distributions grouped by extremes in the 'Processed CNET' cluster metric; red points represent data at $\leq$ 15th percentile (the most labeling agreement and least ambiguous); blue dots are $\geq$ 85th percentile (high H_{cu}). Time-to-first and last letters indicate deliberation time when asked to label a new sound.

in average familiarity/imageability scores as the source becomes more uncertain (c).

Summary

Aural phenomenology rests on a complex interaction between a presumed sound source, the certainty of that source, the sound's acoustic features, its ecological frequency, and its familiarity. We have introduced the HCU400– a dataset of everyday and intentionally obscured sounds that reliably captures affective features, self-reported cognitive features, timing, and free-text labels. Our analysis demonstrates (1) the efficacy of a quantified approach to H_{cu} using word embeddings; (2) the quality of our crowd-sourced likert ratings; and (3) the complex relationships between global uncertainty and individual rating behavior, which offers novel insight into our understanding of auditory perception.

4.4 The Intrinsic Memorability of Everyday Sounds

To understand the relative importance of the gestalt features we collected as part of the HCU400, we created a seven minute aural memory game using these sounds, and paid 4,500 online workers to play it 20,000 times. We then analyzed the relationship between sound memorability and two sets of features– the gestalt features we ascertained as part of the HCU400 dataset project, as well as more traditional 'cognitive salience' features based purely on the spectrogram (using state-of-the-art techniques).

In this work, we show that, while reliable, population level trends in memory and cognition exist (despite the highly individual and contextual nature of attention and memory), these trends are incredibly difficult to explain or predict with our features. However, using a shapely value regression, gestalt features dominate acoustic features when we allow more complex, non-linear mappings.

To our knowledge, this work represents the first general treatment of auditory memorability that combines low-level auditory salience models with multi-domain, top-down cognitive gestalt features. It highlights the challenges we face in modeling and predicting auditory attention and memorability, as well as the complex, contextual, and personal nature of these cognitive processes. It also suggests that the first steps we take here– translating acoustic features into semantic labels, from which we derive representations based on our emotional and symbolic connection to the sound source and it's contextual relationship to other sound objects– is a better approach than state-of-the-art methods that operate exclusively in the time-frequency domain.

4.4.1 Aural Memorability

For a sound to enter our memory, it is first unconsciously processed by a change-sensitive, gestalt neural mechanism before passing through a conscious filtering process [401, 455, 214]. We then encode this auditory information via a complex and variable procedure; frequently we abstract our experiences into words, though we also utilize phonological-articulatory,

visual/visuospatial, semantic, and echoic memory [75, 437].

Different types of memory may also drive more visceral forms of recollection and experience; non-semantic memory, for example, may underpin powerful recollection and nostalgia similar to those reported with music [224].

In this work, we map out the features of everyday sounds that drive their memorability using an auditory memory game. As a recall experiment, we hypothesize that it can provide useful insights into cognitive models for auditory capture and curation. Additionally, we design the task such that it is beyond the capacity of our working and echoic memory and engages long-term memory cognitive processes [273, 265].

4.4.2 What Influences Memorability?

Emotionality is known to have a powerful effect on cognitive processing and memory formation [257]. For music, recall has been shown to improve with positive valence, high arousal sound events [224], though recent research has called the significance of arousal into question [141]. In noise pollution research, the high-level perception of human activity is considered 'pleasant' (more positively valenced) regardless of low-level acoustic features [126]. In general, the emotional impact of a sound is correlated with the clarity of its perceived source [31], though sounds can have emotional impact even without a direct mapping to an explicit abstract idea [346].

For recognition and recall memory exercises, verbalizing a sound or naming a sound (both of which may engage phonological-articulatory motor memory) is the most common and successful strategy [45]. This semantic abstraction has overshadowing effects, though; verbal descriptions can distort recollection of the sound itself, degrading recognition performance without altering confidence [300]. Some researchers specifically isolate and study echoic memory, separating it from naming (a process that doesn't involve outward verbalization) using homophonic sound sources to ensure subjects are not relying internally on a naming mechanism [91]. In everyday life, though, we naturally rely on a complex mixture of echoic

(perceptual), phonological-articulatory (motor), verbal (semantic), and visual memory [75, 437].

In this work, we explore the relationship between both low-level acoustic features and high-level conceptual features with the memorability of sound, in a context that engages long-term memory processes. We set out to test a few important hypotheses, namely:

1. The cognitive processing of sound is similar enough across people that trends in recall across sound samples will be measurable and robust across users.

2. Higher-level gestalt features will be most predictive of successful recall performance. We see from the literature that naming and emotionality have very strong effects in similar tasks– we expect sounds with low H_{cu} (easy to name their source) and strong valence/high arousal to be the most memorable. High H_{cu} (uncertain) sounds elicit weaker emotions, reinforcing this effect.

3. Low-level acoustic feature information will marginally predict memory performance. Gestalt features are not easily mapped to low level feature space (and we expect gestalt features to dominate); however, the literature suggests a measurable, second-order contribution from low-level perceptual saliency modeling.

4. The likelihood of a sound eliciting a false memory will be best predicted by its conceptual familiarity as well as by low-level acoustic features.

5. The context a sound is presented against will have a marginal but measurable impact on whether it is recalled. In other words, we expect emotional and unambiguous sounds to be the most memorable regardless of presentation, but when a sound stands out against the immediately preceding sounds, we hypothesize that it will be slightly more memorable.

4.4.3 Samples and Feature Generation

Audio samples for this test were taken from the HCU400 dataset alongside standard low-level acoustic features [361]. We used default configurations from three audio analysis tools:

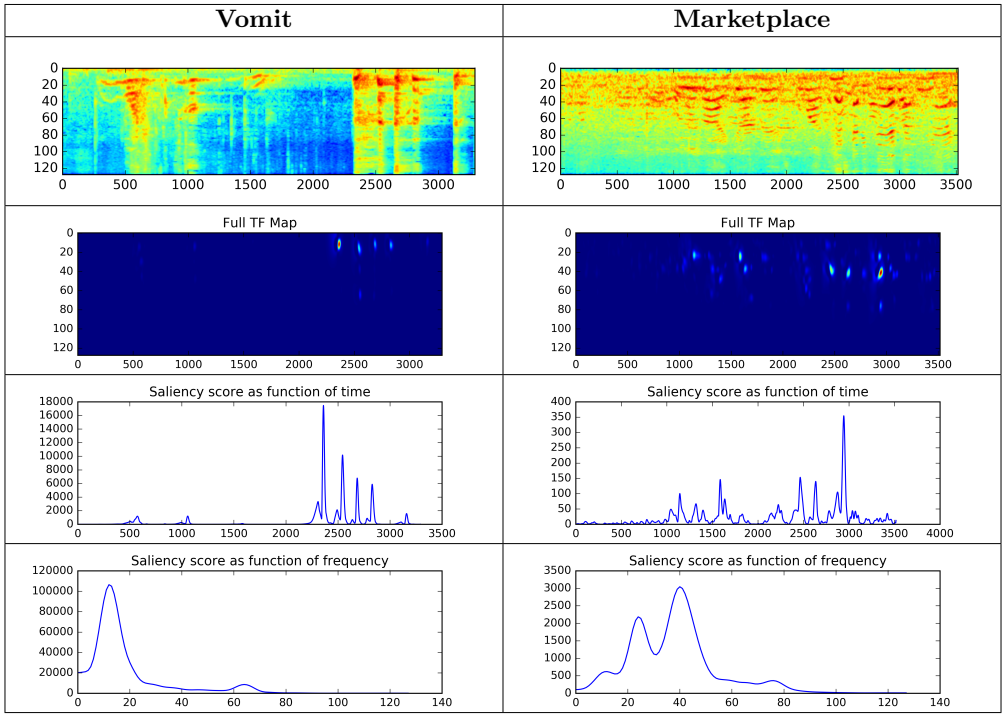


Figure 4-6: A table demonstrating the auditory salience model based on [234] applied to two contrasting audio samples in the HCU400 dataset. The resulting salience scores (bottom) are summarized and used as features in predicting memorability.

Librosa [290], pyAudioAnalysis [168], and Audio Commons [150], which include basic features (i.e. spectral spread) as well as more advanced timbral modeling. We supplement these features with additional summary statistics like high/mid/bass energy ratios, percussive vs. harmonic energy, and pitch contour diversity.

Over the last decade there have been advances in cognitive models that can determine the acoustic salience of sound, inspired by the neuroscience of perception [111, 232]. Here we follow the procedure proposed by [234], applying separate temporal, frequency, and intensity kernels to an input magnitude spectrogram to produce three time-frequency salience maps. Figure 4-6 shows a comparison of temporal salience between two sound samples in the HCU400 dataset with highly contrasting auditory properties. From these maps, we compute a series of summary statistics to be used as features.

High-level, top-down features were taken directly from HCU400 and include causal uncertainty (H_{cu}), the cluster diameter of embedding vectors generated from user-provided

labels (quantifying source agreement or source location), familiarity, imageability, valence, and arousal.

4.4.4 Measuring Memorability

In order to quantify memorability, we drew inspiration from work in [38], which used an online memory game to determine the features that make images memorable. We designed an analogous interface for the audio samples in the HCU400 dataset; this interface can be found at `http://keyword.media.mit.edu/memory`. The game opens with a short auditory phase alignment-based assessment [458] to ensure that participants are wearing headphones, followed by a survey that captures data about where they spend their time (urban vs. rural areas, the workplace vs home, etc). Participants are then presented with a series of 5 second sound clips from the HCU400 dataset, and are asked to click when they encounter a sound that they've heard previously in the task. At the end of each round consisting of roughly 70 sound clips, the participant is provided with a score. Screenshots of the interface at each stage are shown in Figure 4-7.

By design, each round of the game consisted of 1-2 pairs of *target sounds* and 20 pairs of *vigilance sounds*. *Target sounds* were defined as samples from the dataset that were separated by exactly 60 samples– the sounds for which memorability was being assessed in a given round. The *vigilance sounds*, pairs of sounds that were separated in the stream by 2 to 3 others, were used to ensure reliable engagement throughout the task following the method in [38]. Roughly 20,000 samples were crowd-sourced on Amazon Mechanical Turk such that a single task consisted of a single round in the game. Individual workers were limited to no more than 8 rounds to ensure that target samples were not repeated. Rounds that failed to meet a minimum vigilance score (>60%) or exceeded a maximum false positive rate (>40%) were discarded.

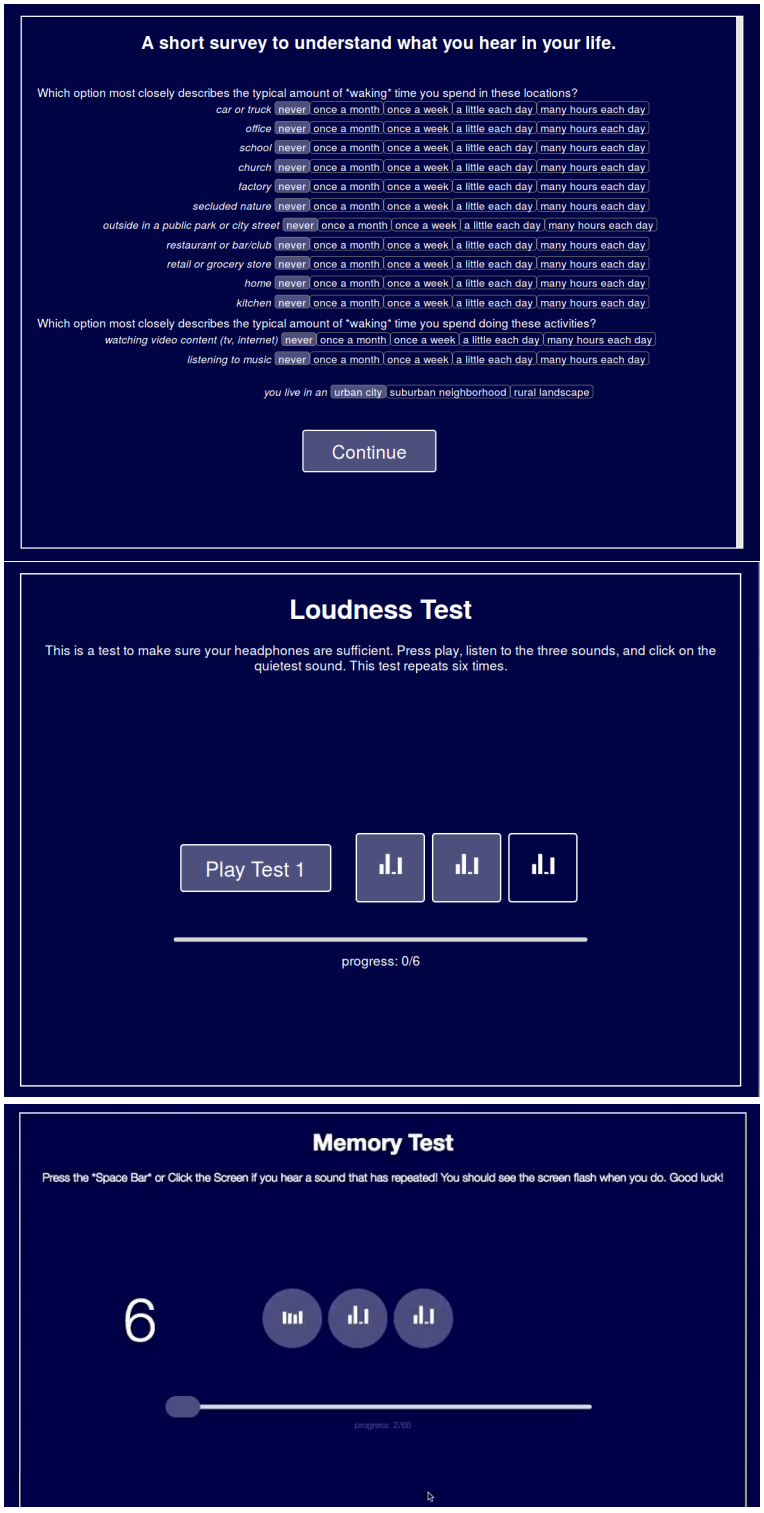


Figure 4-7: Screenshots of the auditory memory game interface presented to participants as a part of our study. The game can be found at `http://keyword.media.mit.edu/memory`.

4.4.5 Summary of Participant Data

We recruited 4488 participants, consisting of a small (<50) number of volunteers from the university community and the rest from Amazon Mechanical Turk. Our survey data shows that our participants report a 51/37/12% split between urban, suburban, and rural communities. We see weak trends in the average time per location reported for each community type– urbanites self-report spending less time at home, in the kitchen, in cars, and watching media on average. Rural participants report spending more time in churches and in nature. Using KNN clustering and silhouette analysis, we find four latent clusters – students (590 users), office workers (1250 users), home-makers (1640 users), and none of these (1010 users). Split-rank comparisons between groups did not reveal meaningful differences in results across user groups; we speculate any differences due to ecological exposure of sounds between environments is not consistent or influential enough at this group level to alter performance.

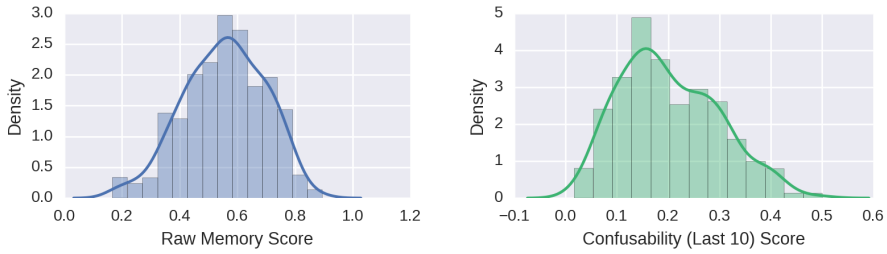


Figure 4-8: Top: A histogram of the raw scores for each sound – they were successfully remembered and identified about 55% of the time on average, with a large standard deviation; Bottom: A histogram of 'confusability' scores for each sound, with an average score of about 25%.

4.4.6 Summary of Memory Data

The raw memorability score M for each sound is simply computed as the number of times it was correctly identified as the target divided by the number of its appearances. However, this does not account for the likelihood that the sound will be falsely remembered (i.e. clicked on without a prior presentation). We additionally compute a 'confusability' score C_{10} for each sound sample, defined as the false positive rate for sounds when they fall

close to the second target presentation (i.e. in the last ten positions of the game). We can thus derive a 'normalized memory score' represented by $M - C_{10}$ (See Figure 4-8 for raw results). In attempting to understand auditory memory, we consider both what makes a sound memorable *and* what makes a sound easily mis-attributed to other sounds, whether those sounds are encountered in our game or represent the broader set of sounds that one encounters on a habitual basis. We therefore model both normalized memorability and confusability in this work.

We confirm the reliability of both the normalized memory scores and the confusability scores across participants by performing a split ranking analysis similar to [38] with 5 splits, shown in Figure 4-9 with their respective Spearman correlation coefficients. This confirms that memorability and confusability are consistent, user-independent properties.

In Table 4.1, we show a short list of the most and least memorable and confusable sounds in our dataset as a function of the normalized memorability score and confusability score.

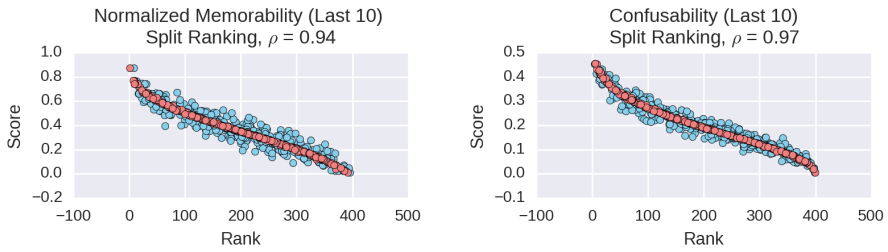


Figure 4-9: The results of the split-ranking analysis for the normalized memorability score and confusability score, using 5 splits; The Spearman coefficient correlations demonstrate the reliability of these scores across study participants, enabling us to model both metrics in the later parts of the work.

4.4.7 Feature Trends in Memorability and Confusability

We consider two objectives – (1) to determine the relationship between individual features and our measured memorability and confusability scores, and (2) to determine the relative importance of these features in predicting memorability and confusability. To address the former, we provide the resulting R^2 value after applying a transform learned using support

Most Memorable	**Least Memorable**
man_screaming.wav	*morphed_firecracker_fx.wav*
woman_screaming.wav	*truck_(idling).wav*
flute.wav	*morphed_turkey2_fx.wav*
woman_crying.wav	*morphed_airplane_fx.wav*
opera.wav	*morphed_metal_gate_fx.wav*
yawn.wav	*morphed_shovel_fx.wav*
Most Confusable	**Least Confusable**
garage_opener.wav	*clock.wav*
lawn_mower.wav	*morphed_335538_fx.wav*
washing_machine.wav	*phone_ring.wav*
rain.wav	*woman_crying.wav*
morphed_tank_fx.wav	*woman_screaming.wav*
morphed_printing_press_fx.wav	*vomit.wav*

Table 4.1: A list of the most and least memorable and confusable sounds from the HCU400 dataset.

vector regression (SVR) for each individual feature. For the latter, we use a sampled Shapley value regression technique in the context of SVR– that is, we first take N random features (N between 1 and 10) and perform an SVR to predict memorability or confusability scores for our 402 sounds and the calculated R^2 of the fit. We then measure the change in R^2 as we append every remaining feature to the model, each individually. The largest average changes over 10k models are reported in table 4.2. This technique is robust to complex underlying nonlinear relationships from feature space to predicted metric as well as feature collinearity.

We find that the strongest predictors of both memorability and confusability are the measures of imageability (how easy the sound is to visualize) and its causal uncertainty. Memorability is dominated by high level, gestalt features, with only one lower level feature ('pitch diversity') in the ten most important features. Low level features, including those derived from the auditory salience models, play a more significant role in determining confusability.

The absolute R^2 values indicate that no individual feature is a significant predictor of memorability by itself. This implies a complex causal interplay in feature space, which we explore further in the set of plots presented by Figure 4-10. In each plot, we show a distribution of feature values for the 15% of sounds that are most memorable or least confusable (blue) contrasted against the least memorable or most confusable sounds (red). We first consider the effect of H_{cu} and valence on memory– low memorability and high confusability sounds exhibit a similar trend of high causal uncertainty and neutral valence (Column 1). ; however, while a fairly broad distribution of valence drives high memorability (eliciting excitement or anxiety), positive valence tends to be a much stronger driver of low confusability. In Column 2, we consider imageability and familiarity ratings, shown to be strongly collinear in [31]. Here, their relationship to memorability and confusability diverge; while both are positively correlated with memorability, *neutral* ratings are the stronger predictor of confusability. This suggests that we are most likely to confuse sounds if they are loosely familiar but neither strictly novel nor immediately recognizable. Finally, Column 3 reveals a discernible decision boundary in low-level feature space for confusability which doesn't exist in its memorability counterpart. The relative importance of low-level salience features, here represented by spectral spread, aligns with intuition– we hypothesize that, in the absence of strong causal uncertainty or affect feature values, our perception of sounds is driven by their spectral properties.

4.4.8 Per-game Modeling of Memorability

The aural context in which a sound is presented, which includes ecological exposure as well as the immediate preceding sounds in our audition task, may influence the memory formation process. The literature supports the notion that, given a context, unexpected sounds are more likely to grab our attention and engage memory in simple experimental settings [380]. To understand this effect in our test, we ran two studies based on a 5 sound context (approximating the limits of semantic working memory) and a 1 sound context (approximating the limits of echoic memory).

Table 4.3 shows the results of a model trained to predict whether the target in each game

Top Predictors for Memorability and Confusability					
Memorability			Confusability		
Feature	R^2	Shapley Δ R^2	Feature	R^2	Shapley Δ R^2
Imageability	0.201	0.126	**Imageability**	0.065	0.078
H_{cu}	0.224	0.125	H_{cu}	0.073	0.078
Familiarity	0.176	0.123	Avg Spectral Spread	0.087	0.078
Valence	0.178	0.120	Peak Spectral Spread	0.037	0.076
Location Embedding Density	0.147	0.117	Peak Energy, Frequency Salience Map	0.059	0.076
Familiarity std	0.103	0.117	**Location Embedding Density**	0.100	0.076
Pitch Diversity	0.084	0.113	Frequency Skew, Frequency Salience Map	0.059	0.076
Imageability std	0.086	0.113	**Arousal**	0.039	0.076
Arousal	0.072	0.112	Peak Energy, Intensity Salience Map	0.044	0.075
Arousal std	0.056	0.111	**Familiarity**	0.045	0.075
Avg Spectral Spread	*0.099*	*0.107*	***Valence***	*0.100*	*0.075*
Timbral Sharpness	*0.094*	*0.091*	*Timbral Roughness*	*0.094*	*0.047*
Max Energy	*0.091*	*0.100*	*Avg Flux, Sub-band 1*	*0.092*	*0.064*
Treble Energy Ratio	*0.090*	*0.020*	*Flux Entropy, Sub-band 1*	*0.091*	*0.061*

Table 4.2: The top performing features from the Shapley regression analysis for both memorability and confusability (gestalt features are bolded); shown are the features ordered by their respective contributions to the R^2 value, with additional features with top performing individual R^2 values appended in italics. The first column indicates the individual predictive power of each feature; the second indicates its relative importance in the context of the full feature set.

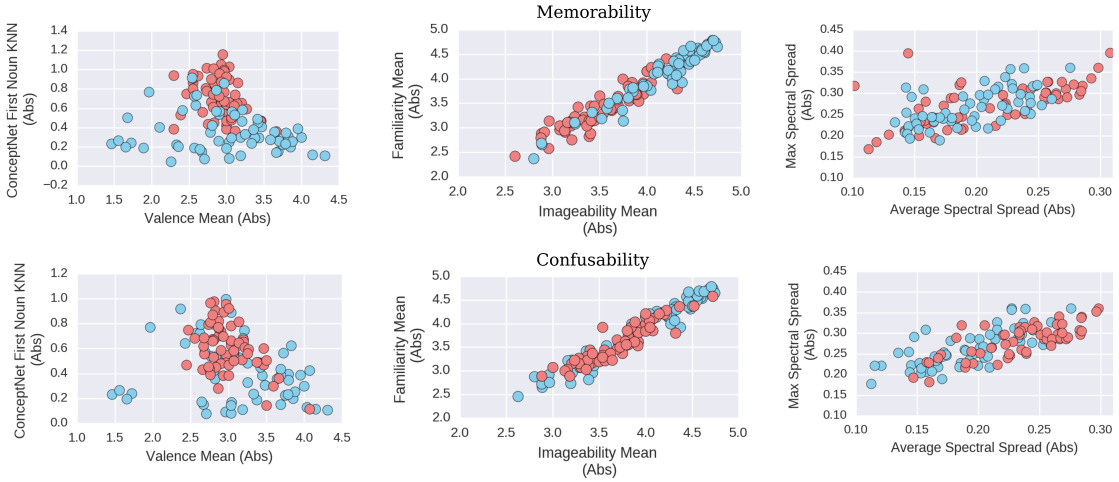


Figure 4-10: Scatter plots showing the changes in distribution of select features based on extremes in memorability (top row) and confusability (bottom row); blue indicates sounds that are most (85th percentile) memorable or least (15th percentile) confusable; red indicates sounds that are least memorable or most confusable.

will be successfully recalled. This model was trained with the most memorable and least memorable sounds only (15th/85th percentiles) with a 5-fold cross-validation process, and results are reported on a 15% hold-out test set.

To begin, a baseline model is trained using the absolute, immutable features of the target

sound. Because there are a limited number of sounds in our dataset relative to the number of games, the feature space is redundant and sparse, and we expect the accuracy of this model to converge to the average expected value over our set of sounds. We then introduce a simple proxy for contextual features– the *relative difference* (z-score) of target sound features with those of the varying sounds that precede its first presentation in each game– to see if our model improves given the context of our 50 most meaningful features (from the SVR analysis; 25 high-level and 25 low-level). In both 5- and 1-sound context cases, however, model performance does not improve as we would expect if the context provided additional useful information.

We also run a classifier to predict memorability and confusability that *only* uses these relative differences in features with nearby sounds. We start with a noise baseline, in which the relative features are calculated using a random, unseen context (these features are still informative as the z-score depends largely on the absolute features of the target sound). We then train the same model with the proper context to assess the difference in performance. There is no improvement when the true context is re-introduced.

This leads us to an insight contrary to our hypothesis – the experimental context of neighboring sounds does *not* exert a measurable influence on our results. While experiential context almost certainly matters, we see no effect in our experimental notion of context. The stable differences across sub-populations in our study are thus unaffected by variations in the immediate experimental context *and* participant ecological exposure (as was demonstrated in the split-rank analysis).

4.4.9 Implications and Conclusion

In this work, we quantify the inherent likelihood that a sound will be remembered or incorrectly confused and confirm that is consistent across user groups. In line with our hypotheses, we show that the most important features that contribute to a sound being remembered are gestalt– namely those sounds with clear sound sources (high H_{cu}), that are easy to visualize, familiar, and emotional. We also show that low H_{cu} sounds that

Memorability Per-Game Models	
Features	**Accuracy (%)**
Absolute + All 5-Sound Context Feats (working semantic)	68.0
Absolute + Top 50 5-Sound Context Feats	69.1
Absolute Feature Only Baseline (˜expected value)	*70.3*
Contextual Only, 5-Sound Context (working semantic)	62.5
5 Sound Context, Noise Baseline	*64.1*
Absolute + All 1-Sound Context Feats (echoic)	68.0
Absolute + Top 50 1-Sound Context Feats	69.5
Absolute Feature Only Baseline (˜expected value)	*70.3*
Contextual Only, 1-Sound Context (echoic)	60.0
1 Sound Context, Noise Baseline	*61.3*

Table 4.3: The influence of contextual sounds before the first presentation of the target on our ability to predict recall across games.

are not familiar or easy to visualize are most likely to be mis-attributed, and low level features play a more important role in predicting this behavior. We see no evidence that these relationships are influenced by the immediate experimental context or variation in ecological exposure.

4.5 Practical Bootstraping with AudioSet

We extend our findings from the prior work to a simple predictive model, bootstrapped with Google's AudioSet, and discuss the implications of our work. It enables more accurate models of auditory attention and memory, and represents a step toward cognitively-inspired compression of everyday sound environments, automatic curation of large-scale environmental recording datasets, and real-time modification of aural events to promote focus, control attention, and alter memory.

4.5.1 Motivation

The cognitive impacts of sound objects are highly subjective [285] and ultimately demand personalized modeling. Here, however, we suggest that even general purpose, non-

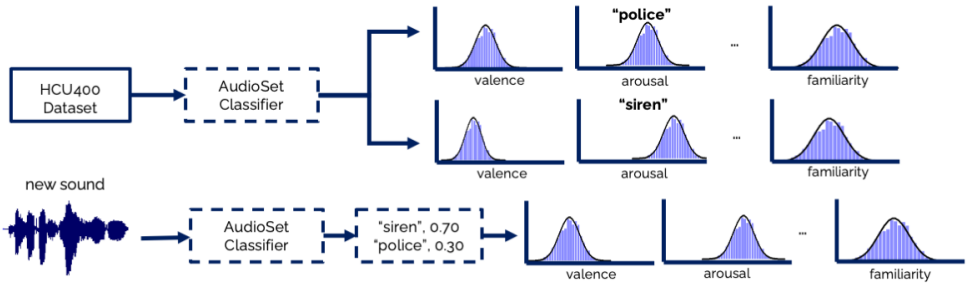


Figure 4-11: We first bootstrap gestalt property scores to all labels in the AudioSet ontology by running a classifier on the HCU400 dataset *(top)*; we then estimate gestalt property scores for unseen audio examples by first predicting AudioSet labels, and then combining the scores associated with these labels, weighted by the prediction uncertainty *(bottom)*

personalized estimators that capture crowd-scale information about intrinsic, semantic properties of sound – often referred to in the auditory psychology literature as *gestalt* properties – can be a useful preliminary step for artistic and creative applications.

We present a simple paradigm for estimating a set of gestalt properties – such as valence and arousal, imageability, causal uncertainty, and memorability – from unseen, real-world audio, using a probabilistic bootstrapping approach that employs an AudioSet [200, 164] classification network as an intermediary. Given limited quantities of annotations of these properties in sound cognition datasets, we can create a robust estimator by first mapping these properties to semantic classes obtained via large, pre-trained networks.

4.5.2 Approach

To illustrate our approach (summarized in Figure 4-11), we consider the HCU400 and memorability datasets presented in [32, 351], and aim to scale the hand-labeled annotations of six gestalt properties – arousal, valence, imageability, familiarity, memorability, and confusability – to unseen audio. We achieve this by building a probabilistic mapping between these scores and the 600+ labels in the AudioSet ontology. To build this mapping, we employ a pre-trained AudioSet classification network[1] to obtain the top k label predictions from each audio sample in the HCU400 dataset. Then, we capture the correlation between each

[1] https://github.com/tensorflow/models/tree/master/research/audioset/yamnet

AudioSet class and gestalt property. To do this, for a given label and gestalt property, we fit a Gaussian distribution to the property scores across all of the sounds with that label, wherein each set of scores (representing the set of human annotations per sound) is weighted according to the network's label uncertainty. While the labels associated with the sounds in the HCU400 dataset are too sparse to fully cover all of the AudioSet labels, we can exploit the existing class relationships in the AudioSet ontology to impute our estimates of the new gestalt properties to the uncharacterized labels: parents adopt mean scores of children, children inherit parent scores (examples of this hierarchy in the AudioSet ontology with increasing specificity include 'music'→'musical instrument'→'plucked string instrument'→'banjo' or 'source-ambiguous sounds'→'surface contact'→'scratch') [200].

To calculate gestalt property scores for unseen audio examples, this process can effectively be inverted: the distributions associated with the top k AudioSet labels are combined– weighted by prediction uncertainty– to obtain mean and variance estimates for the unseen audio. Any number of similar weighting heuristics can be applied, contingent on the application context.

Throughout this approach, we treat the uncertainty of the pre-trained AudioSet model as a proxy for human uncertainty in sound source identification. We use this notion implicitly in the bootstrapping process as we weight the contribution from different instances in the HCU400 dataset by the network prediction uncertainty, mimicking the role of causal uncertainty as the fulcrum between semantic and acoustic processing [159, 161].

The intermediary structure in this approach can be constructed using a spectrum of methods, spanning simple causal intuition derived from auditory psychology literature to rigorous bootstrapping from more extensive datasets onto detailed ontologies. In contrast to traditional transfer or few-shot learning approaches, the structure here has intuitive meaning, and we rely on explicit relationships in label and language space to provide a scaffolding for relationships in cognitive understanding space.

4.6 Implications: Cognitive Audio Interfaces

In a world of rich, complex, and demanding audio environments, intelligent systems— driven by the models of auditory cognition we are developing— can mediate our interaction with the sounds around us. These systems can both enable meaningful, aesthetic experiences and transition design work from humans to computational agents.

In this section, we discuss two promising applications of cognition-informed interface design as future work based on this research.

4.6.1 Parsing Large-Scale, Ubiquitous Audio Datasets

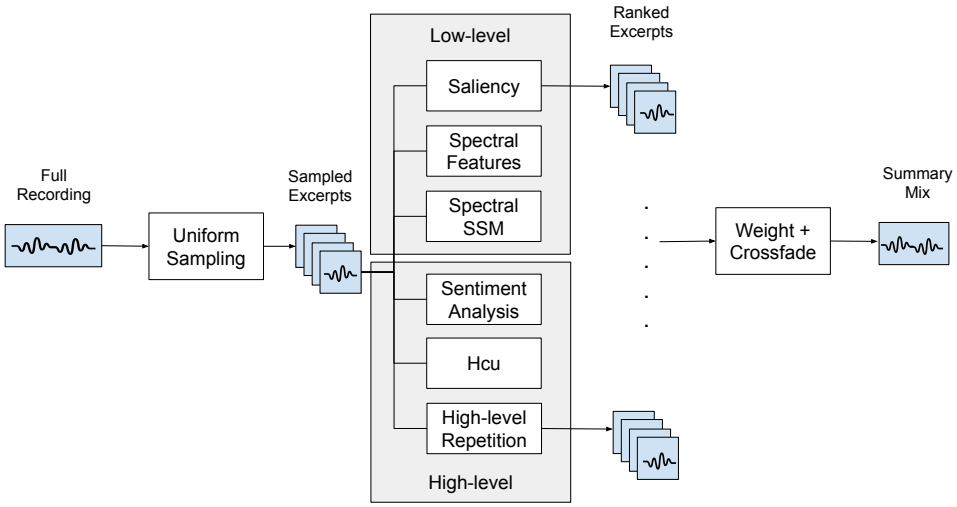


Figure 4-12: A system diagram illustrating the presented approach to audio summarization via cognitively-inspired content curation. A full recording is first uniformly sampled to create a set of thousands of 3-second excerpts; these excerpts are then evaluated and ranked along the high-level and low-level feature axes shown in the center box; finally, summaries are created by selecting excerpts at the extrema of a single feature dimension or a composite 'memorability' feature (derived by weighting the single features according to [348]), and crossfading the samples in chronological order.

As audio infrastructure has become cheaper and smaller, we are now able to collect, stream, and manipulate large sets of audio data from natural environments. The size of these datasets make them impractical for human curation. To surface meaningful content requires

us to extract and isolate sound objects and events, identify them robustly, and infer their meaning or impact.

One such example comes for a major reclamation project of commercialized marshland, the Tidmarsh Living Observatory. Our research group has instrumented this area with hundreds of sensors; twenty-four of those sensors are custom microphones designed for the harsh environment, clustered in different areas of the vast marsh. All of these microphones are streaming their data off-site using a custom API that allows them to be used in real-time. A collection of years of historical audio data can also be easily accessed through this API [284].

In combining wide area sound source separation and localization techniques with deep learning [369, 127], our group can isolate sounds and identify sources from this environment. Armed with labeled sound objects and hours of recordings, we want to automatically generate short summaries of the aural scene.

Our audio summarization project– based on our prior work with gestalt feature estimation– attempts to do just that. It combine gestalt features – based on sound labels and measures of semantic relatedness – with acoustic analysis, to curate ambient audio recordings based on perceptual goals (i.e. 'memorable' summaries for an engaging presentation, 'undistracting' for someone to listen to while studying). We find that the tool surfaces diverse collections of sounds across the long-term environmental recordings that point to different regions of our perceptual outcome space.

4.6.2 Frontiers in User Interface Design

Interfaces that intelligently interact with auditory attention represent an interesting frontier in design.

Another example from TidMarsh is HearThere by Dublon et al. [125]. HearThere users wear a bone-conduction headset that overlays virtual sounds over their natural hearing when they are physically present at Tidmarsh. These virtualized sound sources are actually real

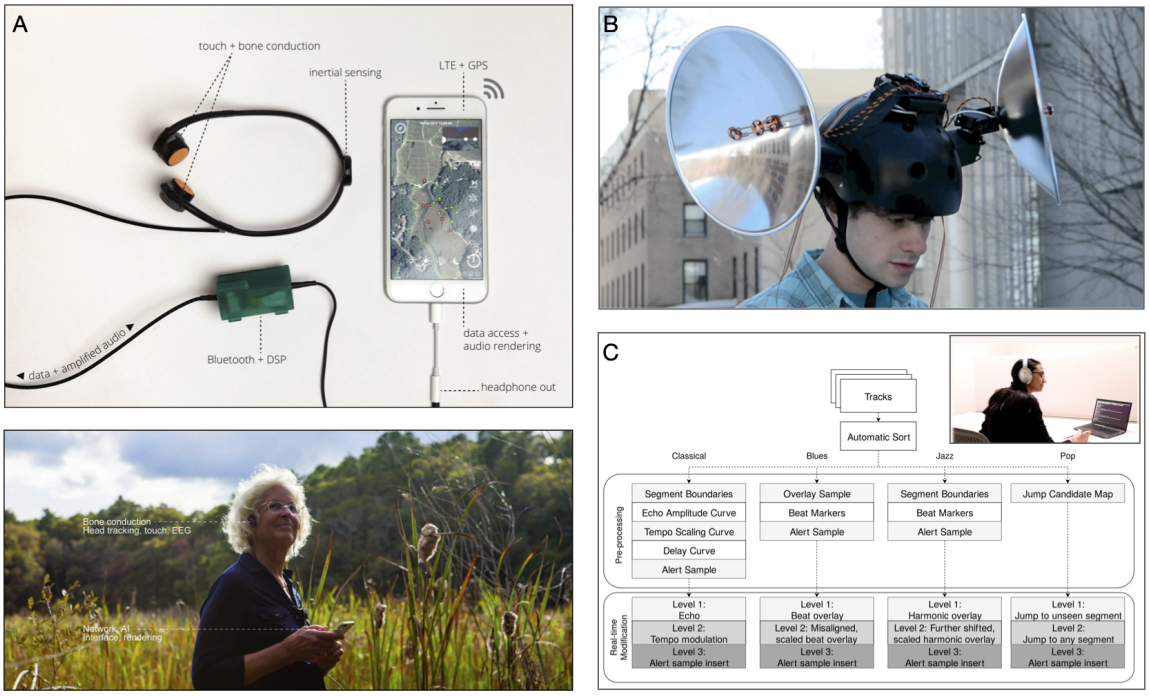


Figure 4-13: **User Interfaces at the Perceptual Boundary**: Dublon et al.'s HearThere project [125] (A) can be used in the marsh environment to naturally extend a user's hearing in a way that requires no conscious effort and is organically managed by our auditory attention. This follows their earlier PhoxEars project [245] (B) which also extended auditory perception through overlay, this time using parabolic microphones on user-controlled motorized gimbals. A final example is the Ananthabhotla's Sound Signaling project [29] (C), which introduces real-time alterations to your music library as a notification tool. These changes are less likely to be noticed by a focused or preoccupied user– fundamentally incorporating their mental state into the design of this notification UI.

sounds in the space– located, identified, and streamed through the Tidmarsh infrastructure. The HearThere headset uses a combination of GPS and head tracking to render these sounds as though they are coming from their real-world locations, but with additional control. You can extend your hearing in distance, or focus on a particular type of sound.

In SoundSignaling, Ananthabhotla et al. introduced a notification platform based on subtle manipulations of your favorite music [29] (adding harmonies to a jazz standard, adding extra layers of rhythm to a blues track, or altering the tempo). It operates on the implicit assumption that attentional load modulates awareness of incongruence, an idea borrowed from Stroop's famous colored text experiments [418] and explorations of auditory and visual switching costs [309]. These modifications aren't noticeable when you're focused, but hard

to miss otherwise.

We can imagine future systems built on a personal attention model that can predict and interact with these attentional characteristics automatically; masking and modifying potential distractors, introducing alerts, and guiding the user's attention intelligently in a closed-loop fashion. The predictions from such a model also may inform robust estimates of user focus; how a user attends to and responds to perceptual stimuli is contingent on their cognitive state. Models of both the user and the environment will drive the sophisticated, intelligent user experiences of the future.

4.6.3 The Future

We believe these and similar modeling efforts will form the cornerstone of audio interfaces of the future; both for next generation off-line compression and search of soundscapes based on the principles of human-perception, as well as for real-time analysis and control of our auditory landscape to direct our attention, promote our focus, and strengthen our memory.

4.7 Summary

In this section, we introduced a general framework to extend our representations of sound objects phenomenologically, by (1) introducing a new, open-source dataset of hand-annotated sound samples, (2) testing cognitive processes with these sounds, and evaluating thier predictive quality, (3) building a first-pass, personalizable model of sound perception using these data as priors, and (4) discussing the implications for future work.

Fundamentally, this body of work a statement about efficient representation. We are, after all, operating on a spectrogram to ascertain sound objects, just as the saliency models that exist purely in acoustic space. In the limit, we might expect a black box auditory perception model to learn the representations and structure we are imposing with VGGish[2] as a necessary intermediate step.

[2]VGGish is a Google model trained on thousands and thousands of labeled audio clips across hundreds of categories that allows us to extract predictions of a sound source from the raw audio.

In some specific, limited contexts, the mapping from acoustic features in a spectrogram to objects in a scene to the cognitive implications of each object may be deterministic enough that we gain little by imposing this additional structure. We should expect, however, that any general treatment– for a sound in multiple contexts, or for an individual across many contexts– will require a probabilistic mapping over semantic features to make accurate predictions. The things that grab our attention are personal and goal-oriented; if you're feeling frightened about a specific threat, what grabs your attention will be different than when you're hyperfocused on a goal in a safe environment. How you feel about a baby crying depends on your mood (sweet, adorable, frustrating, angering, saddening), your situation, and the baby. This kind of perceptual variation we should expect all the time on every scale.

Moreover– in light of time-varying personal differences and contextual factors– the amount of data it would take to train a model to learn a mapping from spectrograms to concepts to cognition as a sub-task in a weakly-biased model is intractable. We must move towards stronger inductive biases operating over perceptually-relevant representations to accurately learn the probabilistic, non-linear, personal dynamics of the relationships that underlie cognition.

Chapter 5

Models of Flow and the Tools to Capture It

We've all lost ourselves in our work, in a great movie, or in a deeply engaging conversation. This intangible quality of what we choose to do is important. We can engage in the same activities– work, play, socialize– with a distracted, fractured mind or with deep, effortless concentration.

For modern psychologists, the name for this type of experience is 'flow state', and it matters– the evidence suggests that the quality of our attention matters more than what we choose to do when it comes to happiness [241]. Moreover, flow states– and not optimism– moderated the effect of the duration of Covid-19 lockdowns on well-being [420].[1]

At work, Gallup estimates 67% of people have completely disengaged from their work (59% 'quiet quitting'), costing the global economy an estimated nine trillion dollars [158]. Harvard Business School's Teresa Amabile describes the solution to this disengagement epidemic with her 'progress principle' concept: "Of all the things that can boost emotions, motivation, and perceptions during a workday, the single most important is making progress in meaningful

[1]These are both large-N, descriptive, survey-based papers. Killingsworth et al. [241] show a large difference in explained variance in happiness when comparing state of attention to activity choice (more than double both within and between persons). Sweeny et al. [420] results are pre-registered and open on OSF. Their findings on the importance of flow are more nuanced, but suggestive.

work" [25]. While we perceive progress when impressive results finally materialize, a better metric– especially for tough, complex, or hard-to-quantify work– shifts our attention from our output to our process. Flow states provide an immediate and intrinsic feeling of progress that is especially important when milestones are infrequent or ambiguous.

Thus, the concept of flow is important for us to understand. According to positive psychologists, it is an integral, experiential part of a well-lived life; a sentiment it shares with the humanistic concepts that predate it (i.e. Abraham Maslow's 'peak experiences'). Unfortunately, the definition of flow is imprecise and heavily criticized [17]– conflating the phenomenological mental state of flow both with its contextual correlates and with its post hoc appraisal. As a result, most researchers end up re-defining the term to suit their task, and the research literature fragments.

In this section, I will review the history of 'flow', it's conceptual critiques, and it's future.

5.1 Csíkszentmihályi's Flow

> The positive aspects of human experience— joy, creativity, the process of total involvement with life I call flow.
>
> ---
>
> *Csíkszentmihályi*
> *'Flow', 1970*

The term 'flow state' was coined by Mihály Csíkszentmihályi in his 1975 work 'Beyond Boredom and Anxiety.' His initial definition featured six characteristics, though a seventh (time transformation) was added shortly thereafter. His definition has slightly shifted over the years, but generally includes the following seven features:

1. The merging of action and skill/awareness
2. Clear goals and feedback
3. Concentration on the task at hand
4. The perceived exercise of control (strength and self-esteem)

5. The loss of self-consciousness
6. An autotelic experience (intrinsically rewarding)
7. The transformation of time

A parallel definition introduced by Csíkszentmihályi has been to identify flow states by their context, a 'balance of challenge and skill' (typically 'perceived' challenge and skill rather than some objective notion of the terms). Below we see the evolution of *this* conception of flow in his work (the x-axis is skill, the y-axis is challenge). Other researchers have experimented with regression models and latent factor models that are based entirely on this notion of balance as the primary aspect of flow state.

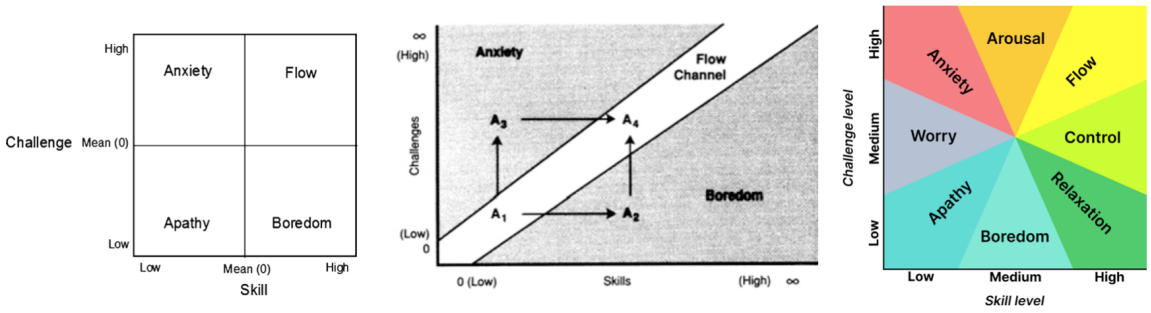


Figure 5-1: Three of Csíkszentmihályi's flow models over the years: the first is his 'quadrant model' (1989), the second is his 'channel model' from 'Flow' (1990), and the third is an updated channel model from 'Finding Flow' (1997).

Csíkszentmihályi's flow definition has mostly settled out of his early work into a nine-dimensional concept– the seven items above with one modification ('clear goals and feedback' separated into two categories) and a dimension for 'challenge-skill balance':

1. A balance of challenge and skill
2. The merging of action and skill/awareness
3. Clear goals
4. Unambiguous feedback
5. Concentration on the task at hand
6. The perceived exercise of control (strength and self-esteem)
7. The loss of self-consciousness

8. An autotelic experience (intrinsically rewarding)
9. The transformation of time

Yet another key aspect of flow states, frequently invoked by Csíkszentmihályi in his prose on the topic, is its connection to optimal living and fulfillment. While this aspect of flow experiences is never explicitly included in a definition, the application of the term in the literature (and certainly in Csíkszentmihályi's work) is almost exclusive applied in aspirational terms to activities that are enhancing to one's life and sense of purpose.

5.1.1 Exemplars

When we think of flow we usually think of high-performance athletes in the zone; Csíkszentmihályi includes games, loosely defined (sports, athletic pursuits, mental, and virtual– things like chess players, rock climbers, and gamers) and artistic pursuits (musicians, artists, and poets). These are 'canonical' examples of flow activity.

According to Csíkszentmihályi's later work, flow specifically *excludes* certain things like watching TV, going to the theater, browsing social media, or losing yourself in a story. He writes "flow always involves the use of muscle and nerve, on the one hand, and will, thought, and feelings on the other" (for Csíkszentmihályi, chess counts because elite chess players train physically).

He thus discounts passive activity. He explicitly criticizes television in his most famous book, pointing out that it doesn't have "the kind of goals and rules that are inherent in flow activities" [99], drawing a distinction between external objects (i.e. television) that structure our mental attention and endogenous control of attention that we exert (i.e. chess). As Keller and Landhäußer summarize in 'Advances in Flow Research', "the activity has to be skill-related for flow experiences to emerge ... That is, activities that are passive in character (such as watching a sunset or taking a relaxing bath) and do not involve a skill component cannot be associated with a flow experience." [239]

5.1.2 ...and Disputes

This criticism marks a departure from some of his own early thoughts on flow, in which he explicitly describes "microflow events such as watching television, stretching one's muscles, or taking a coffee break." As Stefan Engeser argues in the prologue to 'Advances in Flow Research' (forward by Csíkszentmihályi himself) "flow can be experienced not only in typical "flow activities" but in nearly any kind of activity." He discusses flow in non-achievement situations, citing Csíkszentmihályi. "[S]ome individuals mainly experience flow in interactions with others, some in physical movements like walking, some in reading books or watching TV, and some experience flow while watching people walking down the street." Engeser argues that these examples fit the flow definition because Csíkszentmihályi redefines challenges as 'opportunities for action' elsewhere in his work (for exactly this reason).

The ambiguity surrounding passive situations or social situations extends to artistic practice as well– a set of activities where goals are not clear and feedback is not unambiguous. As Charlotte Doyle summarizes in her 2017 article on the topic: [122]

> Interviews with visual artists suggested that in this domain, goals, which are part of problem representations, are not clear (Mace, 1997)... In another interview study, Cseh (2017) concluded that clear goals, sense of control, and unambiguous feedback were not typically part of fine artists' flow experiences.

Doyle makes it clear that what emerges from a creative process is often surprising to the creator, and advocates that this experience of flow should be thought of as a distinct cognitive process.

Finally, there are the set of activities which are autotelic– deeply intrinsically motivating, effortless, and absorbing– that fall even further afield of Csíkszentmihályi's conception of flow than passive activities. These include all manner of addictive, harmful behaviors (that, at least colloquially, most people would label as flow inducing)– gambling and slot machines, facebook and social media, doomscrolling, state of anger or rage– that run against

the narrative of optimal experience. Despite the focus on a well-lived life, flow states are rarely studied in the context of spiritual or religious experience.

5.2 Issues with the Flow Definition.

If we just look at flow as defined by its originator, it is a nine-component concept; sometimes described by just one of the components divided along two axes (challenge-skill balance), only invoked with positive examples that may or may not include passive activities like reading, social activities like debate, or common ritualistic daily activities like walking, making a coffee, or gazing at the sky. Its ambiguity has been a source of confusion, but the critiques run much deeper than the obvious.

5.2.1 Precursors, Consequences, and Covariates.

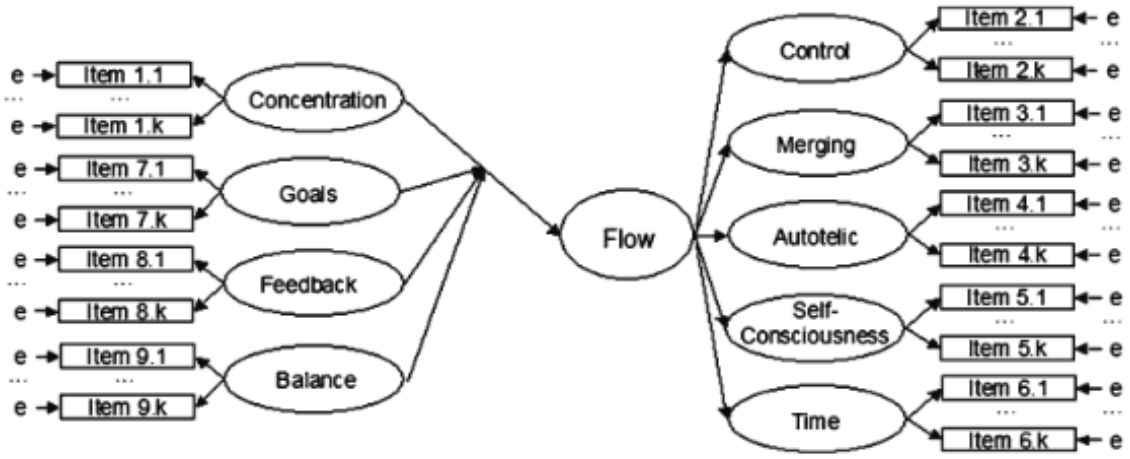


Figure 5-2: Figure 2.8 from 'Advances in Flow Research' [307] by Giovanni B. Moneta suggesting a latent model that separates precursors and consequences of flow.

One of the most common critiques leveled at this nine-component definition of flow is that it conflates antecedents and consequences of flow state with the cognitive state itself.

Antecedents: a balance of challenge and skill

The most common critique of challenge/skill balance is a relatively clear one– the notion of challenge already factors in skill. How cognitively well-resourced and adapted you are to handle a situation– and whether it will demand utter concentration– could easily be reduced to a simple 1-D notion of 'task challenge'. For flow, however, typically researchers wish to distinguish between the laborious, self-aware concentration we might imagine when for a beginner learning a skill with low task fluency, and the automatic, high-performance character we expect of a masterful practitioner in-the-zone. A better conception thus retains the notion of balance, but removes the double reference to 'skill' by replacing 'challenge' with 'task demands.' (This is common practice among top flow researchers) [252].

Even after this fix, there is healthy debate about whether this task structure that we expect to elicit and correlate with flow is actually necessary or sufficient. It turns out there is quite a lot of empirical work to show the contrary– people sometimes have affectively positive, on-task experiences when demands and skill are balanced *without experiencing flow state*. Keller et al. found the relationships between this balance and flow experience was moderated by an internal locus of control or action orientation [238, 236]. Demand/skill balance is not a sufficient condition.

There are also many examples of activities where demand/skill balance doesn't seem to be a necessary precondition either– at least for positively-valenced experiences of deep, effortless concentration that are enriching and meaningful in post-hoc appraisals. In other words, we can have a strong flow experience by all criteria *except the antecedents that relate to challenge and skill* with passive experiences of art, theater, or religious service; in deep, connected conversation; or in moments of awe or meditative depth experienced out in nature or during daily life.

Other task taxonomies– like the mutual information between actions and goal states– have made better predictors of flow cognition than challenge/skill balance in certain contexts [293]. This redefinition, however, opens the door for flow activities that are *not* traditionally considered as such; in fact, the authors stated inspiration for this theory of flow was

the slot machine.[2]

'Challenge/skill balance' describes a set of conditions that should elicit flow, not the flow state itself; as an antecedent, there is strong evidence that it is neither necessary nor sufficient. For certain types of tasks and certain types of people, this condition does imply a significantly increased probability of flow experience. However, even once we leave individual differences aside, challenge/skill balance is not an ideal criteria to separate the full set of tasks that regularly induce flow from those that are not.

Antecedents: clear goals, unambiguous feedback

Some have argued 'clear goals and unambiguous feedback' are redundant criteria because our perception of 'task demand' and 'skill' are already contingent on our understanding of task goals and our progress toward them [252]. Indeed, it is hard to imagine an unclear task with ambiguous feedback for which a user would perceive themselves as 'skillful'– unless, of course, we're talking about artistic practice.[3] There is strong empirical evidence that clear goals and unambiguous feedback simply do not apply for most artistic endeavors (a typical 'canonical' example of flow) [122].

Whether a goal is 'clear' and feedback 'unambiguous' is itself an unclear, ambiguous designation that requires judgement; but art can certainly be exploratory, and its goals can shift in conversation with feedback from the process, fellow collaborators, or artistic output. This precursor to flow once again fails to describe an essential characteristic of task structures that illicit flow-like experience.

[2] As we saw before, the standard definition of flow might be criticized for its flexibility and imprecision; i.e. 'clear goals and feedback' might be extended to art with a semantic flourish (the goal is to be exploratory! The goal is to break new ground! The goal is not to rigidly cling to preconceived notions of what we should take as our goals!); 'challenge' might be redefined as 'opportunities to act' when it fails to account for certain examples. This criticism also applies to the flow theory of mutual information– how we taxonomize tasks and assign information to actions and outcomes is very subjective. These are not falsifiable theories. And as much as I dislike falsifiability as a philosophy of science, it is important to evaluate the plausible explanatory power of these theories without jumping through mental hoops to fit the data to our hypothesis. With enough word-smithing any task could be said to have a 'clear goal' or 'challenge'; the same is true about the way we divide tasks into actions and sub-goals and represent the information therein.

[3] Perhaps also an expert tackling a tricky, novel, poorly specified problem in their domain? Still, task demands and skill seem like the should either be tellingly out of balance or impossible to judge because the task demands are so poorly defined.

The perceived exercise of control

We would expect the exercise of control to correlate heavily with skill in most situations. We might imagine an athlete in a fundamentally unwinnable situation (a star on a bad team) who perceives themselves as meeting the high demands of the task in balance with their skill, though the outcome is outside of their control. Some researchers have argued for the opposite causality; that our ability to achieve flow state during a task drives our self-perception of control [307].

Regardless of whether perceived exercise of control is an antecedent and/or a consequence of flow, it is not the state itself. We can easily enumerate a variety of situations where low control (real and perceived) may actually contribute to flow-like phenomenology. Gambling and slot machines, immersion in theater, media, or art, reading, browsing the internet, ritual, religious experience, awe, anger... there are many kinds of experiences where we experience flow cognition without control. In fact, relinquished control may facilitate flow-like experience, inviting the loss of self-awareness through inaction and an external locus of attention.

Other Consequences

Other aspects of flow are difficult to categorize as core descriptions of the phenomenological state or consequences of it. Certainly time distortion seems to be a consequence of awareness shifting to the task at hand and away from external cues; arguably, whether a task is deemed 'intrinsically rewarding' or 'autotelic' (done for the joy of the means, not the ends) may be a circular post hoc judgement as well.

The remaining three dimensions– describing deep enough levels of concentration that attention to self evaporates– seem to be core descriptors of cognitive state of flow.

5.2.2 Flow Itself

Is Flow a Singular Concept?

Factor analysis of the nine dimensions of flow by colleagues of Csíkszentmihályi show that a single, latent concept does not explain the variance in real-world answers to flow questionnaires well; they suggest that analyses should be run on each dimension of flow individually [220, 4]. In the weakest interpretation this calls into question the validity of the operationalized construct; in the harshest, it calls into question whether the flow concept is truly capturing one latent idea. Other researchers suggest flow should actually collapse down to two dimensions– absorption and task fluency [255, 137].

Is Flow Gestalt?

These questions of representation center around a more fundamental question of the nature of flow– should we think of it as a continuous construct– ever deepening and effortless focus– or as a binary, gestalt one? Csíkszentmihályi wrote "the flow model suggests that flow exists on a continuum from extremely low to extremely high complexity" [96]. Most modern conceptions are moving toward less categorical representations of flow. Moneta– a colleague of Csíkszentmihályi– categorizes flow into two levels, and shows in his work that a third of UK adults experience 'shallow flow' but not the 'deep flow' in their lives [306]. Keller and Landhauber suggest a continuous gradient of flow intensity based on demand/skill balance and subjective task value [252].

The thrust of research is moving in the direction of a consolidated, continuous latent conception of flow (a view that stands in opposition to a nine-factor model). But this is not a settled debate, and indeed its label as a 'state' implies bivalence (you're 'in-the-zone' or you're not). It is reasonable to theorize a categorical, gestalt notion of flow– something greater than the sum of its parts that emerges probabilistically given the proper confluence of state/trait/situation factors. If this is the case, how we operationalize it might naturally require several irreducible, independent dimensions of observation glued together with

probabilistic reasoning.

5.2.3 Is Flow Common?

Using his other major contribution to psychology– the Experience Sampling Method (ESM)– Csíkszentmihályi assessed flow in participants' daily lives using two questions rated on 10-point low-to-high scales (challenge of activity, skill of activity). As part of his methodology, he would z-score each participant such that 'flow states' are any activity with both above average challenge and above average skill ratings for that individual. [98] Assuming normal distributions in response, this results in a flow state designation roughly 25% of the time. Labeled 'the above average hypothesis' by Keller and Landhauber, the method implies flow is abundant in daily life, in opposition to his statements about its relative infrequency (i.e. "**On the rare occasions that it happens**, we feel a sense of exhilaration, a deep sense of enjoyment" [96]).

Most researchers assume flow occurs relatively infrequently and is hard to elicit. Csíkszentmihályi himself compared inducing flow in the lab to "trying to make someone relax in a dentist's chair." [305]. This difficulty makes flow hard to study, and incentivizes more generous interpretations of flow experience as we see with the 'above average hypothesis'.

5.2.4 Is Flow Pleasant?

A central point of debate about flow states has centered around their affective quality– are you happy while in flow, or are you unable to experience emotions at all because you have no sense of self? Here is Csíkszentmihályi on the matter in 1996: "Flow is defined as a psychological state in which the person feels simultaneously cognitively efficient, motivated, and happy" [308]. Here he is again, in 1999: "[D]uring the experience people are not necessarily happy because they are too involved in the task (. . .) to reflect on their subjective states" [97].

In his 2002 book 'Authentic Happiness', Martin Seligman (the other main founder of positive psychology) takes the stance that "it is the absence of emotion, of any kind of consciousness,

that is at the heart of flow" [386]. In 'Advances in Flow Research', Engeser states 'flow is not defined by affective means' [139]. Some researchers, however, focus heavily on emotion when operationalizing flow [283]. While this debate rages on, I believe Corinna Peifer unifies the worldviews nicely [333]:

> Literature provides contradicting statements whether flow is a positive emotional state or if emotions are absent during flow. It is proposed here to rather speak of a positively valenced state, than of an emotion. To experience emotions, self-referential thoughts are needed, which are naturally absent during flow.

Interestingly, the flow physiology literature is split over whether subtle smiling (EMG activation of the facial CS muscle) is a sign of flow or not. It has been demonstrated for musicians, but not for other flow induction tasks like video games. These contradictory results potentially highlight a true difference between artistic flow tasks and other kinds of flow task– for artistic tasks, deep and complete awareness of the affective response elicited through artistic action may be central to flow; for other types of flow task, it is not.

5.2.5 Is Flow Value-Aligned?

As we saw in Csíkszentmihályi's quote that started this chapter, he defines flow primarily as 'the positive aspects of human experience' in line with prior work on 'peak experience' (a similar experience outlined by Maslow decades before as he studied the same idea). Its relatedness to joy, fulfillment, and meaning are consistently highlighted in the literature.

However, it contrasts with prior work by defining focusing on the cognitive state rather than the set of experiences. Maslow defined 'peak experience' as the study of fulfilling activity and what constitutes it– the cognitive state is not included a priori as part of the definition. Flow states– however– are defined by the phenomenology first.[4]

[4]Peak Experiences ask 'Given I have judged that experience as a fulfilling one, what occurred?' Flow instead asks instead 'Given you lost yourself in an intrinsically motivated way in what you were doing, what happened?' The value-aligned assessment of experience motivates the study for the first; the cognitive state motivates the latter.

Much like Maslow, Csíkszentmihályi has clearly been motivated exclusively by peak experiences when he started to study flow; he has kept flow states linked to value-aligned experience throughout his writing in all of his colloquial definitions of flow, though it never appears in his precise ones. Hence the ambiguity around whether flow-state applies to gamblers, addicts, and passive media consumers; whether these experiences are judged as 'peak, fulfilling experiences' or 'inescapable, addictive forces of destruction' has to do with a post-hoc value-/identity-laden appraisal rather than the phenomenology of the experience itself.

There is very little work on flow in these kinds of situations, with only a handful of researchers starting to study 'the dark side' of flow in combat, addiction, online, and while gambling. [381]

5.3 Related concepts.

Flow as a concept sits among many related concepts. In this section, I'll give a brief, high level overview of the ontological landscape surrounding flow.

5.3.1 Altered States of Consciousness and Trait Absorption

Flow is an altered state of consciousness. As such, it shares many features with trance, hypnosis, awe, religious experience, dissociation, depersonalization/derealization, and other unusual cognitive states. Of particular note is hypnosis– a state which motivated Tellegen (who wanted to identify hypnotically susceptible people) to create the Tellegen Absorption Scale (TAS) in the 1970s [425]. This scale weakly correlated with hypnotic suggestibility; instead, it became a measure of trait absorption and is the most common measure of a *trait* related to flow today [281].

5.3.2 Disentangling Flow

When we study flow, there are many other cognitive forces at work that we might need to co-model, control for, or otherwise acknowledge. Modern experimental psychology describes stress, emotion, cognitive load, perceptual load, alertness/arousal, engagement, immersion, attention; a host of related ideas about the state of our mental operation and attention. Additionally, we have personalities, beliefs, and motivations that can structure our mental states– self efficacy, intrinsic and extrinsic motivations, resilience, confidence (well-justified and otherwise), etc.

Empirical work on any one of these ideas reveals the complexity underlying their relationship. For example, Kahneman argues our attentional capacity is drawn from a reservoir of cognitive effort that scales with our level of arousal [73]. Arousal is also a fundamental dimension of emotional state, as well as stress [228]. Isolating any one from any of the others becomes nearly impossible in experimental work; we can rarely claim with certainty that we've only modulated one parameter. We rely on unspoken, simplifying assumptions– that these latent cognitive concepts are independent, or their causal interdependence is weak or controlled for.[5]

5.3.3 Situating Flow

Engeser and Schiepe-Tiska point to precusors of the flow concept as Piaget and Caillois's 'Play', Hebb and Berlyn's 'Optimal Stimulation', White's 'Competence', DeCharms 'Personal Causation', Groos and Buhler's 'Funktionslust', and Maslow's 'Peak Experience' [138].

Flow itself shares much in common with concepts like absorption or immersion (frequently used to describe flow states alongside task fluency); it also shares a lot with engagement, though engagement doesn't imply the consistent depth of attention. In my opinion, the

[5] To illustrate, we may vary our task difficulty to influence flow; that probably means we've also altered the number of perceptual objects, and thus perceptual load; we've altered the amount of processing required, and thus cognitive load; we've placed more or less stress on the user, and made the task more or less enjoyable, etc... to really understand the pure relationship of task difficult to flow, we need to map the entire constellation of tasks that vary over all dimensions, to their latent mental concept, and estimate the interdependence of each pair of cognitive experiences.

best extension of flow is the concept of 'deep, effortless attention' [72]– which targets the cognitive state associated with flow, while dispensing with flow's ambiguity around value-alignment, emotion, and categorical subtext ('acheiving flow' or 'being in the zone'). My preference is to operationalize this notion of deep, effortless phenomenology in the moment, and separate it from (1) an attempt to separate 'passive' from 'active' activities, when in both cases the automatic processing is fully activated regardless of our role in the task structure; which I see as an instinct rooted in Csíkszentmihályi's desire to (2) separate activities based on a post hoc assessments of whether they are meaningful and fulfilling, and only apply 'flow states' to the good ones.

This redefinition– as a type of attention– also properly situates this concept in the broader landscape of work on attention. Attention itself is a rich topic in perceptual psychology and neuroscience (as we saw in the last chapter on auditory attention) with rich taxonomies and debates around early and late processing, selection, and automaticity [24]. These insights are quite relevant to the ontological debate surrounding flow; yet models of attention are rarely discussed in flow contexts.

When it comes to the other side of 'flow'–understanding the nature of fulfilled living– Maslow and the humanistic psychologists were the first to introduce many of the modern concepts surrounding flourishing (and in my opinion, psychology's best). Maslow's treatment of peak experiences (and later plateau experiences) powerfully speak to the ephemeral quality of experience alongside the cognitive postures and habits that structure a well-lived life. His work, of course, rests on the back of millennia of philosophers and religious leaders, who have tackled eudiamonia, hedonia, and spirituality; the kinds of experiences, beliefs, and attitudes that weigh in the balance of meaningful living. There is something to be said for the way we approach our daily tasks; the joy of mastery, progress, and execution; the state of total connection to our craft, our family, or our world. States of effortless attention matter, but they are only a small part of the answer to a larger, eternal question.

5.4 How do we measure flow?

The confusion around the flow concept has resulted in a fractured literature; one review found 24 distinct implementations of flow in just 42 articles [17]. There are literally dozens of instruments measuring flow in different contexts and different languages. In mainstream flow research, however, there are only a few general purpose methods that have become most dominant. These methods were used by Csíkszentmihályi or endorsed by him [312].

- **FQ** One standard technique to measure flow experience in general is with the Flow Questionaire (FQ), a series of descriptions for which participants check 'yes' or 'no' to indicate whether they've ever experienced something similar. This was used i.e. to test how many people had ever had shallow or deep flow experiences by Moneta [306], and is generally regarded as a good instrument for assessing flow prevalence.

- **ESM** Csíkszentmihályi used his Experience Sampling Method (ESM) to assess flow in participants' daily lives, by paging them at random times to answer two questions rated on 10-point low-to-high scales (challenge of activity, skill of activity). As part of his methodology, he would z-score each participant such that 'flow states' are any activity with both above average challenge and above average skill ratings for that individual [98]. This method has been criticized as it result in roughly 25% of daily life labeled as 'flow' on average [17].

- **FSS, DFS, Short FSS** The Flow State Scale (FSS) is a 36 (5-point) Likert question survey in which four questions are mapped to each of flow's 9 dimensions developed by Csíkszentmihályi's colleague Susan Jackson for sports applications [221]. After critique by Doganis, Iosifidou and Vlachopoulos [120] the FSS was refined to the FSS-2 with five questions replaced. This version of the scale was validated in music-playing and education contexts, to make it a general use flow scale [222]. The FSS-2 is also adapted to measure flow trait and called the Dispositional Flow Scale (DFS and DFS-2) by simply changing the tense of the questions so they refer to a participants past experience in general instead of their current activity. In the 2008 paper introducing these scales, Jackson also introduced short versions (the short FSS-2 and short

DFS-2), which are 9 questions instead of 36, and still map to the original 9 axes of Csíkszentmihályi's flow.

- **FSS** Another general purpose scale, called the Flow Short Scale (FSS– yes, confusingly, sharing the same abbreviation as the Flow State Scale), is the only scale developed outside of Csíkszentmihályi's direct orbit to have gained widespread adoption (especially in Europe, where Flow research has been more popular). This scale was developed initially in German [358], though it has been translated broadly. It includes 10 (7-point) Likert questions (compared to the short FSS-2 9 questions). Though the questions are similar, the FSS projects its questions down to two latent dimensions for analysis– Absorption (4 questions) and Task Fluency (6 questions).

These scales are not without critique [258], though most of the criticism follows from the critique of the flow concept itself rather than the instruments. These techniques tend to be unruly (9 irreducible dimensions to analyze?) or improverished (two questions that suggest an unrealistic amount of experience is in flow). With so much dispute about the core concept itself, there is certainly room for improvement.

5.5 Improving Flow Estimation

One approach to improve the flow definition is to divide the concept back into the two separate areas of study from which it owes its origins– deep, effortless attention (the phenomenology of experience) on one hand, and its interaction with fulfillment (a la 'peak experiences') on the other. This decision leaves us with a question about how to properly operationalize the concept, given the typical questionnaire has several questions that either (1) relate to antecedents for challenge/skill balance that are not integral to the core phenomenological state of flow, or (2) probe an activity's importance or reward based by post hoc, value-laden assessments that have little to do with the experience in-the-moment. These criticism are in line with a new but growing movement to redefine flow-like states as simply 'deep, effortless engagement' [72]; stripping it of many of the nine features that currently define it.

In the following section, we propose two improvements to standard flow measurement. We can improve our surveys to more accurately capture user-experienced flow, especially as they experience it unfolding over time– a dimension of flow experience that has been largely overlooked in standard survey methods. We can also shift our focus to include a priori trust in bio-behavioral indicators of flow states.

5.6 Survey Improvements

5.6.1 The Standard Flow Survey

Regardless of the survey questions we choose, we're left with self-report ratings of the experience itself– self-interrogation of a state defined by its dissolution of self-awareness– and its consequences (i.e. how did time feel?). When we use this method to measure flow, we (1) have a number (i.e. 3 out of 5) with no notion of uncertainty that represents flow state over a multi-minute interval. Moreover, self-report is subject to biases (demand characteristics, memory 'peak-end' effects, and participant self-awareness).

We can improve the survey process alone in several ways. The typical flow study asks nine specific questions (i.e. 'I feel I am competent enough to meet the highest demands of the situation.') or two questions ('challenge/skill of activity') instead of one simple one ('Did you experience flow?'). It suggests we either don't trust people to understand the concept or we expect a more accurate insight by asking several diverse questions instead of one.

If we educate a participant with examples and clarify exactly how we'd like to operationalize the definition, this question seems as reasonable for self-interrogation as the other standard questions. Moreover, given the literature demonstrating that these surveys merely correlate with an increased likelihood of flow state rather than indicate the core state itself, this seems to be a better (only) way to get a true self-assessment of flow state. We introduce this self-interrogation to our surveys.

While asking about a gestalt concept like flow implies reaching for a higher level of abstraction than usual in our survey, we might favor lower levels of abstraction with other aspects of the standard questionnaire. For instance, in typical practice, we ask secondary questions about time perception ('how did it feel?') instead of interrogating time perception directly by asking for duration and certainty estimates.

We might hypothesize, however, that the answer to 'how time felt' could be self-interrogated in multiple ways. The first is contingent on two estimates– how long you believe has actually elapsed vs. how much of that time you felt elapse (i.e., we could imagine your answer to 'how time felt' would drastically change if it is revealed to you that 2 hours had passed when you thought it had actually just been 15 minutes). Alternatively, we might imagine someone answering this question with a pure self-assessment of certainty– i.e. how accurate do I feel I can be in my estimate of the current time? In this model, they would be unsurprised to find out it had either been 15 minutes or 2 hours, and this 'sense of certainty' would not change their rating with knowledge of how much time has actually elapsed.

This seems to be an opportunity to improve our understanding of the relationship between the flow state and dimensions of time experience– by explicitly asking about certainty, we also have an opportunity to integrate some preliminary notion of stochasticity into our

representation of flow instead of treating it as a point estimate. We thus add a quantitative slant to time experience by asking participants to (1) guess the time, (2) guess the duration on task, and (3) estimate their certainty in both cases.

5.6.2 The Peak-End Rule

In 2008, Kahneman demonstrated an interesting effect– that over short painful sessions of hand in cold water, people would state a preference for a session that had the exact same pain experience as another, but added a short duration of less pain at the end [119]. This experiment shows that memory does not simply integrate over an experience; our recollections and judgements are filtered through a memory process that is biased towards memorable moments– hence 'the peak-end rule.' The peak-end rule has been replicated robustly over many tasks [209] and time-scales [415].

Unfortunately, the rule is really more like a 'heuristic'– though the peak/end of an experience is better correlated with final ratings than averages, the relationships are more complex than a clean weighting of peak/end experience– for example, sometimes the end doesn't seem to matter much at all [345, 297, 165].

This insight about how a survey result applies to the task under study points to the lack of consideration for the dynamics of time in flow survey results. We expect flow states to take some time to access, yet we have no sense for how long this takes to achieve nor how fragile that depth of focus may be.

5.6.3 Adding a Notion of Time

In our work, we use the standard short FSS-2 to benchmark against other literature. Because this survey asks questions as they pertain to the entire interval, the peak-end rule would suggest we should assume that they apply only to salient moments during the interval. In other words, we should expect that that judgment does not map to the entire interval or its average, but instead a few unknown timeslices of the experience.

Mapping these overall judgements back to the interval can be difficult, since we don't know how they apply in time. To interogate the time dimension of flow, we introduce questions that specifically request data about the most salient moment of flow (what was the *peak* depth of flow you experienced?) and durations (how long did it take you to get into flow? What percentage of the time were you in flow? Was it one continuous experience of flow, or many small experiences?). This separation of intensity and time helps disambiguate the estimation process a participant may use in their overall ranking.

As part of our experimental work, we also asked participants to draw their flow experience over time. Surprisingly, people reveal a diversity of temporal experience through these drawings, especially in the face of different task structures. It's now quite easy to analyze these kind of data, fit functions to them, and utilize them as priors. How to properly interpret and utilize this gestural data about a time interval is an open question.

5.7 Behavioral and Physiological Measures in Context

When we think of the people we know, we can usually tell whether they are in a deep state of focus without surveys. At work, they might be very still and locked on to what they're doing with a contorted but relaxed facial expression. They might have a repetitive behavior that they engage in unconsciously like shaking their leg or clicking a pen. For a musician, their expressions may distort in line with what they're playing in a way shows deep connection to the sounds they're producing, with a corresponding lack of focus on how they look to others, their visual environment, and their physical body. Unlike a worker in the zone, their eyes are likely closed– the last place we'd expect them to look is at their hands– and their physical movements exaggerated in line with the rhythm.

In all cases, interrupting them might be very difficult to do– their threshold for noticing these distractions is high in these moments; and their corresponding reaction is typically one of heightened shock. Observing this pattern– a person oblivious to relatively obvious external stimuli and their corresponding surprise when they recognize it– is the normal

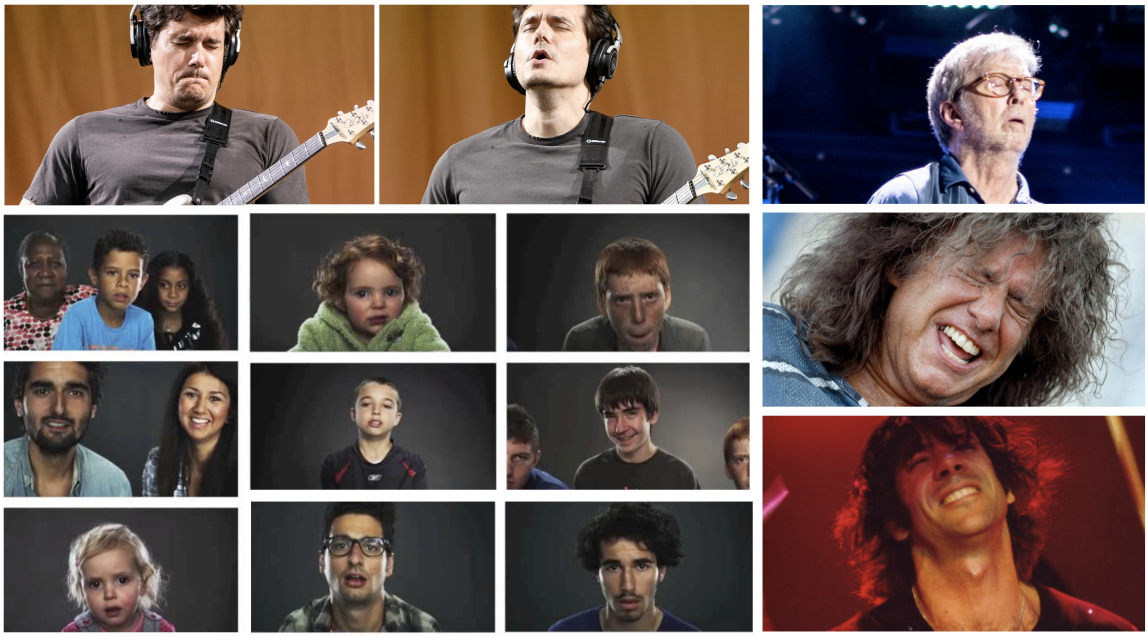


Figure 5-3: Diverse flow faces; we see Musicians John Mayer (top-left), Patrick Metheny (middle-right), and Christian Vader (bottom-right) furrowing their brows and contorting their faces; Eric Clapton, in contrast (top-right), has an unusually relaxed expression. All of these musicians have their eyes closed. The images in the lower left are from Robbie Cooper's 'Immersion' art project, showing people playing video games. We see a different kind of facial expression here; eyes locked, with expression-less that may be relaxed or may be contorted. Blink rates seem to slow, head motion is either non-existent or shifts unusual as gaze remains unbroken. Despite the differences between musicians and gamers, the subtle cues suggest a similar utter focus on the percept being acted upon. Photos of Metheny/Vader courtesy of Joe Paradiso.

mechanism we use to gain confidence that the subtle behavioral and physiological cues we attribute to that individual's focus are correct.

We can model our formal process of reasoning about flow states off of this intuitive process we utilize in our lives. We don't use surveys to build our models of other peoples' attentional states, yet our mental models are typically very good. There are a few important features of this anecdote: (1) models likely need to be heavily personalized and may vary by activity; if we care about understanding real instances of natural flow, we need to study it idiographically, with real activities, in real contexts, (2) we have access to other powerful indicators of attentional depth that we should trust a priori (as much we trust survey results), and we can combine them with surveys to infer the underlying cognitive state. To formalize these improvements, we suggest a move toward:

- **Field Measurement** Csíkszentmihályi studied people in their normal lives with surveys; he compared inducing flow in the lab to "trying to make someone relax in a dentist's chair." [305]. Flow is rare; to study it effectively requires tools that people can live with. As far as flow research goes, so far, that has only involved surveys and pagers.footnotePagers are used as part of 'Experience Sampling' to simply deliver short surveys at random moments, and thus mitigate the biases of recollection.

- **Psychophysiology** In the lab, several people have studied or put forth psychophysiological theories of flow. [333]. These usually follow from Dietrich's hypo-frontality hypothesis [116] (downregulation of prefrontal cortex because attention is effortless) with EEG/fNIRS or the hypothesis of a distinct physiological profile for flow. Preliminary fNIRs work has shown no evidence of hypofrontality [192]. There is contradictory evidence for facial EMG and EDA; possible trends for salivary cortisol, and general corroboration for decreased HRV [240, 283]. Much of this work is preliminary; all of it relies on simple survey results compared against population level trends in summary statistics. Psychophysiological modeling in this space is still relatively underdeveloped.

- **Behavioral Cues** An under-utilized tool for studying flow states, there are many be-

havioral cues we can use to infer depth of focus (much like we do as people). Obviously, someone in flow is not task switching, self interrupting, and shifting their attention in an overt manner. Beyond these obvious indications of inattention, someone's stillness, blinking, and relaxed face posture can indicate the dissolution of self-awareness; they also stop responding to distracting stimuli– until, of course, a sufficiently substantial interruption takes them out of the moment (large startle responses are a good indicator of flow in the moment before). We can tap into these common sense signals to estimate flow.

Our goal is to move from (A/B) in Figure 5-4, where we treat flow as deterministically and perfectly linked to one number over the entire interval duration, to (C), where we use multimodal insights to estimate *the probability* that they are in an underlying flow state at a specific moment.

5.7.1 Toward Data Generating Theories

In the world of (A/B) above, researchers put forward variations of the survey in line with their beliefs. These definitions are justified by intuition, rarely compared, and largely tautological in the contexts where they are applied (there is no arguing with a particular definition; it simply represents a new variant of the concept that highlights certain features the author deems important). This is problematic, because the expansion of practical definitions stands in direct opposition to theory-building and scientific abstraction. Moreover, with so many definitions in practice, we can't easily identify which are the best abstractions.

Comparing and consolidating definitions is a necessary first step; but in the context of current practice, this is difficult because the theories are not data generating. If we wish to compare them, we must assume data-generating features about them; for example, in the contexts for which we a priori expect the definitions to apply, the best one gives more consistent/high ratings compared to others, maximizes the variance, captures a presumed feature of the underlying distribution, or minimizes overlap with other established concepts. Notice that none of these claims are encoded in the definitions themselves.

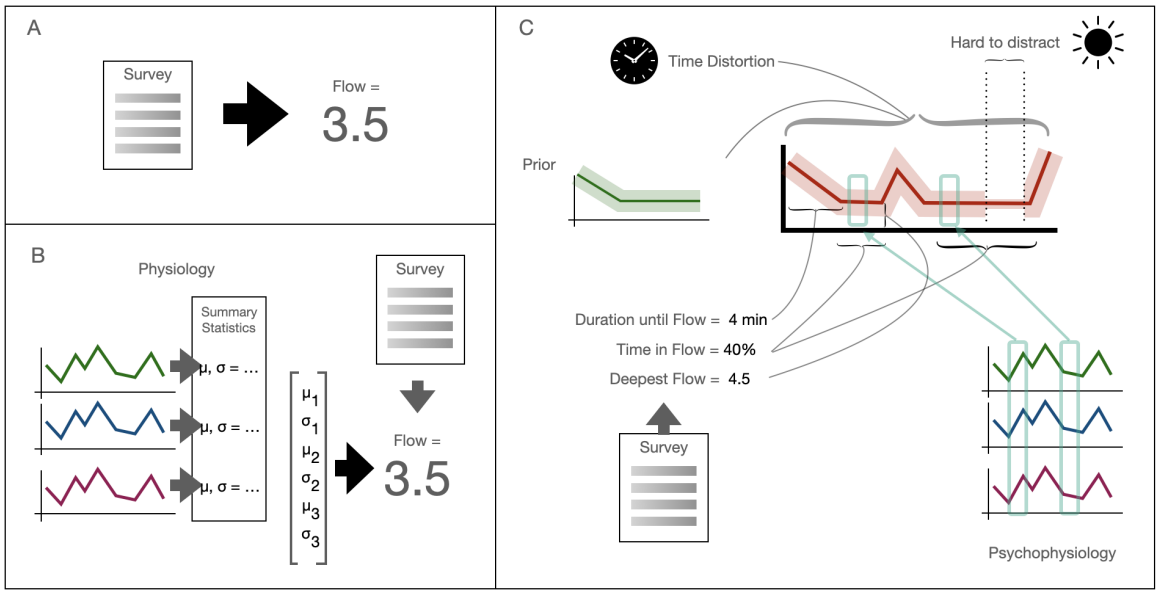


Figure 5-4: A Representation of Flow Estimates. (A) Surveys; they don't capture a notion of uncertainty for the interval that was ranked other than perhaps something like Cronback's alpha for internal consistency. More or less, you have '4 out of 7' flow for a task, unidimensionally or multidimensionally. (B) We try to learn what physiology represents flow states. We treat the state as known– it is once again the singular number reported by the user, the 'ground truth', for a given task. We typically will then treat the entire task as 'in flow' or 'out of flow' as labeled, and compare summary statistics to the number. More sophisticated papers may apply more intelligent processing of the time-series data. (C) Our proposal is to learn a probability distribution per timeslice during an interval, and adapt the survey questions to give insight into the time-varying nature of flow and attention. We also apply our collected data intelligently to the relevant time steps– if we introduce a small distractor that goes unnoticed, we can infer something about that brief period. If the user incorrectly estimates the passage of time, we expect that to reflect an integrated estimate of their experience. If they give a survey response, we expect that to apply to some moment of peak salience, not the entire interval. Additionally, instead of treating survey responses as ground truth, we treat them as strong suggestions; we co-learn all relationships. When multi-modal signals all agree, we have more faith in the ones that do and trust them more. Thus, if someone has a clear physiological indicator that they are focused (i.e. unusual stillness while working), we can infer latent focus state from that *just as much* as we infer that the physiology is representative of flow. We move improve (B) with a stochastic, time-varying notion of flow for which survey data is treated as one of many noisy sources of data for the underlying state.

The real definition of flow is this data-generating claim we use to compare it. In other words, we see two kinds of statements that rarely intersect in the literature– *'Flow is X.'* (i.e. the answers to these survey questions) and *'We expect Flow to have features Y under conditions Z.'* (i.e. a good definition of flow will be one that gives high ratings for musicians and gamers). Instead, we should say *'Flow is the thing that has features Y under conditions Z.'* Our definitions are theories; our theories need to be data generating so we can compare them.

When we move to multi-modal, probabilistic representations of flow, we are faced with a similar proliferation of possible theories as we see under surveys– what are the underlying relationships between these measures? Which do we trust a priori, and how much? In contrast to the survey-based definitions however, these definitions explicitly encode the data generating process about how flow manifests across all of the measures for which we have access. It is easy to compare and improve theories based on how well they explain the real distributions of data we capture.

This forms the basis of our contributions to flow research. We attempt to (1) base our estimate of flow on self-report, physiology, and behavior; (2) we build tools to capture these signals in the course of normal daily living; and (3) we modify our use of surveys in several important ways such that they give us insight into the time-varying nature of flow over a task interval. Moreover, we refrain from treating them as 'ground truth'– which they are not– and instead treat them as one of many noisy estimates from which we can estimate the true latent flow experience. When our predictors agree in common sense ways, we have higher confidence; when we have higher confidence, we learn about the trustworthiness and consistency of our predictors.

We can thus create explicit, data generating theories of flow to describe the underlying relationships between these indicators in the language of PPLs. Armed with a large dataset of naturalistic, bio-behavioral flow data, we can compare the our theories against real world data and converge on the best definitions.

5.8 New Wearables for Flow Estimation

The following sections introduce wearables I've built to study flow in the wild.

The first, Equinox, is a watch designed to integrate into someone's life as they hide the rest of their clocks– when they need to check the time, they have to first guess it with this watch. It enables the study of time distortion in daily life; we show that– despite the rigidly clocked nature of modern life– people readily experience time distortion during their day, and there is variation we might take advantage of when we try to estimate flow naturalistically. Based on the literature, we expect time estimate error to accumulate relative to attentional/emotional state over an interval.

The second, Feather, is a leg-band that pokes the user at various levels of intensity throughout the day. It drops down to a barely noticeable level and slowly gets stronger until the user notices it. In initial testing, it seems that the threshold at which people notice a small distraction fluctuates throughout their day; moreover, this gives an interesting and accurate point estimate of focus level.

The final, Captivates, is a pair of smartglasses really targeted to measure head stillness and blink rates, which are common sense indicators of deep flow (see again i.e. Figure 5-3 Robber Cooper's 'Immersion' art piece). Captivates also allows an intervention to test awareness/vigilance– an LED in the periphery of the glasses will infrequently change colors from green to blue. This change is slow enough to be 'change-blind'– not to draw attention to itself. By looking at how long it takes a user to notice the color change, we have yet another quantitative behavioral insight– this time, good assurance that the participant is in a state of flow just before and during that period should they miss the change; and good assurance that they're not particularly focused at that moment should they catch it immediately. Preliminary tests show delays up to 15 minutes before noticing.

Figure 5-5: David sporting the wearables described below.

Figure 5-6: The custom 'Equinox' watch can notify the user, monitor ambient light levels, and collect time estimates, duration estimates, or experience sampling survey data all day long. In our initial test, we ask users to guess the time in order to check the time.

5.9 Wearables for Flow Estimation: Equinox

5.9.1 Introduction

Our perception of time varies with our experience. Time flies when we are deeply engaged in 'flow' (i.e while playing music or sport) and slows when we are highly aroused or stressed (i.e. during a car accident) [139, 411]. Research has shown significant time distortion when we are highly engaged; under extreme states of stress or emotional arousal; and when environmental cues are removed. By measuring these time distortions, we can thus (1) infer someone's cognitive experience and state of deep engagement [70] and (2) evaluate design (time perception has become an area of increasing focus within HCI [318, 263], and shaping time perception is frequently an explicit interaction design goal [460]).

Time distortions have been studied in controlled laboratory situations (like during a video game or while looking at emotionally evocative images) and in specific, extreme situations (like cave exploration or free-fall) [370, 363, 314], but have not been studied in daily routine. Clocks are ubiquitous, which makes the phenomena difficult to capture.

This research explores the possibility of extending time perception research to naturalistic settings. We present a new wearable called 'Equinox' specifically created to enable daily time perception research, and tested it with a small set of users to understand the feasibility of ecologically valid time perception measurement. The contributions are as follows:

- We present a new watch interface specifically designed for time perception study. To our knowledge, this is the first intervention targeting the study of time perception across normal daily experiences.
- We present a novel time perception time experimental paradigm. In order to integrate into modern life, Equinox asks participants to guess the time when they need to check the time (we refer to this guess of the clock state as a 'clock-time estimate'). To our knowledge, clock-time estimates are a novel approach to study perceptual distortions in the time perception literature; time perception literature exclusively uses intervals or durations.
- Equinox was evaluated by seven initial users over 31.8 hours in their normal lives (playing sports, working, etc). This data provides insight into the feasibility of naturalistic time perception study. We aim to answer a few questions: (1) Can we collect meaningful time perception data at all during a user's normal, highly scheduled life, and on what time scales? (2) Do people experience meaningful and varied time distortions on a daily basis? (3) Are there obvious trends in the data that suggest clock-time estimates are a feasible or valid way to approach time perception research?
- Finally, we summarize our lessons learned to inform future design work.

5.9.2 Background

William James famously suggested that 'varied and interesting experiences' feel short in the moment and long in retrospect [223], and the data has borne him out. Prospective time-keeping (where the user is informed ahead of time that they will be asked to estimate duration) are usually more accurate and less compressed than retrospective assessments

[71, 202]. There is less research on longer time estimates (hours to days), but the studies that have tested these intervals suggest similar trends [70, 433].

Time perception is usually measured prospectively by estimating, producing, reproducing, or comparing durations [185, 430]. While production has been used to assess perception of intervals on the order of minutes (i.e., play this game until you feel 10 minutes have passed); for longer durations, estimation is the most common and practical strategy.

Influences on Time Perception: Arousal, Attention, and Environment

Emotional state alters time perception [432, 124], but recent work suggests the mechanism is mediated by arousal, a.k.a. the intensity of a felt emotion– for instance, sad faces do not have the same effect as frightened ones [123]. Research in depressed patients corroborates a relationship between time perception and arousal [430]. Increased arousal during short, adventurous activities slows time up to 30% [411].

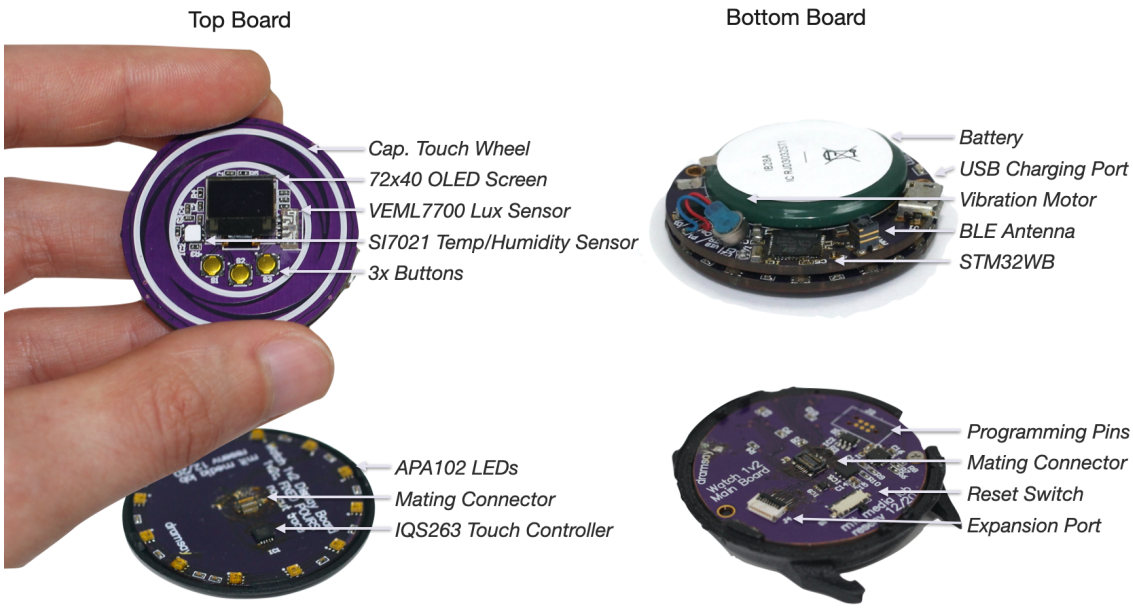
Deep focus is another cognitive state that alters time perception dramatically. Time estimates are 40% underestimated during hypnosis [404], during video games [370, 314], and in flow [191]. It is common to see this relationship exploited to interrogate video game immersion [314, 363].

Environmental cues also can alter our perception of time. Time spent building small scale models of i.e. trains or ships appears to compress time [299]. Moreover, light-deprived cavers will drift towards a 48 hour circadian cycle and underestimate their time underground by about half [190].

Models of Time Perception

While there are challenges unifying a theory of time perception with the variability introduced by memory, duration, and task structure[6], the 'pacemaker-accumulator' model first

[6]Take 'Vierodt's Law'– commonly reported in time perception literature across durations– that people have a central tendency when making time duration estimates, and will overestimate short durations and underestimate longer ones. This effect has recently been called into question based on systematic methodological errors in time perception research. [176]



Deconstructed Equinox to show the PCB layout and mechanical design.

Figure 5-7: The two PCB design of the Equinox watch. This design sandwiches two boards together, so that the top board can support a large, flat touch interface. Individually addressable LEDs emit light around the mid-plane of the watch.

proposed by Treisman in 1963 is still dominant [435]. This two-part model is comprised of an internal clock (influenced by arousal/environment) generating ticks that are counted by an accumulator (influenced by attention) and filtered through memory. Some intrinsic neuro-scientific models [124, 423] support this conception, though other modality-specific models are gaining traction (i.e. 'dedicated' neural circuits just for motor prediction and timing) [48, 423].

5.9.3 Equinox

A watch interface is a minimally disruptive, well-received interface for psychological experiment [198] and is one of the only ways to preempt phone-based time checks. Equinox (Fig. 5-6 & 5-7) is a custom watch interface that captures white light/overall light levels, temperature, and humidity; it syncs with a phone over BLE and keeps track of the time accurately. It provides a 72x40 OLED display, three tactile buttons, twelve LEDs, a vibration motor, and a capacitive touch wheel for user interaction. The watch com-

municates to a cross-platform native smartphone application, and authorizes and uploads data to a Firebase database instance which synchronously supports multiple watches and users; it is also programmed with a range of interactions including duration estimates, clock-time estimates, and experience sampling. Equinox is an open hardware design, with PCBs available at https://oshpark.com/profiles/dramsay. A bill of materials, custom firmware code, mechanical design files, and the companion application code are all available at https://github.com/mitmedialab/equinox.

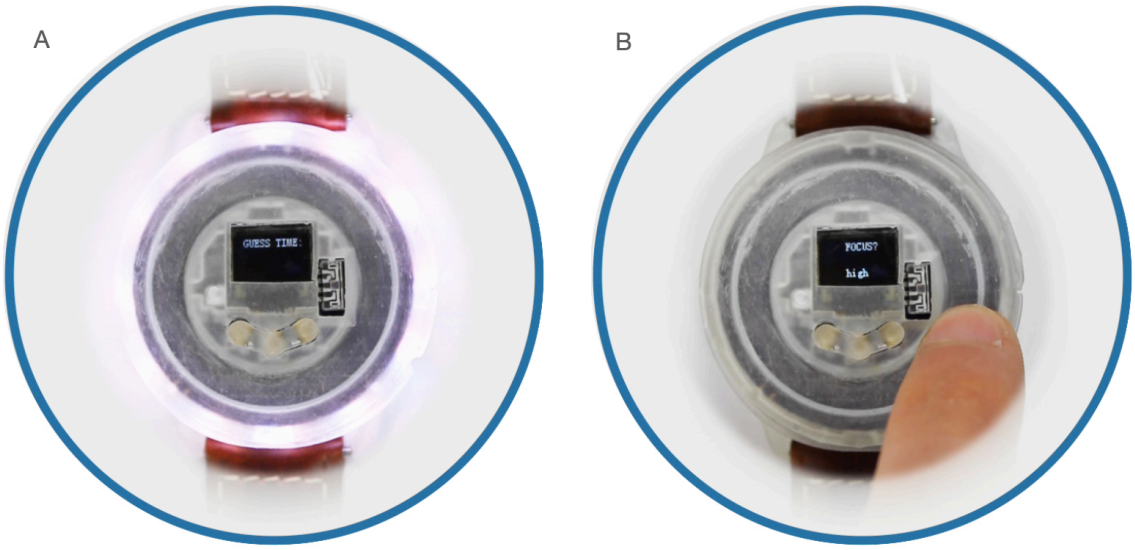
Design Considerations

Sensor selection and UI design considerations in Equinox are driven by the need to adapt laboratory time perception measurement instruments to daily life, embody them such that they're usable and reliable, and to collect secondary environmental data that might influence timing judgement.

We prioritized measuring time perception over minutes to hours, to ascertain a picture of time distortion across a full day, with interaction support for both duration measurements (*where the user starts the watch, and then estimates the elapsed number of minutes when prompted*) as well as clock-time estimates (*guess the time when you need to check the time*). The device provides enough UI degrees of freedom to prototype any major time perception task.

We included secondary measures of temperature and humidity because they can indicate indoor/outdoor context switches [105], influence arousal level [186], and provide an ambient baseline for additional skin-temperature arousal monitoring techniques [114]. We also include ambient light measurements, as peripheral light can be a powerful indicator of time passing [422, 190]. These measures give us insight into arousal and environmental cues– the two main factors outside of engagement– that might affect time perception, especially when paired with additional wearables [84, 349].

We want to integrate time perception measurement into a user's life in a way that is minimally disruptive and practical. Users need to check the time frequently; moreover, almost



Equinox interaction design examples.

Figure 5-8: Two possible interactions with Equinox. (A) shows the watch vibrating and lighting up to prompt a user to guess the time, (B) shows an example of an ESM survey where the bottom of the touch dial maps to a 5-point likert scale rating. For our main exploratory analysis, we allowed the user to initiate a 'guess time' interaction, and displayed the current time immediately after their guess.

every screen-based UI immediately greets users with the current time, making it difficult to hide explicit time cues from users. We designed our interaction as a minimally invasive, easy to bypass interaction that provides the current time immediately, with minimal overhead, and which can be easily dismissed and reset should the user receive an external cue.

Our desire for minimal disruption during field use also informed our engineering goals– all day battery life, responsive and precise touch UI, data caching when operating without a nearby phone, accurate time-keeping, and easy charging.

Finally, we had a desire for researcher usability, enabling flexibly prototyping of various time interactions, seamlessly data collection, and easily integration with a wider wearable ecosystem. Hardware and software transparency is integral to physiological monitoring [319]; this platform affords hardware upgrades and extensions, alongside easy cross-platform app development with a single toolchain. This makes Equinox a useful, open-source prototyping platform for others.

5.9.4 Evaluation

Preliminary Stages

We started with a preliminary engineering evaluation to check that all engineering systems were successfully capturing a range of realistic conditions (i.e. direct light, cold temperatures), that data was correctly cached when out-of-range of a BLE master device, and that the touch controller provided smooth, single-minute resolution data entry. Battery life was >10hrs in all tests, and clock drift never approached 1 minute (our lowest resolution).

We performed short interviews with potential users, to inform a few preliminary design decisions before running our initial usability study:

- **Time Intervals.** After building the watch to support up to 15 second steps, we decided on 1 minute resolution, based on how frequently we expect people to check the time.
- **Clock-time instead of Duration.** Duration measures require the user to initialize a timer, and should be rendered invalid if they see a clock (a peripheral time cue) during that test. We decided against this interaction because (1) it puts an initialization burden on the user, which they must remember and which primes them prospectively (retrospective measurements are preferable because they show more time dilation), and (2) it creates a parallel, secondary timing task from the primary user timing task of tracking their daily schedule. This adds a lot of friction, and room for unwitting bias from checking the clock in parallel. We instead decided to integrate the time-perception measurement into the time-monitoring process people already engage with. We use the natural instinct to check the time to capture measurements; it also makes 'mistakes' more clear to users (glancing at a clock during a task where you estimate the clock-time obviously invalidates that interval; this is less obvious for duration estimates).
- **Expectations of 'Mistakes'.** Participants were willing to put post-it notes over clocks around their workspace and switch to push calendar notifications to alert them

for meetings. We still expect many intervals throughout the day to be invalid because of scheduled cues or clocks; thus, we designed an interaction to accommodate these frequent 'failures' with the tap of a button.

- **Natural Breaks instead of Interruptions.** Instead of interrupting the user to initiate a clock-time estimate after a random interval with a vibration and light alert (Figure 5-8), we decided to allow the user to use the watch as they normally would to check the time, but with a short preceding interaction (either guess the time, or hit a button to indicate you'd seen a clock or received a reminder since you last looked at your watch). This interaction is easy to bypass if the user is in a rush; it also maximizes the interval lengths we can study naturalistically, and gives us an additional signal into user stress and focus (i.e. we expect stressed/distracted users will check the time more frequently).

Our final interaction was designed with speed in mind, knowing how frequently people must check the time. When the user touches the dial of the watch, it enters 'guess time mode'; releasing the dial locks in that guess and quickly reveals the actual time. This interaction requires a single gesture, without buttons. Users who have noticed the time elsewhere can instead hit a button that will bypass the 'guess time' interaction, show the time, and reset the interaction.

Primary Exploratory Study

Using this interaction, we collected 32 hours worth of data with 7 initial users (age 25-34) in their normal, naturally-lit workspace, with no constraints on work tasks. Fig 5-9 shows the resulting 59 time estimates taken between 7 min and 1 hr 51 min after last checking the time (mean=33.5 min). No users reported major input errors, though some reported surprise at their accuracy (both good and bad).

Our data shows that average time estimate error was 12.5 minutes (after accounting for off-by-one-hour mistakes). Errors heavily favors time compression at an average 7.4 minutes

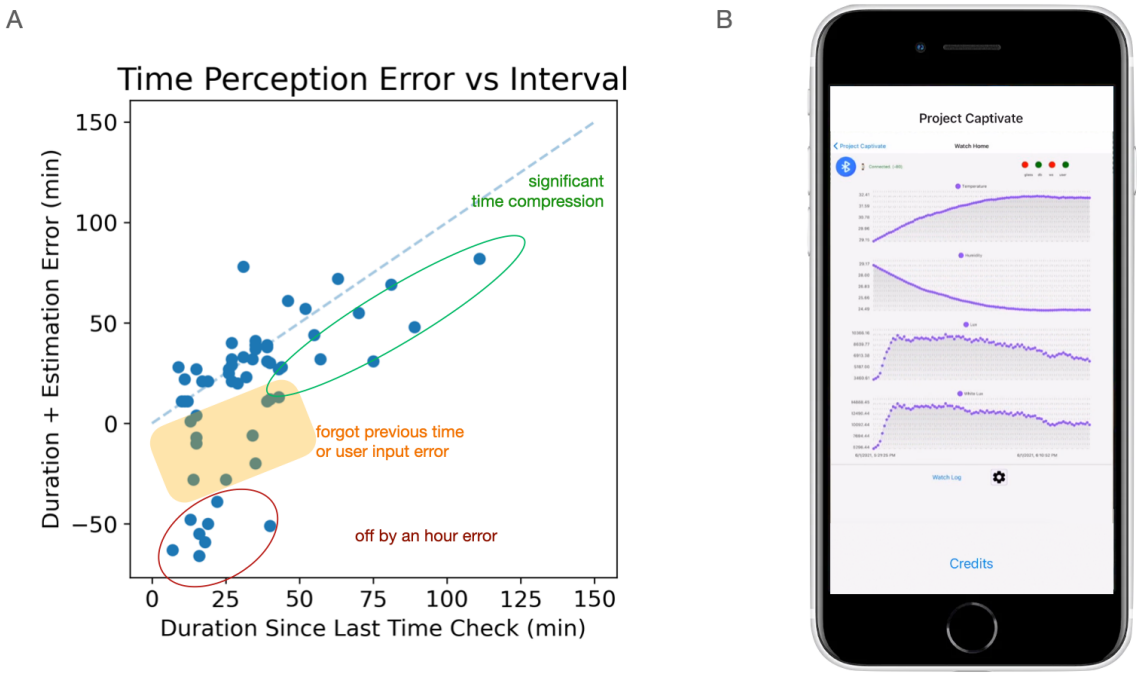


Figure 5-9: Resulting Data from Equinox study. (A) compares the duration since the last time a user checked their watch (x-axis) with the duration their estimate (so if they checked their watch at noon, and then guessed it was 12:15 at 12:20, we'd see a point at (20min, 15min))– accurate guesses fall along the dotted line. (B) shows continuous light level (white and full-spectrum), temperature, and humidity data displayed in the custom iPhone application, which synchronizes data with a secure online database. These estimates are during normal daily activity and represent a range of focus states.

short (22.1% of the actual duration). These time errors reasonably match expectation of retrospective, duration-based time perception lab research [404, 411]. For durations over an hour, estimates tend to favor 5 minute increments.

Over twenty of our estimates were after 5PM; a comparison between daytime and evening time estimates showed no meaningful differences, though this data is preliminary (evening errors = 10.8 min ± 10.9, day errors = 13.4 min ± 11.0, average time since last check = 31.0 vs 34.9).

5.9.5 Discussion and Future Work

In this section, we presented a novel, open source interface for naturalistic time perception measurement, and validated some basic assumptions by using it in a 32 hour pilot study. We were able to collect high quality data over the course of the study, which validates that (1) naturalistic time perception measurement during a normal workday on the order of 10-45 minutes is possible and (2) users experience significant and measurable time distortion in normal life that we can capture, especially as intervals grow.

This preliminary data shows no obvious major differences between clock-time estimates during unscheduled, dark evening hours compared with the workday, which suggests the influence of peripheral cues may not be as significant a factor as we had initially anticipated.

The raw data reveals some fascinating trends. We would expect long duration, 'in-the-zone' work periods to have some of the most significant time compression, which they do. However, clock-time estimates have unique errors compared with typical duration techniques. Three users made off-by-an-hour errors, which suggests they were conceptualizing the task as a prospective duration task ('how many minutes have passed?') and focused only on the minute value.

Another three users made estimates that were not off-by-an-hour, but still greater than the interval since the last time they checked the watch (i.e., checking the time at 10p, then guessing at 10:30p that it's 9:57p or 11:02p). Given the frequency of this phenomena,

we hypothesize that users don't memorize the time precisely (i.e. 1:43PM), and instead conceptualize its meaning (it's a early afternoon and I still have a while before my meeting) in a way that can be very forgettable if they have no incumbent commitments.

These errors raise some interesting questions about how to properly handle the data we receive; clock-time estimation may not directly map to a rigid prospective/retrospective paradigm. With continued use and habituation we might expect 'more retrospective' assessments, though the overall trend between focus and time distortion in both cases is similar. These data elicit interesting questions about how we conceive of time in our daily lives– how should we interpret a failure to internalize the time when glancing at the clock?

Our work suggests ecological data is varied, insightful, and worth capturing. It also raises questions important to future naturalistic study specifically and time perception research generally. Future studies will collect more data; pair the watch with other wearables; extend the work to include experience sampling of related cognitive phenomena; and more rigorously evaluate environmental covariates.

Most importantly, we plan to further explore clock-time estimation as a technique, and compare it more robustly to duration-specific estimates. Future work will focus on further understanding the mental processes behind clock-time estimation and how they differ from and interact with standard time perception research practice. This question is central to naturalistic time perception work, as clock-time is central to naturalistic living and will always confound time perception research in the wild.

Moreover, the common errors we find in wall-clock estimates suggest variability and complexity in individual approaches to clock-time estimation, and suggest that we frequently fail to internalize (hold in working memory or shift to long-term) exact numerical clock-time representations. Beyond time distortions, our data suggests the frequency with which a user checks the time and forgets the time may also provide quantitative insight into user phenomenology.

5.9.6 Conclusions

We present an open-source platform for the naturalistic study of time perception. Equinox has successfully demonstrated the feasibility of this agenda. With it, we measured high levels of time distortion in our participants' daily lives, alongside a new method for time perception measurement– the clock-time estimate.

We plan to use the lessons learned from this exploratory study to refine and expand our data collection techniques. Naturalistic studies provide an opportunity to vastly scale data collection compared to lab-constrained designs, which will be crucial to improve our underlying models of time perception. Ultimately, tools like Equinox will allow us to infer engagement and emotion from time distortion in a quantitative way so that we can evaluate the design and impact of our tools and our environments to increase the depth of our engagement– and thus our happiness– with our daily experiences.

5.10 Wearables for Flow Estimation: Feather

Feather is a leg-mounted interface vibrates over a range of intensities at the threshold of awareness, slowly increasing its strength over time. When the user notices a stimuli, they indicate it by pressing a button, and the system records the corresponding vibration intensity. We test whether there are significant enough fluctuations in the threshold of exogenous attentional capture with Feather to usefully inform a model of user cognitive state estimation with a small set of initial users in real-world settings.

5.10.1 Introduction

The depth of a person's attention in the natural environment is difficult to measure. Many domain-specific techniques have been suggested [327, 235, 429]; these, however, typically rely on self-report scales, which have considerable systemic error [165] or focus on overt attention (i.e. eye gaze) instead of underlying depth or quality [197].

Images of the Feather leg-strap prototype.

Figure 5-10: The 'Feather' device is strapped to the leg to monitor how intense a vibration is required to exogenously draw the user's attention away from their task in naturalistic settings. It can easily be hidden under the user's clothes. When the user senses a vibration, they indicate it by hitting the 'surprised smiley' icon.

At the same time, our attention shapes our perception in ways that have been been leveraged to inform notification design (i.e. you won't notice a notification designed at the threshold of noticing if you're deeply focused) [30, 216]. In this section, we invert that logic to create a novel methodology for quantifying engagement. The contributions are as follows:

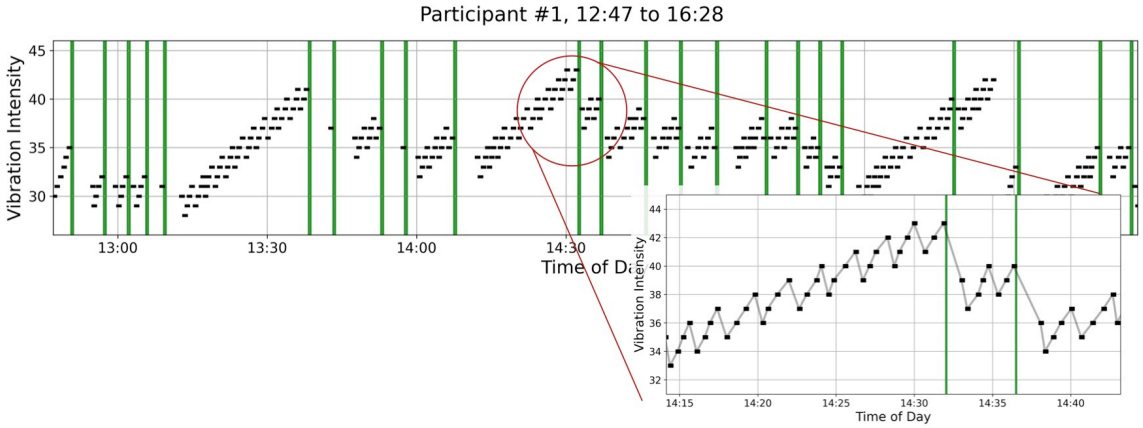
- The design of a device to make inferences about user cognitive state by measuring fluctuations in the strength of a tactile stimuli required to capture the attention of the user ('Feather'). To our knowledge, this is the first intervention to use variations in our threshold for noticing to inform a model of cognitive state.
- A prototype that can be used for naturalistic study in daily life, and run a pilot user study with 6 participants who spent a total of 26.4 hours testing this device while at their desk.
- An evaluation of the initial data with the goal of (1) assessing whether these fluctuations meaningfully capture information about the cognitive state of the user and (2) informing future interface design.

5.10.2 Background

Interactions of Mental State and Design

Notifications can be explicitly tuned to balance noticeability and distractability [244]. Two interventions take advantage of this research to design notifications just at the threshold for capturing attention– aurally by modifying background music [30], or visually at the edge of the visual field [216]. These notifications are designed such that fluctuations in a user's cognitive state alters the likelihood of noticing the stimuli– deep focus will proceed uninterrupted.

Compared with other modalities, tactile stimuli are very successful at capturing attention even during active tasks [367], and have been used extensively in the peripheral notification design literature [417, 337].



Representative data from a session with the Feather device.

Figure 5-11: Representative data from a user wearing the device. Green lines indicate that the user has noticed the stimuli; black marks show the stair-stepping increase of vibration intensity that start 1-5 minutes after the last noticed event. There are a few long sections without a notice, as well as a few sections where the user continues to notice lower and lower vibration intensities in rapid succession, consistent with the idea that focus state is altering this threshold.

Related Measurement Techniques

The interaction between mental state and responding has been used to infer cognitive states in the peripheral detection task (PDT)– once an ISO-standard road test to categorize cognitive load while driving [462]. In the PDT participants press a button when they notice a light in their peripheral vision; a newer extension of this task replaces the light with a vibration on the shoulder (tactile Detection Response Task or tDRT) [414].

tDRTs provide empirical support for the use of tactile cues as a cognitive indicator, but are tailored for high cognitive load task situations. The test stimuli are fired rapidly (every 3-5 sec), at constant intensity, long duration (1 sec), and reliable noticeability (>95% in practice– reaction times are the primary tDRT measure). Furthermore, tDRTs do not need to disambiguate noticing and reacting (response conflict is a known effect) [462]. This rapid and consistent dual task paradigm is a poor fit for the naturalistic study of focus.

5.10.3 Feather

Feather is based on the Pavlok, a commercial device designed to induce pain with electric shock (50-450V, 4mA, 100 steps of intensity) and to deliver tactile vibrations across a range of intensities and durations (using a low-power SMD cylindrical ERM motor; likely the NFP-RF2323, which delivers 0.4G.)

Preliminary Design

The main considerations for Feather's design are (1) the ability to present finely resolved stimuli at the threshold of noticing, (2) consistent and comfortable long-term use in naturalistic settings, and (3) an interaction that allows the user to forget about the secondary task and accurately probe the intensity required for exogenous attentional capture.

Our initial testing revealed that leg-coupled vibrations in the range of our device transition through a range of intensities that match the motor's operating regime to the threshold where vibrations are difficult to notice even with full focus. Furthermore, leg movement is minimal, which provides a more consistent contact and reduces the perceptual noise from motion, (besides supplying a discrete place for comfortable, long-term use). Based on these initial considerations, we created the form factor seen in Fig 5-10.

Our initial tests also included electrostimulation at very low levels. Participant concern around the pain of shock (even at the threshold of noticing) induced anxiety and hypervigilance. Though we were able to find a region of stimulation that passed from unobservable to noticeable, the perceptual transition was sharper than with vibration. For these reasons, we focused exclusively on vibration.

Calibration Tests

Our first characterization tests were to estimate the precision of the system across trials. We fixed a MPU6050 accelerometer to the device and captured acceleration data at each

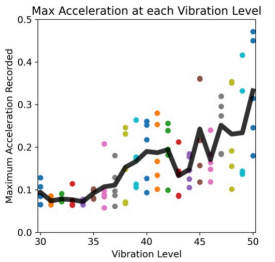
vibration command level (range from 1-100, though the lowest practical value is 30) over five trials. Peak acceleration data for each trial is shown in Fig 5-12, with a line to indicate the average peak acceleration per level.

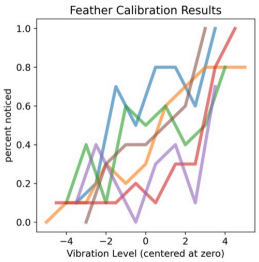
The data reveal a concerning and highly variable peak acceleration through the command levels of the Pavlok device. This is likely due to the fact that we are driving this motors at the limit of their design (even the smallest vibration motors are built to be create unambiguous vibrations), and it seems the initial torque required to pulse the offset mass varies depending on how it is resting.

Peak acceleration does not map directly to perception, though; we could consider average acceleration and duration, and we haven't factored in the quality of the contact with the skin and user perceptual acuity. To fully characterize the uncertainty introduced by the system, we had six participants rapidly test on a narrow range of levels around their threshold for noticing. Each level appears ten times; users were instructed to indicate when they felt a vibration. Placement has a large impact on test results; further back on the calf (more muscle contact) is less sensitive and more variable than further forward (coupling to the bone).

The results after proper placement are more encouraging— on average there is an increase in the likelihood of noticing a stimuli by 9.5% ($\pm$ 2.5%) with each increased intensity step when the user is focused; it takes 6.5 steps for the average user to transition from rarely noticing to consistently noticing a stimuli ($<$20% to $>$80%) when focused. This is true across a wide range of thresholds, which vary across individuals (41$\pm$9 on a scale from 1 to 100).

While this is not a perfect mechanical system for our task, it is sufficient to test for cognitively induced differences in noticing throughout the day. To improve the quality of the readings, as the stimuli increase in intensity, we test at each level 3 times; this improves the odds of a high peak acceleration (which we can see in 5-12 follows a linear trend and a consistent increase in vibration intensity); it should mean our calibration transition range should also represent the worst case (i.e. the 80% odds of noticing one pulse gives 99.2% odds when it is presented three times).





Participant #	1	2	3	4	5	6
Avg Noticed Intensity	37.4	42.0	74.9	46.5	56.4	44.5
StdDev Noticed Intensity	2.4	7.2	10.3	2.9	12.4	2.9
Max Noticed Intensity	45	61	90	60	73	49
Min Noticed Intensity	31	31	61	36	42	39

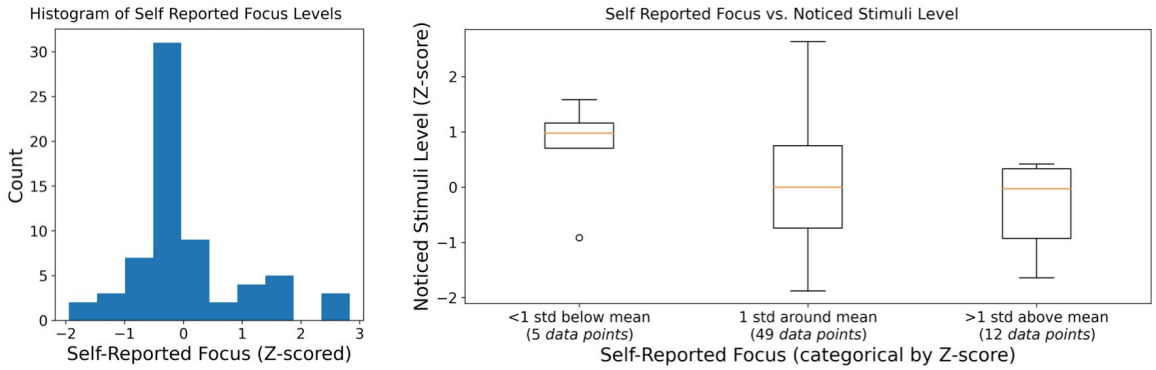
Calibration data and summary results.

Figure 5-12: (Left) Peak acceleration measurements from the motor when commanded at various intensities; dots represent raw measurements, the line represents the average. There is quite a bit of spread here, though the highest acceleration when as we continue to draw samples at each level follows a linear trend. Peak acceleration does not capture all the features of the input waveform that interact with perception (i.e. average intensity, duration). (Center) To characterize this noise in practice, users were asked to do a calibration procedure where they were rapidly presented with a random string of intensities around their threshold of noticing. It takes users an average of 6.5 steps to move from a level where they notice less than 20% of the stimuli to more than 80% when they pay attention. Threshold estimates are thus probabilistic, but informative; to capture better quality information, we present each level of stimuli 3 times as they increase. (Right) Summary data from the user study, collected over 26 hours with six participants. Most participants have a pretty large spread in noticed intensity even over a single working session; for some, like participant 4, the data indicate a real effect of cognition on the data (i.e. a tight standard deviation with large outliers).

In initial testing, the best method we found for robust placement was to run quick tests with the user while they move it from the front of the leg working backward, such that it is both comfortable and operating in the right regime of sensitivity. We followed this process for the remainder of the task. During this process, multiple participants suggested they would like a question mark button or said they struggled to tell a vibration from their blood flow. Many volunteered comments in the initial testing along the lines of "yes I felt that, but there's no way I would notice that if I wasn't paying really close attention to my leg."

Pilot Test

Six users wore the feather for a total of 26.4 hours while going about their normal workday. Participant are instructed to hit the surprised smiley button when they notice a



Images of the Feather leg-strap prototype.

Figure 5-13: The distribution of z-scored, self-reported focus based on self-report during our test (left; based on a 5-point likert scale very distracted/average/very focused), as well as a comparison between these self-reported values and the level of stimuli that grabbed the user's attention (right). The dataset is small but suggestive that as people start to focus, they become more aware of distractions. We posit that for deep levels of focus, the threshold starts to increase again (an inverted-U shape); to capture this will require a larger dataset taken especially in deep states of focus.

stimulus, and then asked to self report their level of focus prior to the interruption (1-5, very distracted/slightly distracted/average/focused/very focused). The pattern of stimuli (a stair-stepping pattern) was designed to (1) present each level of intensity three times, helping to reduce noise from stimuli variability, (2) remove simple repeated/increasing patterns of intensity or timing information that could prime responses, and (3) match the last intensity that was noticed after 8.5-12.5 minutes so the user can re-focus (1-5 minute break, followed by vibrations every 15-45 seconds). This pattern- alongside typical results for a participant- can be seen in Fig 5-11; summary results are in table Fig 5-12R; self-report data is summarized in Fig 5-13.

5.10.4 Discussion and Future Work

The data reveal that fluctuations in a user's threshold of noticing a tactile stimuli exist well beyond what can be easily explained by normal variation from the calibration phase, especially given that stimuli repeat at each intensity three times during the sequence. It corroborates that Feather is capturing cognitive effects in this behavior signal; this interpretation is reinforced by patterns of noticing in Fig 5-11; it's also reinforced in the unsolicited,

qualitative comments volunteered by every participant (i.e. 'it's so faint I would never notice that if I was working', 'I definitely missed a few', 'I wish there was a question mark button because I thought I might've felt it', etc).

Self-report data is weakly informative; individuals almost always select 'average' when describing their state of focus. To compare self-reported focus with stimuli intensity, we separated the self report into three categories ('low focus', 'average', and 'high focus'). We anticipated a Yerkes-Dodson relationship between self-reported focus ratings and distraction threshold- i.e. as you start to settle into focus you become more easily distracted, and then as you get deeper into focus you become harder to distract. The preliminary data shown in Fig 5-13R implies a simpler relationship, however; higher levels of self-reported focus correlate with higher sensitivity to distraction. Future work will explore whether deeper states of focus show the expected reversal (these states may be missing from this test; flow is elusive), and explore how personal/contextual this relationship might be. It may also point to a requirement for a longer delay between vibrations– the designed 10.5 min average cycle time may simply be too short (the cognitive cost of an interruption is estimated to be >23 minutes [279], but the trade-off with data collection is large).

This exploratory pilot study points to interesting future work. This technique shows tremendous promise in capturing quantitative data about fluctuations in distractability; however, there are several challenges to overcome to improve this system. Future work must start with fast and minimal calibration methods that ensure proper, repeatable placement and fit. Long-term comfort and habituation are also unknowns with the current design. This interface may also benefit from a secondary wearable that measures stillness (or another strong correlate of distraction) to corroborate general distraction level and the contrast between the induced vibration and natural movement [84]. A secondary wearable like a smartwatch would also facilitate easier data entry (rather than having an app with you). Most importantly, repeatable, precise stimuli and mechanical coupling should be improved; what can't be addressed must be modeled probabilistically, favoring topologies that are robust to interpersonal/intersession variability.

5.10.5 Conclusion

Feather represents a novel interaction that uses fluctuations in perceptual acuity to infer a user's mental state. We characterized the system and demonstrated its ability to capture variation in noticing that is best explained by corresponding variation in cognitive state; we also point to future refinements of this promising method.

In the near future, probabilistic combinations of this technique with other, similar bio-behavioral strategies will give us a much better picture of a user's psychological state, and thus on the causal forces that alter it. Feather represents a step towards empirically motivated design and closed-loop, adaptable systems that will inspire deep focus and engagement in their users across real-world settings.

5.11 Wearables for Flow Estimation: Captivates

Captivates are an open-source smartglasses system designed for long-term, in-the-wild psychophysiological monitoring at scale. Captivates integrate many underutilized physiological sensors in a streamlined package, including temple and nose temperature measurement, blink detection, head motion tracking, activity classification, 3D localization, and head pose estimation. Captivates were designed with an emphasis on: (1) manufacturing and scalability, so we can easily support large scale user studies for ourselves and offer the platform as a generalized tool for ambulatory psychophysiology research; (2) robustness and battery life, so long-term studies result in trustworthy data over an individual's entire day in natural environments without supervision or recharge; and (3) aesthetics and comfort, so people can wear them in their normal daily contexts without self-consciousness or changes in behavior.

Captivates have been validated for a small set of beta testers. We also interrogate the impact of our aesthetic design; corroborating that our additional design and miniaturization effort has made a significant impact in preserving natural behavior.

Figure 5-14: Captivates Final Design.

There is tremendous promise in translating psychophysiological laboratory techniques into real-world insight. Captivates serve as an open-source bridge to this end. Paired with an accurate underlying model, Captivates will be able to quantify the long-term psychological impact of our design decisions and provide real-time feedback for technologists interested in actuating a cognitively adaptive, user-aligned future.

5.11.1 Related Work

Non-contact Physiological Measurements related to Cognitive State

We seek to explore the space of sensor systems that do not require direct contact to the skin, a physically less-invasive technique that results in a more natural user experience and a more robust system for measurements on-the-go. Contact-based sensor techniques are a popular method for measuring physiological signals, but these techniques present challenges in dynamic environments. Contact-based methods include electrocardiogram (ECG) for measuring the heart, electroencephalogram (EEG) for measuring the brain, electrooculography (EOG) for measuring eye movement, and electrodermal activity (EDA) sensing for measuring changes in skin conductance. All of these methods have been applied to cognitive state estimation [413, 463, 166, 251, 246] typically under controlled conditions. Oftentimes, researchers struggle with noise artifacts from internal and external sources [256, 102, 443]. Contact sensing is not ideal for measurements throughout a person's dynamic day because of the mechanical skin-electrode coupling, which is highly susceptible to noise artifacts from user motion [90]. By focusing on non-contact techniques, we have greater flexibility for fit and comfort, and our research can be generalized to systems that are not wearable and do not require physical skin contact.

Face Temperature

As summarized in Table 5.1, several studies have found that our face temperatures vary with internal state changes (e.g., fear, joy, anxiety, etc.). For example, in [15], it was observed that nose temperatures decrease with the increase of cognitive load due to blood

flow restriction when autonomic nerve activity increases. Cognitive load was modulated using a series of reading comprehension and Stroop Tasks (i.e., matching color of a word with the word itself) at varying difficulty. Thermal camera measurements in that study required a fixed distance and perspective of the user; for mobile use, similar resolution would require multiple thermal cameras in each location, is prone to occlusion, and is cost prohibitive.

Emotions								
	Stress	Fear	Startle	Sexual arousal	Anxiety	Joy	Pain	Guilt
Regions								
Nose	↓	↓		↑		↓		↓
Cheeks			↓					
Periorbital			↑	↑	↑			
Supraorbital			↑		↑			
Forehead	↓↑	↓		↑	↑		↓	
Maxillary	↓	↓	↓				↓	↓
Neck–carotid			↑					
Nose	↓							
Tail		↓					↓	
Fingers/palm		↓					↓	
Lips/mouth				↑				

Table 5.1: Overview of the Direction of Temperature Variation in the Considered Regions of Interest Across Emotions from [215].

An alternative is to fix a face temperature sensing device onto the user. In [301], the researchers fixed a passive infrared radiation sensor (i.e., thermopile) to a tethered, glasses-like wearable to measure nose and forehead temperatures to control for movement. Forehead temperatures are used as a baseline since it is believed to remain stable relative to internal temperature, and face temperature varies with other internal and external events (e.g., convection due to wind). However, forehead temperature was demonstrated to increase with increased cognitive load in [15]. Pompei [342] suggests the temple region provides a better estimate of a person's internal body temperature given its proximity to the large superficial temporal artery. Thermometry based on this assumption is used in hospitals, as it approximates the accuracy of more invasive rectal techniques– the most accurate method for measuring internal temperature in a clinical setting [195].

Blink Rate and Eye Gaze

As described in [408], there are four types of eye blinks: reflex blinks, voluntary blinks, non-blink closures, and endogenous blinks. Endogenous blinks– blinks that are not consciously

initiated or due to any "identifiable eliciting stimulus" – are theorized to relate to information processing, cognitive load, and task demand.

In [256], test subjects reduced their blink rate under higher cognitive load. Furthermore in [316], eye blink rate correlated with task difficulty over arithmetic tasks (blink rate decreases as the difficulty of the task increases).

Others have studied blink rate as they relate to creativity or emotional shifts, and blink magnitude has been linked to attentional control and alertness [326]. Moreover, spontaneous blink rate drops precipitously when using screens [33] and blink patterns change in response to visual task demands [206]. There has also been significant effort to relate blink rate with dopamine dis-regulation and the study of addiction, though evidence for this specific theory appears weak [103].

The evidence suggests that blink rate and blink magnitude are both powerfully tied to mental experience and poorly characterized. While it's uncommon to find blink rate sensors in existing wearables, a few have been reported in the literature. In [112], the authors detect blink rate and magnitude based on changes in near infrared reflectance using a near infrared diode/phototransistor pair with 85% accuracy.

Head Motion Dynamics

Head motion is an indicator of cognitive effort and attentional control. Cognitively strenuous tasks result in less head movement, and patterns in the data could be used to discriminate common tasks and moments of task shift [276]. Subtle head motion dynamics can predict ADHD in infants [447]; it has been used to measure attention in classrooms and in meetings [428, 347]. Head motion has also been used to predict speech prosody and heart rate [100, 40]; mirroring of head motion between patient and psychotherapist may also help predict patient outcomes [352].

Gaze is also a powerful indicator in attention research. Apart from directly linking head dynamics to cognitive state, Stiefelhagen et al. were able to estimate gaze direction from head pose with 89% accuracy [412]. Given some a priori knowledge of the environment, it

is possible to predict fixation points using head pose and 3D location; traditionally, this is accomplished with computationally demanding systems like camera-based eye tracking, which can also result in very awkward wearable systems.

Psychophysiological Data in Non-Laboratory Settings

Contact-based measuring techniques for cognitive state estimation are popular, especially for the consumer market. Several products exist that claim to help improve your attention or learning, and aid in relaxation in natural settings. Muse [8] is an EEG device that is designed to help train in meditation, but also claims to "make it easy to access and use brainwave data, inside and outside the laboratory and in real world environments." Muse has made practical trade-offs for real-world use compared to usual EEG systems, with dry electrodes that do not require additional conductive gel for better coupling [149], but motion artifacts remain. Regardless, Muse no longer directly supports access to the raw data [373]. There are other open-source methods of collecting raw EEG signals, but most of these designs aren't productized for on-the-go applications [9].

There are several commercial smart eyewear devices that include access to sensors data. Google Glass is one such design that has gained traction across the academic spectrum, from analyzing head motion and blink frequency for activity recognition [218], augmented-reality based indoor navigation [355], and even surgical applications [452]. Google Glass was successful as a research platform because it was the first such scalable platform that could survive a user's regular day, and it provided a software development kit (SDK). Unfortunately, consumer adoption remains elusive and its design is met with some social push-back. Recent consumer smartglasses attempt to fix this shortcoming, taking their design in a glasses-first direction. Vue [13] is a pair of glasses for activity tracking that offer wireless bone conduction audio for discrete listening. They are designed to look like your average pair of glasses, with all the electronics densely embedded within the plastics. Focals by North [6] takes a similar approach by offering activity tracking and a projected display onto one of the lenses for notifications and short information intake. The display in Focals is not visible from the outside world, allowing the glasses to retain a traditional design

aesthetic. Both the Focals and Vue are sensor-rich systems and follow a more consumer-friendly design approach but, unfortunately, offer no ability for researchers to grab any of the raw sensor signals or to interface their own sensors.

To the best of our knowledge, J!NS MEME [14] is the only commercial platform that takes a design-first approach for smart eyewear while also benefiting researchers with an SDK (though their software applications are not currently supported outside of Japan). The platform includes a 3-axis gyroscope and accelerometer (420 USD) and optional EOG sensor (1050 USD). The EOG-version of the device has been used by researchers for monitoring fatigue and drowsiness levels, but has been criticized for poor sensor performance and a lack of noise robustness (i.e. when talking or moving) [366, 421]. As explained in subsection 5.11.1 and further supported by these studies, contact-based sensing includes additional data challenges for ambulatory measurements. Furthermore, available gyroscope and EOG sensor data are processed through proprietary algorithms before exposed to the researcher. J!NS MEME is a clear step in the right direction, but its lack of support outside of Japan, proprietary obfuscation of raw data, and high price make it difficult for researchers to adapt it to their custom applications.

More general open-source hardware acquisition platforms do exist in literature that allow researchers to interface multiple physiological sensors. In [80], a device is presented that uses a non-contact radiation measurement for temperature and electrode-based techniques for ECG and EDA. Their results are promising, however their system was designed for lab use (non-wearable and tethered), similar to OpenBCI. This is often the case for these platforms— if they aren't commercialized, there is little incentive to complete a robust design for other researchers to use.

5.11.2 Design Principles

Given our motivation to build novel physiological sensing into a platform that would enable large-scale, long-term study of natural contexts, we distilled our objectives down into a set of design principles. These considerations were weighed with the help of various traditional

and smart eyeglass manufacturers and product designers in the United States, Shenzhen, China, and Seoul, South Korea.

Scalability

We chose a radio transceiver that supports multiple radio protocols for easy integration into different network architectures such that Captivates could serve as the foundation for a larger portfolio of responsive systems. It is designed to support meshing with similar devices over Google's Openthread stack, in which each node has enough on-board resources to serve as an active mesh intermediary in addition to the resources required for simple edge computing. Mesh protocols are advantageous to future Internet-of-Things (IoT) adoption because they sidestep support and troubleshooting related to individual router infrastructure.

Extensibility

It was a major goal to enable other researchers to use Captivates for their own experiments without redesign work. We made the glasses easily reprogrammable without disassembly or rework, and fully open source, so anyone may utilize the device. We also included enough additional connectivity to simplify integration with outside sensing platforms, including support for multiple popular network stacks like Bluetooth.

Design for User Comfort and Signal Robustness

Since this system is intended to be worn throughout the day, across static and dynamic activities (e.g., sitting, walking, etc.), our sensing modalities were optimized for comfort as well as consistency across activity and environment. As a result, we avoided sensors that require significant contact force (uncomfortable for long term use) or skin adhesion (contact changes over time). Furthermore, electrode-based techniques that require this type of coupling often have reduced signal integrity in long-term, mobile applications due to the dynamic mechanical stresses at the electrode-to-skin bond during movement, and natural

variation of other confounding electrical sources, i.e. cardiac activity, ocular movements, eye blinks and muscular activity [181, 443, 364, 102, 95].

Manufacturability

After considering the appearance and comfort of the glasses, we optimized the system for easy production and assembly. The design balances the part count and overall assembly time with ease of modification and extensibility. Our goal was to have a system that is geared toward mass production, allowing us to cheaply scale production of plastic parts and assembled printed circuit boards at a cheap cost. In addition, we also wanted the design to be 3D-printable to allow users to easily produce small batches with custom modifications.

5.11.3 System Design

In this subsection, we discuss the technical system design. Our goal is two-fold: first, to highlight engineering work beyond typical prototype design that went into making these more product-like, in service of of robust real-world operation and design. Secondly, to introduce the structure of the technical system for researchers who might want to modify it. The design is open-sourced and available on our website at `https://captivate.media.mit.edu/resources.html`.

We divide the system into three logical parts: the electrical, mechanical, and firmware. A summary and comparison of our final system with other popular smart eyeglass systems is shown in Table 5.2.

Sensor Selection

Based on the design considerations outlined in subsection 5.11.2, we cross-referenced the recent literature for physiological indicators of cognitive load and attention. Much of the recent research is focused on frontal lobe electroencephalogram (EEG). There are several

Design Target	Captivates	Google Glass Enterprise Edition 2	Focals by North	J!NS MEME ES
Weight	44g	46g	72.57g	36g
Battery Life	10 hours	<8 hours (depends on usage)	18 hours (intermittent use)	12 hours
Battery Recharge Time	3 hours	1 hour (fast charge)	2 hours	2 hours
Ruggedization	None	Water and dust resistant	Water and dust resistant	None
Sensing Modalities	•Nose and Temple Temperature •Touch •Blink Rate •9-axis IMU •3D Location	•Microphone •Touch •Camera •Gyroscope •Accelerometer •Magnetometer	•Microphone •Ambient Light •Gyroscope •Accelerometer •Magnetometer	•Accelerometer •Gyroscope
Actuators	•Eye-facing LEDs •Externally-facing LEDs	•Display Module •Mono Speaker, USB audio, BT audio	•Display Module •Mono Speaker	None
Price	$200 (BOM Cost)	$1000	$599	$420

Table 5.2: System Specifications and Comparison

initiatives already exploring this technique, including BrainCo [7], Muse [8], and AttentivU [247]. Additionally, as described in subsection 5.11.2, electrode-based techniques aren't well suited for long-term use and during mobile activities. For these reasons, we decided to refrain from this modality.

Furthermore, contact-based sensing of this type always requires a physical, head-mounted wearable. One of our criteria for sensor selection was the ability to approximate our measurements with off-body techniques. Any insights from this research can thus be generalized (in specific environments) without the use of smartglasses per se– unfortunately, this kind of contact-less physiological sensing infrastructure is too cumbersome and expensive to scale across the many real-world environments visited by even a single user. Wearables present the best opportunity to rapidly scale data collection across users and contexts.

We converged on a non-contact temperature sensor (i.e., thermopile) on the nose, as nose temperature correlates with cognitive load [15]. However, since nose temperature is also influenced by external stimuli (e.g., air flow), a baseline temperature measurement is required. For this we added a second non-contact temperature sensor at the temple, which should not see rapid fluctuations during internal state changes [342]. The second sensor we chose was an eye blink sensor, as cognitive load influences blink rate [256, 316]. To accurately measure blinks, we are using an IR emitter/receiver pair since it is relatively low-power compared to cameras and can detect blink rate and intensity [112]. We also include an IMU, which allows us to capture a person's activity level and head pose [217]. Head pose

is correlated with gaze direction, which is important to determine what a user is fixated on under controlled conditions [412]. Lastly, we included VIVE Tracking System compatibility, so that off-the-shelf hardware can provide each pair of glasses a 3D location estimate, ideal for events with crowds and/or monitoring individual movement in a room. This can also be used for tracking fixations based on head pose and 3D localization data. Finally, the system also includes on-board user-facing LEDs that can subtly alert the wearer to any notifications (e.g., take a break, alertness is dropping, etc.) and influence their attention, as well as externally-facing LEDs for creative use in performance.

Electrical System

Circuit Layout

In order to maximize the amount of circuitry that can fit into a traditional eyeglass form factor, we spread the electronics across all faces of the glasses. This requires a three PCB-assembly design, one on either side and a flexible one running across the front that connects both side electronic assemblies. The side PCBs are standard FR4-core, at 0.8mm thickness to minimize the overall temple arm thickness. The front flexible circuit is the most novel piece of circuitry that we designed, as it conforms to the contours of the eyeglass's front plastic housing, folds down into the nose piece, and robustly handles dynamic bending within the hinge.

One of the most challenging issues to overcome with this system was how to incorporate the sensors around the nose into the electrical and mechanical design. From a design perspective, the easiest approach is to solder the near-nose sensors to the flex circuitry through a set of wires, but this method significantly lengthens and complicates the assembly process. Instead, we matched the bends of the flex circuit to that of the mechanical housing around the nose, creating two tails of flex circuitry that included the temperature and blink sensors. During installation, these tails fold into the grooves of the plastic housing without any additional soldering. The final manufactured flexible circuit is shown in Figure 5-19.

Processors

A block diagram of the entire system is shown in Figure 5-15. For the main processor, we use a dual-core microcontroller (MCU) that has an ARM Cortex-M4 processor for the user application and an ARM Cortex M0+ processor for the network stack. This processor allows us to treat the MCU as an application processor, without concern for network stack overhead. It supports quick embedded firmware development and increased versatility for applications that don't require the network component. It also supports three major network stacks (i.e., Bluetooth 5, Zigbee 3.0, and OpenThread). The chip is thus a versatile choice, minimizing the need for hardware redesign in a clean and simple electrical design that integrates this functionality into a single discrete package.

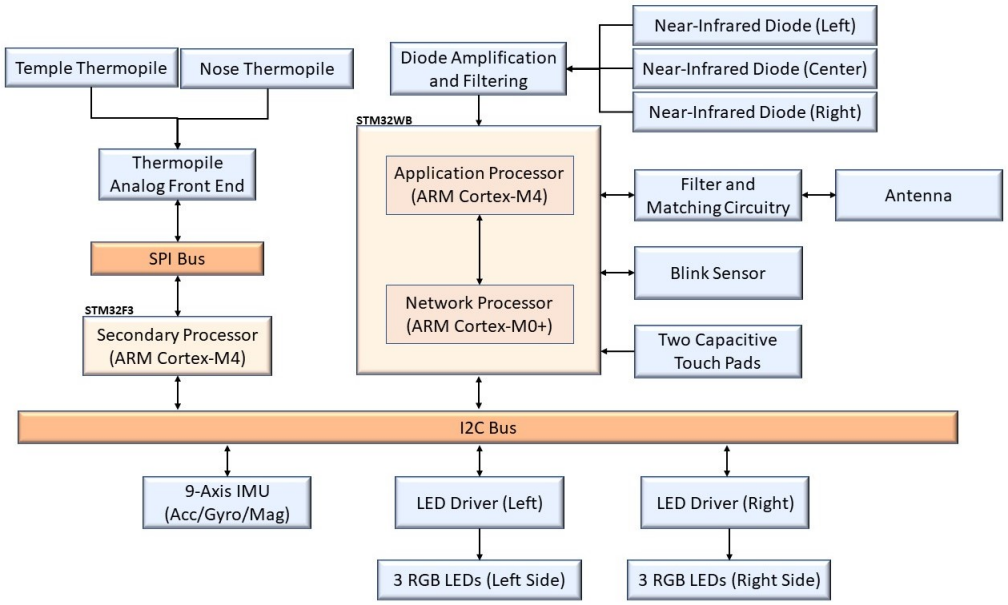


Figure 5-15: System Diagram

The dual-core MCU exists on one of the two-sided PCBs, which are separated and connected by the flexible circuitry. Given that sensors exist throughout the glasses, and that the long separation between the rigid PCBs can lead to signal integrity issues, we added a secondary Cortex-M4 MCU on the opposite side of the glasses. This secondary MCU manages the sensors nearby, applies our preprocessing algorithms to those sensor streams, compresses the data, and sends it to the main processor when ready or requested. While this increased the software complexity, it allowed us to reduce the number of conductors passing through the hinges and across the length of the flexible circuit.

Power Architecture

Our system utilizes two batteries for better weight distribution, but this approach increases the complexity of our power architecture and charging circuitry. To handle possible cross charging and rapid aging from physically separated batteries, we used a power multiplexer (MUX), which powers 3.3V, low-quiescent, linear regulators, one on each side of the system, and also directly powers the two LED drivers located on the flexible circuit.

Temperature

The thermopile circuit consists of two analog thermopiles, one faced towards one side of the user's nose and the other faced towards the user's temple. These thermopiles are then connected into an analog front end, which amplifies and filters the signal before being converted by the secondary processor's 12-bit analog ADC. Once a set of readings is collected by the secondary processor, the main processor is alerted over an interrupt line, and after acknowledgement, the data is transmitted over the I2C bus to the main processor.

Blink

The blink sensor is an IR emitter/receiver package that showed promising results in [112] on detecting blink events. We experimented with different blink detection locations and found that putting the sensor on the bridge of the nose, facing towards the eye, yielded the best performance. The near-infrared diode that emits light can be modulated through the main application processor to better dynamically adjust the reflected power seen by the receiving diode and enable synchronous detection. This feature is important for environments that have external near-infrared light sources (e.g., sunlight) that can saturate our blink sensor.

3D Localization and Pose Estimation

For the 3D localization, we are using two VIVE base stations [5] as our signaling sources in the environment in order to triangulate the glasses in 3D space. Each base station emits a vertical and horizontal near-infrared sweep, alternating between both sequences and between both base stations. VIVE states that a less-than 2mm accuracy and a range

of 16 feet can be achieved with their system, but based on preliminary testing we believe our tuned custom system can detect the base station sweeps at a range of at least 30 feet.

The hardware architecture for our receiver consists of three near-IR sensitive diodes, one on each face of the glasses. The search criteria for choosing which type of diode to use consisted of finding a surface mount component that had a slim form factor but a large enough receptive area that also offered adequate responsivity to capture the momentary sweep. We evaluated six types of diodes and found that the VBPW34FASR by Vishay performed the best and conformed to our design constraints. The three Vishay diodes in our circuit face in different directions (see i.e. Figure 5-14 top and bottom for an example of the front facing diode) and are then tied together (i.e., summed) and passed into a transimpedance amplifier to convert the diode-generated current into a voltage. The reason for the summation is that we wanted to treat the system as a point mass that has a large FOV so that as a person turns their head, at least one diode will still be exposed, increasing the reliability of the system. This also allows for a simpler analog front end with the trade-off of decreased accuracy, since the diodes are not spatially co-located but are treated as such in software.

For head pose, we chose a digital 9-axis IMU that connects to our I2C bus and offers various internally computed metrics, such as activity classification and a quaternion matrix for estimating head pose. Software calibration was done to align the IMU's reference coordinate system to the glasses coordinate system for accurate head pose estimation.

Other Components and Interactivity

Our system offers a few other components for easy usability, interactions, and actuation. In order to easily program our system, we wired the USB micro port to both the battery charging circuitry and to the single-wire programming interface for the main application processor through a modified USB cable and an ST-Link Programmer— this can be done by anyone with just a programmer and a spliced USB cable. In addition, the system has two capacitive touch points on the left arm that are sensitive enough to identify taps and both forward and backward swipe gestures. The sensitivity of the touch points can be adjusted

in software to tune for specific applications. Additionally, there are six RGB LEDs, one facing towards either eye, two on the top, and two facing forward. The LEDs toward the eye are intended to be used as notification [92] while the other LEDs can be used for crowd applications (e.g., a concert where the LEDs are actuated by the person's physiological signals). Our electrical design features several GPIO breakout pads scattered throughout so to make retrofitting for other applications easier.

Mechanical Design and Assembly

The mechanical design of the glasses required several iterations as we converged to a final form factor. As outlined in subsection 5.11.1, we wanted to choose a design that looks less like a research prototype and more like a traditional pair of glasses that is acceptable to wear in public contexts. Figure 5-14 shows the front view of our most recent design. Openings above the bridge of the nose and on either side of the glasses expose the near-IR diodes to the incident light emitting from the VIVE base stations. These openings may be covered with near-IR transparent plastic, but we've left them open for easy debugging. These openings can be covered or completely omitted for applications that do not require the 3D localization feature.

Hinged Design

Our early, hinge-less designs were fragile, difficult to transport, and less socially acceptable (subsection 5.11.4). In conversation with glasses designers, we learned the importance of hinge compliance for comfort. Our final version features a custom hinge with circuitry running through it. Our flex circuitry is preformed (Figure 5-17) and inserted into the front housing, resting on internal raised plastic features. The flex circuit is then sealed by a piece of plastic acting as the back plane. The hinge preserves a wide bending radius in the flex PCB in open and closed positions, as well as plastic guards to protect it from direct exposure to the external environment.

Light Pipes

Figure 5-16: Side View of Captivates.

The glasses enable concert and festival applications as well. We included several RGB LEDs; one on each brow lights up the entire top ridge of the glasses.

We first attempted to do this by shaping polymethyl methacrylate (PMMA). We worked with a manufacturing facility in China to laser cut these PMMA pieces and chrome-coat three of the four sides to maximize the light emmission on the top-face. Although this worked, we found it cost prohibitive and settled instead on side-glow fiber optic cables (Figure 5-18).

Apart from the brow lighting, we also have lighting at two points on the front face of the glasses where some traditional glasses have rivets. To increase light visibility, we added commercial-off-the-shelf PMMA light pipes with fresnel lensing on one end so that the emitted light is less directed, offering better angular visibility.

Battery Placement

Battery placement in the glasses is important for weight balance, comfort, and power capacity (batteries are usually the most dense component in smart eyeglasses). The ideal

Figure 5-17: Pre-bent Flex Circuitry

Figure 5-18: Illuminated Brow of Glasses

method is two identical batteries, one on either side, to better balance the weight at the cost of complexity in power distribution architecture and charging circuitry.

For our design, we have two, 150mAh, batteries on either side of the glasses, running in series with the side printed circuit board. The circuit boards are designed so that when assembled, the width is uniform across the batteries and across the circuit board assemblies for a streamlined design (Figure 5-19).

Figure 5-19: Deconstructed Captivates.

Firmware Architecture

For a system of this complexity a low-level operating system is necessary. We decided to use FreeRTOS, a lightweight real-time operating system that is capable of being run on microcontrollers and small microprocessors which makes it suitable for this project [11]. This operating system is also supported by ST Microelectronics, the manufacturer of our processors, and has several additional tools written by ST Micro for easier system integration. FreeRTOS allows for the encapsulation of individual blocks of code into threads that can be triggered by a variety of sources (e.g., hardware interrupts, timers, other threads, etc.) and has various methods of passing messages across threads (e.g., queues, semaphores, mutexes, etc.).

There are numerous threads running in our system controlled by a processor's master thread. The radio communication thread or inter-processor thread can enable certain sensor channels, actuators, or send or receive messages. Once enabled, sensor channels operate autonomously. The direct memory access (DMA) controller is used for blink sampling (1kHz) such that the thread only awakens once a second to pre-process and pass a pointer to the

master thread. The digital IMU is interrupt driven, as is the secondary processor (it buffers 10 samples from each thermopile sampled at 10Hz and waits for an acknowledgment before sending the data over I2C).

3D Localization

The 3D localization code was adapted from [389]. Every time one of the near-IR diodes is saturated by incidence light (i.e. rising or falling edge), an interrupt is triggered which results in a callback recording the exact timestamp of the event in microseconds. These timestamps are then fed into a thread that classifies if the event was part of a sequence (i.e., two flashes and a sweep) or if the event is noise. If part of a sequence, the code catalogs all the sweeps and calculates an estimate for the 3D location of the device. Since the code requires the global coordinates of the VIVE base stations, they can either be hardcoded into the software or updated via the server.

Networking

Our choice for flexible networking is an open-source version of Thread [12] called OpenThread [10]. It "is an IPv6-based networking protocol designed for low-power Internet of Things devices in an IEEE 802.15.4-2006 wireless mesh network, commonly called a Wireless Personal Area Network (WPAN). Thread is independent of other 802.15.4 mesh networking protocols, such a ZigBee, Z-Wave, and Bluetooth LE" [10]. OpenThread has a variety of features that makes it suitable for our system. One primary advantage is that it's a protocol that is seeing wide adoption across a variety of platforms, and since its being supported by Google, it's integrated into many of their home devices— in theory, our system could use Google Home nodes as hops in our network. It's also scalable to hundreds of nodes, making it ideal for future integration of our glasses into an ecosystem of sensors and actuators. In addition, it's IPv6-based so integrating it with other IPv6-based systems is simple and doesn't require a network translator, making it appealing for a "future-proof" implementation. There also exists Contiki [128], another low-power IPV6-based protocol with evidence that it has improvements in latency and packet loss rates over OpenThread [136]. However, [136] was just an introductory comparison between both protocols, and more work needs to

be done on measuring the power efficiency and performance of both protocols. We chose to stick with OpenThread due to the native support ST offers for our network processor and the active open source community around it.

Captivates can also trivially run a standard BLE stack implementation.

Production and Extensibility

We opted to manufacture our plastics using Nylon PA12 through a selective laser sintering (SLS) 3D printing process using an external vendor since Nylon is much more resilient and robust than other 3D printing materials while SLS printing offers a more uniform surface finish. However, we were also able to print all of our mechanical components with consumer-grade 3D printers (e.g., Formlabs Form 2, Prusa) so that is an avenue that is available for quick experimentation.

Our two rigid circuit boards were designed with specifications that fit most common circuit board manufacturers' capabilities (i.e., 4/4mil traces, 4 layers, 0.8mm board thickness, 0.2mm minimum hold sizes, and 1oz copper). The front flexible PCB is the only design feature that has a custom design feature, specifically having transition regions between 2-layers and 4-layers. We worked closely with a manufacturer in Shenzhen on producing the boards so it is possible other manufacturers may be able to support this feature. Otherwise, you can avoid the transition altogether and just make a 4-layer flexible PCB which would trade-off robustness for added simplicity. For circuit board assembly, we were able to assemble a few sets by hand in our lab but many of the components are small and the front flexible PCB has 2 ball grid array (BGA) components, so if lacking the soldering capabilities and expertise, an outside assembly service is recommended.

For researchers wanting to add additional sensing modalities, we've added a few thru-hole connections on the PCB located on the right side that provides 3.3V power, ground, and connections to 3 general-purpose input/output pins on the secondary MCU, equipped with access to the I2C communication and analog-to-digital conversion (ADC) peripherals. If some of the existing sensing subsystems are not needed for an application, those sensors

can be removed and the existing connections can be re-purposed. One example would be to remove the blink sensor and use the existing power and the ADC lines to sample the signal of a different sensor. To make this easier, we've added various thru-hole regions on the front flexible circuit that wires can be soldered to. Another method that can be used is removing the PCB with the secondary processor and reusing our existing power circuit to design a new board with other sensing modalities.

All the electrical and mechanical design files are linked on our website, as well as further information on our progress with the system: https://captivate.media.mit.edu/.

5.11.4 Validation

System

We calibrated the antenna circuit using a vector network analyzer (VNA) while the PCBs were in the eyeglass frames and on a users head. This approach is warranted since any mass near the antenna can attenuate the signal and distort the balance of the circuit. We also calibrated our 3D Localization System by tuning our circuit to barely saturate at 30-feet from the VIVE emitter, ensuring our signal-to-noise ratio is optimal for that range. In the next subsubsection, we describe our detailed calibration of the temperature sensors.

Temperature

To estimate cognitive load, you only need to measure relative temperature differences from the baseline [15] so thermopile calibration for absolute temperature approximation is less important. However, for applications that require an accurate estimation of absolute temperature, a thorough calibration should be performed on each individual thermopile since different resistive losses and noise artifacts exist based on the location of the sensors.

From [426], it is recommended that we account for heat flow other than radiation, such as convection and conduction from nearby objects, in Stefan-Boltzmann's Law. [233] We calibrate our thermopiles using this equation:

$$T_{object} = \sqrt[4]{T_{reference}^4 + f\{V_{TP}\}/A} \tag{5.1}$$

where

$$f\{V_{TP}\} = (V_{TP} - a_0) + a_1(V_{TP} - a_0))^2 \tag{5.2}$$

where V_{TP} is the thermopile voltage, T_{object} is the temperature at the active junction, and $T_{reference}$ is the temperature at the reference junction [233]. The constant, A, is a product of the thermal resistance of the thermopile, number of thermocouples within the device, the Seebeck coefficient, net emissivity between the object and the device, Stefan-Botlzmann constant, and field of view (FOV) of the device. In most cases, this constant is approximated for during calibration by varying the temperature of an object and of the thermopile housing, while placing calibrated thermocouples throughout. a_0 and a_1 are terms to solve for during the calibration process.

For our system, to estimate each thermopile's reference temperature, $T_{reference}$, each thermopile has a thermistor built-in with a manufacturer-provided characterization [26]. To calibrate, we took a Peltier module that converts a voltage potential into a temperature differential and using a thermal compound, embedded a calibrated thermocouple between the Peltier module and a thin piece of laminate material. We then placed a calibrated digital thermopile right next to our analog thermopile as a reference. Finally, using thermal compound again, we attached a thermocouple to the housing of the thermopile to estimate the accuracy of the built-in thermistor.

The test was performed over 10 different temperatures for both the nose and temple thermopiles, ranging from 60-115 degrees Fahrenheit, the average temperature ranges expected across the surface of a person's skin. For each iteration, we waited a minimum of 5 minutes to ensure the temperature of the reference object has stabilized. After the experiment, we used non-linear least squares to solve for the constant A given V_{TP}, $T_{reference}$, and T_{object}. This allows for a suitable approximation for temperature within our ranges but assumes constant emissivity across people so a more thorough calibration is warranted where tighter absolute temperature tolerances are required. (Though emissivity varies across people, it is

	A	a_0	a_1
Nose Thermopile	7.80e-10	-2.31e-01	3.61e-03
Temple Thermopile	4.21e-10	-3.62e-01	8.31e-02

Table 5.3: Thermopile Calibration Values

not significantly correlated with skin pigmentation [82] or sweat [86]; it does appear to vary with age [323].) Table 5.3 shows our results from the least-squares approximation; these values compensate for differences in resistive loading for each thermopile (e.g., variable trace length), in addition to calibrating the thermopiles for absolute temperature measurements.

Real World Evaluation

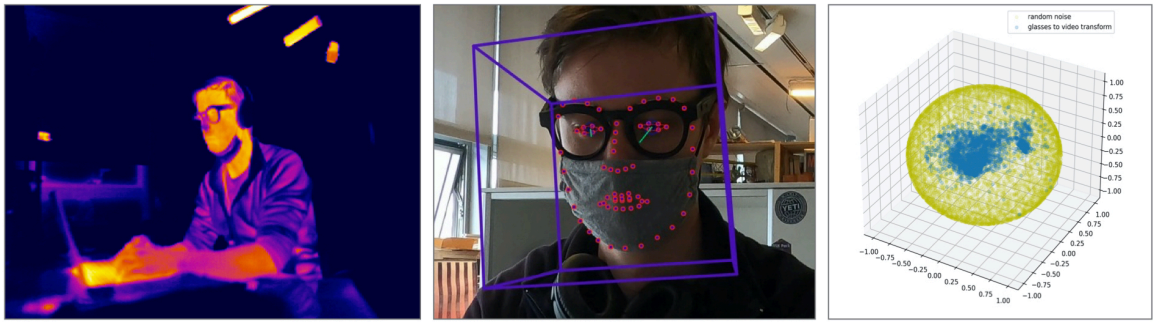


Figure 5-20: Thermal imaging (right), the OpenFace video analysis tool (center), and a comparison of orientation data between the glasses and extracted from the video (a tighter blue cone is better).

We validated the system in a pilot study in which 5 participants (2 female/3 male, ages 26-34) wore the glasses while being videotaped using a GoPro set to 60 FPS/1080P. Participants were recorded in naturalistic environments (their offices or homes) working as normal. Over 45 hours of data were collected in this way (>9 hours per participant on average).

This data has high ecological validity– participants have a range of face sizes and morphologies, studies were done at various points in the day from 10 AM to midnight in various lighting conditions and environments, and gaze and blink patterns varied dramatically over participants and over tasks. One participant made infrequent but clear shifts in visual attention between their screen and a desk; another participant shifted gaze frequently

between a phone, a laptop, and a television screen; a third was making frequent large gaze shifts across a large monitor. Recorded activities were screen-based because we needed long-term, consistent and clear video recording of participants' faces. Blink behavior across participants varied dramatically, from an average of 11.3 blinks/min to 62.1 blinks/min over individual participants (31.0 blinks/min average across users).

We performed three basic comparisons to ensure our glasses data was of high quality. For two of these comparisons, we used the OpenFace machine learning classifier tool on our videos as a reference point (center of Figure 5-20) [42]. OpenFace uses trained machine learning models to calculate facial landmarks that are then used to estimate facial muscle movements (e.g., blink, smile, eyebrow raise, etc.). OpenFace struggled with our naturalistic videos– especially due to the framing and camera angles that we were forced to use when installing them in participants' real workspaces. We attempted to install the camera in a way that the eyeglasses do not occlude the eye area of each participant from the camera. For two participants, OpenFace's self-reported successful classification rate was between 25 and 40%. With manual cropping and editing, we were able to improve these classification rates to between 65 and 85% per participant. For all further comparisons, we utilized contiguous video subsections longer than 200ms (the duration of a blink) where no more than 2 consecutive frames were dropped by the classifier. Time alignment between the video and glasses data was performed by looking for common minima in an extracted blink signal cross-correlation across several of these video subsections. This time alignment was verified visually with a data-video overlay.

1. Thermal Comparison

We did a few simple comparisons between an Optris PI400 infrared camera (382x288 pixels, .08°C sensitivity, and ±2°C accuracy), pointed at the side of the face where we are collecting thermopile data, to ensure proper functionality of our thermopiles in realistic settings (as seen in Figure 5-20L). This basic comparison shows agreement between the Optris and glasses data, within the error of our Optis measurement. While our data is aligned, even these high-cost IR cameras do not have the pixel resolution to precisely capture 5mm^2 patch of skin underneath our thermopiles at a distance that still allows for naturalistic participant

behavior (the entire nose is blurred over 20 pixels at the demonstrated resolution and distance, and this effect of this spatial averaging varies with subtle participant motion). We expect higher resolution, and more spatially consistent data from our thermopiles than an IR camera can provide; however, this did serve as a reasonable verification that the calibration process described above was functioning properly.

2. IMU Comparison

IMU data was referenced against head orientation as extracted from OpenFace. OpenFace uses a world coordinate system at the frame rate of 60Hz, while our IMU data is referenced to the head and captured at 10Hz. To check for agreement, we interpolate the IMU data to our IMU timestamps and calculate the orientation transformation required to map the OpenFace orientation to the IMU orientation at that moment. Figure 5-20 (right) shows these transformations for each time-step over several minutes for Participant 5. Ideally, the mapping between the IMU and OpenFace data would consistently fall at a single point on the unit sphere (one reading maps exactly to another with the same coordinate frame transfer). If the IMU data and the OpenFace data were uncorrelated, we'd expect the yellow unit sphere pictured. Naturally, we get a cone centered around the coordinate transform, which shows the combined error in measurement between the IMU and the OpenFace classifier. The error we see in this cone has a standard deviation (in Euler angle °) of [20.4, 9.8, 22.2]. This demonstrates a rough agreement of motion from one measurement to another across time. Participants often move around their environment in the study, and the quality of the OpenFace orientation extraction as a face frequently translates forward and backward is un-characterized; however, this check combined with other basic checks for drift leads us to believe the IMUs are functioning within the manufacturer specifications of <5 degrees of absolute error at all times, capturing transient behavior that is correlated with activity, physiological state, and mental state. We expect the IMU– with internal drift compensation– to perform more robustly than video-based methods.

3. Blink Comparison

To ensure blink data from the glasses was reliable, we designed a basic, micro-controller-friendly algorithm to extract blinks from the IR signal received by the glasses. This algo-

rithm not only captures blink timestamps; it captures useful metadata about each blink that has been shown to correlate with alertness [21]. The algorithm works as follows, and was applied to each participant without personalization (See Figure 5-21 for representative data):

1. Low Pass Filter the raw 1kHz IR data (light orange) using a 50-point moving average filter (green).
2. Take the first derivative of this signal (red) and smooth it with a 25-point moving average filter (purple).
3. Using a fixed threshold of -0.05, wait for 100 peaks in the first derivative signal that dip below this. At this point, we predict the user is wearing the glasses.
4. Record the next 100 minima below -0.05. We know from preliminary testing that >10% of these are almost assuredly real blinks, and >10% are almost assuredly not, across all users and conditions.
5. Calculate the average value of the 10 shallowest peaks (a noise metric close to -0.05) and an average of the 10 largest peaks (known true blinks) from these minima; calculate a threshold value a quarter of the distance between them, just above the noise level.
6. Compare the running first derivative signal against the threshold; whenever it drops below this value, we consider a blink to have begun. For the next 175ms, we monitor the first derivative and capture the minimum eye close velocity (the negative peak), the maximum eye open velocity (the corresponding positive peak), and the difference in timestamp between these two (a blink duration). This search window and duration are shown when the blink detector triggers in a faint brown (search window) and dark brown (extracted duration).

Initially, we planned to use OpenFace [42], a common standard reference for face tracking, as our baseline for blinks. We quickly realized that OpenFace struggled to locate blinks

properly within our data, even after hand-cropping the videos to increase classification success. Moreover, OpenFace's 'confidence' metric did not directly map to the blink signal– when participants gaze downward at a desk or a phone, they can appear almost as though their eyes are closed. This is a situation where the head remains confidently 'tracked' but blink detection is very difficult for the classifier.

To overcome this, we hand-labeled over 2 hours of blink footage, capturing 2586 blinks (>500 per participant). These labels are done by watching footage (in slow motion for fast blinkers) and tapping in time on a keyboard– as you can see from the blue outlines of hand annotations in Figure 5-21, they are not always perfectly accurate. To handle these timing errors, we use a wide margin for error of several hundred milliseconds when estimating the quality of OpenFace and our glasses data.

We hand-labeled only footage that had been pre-selected for its high confidence from OpenFace but we were still forced to skip about 15 min of footage that contained subsections that were difficult for labeling (i.e., the head was at an angle such that blinking was occluded). Our results for OpenFace blink detection thus overstate the quality of OpenFace blink detection based on the provided confidence metric, since we've thrown away about 1/5 of footage that contained some of the most difficult and ambiguous 'confident' OpenFace blink data.

For our glasses, while the IR emitter is normally on to ensure a strong received signal, we designed the circuit to latch the emitter off in bright IR conditions to prevent diode saturation. We did not encounter this condition in our testing (though many of our indoor environments had bright ambient light from nearby windows); the received signal was consistently in a good envelope. We did, however, detect a couple of small instances of bursts of environmental IR pollution from other electronics. These instances are easy to identify because they cause consistent and repeated blink detection at a constant rate over a short period. We removed an additional 2-3 minutes' worth of data from the two hours of our recording because of this issue.

Our results are listed in 5.4. The glasses algorithm outperforms the OpenFace classifier by a wide margin, even for a relatively constrained naturalistic test with a generic blink detection

algorithm. Breakdowns in classification accuracy are quite different for both algorithms. OpenFace struggles when people are looking down (i.e. at their phone), which they do a lot in natural settings.

Ironically, participants with rich IR glasses data streams perform the worst with our blink detection (lower participant as compared to upper participant in Figure 5-21). This data captures more secondary eye movements– it is marked by good IR coverage of the eye, high contrast between pupil and cornea, and lots of attention shifts and saccadic movement– and these details can significantly alter reflected light, which engages the blink detector.

In the raw IR data, we can see characteristic signals of a blink, but we also see the opportunity to extract gaze/attention shifts (looking down at a phone or desk vs up at a screen) and even saccadic eye movement in some cases. With a more complex future algorithm that incorporates metadata about the blink, signal shape, and/or classifies other eye movements like gaze shifts, we should be able to do an even better job than this baseline technique.

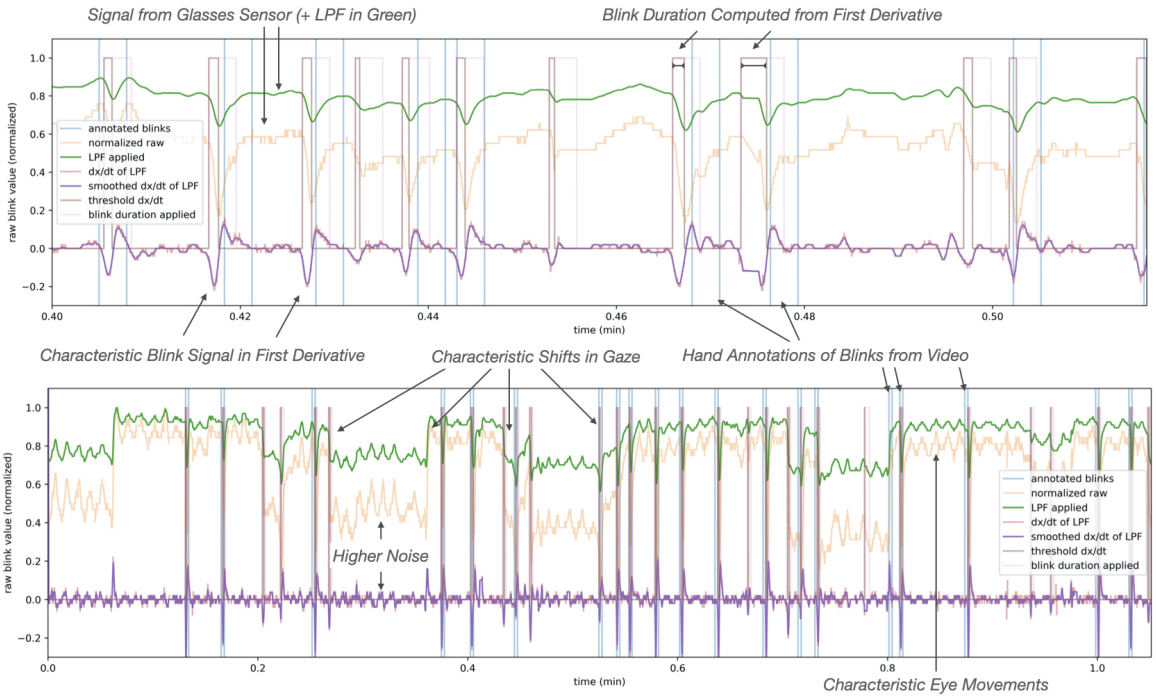


Figure 5-21: Representative blink trace data from Participant 3 (bottom) and Participant 5 (top).

These early results have been further corroborated in a separate pilot user study with

	OpenFace		glasses algorithm	
participant #	*sensitivity (%)*	*false discovery (%)*	*sensitivity (%)*	*false discovery (%)*
1	38.4	21.7	91.9	5.0
2	53.3	58.0	88.7	33.8
3	67.7	19.1	62.1	18.2
4	75.9	37.9	85.2	19.9
5	68.9	44.4	80.3	14.7
overall	**61.2**	**38.8**	**81.6**	**18.2**

Table 5.4: OpenFace and Glasses Algorithm Success for Detecting Blinks Across Participants.

a further 6 participants, in which each user performed half an hour of cognitive tasks (IRB protocol #2001000083) alongside video recording. Initial results show comparable performance to off-body systems across subsystems and users in this test as well.

These results give high confidence along a few major dimensions: first, our sensor geometry is designed well enough to accommodate a diversity of head sizes and eye/nose angles. Secondly, we see useful data capture with the glasses that outperforms high quality and well-controlled external references. We anticipate noise due to motion, temperature acclimatization, and external IR sources to become more pronounced under social, ambulatory conditions. These initial findings support the idea that large scale, naturalistic data collection are a fundamentally different challenge than typical highly-controlled, heavily-repeated laboratory experiments, and that Captivates is positioned well to extend our research into these challenging, novel scenarios.

Design

As we've outlined above, a major objective of the Captivates project is to create a pair of glasses that significantly improves on aesthetics and comfort, while maintaining all day robust performance. As much as possible, we wanted to build something that would not alter behavior or interpersonal interaction; a pair of glasses that had the highest likelihood of fading into the background over the day.

To test this hypothesis, we purchased several other pairs of smartglasses for comparison

and ran some initial surveys. The glasses we used are shown in Figure 5-22; they include well known smartglasses: Google Glass, which includes a suite of functionality, including an outward facing camera; Bose AR, which simply plays audio; the Snapchat Spectacles, which can take a picture on demand; and the GoVision video recording glasses. We chose these glasses as reference points because they bound interesting corners of our design space– all include electronics, but the Bose AR hides it completely, designed for broad appeal by a large corporate team with access to market research. Google glass embraces a tech-futuristic aesthetic; our prototype could be (generously) construed this way as well. Spectacles are well-designed, but have outward facing sensors like Captivates.

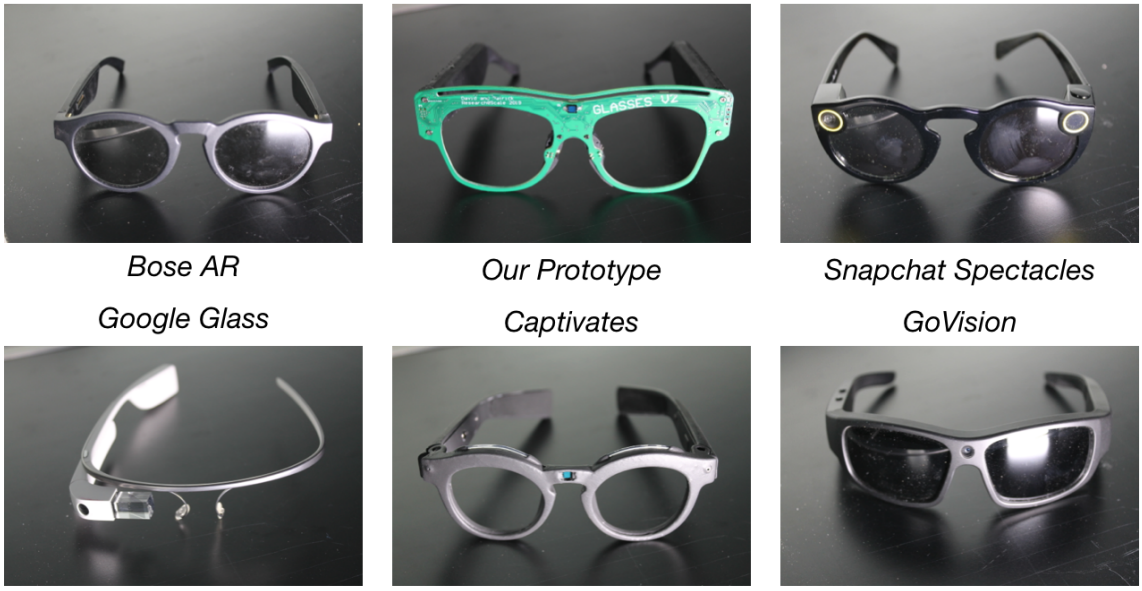


Figure 5-22: Smartglasses Compared in Survey

To compare against these products, we included our initial prototype (same functionality, unoptimized for aesthetics) and our final design. Our survey population included 101 participants; 63% female, and of all participants, 50% wear glasses daily and overall, skewed slightly towards 26-35 year-olds (57%). In addition to evaluating each pair of glasses, we rated participants on the SCS-R scale for Public Self Consciousness and Social Anxiety, to see if there were trends mediated by these psychometrics. Distributions along these axes were unimodal and no clear differences in ratings appear for high vs low groups in either category.

The results, shown in Figure 5-23, reveal some interesting trends. In the survey, a neutral option was also available to participants. We opted to exclude neutral ratings in these charts to highlight trends in strong reaction to each pair, but each rating sums to a count of 101 with neutrals included.

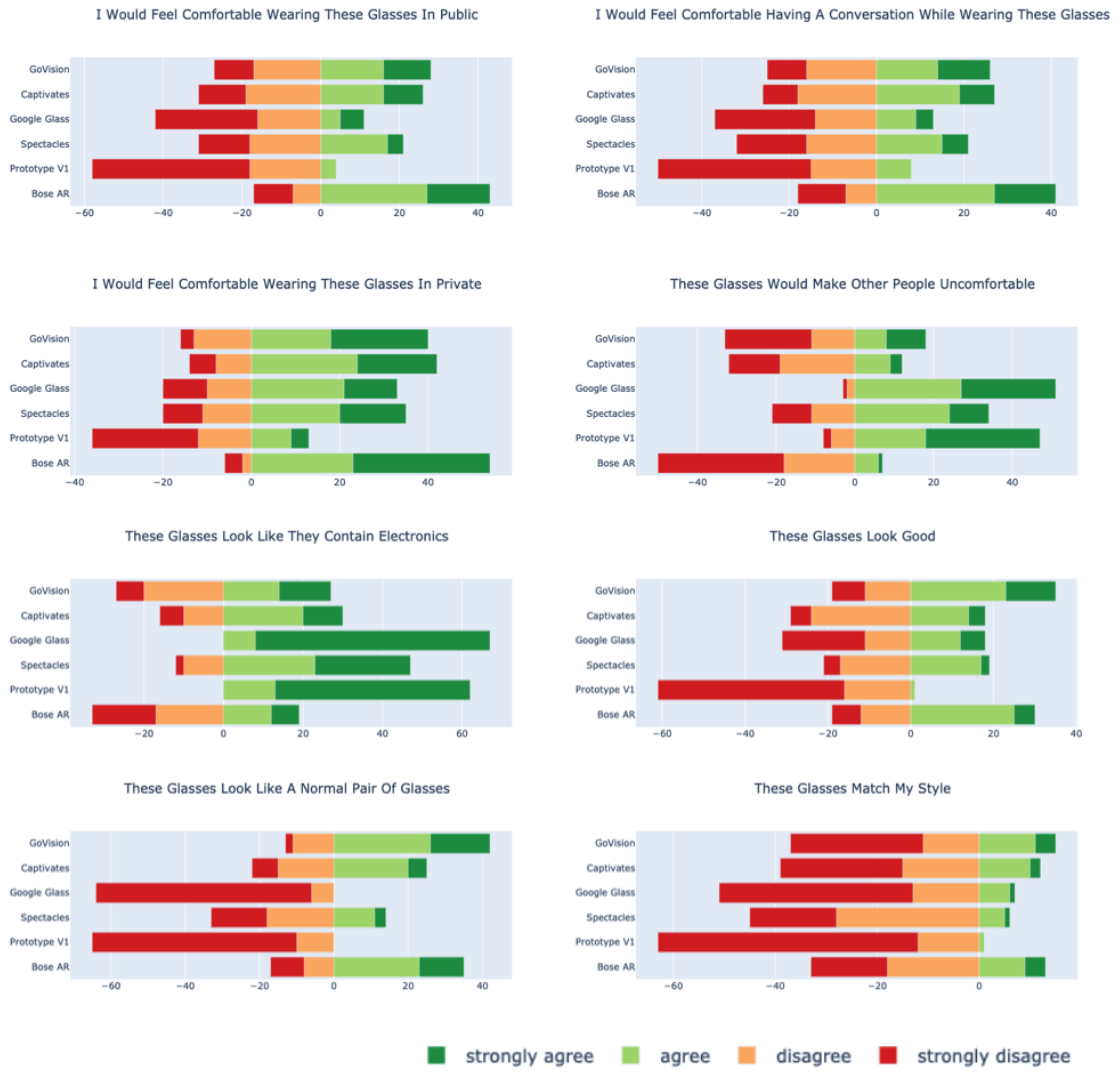


Figure 5-23: Glasses Survey Comparison Results

First, we notice that all of these glasses fare pretty poorly on style– even the most positively reviewed pair here received only 20% favorable ratings (Captivates received 18%). This is not surprising after our conversations with glasses manufacturers; glasses are a style accessory, and it's quite difficult to make a single pair that is generally well-liked. Our

initial prototype performed notably worse than the rest, both as ranked for personal style and in general appearance. The final Captivates design performs comparably to these other products.

The Captivates glasses actually do quite well compared to the others along self-assessed behavioral dimensions; they perform quite competitively with scores of comfort wearing them in public, private, and during conversation, as well as in assessment of discomfort to others (only outperformed by the Bose AR glasses). This is a truly significant shift from the rankings of the prototype (where an overwhelming 85% are uncomfortable with the idea of wearing them in public or while conversing), and clearly justifies the increased effort if our goal requires us to minimize changes in social behavior, self-consciousness, and stress or anxiety.

These behavioral ratings do not follow from the style ratings; they instead appear more closely linked to notions of social norm violation (do they look normal, and is it clear that they encapsulated electronics). Along these dimensions, the Captivates fares quite well, alongside the Bose AR and the GoVision frames.

A reasonable conclusion is that self-assessed comfort in social situations is dominated by the stigma surrounding social norms (video recording and electronics) compared with personal fashion or style match, as long as the style is somewhat conservative and doesn't grab attention. That is a positive result; it validates the notion that a generic pair of glasses doesn't alter self-assessed behavior as long as its style is within the norm (there is minimal advantage to trying to tailor the glasses to individual style for research). It also validates the notion that going from an obvious prototype that is outside the norm and has obvious electronics to something that looks 'normal' and integrated has a huge impact on self-assessed behavior.

There is still obvious room for improvement. The Bose AR are the best anchor point in this survey, as they include minimal electronics that are not visible, they have no electronics or unusual features on the face of the glasses, and they've had commercial success. A strong majority of people (58%) are happy to wear these glasses in public and in conversation without perceived social stigma.

Our glasses still have room to grow to compete with Bose AR. That said, we have integrated significantly more sensors throughout the glasses, including in difficult to ignore areas of the front face. If we take a look at open-ended comments surrounding the Captivates frames, most of the critiques are centered around four main themes: about 18% noted either bulkiness or the exposed IR diode in front as chief dislikes; the next most common complaints (roughly 7%) noted the surface finish or the lack of lenses.

5.11.5 Conclusion

With the Captivates project, we set out to build a system for long-term, in-the-wild psychophysiological monitoring at scale. We integrated many physiological sensors into a glasses form factor, particularly in challenging locations around the bridge of the nose. Our area of interest– naturalistic psychological measurement– dictates that our wearable be unobtrusive and socially acceptable. With that in mind, we focused heavily on an aesthetic that would minimize social anxiety and stress, as well as optimizing our comfort, weight, sensor selection, and battery life to let the user put them on without thought and forget about them all day long.

Our validation shows success in data quality and success in aesthetic design, in-so-much that people self-report that they would feel comfortable across public and private interactions, including conversation, while wearing them. It also validates that this was absolutely not true for our initial prototype, which 85% of people would not feel comfortable wearing in public, and demonstrates the value in our additional engineering effort to enable naturalistic study.

We believe to make meaningful progress on psychophysiological modeling, we not only need to move to naturalistic settings, but collect vast amounts of data. Once again, Captivates was designed to hit this mark– it is a scalable and robust solution, enabling large-scale, long-term studies with minimal administrative oversight. It is also open-source and available for other researchers to use and adapt.

Given these intentions, Captivates unique contributions are as follows:

- An open-sourced, manufacturable and scalable platform that is available for other researchers to use, audit, and modify. Captivates enable researchers to leverage a hardware platform for their specific use cases, and can enable large-scale, trustworthy data sharing on physiology and mental experience without proprietary algorithms or APIs common to productized platforms.

- A product-like, comfortable, and aesthetically neutral design that can be used in real-world settings without altering naturalistic behavior. This is crucial for studying this physiology under as-realistic conditions as possible, and prototype-like designs alter self-reported behavior. These improvements require significant additional engineering effort.

- A robust form factor that is easy to use and lasts all day on a single charge. This is crucial for scaling trustworthy data collection without administrative overhead, and makes the glasses less intrusive into a participant's daily experience. Large scale data will be necessary to make any generalizations given the high levels of noise and uncertainty associated with real world settings.

- A platform to study understudied physiological indicators of face temperature, subtle head motion, and blinking in natural settings. These have been demonstrated in lab contexts, but naturalistic analysis of these indicators is novel. Furthermore, our comparison against standard camera-based techniques shows we significantly outperform them for blink tracking and spatially consistent temperature measurement, even in a relatively uniform working environment. Heart and respiration rate estimation from our head mounted IMU data should be possible as well. As a stand-alone device, or as part of a broader physiological measurement suite, Captivates serves as a power research tool for in-the-wild physiological and psycho-physiological data collection and modeling.

In order to accomplish these goals, we spent a summer in China working with manufacturers and glasses designers, partnering with companies, and worked with sourcing agents. We pushed our design towards product and manufacturing in a way that is uncommon within

academic research, but which we believe will become more prevalent in the near future. The ability to manufacture and quickly scale hardware design opens up opportunities for new research techniques. Captivates leverages this unique opportunity to scale data collection.

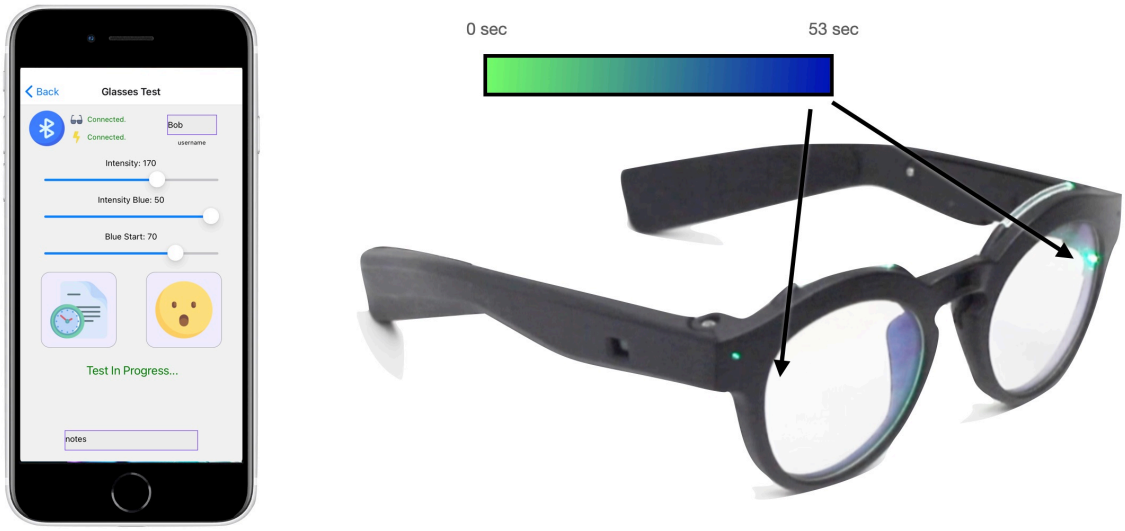
Captivates opens the door for large-scale data collection of physiological indicators, in a way that will enable better predictions of mental state in natural contexts. Our initial real-world pilot study yielded promising results and further data collection is planned to fully validate the platform. The immediate future will be spent collecting data and using that data to engineer and validate psychophysiological models that can estimate mental states accurately.

Success with these models would bring us a long way towards translating psychophysiological laboratory techniques into real-world insight. Captivates serve as an open-source bridge to that end. A strong model will enable Captivates to quantify the long-term psychological impact of our design decisions and provide real-time feedback for technologists interested in actuating a cognitively adaptive, user-aligned future.

The Captivates system is a key enabling piece of future responsive, adaptive environments. We envision an ecosystem of such technology, including dense arrays of distributed sensing and better actuation of shared spaces. With the Captivates project, we are one step closer in understanding the impact of our designs on human experience, and thus one step creating truly supportive, responsive environments.

5.12 Wearables for Flow Estimation: Peripheral Light Interaction

In the previous section, we looked at the Captivates platform, which introduced and validated many useful physiological sensing modalities for on-the-go flow estimation. Using the LED mounted in the periphery of the thetse glasses, in this section we introduce a novel behavioral interaction to measure attention dynamics in naturalistic settings. Our contributions are as follows:



Images of the Smartglasses platform and the Interaction Design used in this experiment.

Figure 5-24: The Captivates Smartglasses platform presented in [84] is used to measure self-interruption behavior. Two symmetrical peripheral LEDs slowly change from green to blue over 53 seconds; when the user notices the change, they indicate it by hitting the 'surprised' emoji in the companion application (which is left open on the table near them). These transitions are spaced 10-20 minutes apart to give the user sufficient time to achieve a state of deep focus on their primary task. The change is slow and gradual to minimize the probability of drawing attention to itself.

- We designed and created a novel, wearable, and discrete interaction to measure self-interruption dynamics in daily life. This interaction is built on top of Captivates; the glasses present a slow color change in the periphery of the user's visual field, gradual enough that users' attention will not be drawn to it. When the user notices a shift in color because they have scanned their environment, they indicate that with a companion application.

- We ran two small pilot studies to test the efficacy of this interaction and demonstrate success in creating an interaction that is 'change-blind' to study participants and captures data related to their state of engagement. We discuss and analyze these preliminary results.

- We discuss future applications and design considerations of this interaction for studying naturalistic attention dynamics and for quantifying the immersive success of de-

signed experiences.

5.12.1 Background

We live in a world with constant notification and interruption that fractures our attention. Workplace interruptions occur every 4-11 minutes; 70% of the roughly 90 emails we receive in a day are opened within 6 seconds [280]. External interruptions make us more likely to self-interrupt in the hour after they occur [101].

Fortunately, notifications can be designed to to be more or less distracting from our primary task [244]. Researchers have prototyped notifications to allow deep focus to proceed uninterrupted by designing them right at the threshold for capturing attention. The user's attentional state is the primary factor that determines whether they notice an aural cue (by modifying background music) [30] or a visual one [216].

At the other end of the design spectrum, some cues are impossible to ignore. Motion and luminance changes powerfully capture attention [343]. When motion cues are masked– including using gradual fades– large visual changes are difficult to spot, regardless of the magnitude of change in contrast or color [391]. This phenomena is known as 'change-blindness'. Moreover, research on 'inattentional blindness' supports the idea that we are blind to large, obvious visual changes in the center of our visual field when focused on a task [225].

Some distractors are more perceptually salient than others. However, even given a task, we have no accepted standard to measure our sensitivity to chromatic or luminous changes under varied ambient lighting conditions; the best models estimate cone absorption in the retina [76, 456]. Ambient light levels mediate visual detection through pupil dilation; a smaller aperture gives better spatial resolution while a larger one gives higher signal to noise and more light [325, 130]. Large pupils may thus give an advantage to detection of faint peripheral stimuli [282]. A simple relationship between low-level perceptual changes in brightness or contrast and attentional capture is impossible, though; for example, luminance

changes characteristic of a new object in a scene will capture attention more readily than twice the difference at a less-surprising location [403].

Many factors influence our perception of task-irrelevant, peripheral distractors. Motion and luminance differences– especially unexpected ones close to our visual focus– are most likely to grab our attention. While ambient conditions likely effect our ability to discriminate peripheral stimuli, this is likely to be a second-order effect relative to our psychological factors like expectation and attentional state.

Measurement Techniques

Mainstream measures of engagement rely on self-assessments [139] that are biased by demand characteristics and peak-end memory distortions and fail to capture uncertainty well. One more quantitative approach to cognitive state inference is the peripheral detection task (PDT), an ISO-standard test to measure cognitive load while driving [462]. This test has been updated to use a head-mounted peripheral light as a secondary stimuli while driving (hDRT, or head-mounted detection response task) [414].

In the hDRT, a red LED is turned on rapidly (every 3-5 sec), at constant intensity, long duration (1 sec), and reliable noticeability (>95% in practice– reaction times are the primary hDRT measure). When the participant notices this light, they hit a button; response latency is the primary measure of task load.

DRTs do not disambiguate between noticing and reacting (there is evidence of physical response conflict); also, due to the consistent nature of the task, participants treat it as a secondary goal for which they strategically allocate attention [462]. While useful to characterize multi-tasking load under intense driving conditions, this rapid and consistent dual task paradigm is a poor fit for the naturalistic study of deep focus.

Participant #	1	2	3	4	5	Overall
Low Warm Light *(avg (std))*	4.6 (1.4)	5.2 (1.5)	2.5 (1.1)	9.0 (1.2)	5.9 (1.5)	**5.4 (2.5)**
Bright White Light *(avg (std))*	4.6 (1.5)	6.3 (2.4)	5.5 (1.4)	8.1 (0.6)	6.1 (1.1)	**6.2 (1.9)**

Figure 5-25: To test the effect of variable lighting conditions on perceptual acuity of the LED color shift, we used a specially designed lighting room set to two extremes; a bright white light (high blue light spectrum- Figure Left) and a low warm light (Figure Right). Participants were asked to look at an 'X' on the wall and pay attention to the LED color change in their periphery, indicating when they noticed the transition in the application. Participants were exposed to ten transitions in each lighting condition; the order of exposure to each condition was counterbalanced across participants. Results from these calibration sessions (in seconds) are shown in the table; the transition was sped up 3x from the typical intervention to 17 seconds to maintain attention.

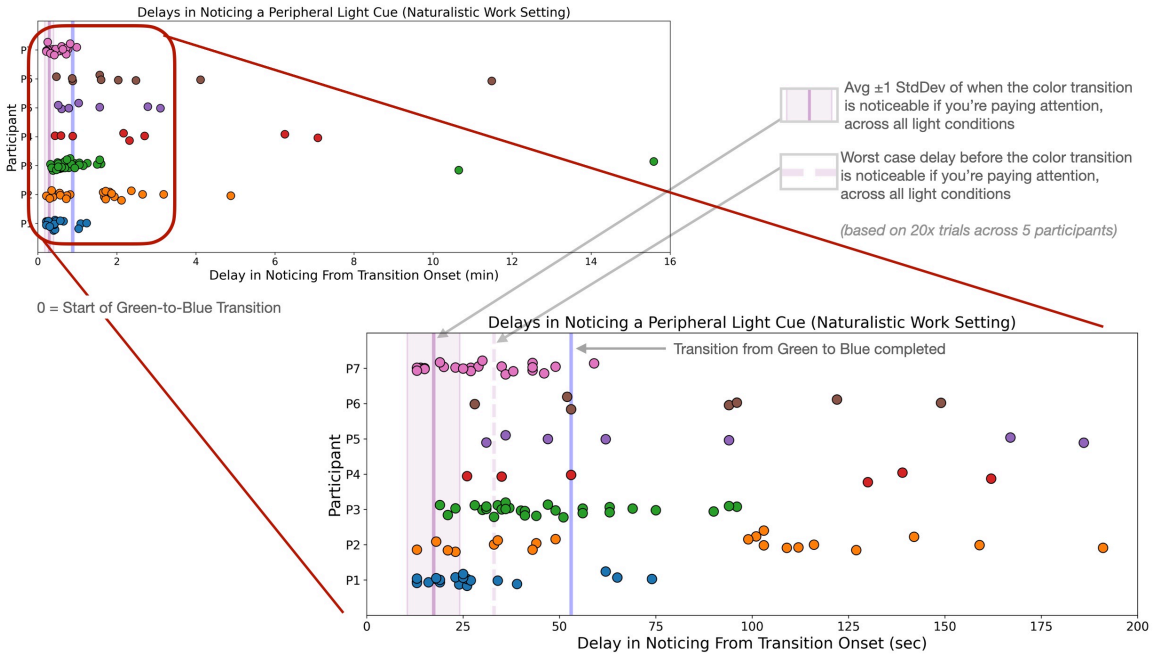
5.12.2 Captivates Interaction

We design a new interaction using the two symmetrical RGB LEDs mounted in the outside peripheral vision of the user, controlled by LED driver ICs that enable ratiometric and logarithmic PWM dimming. Figure 5-24 details the intervention design– when the user notices a subtle, gradual change in color of the LED (which occurs every 10-15 minutes) they hit a button to indicate it. Their delay is an indicator of the frequency of their self-interruption. The specific design decisions were based on an iterative design process to isolate a subtle, gradual, but distinct peripheral change. The glasses include 300 mAh battery capacity that allows the intervention to run continuously for over 7 hours per charge; it is also a socially acceptable form factor that enables daily use in normal life without inducing self-conscious behavior.

Calibration Test

For our initial test, we measure (1) when during the transition from green-to-blue an attentive user notices the change, and (2) how much variability various lighting conditions introduce. Five users sat in a closed rectangular room outfitted with a full Color Kinetics lighting system (programmable with color temperatures between 2300K and 7000K). While looking at an 'X' on the wall in front of them, they indicated when they noticed the light in their peripheral vision shift colors. To maintain attention, the transition was sped up 3x from the typical intervention to 17 seconds. They repeated this task twenty times with random delay– ten under bright white lighting conditions and ten under low warm lighting conditions (counter-balanced). Low warm lighting is most favorable to the task: dilated pupils give the best peripheral detection and the most contrast with the background (warm light has very little of the target blue wavelength). In contrast, the bright white light has much less contrast in the blue spectrum and constricts the pupils.

Results are shown in Figure 5-26; there was not a significant difference in the participants ability to notice the color shift across light conditions. Across all 100 trials, the best and worst cases for perceiving the change were between 12% and 65% through the transition.



Summary results from preliminary testing.

Figure 5-26: Initial results from seven participants wearing the interface in natural work/social settings for a minimum of 2 hours each (19.3 hours captured total). Data points represent the delay from a transition starting to when the user noticed it. Included in the plot are indicators of when the transition becomes observable based on the calibration data (purple), as well a worst case indicator over the 100 calibration trials (dotted– this represents the worst case moment, across lighting conditions, that we'd expect the transition to become obvious if the user was paying attention.) The slow transition completes at 53 seconds (blue). We see several data points in which users didn't notice the transition for many minutes after it was completed (top left); a zoomed view of the first few minutes after transition onset are shown in the bottom right. As expected, noticing delay follows a roughly log-normal distribution for each user.

Primary Test

In our main test, seven users experienced the intervention in Figure 5-10 as they went about their daily lives, with an open iPad within reach to indicate their awareness of the LED changing. We collected over 19 hours of data, with transitions spaced 10-15 minutes after any indication to allow the user to regain task focus. The data are represented in Figure 5-25. The longest a participant went without noticing the light change was 15 minutes; 35% of the transitions were noticed after the transition had fully completed, with the vast majority significantly delayed (74% of those were >30 seconds after the light change had completed). Four of the seven users had delays of 5 minutes or more in noticing at least once, indicating a deep state of focus on their primary task.

5.12.3 Discussion and Future Work

Whether or not a participant notices a gradual, peripheral color change is dependent on environmental lighting and attentional state. We designed an intervention with the hope that users would be change-blind to it. Our initial results imply minimal variability in noticing for vigilant user across extreme ambient lighting conditions; this variability is small compared with the delays we find as users engage in their day. Our results support our hypothesis that this intervention quantitatively captures useful insight into self-interruption and environmental awareness in real world settings.

In our future work, we will collect more data, and build probabilistic models to make strong inferences about an individual's attention based on it. Future experiments may introduce longer delays between transitions to allow users more time to achieve a deep state of focus; this trade-off between quantity and quality of data is unstudied. Though it appears the effects of ambient lighting are minor (and can be modeled as noise), we also haven't tested for interaction effects between attentional state and ambient lighting. Future work may extend our understanding of ambient lighting conditions using tightly controlling lighting or by measuring ambient light levels with an additional wearable.

5.12.4 Conclusions

In this section, we presented a novel interaction design to study self-interruption and immersion across naturalistic contexts. Users indicate when they have noticed a gradual, subtle, peripheral light cue change– a transition that only happens every 10-15 minutes. In two small pilot studies we have demonstrated that ambient lighting levels minimally affect perception of this transition, and a significant portion of these changes are unnoticed for several minutes by our users in naturalistic settings; attentional state is thus likely to be the primary causal precursor of this data, with a large impact on the measured delays.

We believe this interaction can improve our understanding of the dynamics of human attention, and provide quantitative insight into design's impact on the creation of immersive, engaging, meaningful experience. This approach represents a quantitative move forward toward personalized, adaptive, and empirically grounded immersive design.

5.13 Limitations and Future Development

This chapter introduced several, novel hardware prototypes– a set of tools to evaluate flow states naturalistically and bio-behaviorally. While these interfaces each include (1) an all-day battery life and (2) an unimposing form-factor so we can collect naturalistic data, the work is largely exploratory– testing new bio-behavioral principles in real-word contexts at a foundational level. Our goal has been to answer basic questions about whether there is interesting variance in the signals we collect that we expect to be useful for psychological inference. These studies feature small test populations; this data is very time-consuming to collect (a handful of data-points a day per user, maximum) and fast design iterations are important.

Having established more work is warranted, a major contribution of this work is in charting the course for future development. Each of the current prototypes includes sources of noise that could be better controlled; each requires a relatively sophisticated underlying model

to operationalize it effectively. Below are specific limitations and opportunities to improve for the devices introduced above.

5.13.1 Equinox

Equinox represents, to our knowledge, the first attempt to quantify time distortion in normal, daily life, and the first use of 'clock-time' estimates in time perception research. Our preliminary results suggest normal people experience significant daily distortions in time perception; also, we might reasonably expect people to work for intervals of about half an hour on average without checking the time in the normal course of living– sufficient time to test the experience of time. Our data is also suggestive that duration estimates and clock-time estimates have different underlying mechanisms. No firm conclusions are possible from such a small dataset, but the initial results are encouraging.

To utilize the data from these estimates requires an underlying model that ties cognitive states to time perception. As a first pass, it's possible to adapt Treisman's pacemaker-accumulator model; however, our initial data (and much of the philosophical debate around time perception) suggests this model is almost certainly inadequate.

Future work needs to dive deeply into the underlying mechanism for time distortion measurement. Given the early test data, we hypothesize that clock times are not conceived and remembered as explicit time values, but experienced in rough terms related to the time left before the next time explicit commitment for that day. Future versions could integrate the key time-keeping tasks of the day (i.e. calendar integration, to make sure you are on-time for commitments so you don't have to check the time); the interface would then be explicitly aware of key time landmarks for the user throughout their day.

Another major question involves individual variation in time perception; whether some people are naturally imbued with a better 'sense' of time than others, and whether this 'sense' improves through implicit feedback when interacting with an intervention of this kind. (Even when the feedback is hidden, they will get *some* objective feedback about the

time of day in order to allow them to live a normal life.) These are open questions that require large-scale, long-term data collection to answer in a definitive way.

A final limitation comes from the pervasive nature of time-keeping. Each incoming email and text is timestamped. Games frequently inform the user how long each round has been, or include countdowns; movies and videos highlight exactly how long they will last and how much time has elapsed; we frequently consider at this information when selecting content. Every person has a large, visible clock on their phone, their watch, and their computer; you cannot ask them to hide these devices. Time cues– varying from the explicit (a clock or a timestamped email on a day with relatively consistent messages) to the implicit (this video is 12 minutes long)– are embedded everywhere; it's difficult to account for the influence of these time cues in natural time estimates, and also difficult to formalize a process that accounts for all of them.

5.13.2 Feather

Feather is a device to test variation in threshold of noticing in very subtle vibrations on the leg over a long interval. Once again, while the data is strongly suggestive that some of the variation we've capture is attributable to attentional state, this is preliminary work and the data is insufficient for specific conclusions about the underlying psycho-tactile relationships. There are a few major challenges to the design of this interface that should be addressed in future revisions:

- Repeatable placement. Our sensitivity to vibration varies greatly over the body; this is particularly true on the leg. It is difficult to draw conclusions across sessions, and the initial calibration process to find an ideal placement is very important.
- Repeatable vibration. With this prototype, we used off-the-shelf vibration motors which are designed to create powerful vibrations in a small package; we are driving them at the lower limit of their ability and there is variability in the resulting delivered peak force as a result. In this work, we attempt to account for this by calibrating

the entire system as a black box and treating the values probabilistically; future work should tackle this problem head on with a custom vibration motor design that operates at much lower intensity, or intentionally dampens the motor such that the range where it operates intersects our tactile threshold.

- Psycho-tactile linear ramp. The perceived intensity of a vibration is a complex psychophysical relation; we would like to have precise control over vibration intensity and increase it linearly in perceptual space. In this work, we stepped up our intensity at the finest resolution we were given, which was designed by Pavlok to have this property (roughly). This can and should be more precisely designed.

- A Model of Habitation. For this to work, the motor needs to press against the skin. We might reasonably expect a habituation effect (lessening of tactile sensitivity) from wearing the device all day. We need more data to measure and account for habituation over the day, perhaps including force sensing/contact sensing to make sure contact and placement are consistent.

- Motion Artifacts. Detecting a vibration will be more difficult if motion is happening on the leg. Most of the cases we care about imply stillness; however, this is another confounder we should consider in future modeling efforts. Future versions should include an accelerometer to account for leg motion.

- Sensory Confusion. Several people reported uncertainty about whether they felt something or not during calibration and during the experience. In the first iteration, they were told to simply wait and hit the button when they felt totally confident that they had felt a vibration. These moments of uncertainty could be better captured and integrated.

Despite these challenges, our first pass interface shows significant variance in reported threshold values well outside of the region we would expect given initial calibration on the user– we have a strong baseline for the limits of what they were able to perceive when they were focused. This variation cannot be explained by habituation effects or placement because the reported thresholds are not at all monotonic, as we would expect with baseline

drift. The preliminary data thus shows reasonably strong evidence that psychological state is represented in the data. This is the strongest reasonable claim from the exploratory data collected; future work on tactile thresholds for psychological insight will require significant engineering effort into more precise and repeatable hardware and a stronger underlying model of tactile perception that incorporates psycho-physical principles alongside a model of habituation and placement.

5.13.3 Captivates LED interaction

Captivates are used with the peripheral LED to track noticing behavior. This interaction relies on minimizing the attentional process which draws your attention to the light; successfully making the color change 'change-blind'. We have controlled for the primary driver of attentional processes in visual scenes– changes in luminance and motion (effectively, something new entering the scene or a sudden shift in a light's brightness); the delays we see in noticing and the small variability in noticing in controlled tests over a wide variation of external light conditions suggest we successfully achieve this.

It is possible that aspects of the light stimuli itself, its color transition, its contrast, or its placement in the visual field makes it more likely to exogenously grab attention (something about it calls the user's attention to it) vs the endogenous process we're interested in. It is possible that situations where the peripheral light enters into central vision or passes in and out of the visual field (depending on face geometry, individual peripheral vision sensitivity, and task-dependent gaze shifts) and the luminance contrast of the LED with ambient light increases the probability of an exogenous attentional effect.

This suggests that shorter recorded intervals may not be as strong an evidence of very low focus state (i.e. we could attribute the short interval either to their focus or a really obvious transition) vs the longer intervals are for high levels of focus.

In practice, this techniques appears to be relatively robust to varieties of realistic external lighting conditions; for example, when we compare LED notice durations in the lab (highly

controlled lighting in our 'lighting lab') to LED notice durations at home (highly uncontrolled lighting) we see similar distributions; similarly, we don't see major differences in the distributions of Tetris (no gaze changes– the game is played on a small screen in the central visual field) compared with other activities. These are all very preliminary data and merely suggestive. Future improvements include adapting to external light levels, modeling gaze/LED visibility, and collecting more data to inform a model of these effects.

5.13.4 Captivates In General

Even with the huge development effort behind Captivates, there are still many sources of noise and room for improvement in all of the sensing modalities. The dual-processor design can be a challenge for development. Newer revisions simplify the structure and replace many of the analog traces to minimize noise, improving thermal and motion sensing. Offline blink processing is a challenge, as gaze changes and external lighting also appear in the data, and face geometry can have a large impact on the signal. There is more work to be done to model the blink signal, combining it with head motion data and co-modeling gaze/saccades where there is evidence for them in the data. Future versions of the glasses should incorporate other modalities of interest, including pupilometry and secondary measures of facial expression.

Chapter 6

Measuring Flow Across Contexts

In this chapter, I review an experiment I conducted on flow states across 22 participants.

6.1 Prior Work

6.1.1 Flow Induction in the Lab: Tetris

Flow studies are typically correlational and survey based; few researchers have attempted to induce flow in the lab or study its physiology. [305] When experimental work has attempted to look at flow, video games are the most common method, specifically Tetris. [240, 305, 236, 237, 304, 328, 107] Moller et al. write in [305] that work with Tetris "...represented an important and significant advance toward developing a reliable induction procedure for inducing flow experimentally." (Moller collaborated with Csíkszentmihályi himself on flow induction in the lab using Tetris. [304])

These studies typically present the game under three difficulties– easy, adaptive, and hard– with the assumption that adaptive will produce flow. In the earliest of these video game studies (which used Pac-Man instead of Tetris) gameplay at each difficulty setting lasted 5 minutes each [357]. For Tetris, the reported task durations are 3-5 minutes, 3 minutes,

6 minutes (2x), 8 minutes, and 15 minutes [107, 236, 304, 268, 192, 237]. In these studies, flow is usually measured with the Flow State Scale (FSS), [305, 268, 192, 242] though [236] use one simple question as their measure ('To what degree did the demands of the game match your ability?'). We repeat the use of this measure in our study and use durations commiserate with the longest reported in previous literature– itself twice as long as the average– with one of our conditions almost twice the longest reported duration (25 minutes).

The adaptive strategy used with Tetris to induce flow is novel for each researcher. In [236] the heuristic was to increase the speed if players complete more than four lines and decrease it if players complete less than three lines over thirty tetrominos. In [304], they used a similar adaptive strategy in a baseline session to judge player skill, and then set the speed at a fixed level for their test conditions. In [268], a much more sophisticated dynamic strategy is used [267]: every ten moves– depending on board height, holes, and score– the distribution of tetrominos will either be 'random' or 'helpful', where helpful pieces are based on past game analysis of which pieces are easiest to handle for a player at this skill level.

These sophisticated strategies do make a measurable difference compared to normal Tetris in short duration play (3-5 minutes), as measured by the short Flow State Scale; however, the differences are relatively small, and no difference in FSS ratings between adaptive techniques and trivially easy settings appear at this timescale (n=40, with likert-scale ratings out of 5.0 the average 'adaptive' score was 4.0, 'easy' score was 4.0, and 'normal' score was 3.8) [268]. This appears to corroborate earlier findings that easy conditions are similarly effective as adaptive conditions, which are moderately more successful at flow induction than overtly difficult ones, as measured by challenge/skill balance on slightly longer time-scales [236].

6.1.2 Physiological Analysis of Tetris

The literature looking at the physiology of flow is even sparser, as it is contingent on flow induction. As we saw in the last chapter, general flow physiology shows conflicting evidence for facial EMG, breathing, and sympathetic activation in flow states [240, 333]. Studies of Tetris specifically show larger respiratory depth, decreased heart rate variability

[192], and high salivary cortisol [237]. fNIRs data doesn't show reduced oxygenation in the prefrontal cortex while playing Tetris, as hypothesized under hypo-frontality [192]. It has been conjectured that video games generally introduce relatively higher levels of stress and cognitive load compared to other flow tasks, complicating the interpretation of these results.

Physiological analysis techniques vary widely in the literature; for instance, in the literature above, HRV is captured using four different methods– the raw IBI intervals, averaged; the RMSSD (RMS of difference between adjacent intervals); the low and high frequency power of its FFT; and the $\log_2()$ transform of these power values. After selecting a summary statistic, one value per user and interval is usually aggregated and compared using a standard statistical test. At this point, techniques once again diverge; in some cases a statistical test that rests on an assumption of normality is applied, typically after a check (i.e. t-test, Tukey HSD) [192, 283]; in other cases, non-parametric tests are used (i.e. Friedman's ANOVA) [237]; in still other cases, linear models are fit to predict self-reported FSS [242, 106]. The unifying feature of this body of work is a single summary statistic per physiological indicator per interval, compared against a 'ground-truth' FSS score.

6.2 Methods

6.2.1 Main Conditions

Each participant wore the Empatica E4 on their non-dominant wrist, the Captivates Smartglasses, and the Equinox watch on their dominant wrist (as described in the preceding chapter) and answered survey questions about their flow states. The main conditions were (C1) in the lab, playing Tetris; (C2) in the lab, doing an activity of their choice that typically induces flow; (C3) at home, playing Tetris, and (C4) at home, repeating the flow activity of choice. For each of these conditions, they were left to do the activity for 20-30 minutes, giving 1hr40min of data collection per person across these four conditions on average.

These occurred over two sessions, a lab session (S1) containing a short demo and familiar-

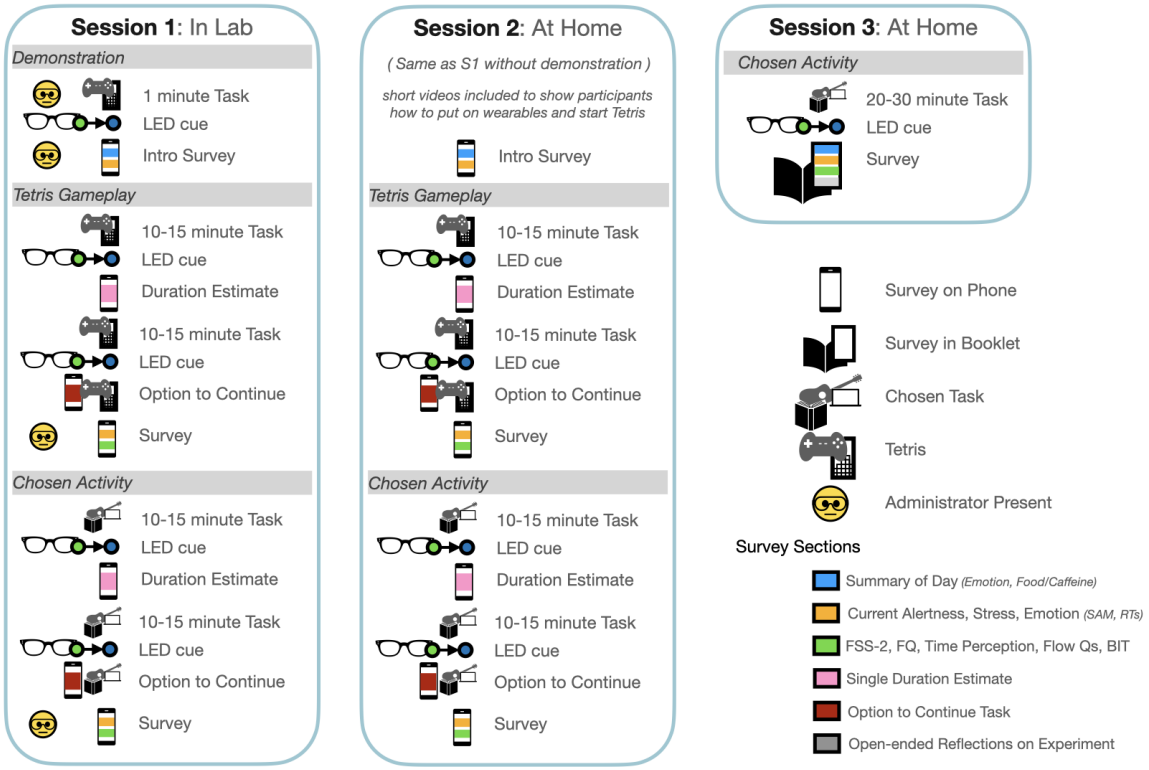


Figure 6-1: A summary of the main study conditions measured across three sessions.

ization session with the Tetris game, followed by C1 and then C2; and a home session S2 that simply included C3 followed by C4. Notably, the experimenter is not present for the home session; the experiment is designed to be portable and self-administered (the application includes many short videos walking the participant through the setup). In S1, the initial part of the experiment includes (1) interaction with the experimenter to guide the participant through putting on the wearables and (2) a 2 minute demo of the interaction paradigm while playing Tetris, with a short example of the peripheral LED change so the user gets comfortable with Tetris; (3) the experimenter also checks on the user in between the Tetris and Flow tasks. These three additions are not part of the home session; at home, the experiment starts with short videos walking the user through how to put on the wearables instead. Otherwise the structures of lab and home sessions are identical.

To signal the end of each condition, the LED in their peripheral vision slowly transitioned from green to blue (as described in the previous chapter)– how long it took for them to

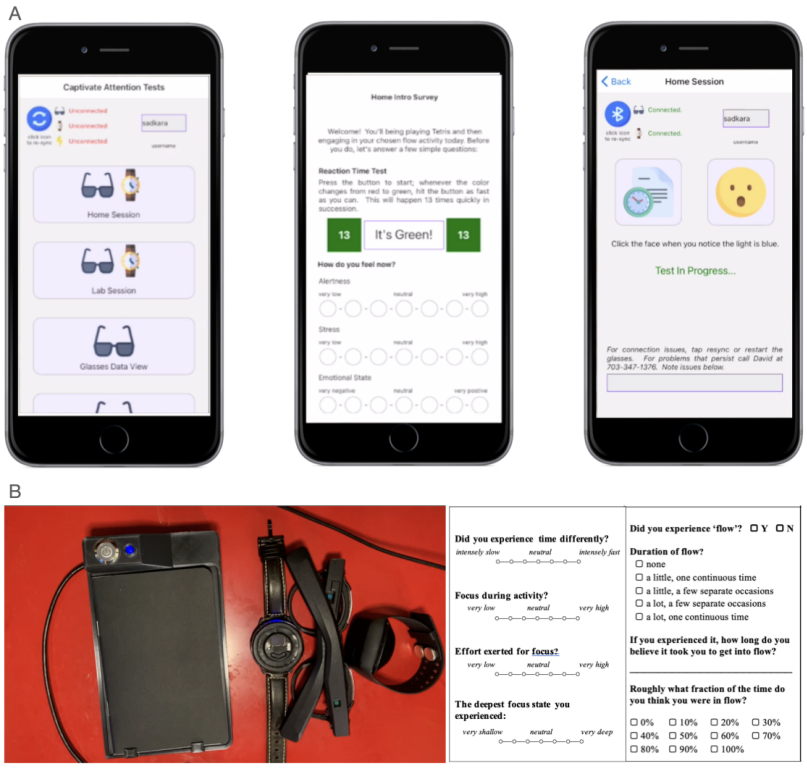


Figure 6-2: Screenshots of the companion application used during the first two sessions (A)– we see the main screen where we can select a session type, a typical survey screen, and the screen showing when the user is engaged in the task (they indicate they've noticed the LED color change by tapping the surprised face button). (B) shows the new interface for the final task; no companion application is required, just this box. To end the task, the watch is touched and the survey is completed in a the booklet that sits on top of the black data collection device shown.

notice this cue was recorded; they also had the option to continue the activity before filling out the end-of-condition survey. Additionally, an LED change occurred between 10 and 15 minutes into the condition, triggering one additional question when it was noticed ('how long do you think it's been?') A total of eight LED changes are registered over the four conditions– two changes per session.

The lab session took place in a quiet, windowless white room. Noise canceling headphones were provided, without music but with sound effects for Tetris. The user could recreate their task of choice however they saw fit to improve the likelihood of flow state, including listening to music.

Self-selected flow activities included a variety of personal video games, writing, reading, coding, designing, problem set solving, video editing, crocheting, djing, and drawing.

6.2.2 A Final Session

Participants were also asked to engage in one additional flow condition at home in a separate session (S3, C5) with their chosen flow activity. In this session, there is (1) no preceding Tetris condition, (2) no preceding surveys, (3) no mid-task interruption (i.e. the condition lasted 20-30 uninterrupted minutes), and (4) no companion phone device (parts 1-4 required a smartphone open nearby for participants to indicate they noticed the color change and to administer surveys). Instead, a small black box with no screen was provided to capture data; a booklet was provided for the survey; and the Equinox watch was used to indicate when the task had started and when they noticed the LED (which the participants could decide). The goal of this final intervention, besides corroborating the home flow data in C4, gives us a window into the task physiology when the task is uninterrupted for much longer and the visual and behavioral cues corresponding to study participation are minimized.

The test is ordered (1) lab Tetris, (2) lab flow activity, (3) home Tetris, (4) home flow activity, (5) home flow activity (long) instead of randomizing order over conditions for two reasons. The first is pragmatic– it's easier to bring people into lab and experience the test with the administrator nearby before sending them home to repeat it alone. The second is to structure the impact of novelty effects in line with the analysis– any discomfort with the lab setting and the wearables will be most impactful in the lab/Tetris condition and least impactful at home with a natural task. A major goal is to compare the typical lab study (which contends with these novelty effects and structured tasks) to naturalistic, at-home flow task (ideally unaffected by the novelty effects we mentioned).

6.2.3 Tetris Selection

Usually, the version of Tetris selected for in-lab flow induction is an open-source implementation where the tetromino drop rate can be set. In our testing of $\tilde{1}0$ different versions

of Tetris, its ability to elicit a deep flow has many intangible factors. We selected an implementation of Tetris called 'Falling Blocks', which is (1) highly rated in the app store (social proof of engagement), (2) doesn't include ads, breaks, timers, complicated menus, or unusual gamification/variation on top of the game (most other implementations), and (3) works with a high-quality external X-Box controller. The goal was to have a high quality, mobile setup with responsive, deterministic controls so that the study could be easily recreated at home as it was experienced at lab.

Previous work on adaptive gameplay has shown slight increases in flow induction compared to normal Tetris in 3-5 minute sessions, though the differences are not large, and adaptive play does not improve FSS ratings compared to trivially easy play [268, 236]. In our flow test, Tetris sessions last for much longer (roughly 25 minutes), but follow the normal ebb and flow of Tetris gameplay; starting relatively easy, getting harder until the player loses, and then starting over.

Our task structure is faithful to the gameplay structure that has lead to Tetris's popularity. This decision means we can select an unmodified, high quality version designed by a professional game designer instead of creating our own version of an open-source implementation. A professional game improves the odds of flow for reasons outside of the tetronimo drop speed (drop speed is typically the only design feature of interest for flow research). Based on the available data, trivially easy Tetris results in scores similar on the FSS as compared with a tuned level of challenge. Our version of Tetris should start 'trivally easy' for a few minutes, and slowly walk through a region of optimal challenge before getting too difficult.

We can compare the first few minutes of gameplay for each session with more standard flow induction in the literature; it is the same duration and in the 'easy/adaptive' regime. While the never-ending, tuned version might work well in short play-sessions, it removes fundamental aspects of the task structure over longer time-sessions that could contribute to flow state overall; namely, a period of relative ease to practice and master the controls and strategy, internalizing them and improving; and a period of optimal challenge that pushes you to your limit, rewarding you when you achieve. As we've seen from the enduring success of the game, alternating these periods can lead to a highly engaging overall experience.

6.2.4 Surveys

The entrance survey, completed before the lab session, includes (1) the Tellegen Absorption Scale (TAS), (2) the Flow Questionaire (FQ), (3) basic demographic information, (4) descriptions of experience with flow in the past, (5) the Brief Inventory of Thriving (BIT), and (5) questions regarding familiarity/enjoyment with video games in general and Tetris specifically.

At the beginning of each of the two main sessions, surveys captured self-reported state of mind, caffeine/food intake, and exercise. Surveys that bookend each condition included self-reported emotional states (the self-assessment manikin), stress, alertness, and reaction time speeds. Surveys at the end of each condition included the short FSS-2, the FQ applied to the task, the BIT, questions about time perception (including 'how long do you think it's been? What time do you think it is? How certain are you?'), and new questions about flow introduced in the last chapter (including 'did you experience flow? If so, how long did it take you to get into flow? What percentage of the time were you in flow? Once or many times? Please draw your focus over time').

After the final session, overall interview style questions about the experience were included, asking the user to provide their own insight about home vs. lab environments, flow experience, and the impact of the wearables and study on their state of mind.

6.2.5 Physiological Feature Extraction

Proper feature extraction is essential for fair cross-context comparison. We focus on intra-individual analysis, thus we don't introduce pre-processing steps intended to remove individual differences (z-scores, etc).

6.2.6 Blink Processing

Blink data is processed using the algorithm described in the earlier Captivates work, which bandpass filters the signal to capture frequency content relevant to blinking, takes the

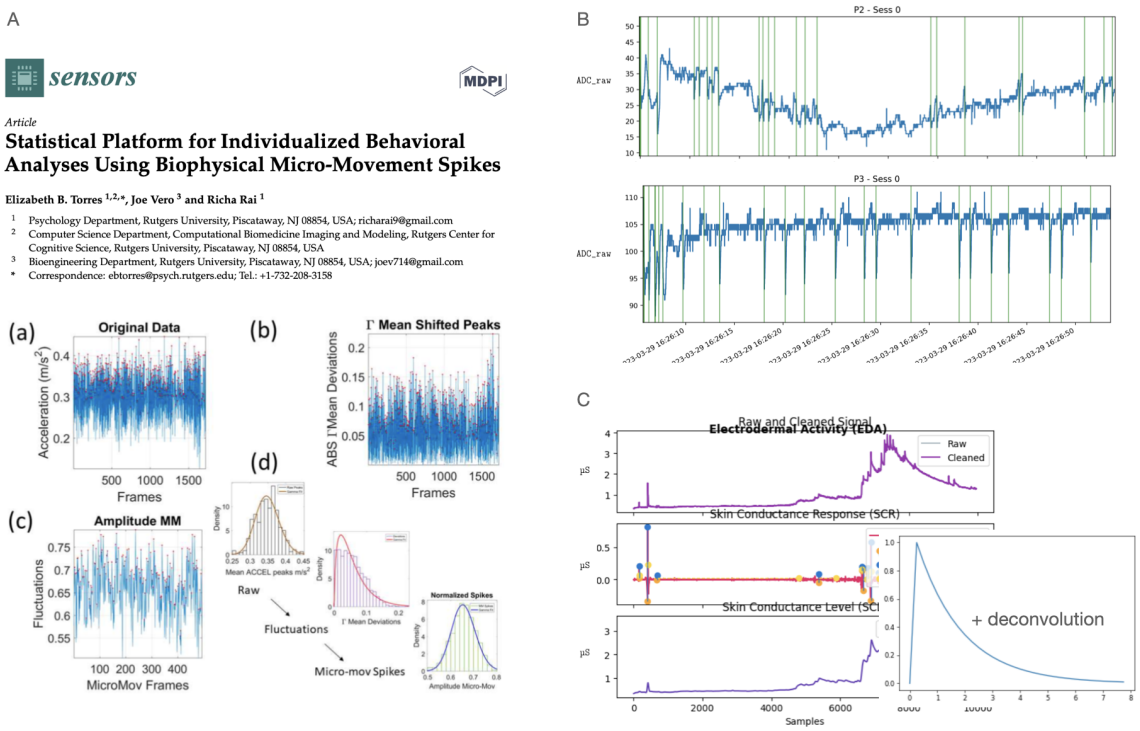


Figure 6-3: Examples of feature processing: (A) is a technique for extracting micromovements from [434], (B) shows blink identification for an ideal case (lower) and a more challenging case (upper); (C) shows traditional SCL/SCR extraction and the canonical SCR response model used to deconvolve the raw signal to approximate iSCR.

derivative, finds peaks, and matches them to the duration we expect from standard blinks. In the original paper, we were able to identify about 80% of blinks correctly this way; there is, however, interpersonal variability in the quality of the results based on face geometry, ambient sunlight, and noise from head motion.

Fig 6-3 shows an example of extracted blinks from an ideal participant and a non-ideal one; blinks are easy to separate in the first case, and more ambiguous in the second. In addition to blink intervals and counts, this script also extracts blink velocity, which has been shown to correlate with arousal in certain contexts [226]. We thus use the number of blinks over an interval (NUMBLINK) and the blink velocities (BLINK_VEL) as our two main features extracted from the raw blink signal.

6.2.7 Electrodermal Activity

EDA is typically divided into tonic (Skin Conductance Level- SCL) and phasic (Skin Conductance Response- SCR) components as a first step before any comparisons. We use the Neurokit 2 library to separate the E4's EDA measurement into these two components as described in [63]. Absolute differences in SCR contain useful information [145, 336]; we thus keep the tonic component in absolute units for comparison.

The phasic component of EDA– SCRs– exhibit a characteristic shape in response to stimuli [36]. In our analysis, we use a standard peak-finding algorithm to identify SCRs and extract the number of presumed SCRs over an interval (SCR_NUMPEAKS) and the distribution of their amplitudes (SCR_AMP). To generate a more general estimate of underlying autonomic activity, we also generate an integrated SCR (iSCR) by deconvolving a prototype SCR response function (from [36]) with the SCR component of the EDA signal in the frequency domain as discussed in [199]; the result is a time-series estimate of underlying autonomic activity.

6.2.8 Cardiac Features

The E4 provides an estimate of heart rate (HR) derived from the wrist PPG sensor, which we use as a feature directly. Additionally, we feed the raw PPG signal into the peak-finding algorithm described in [134] to estimate inter-beat intervals. This signal is then transformed to RMSSD (RMS of successive differences), a technique which has been used in prior flow research to show successful flow state discrimination [237], and which is validated in the literature as a strong summary statistic for heart-rate variability on timescales as low as 30-60s [388].

6.2.9 Temperature

In line with prior work on the psychophysiological predictive work on facial temperature [16], we take the difference between nose and temple temperature readings as our primary feature (NOSE_DIFF). We also include the raw wrist skin temperature readings as a feature (SKIN_TEMP); in line with prior work, the average skin temperature values may provide similar insight to SCL over short duration (30 second) time slices [372]

6.2.10 Movement

Our dataset includes raw 3-axis accelerometer readings (linear acceleration in Gs) at the head and wrist, and raw 3-axis gyroscope readings (angular acceleration in rad/s) at the head. Our goal with movement features is to identify (1) stillness and (2) repetitive, fluid motion; both could be signs of deep focus. We divide our signal into three main feature types for this task:

- The raw magnitude of the acceleration and rotation vectors (HEAD_ACC, HEAD_GYRO, WRIST_ACC). For the gyroscope, this is simply the Euclidean norm; for the accelerometer values, this is the Euclidean Norm Minus One (ENMO), a common technique to transform 3 axis accelerometer data into an estimate of motion [439].

- 'Micro-movements' as identified in [434]; this technique involves fitting a Gamma distribution to the accelerometer and gyroscope signals, mean-shifting them, and then looking at the distribution of secondary peaks (their absolute difference from the average– which itself follows a gamma distribution). (HEAD_ACC_MICRO, HEAD_GYRO_MICRO, WRIST_ACC_MICRO)

- An experimental 'self-similarity' feature. A metric for repetitive motion is not easy to find in the literature. As a first pass, we split the raw 3-axis values segmented into 1-minute intervals and create two feature vectors– one is the kurtosis of the cross-correlation of this time slice with each of the other timeslices (so each entry in the vector is a single value representing the 'peakiness' of the cross-correlation of the current timeslice with one other timeslice, repeated for all timeslices in a session), and (2) an autocorrelation of the motion signal across the 3 axes.

6.3 Hypotheses

The structure of this experiment is designed with two main goals: to evaluate whether current experimental paradigms are sufficient to model naturalistic flow states, and to serve as an exploratory starting point for probabilistic, bio-behavioral theories of flow. With respect to current methodology, we hypothesize two main dimensions in which current methodology might be lacking: we hypothesize (1) the time-resolution of current techniques is insufficient to identify a physiological signature of flow states; we also hypothesize (2) that physiological results do not generalize across environments and tasks at the individual level.

6.4 Results and Discussion

6.4.1 H1: The Influence of Time

In all previous work on flow state physiology, questions are structured such that they apply to the entire task interval. In our work, we introduce a second dimension– moving to time and depth estimates: (1) How long did it take you to get into a flow? (2) What fraction of the time were you in flow? (3) How deep was your deepest moment of flow? (4) Did you experience it in one continuous interval or several small intervals? (5) Please attempt to draw and describe your depth of focus over time. This addition gives us important insight that was previously obscured when asking participants to combine these dimensions themselves in the form of an overall judgement.

As we see in Figure 6-4, while a significant minority (35% in lab, 6-4 top-right) report relatively fast flow induction ($\leq$ 3 minutes), the vast majority are longer than 5 minutes (65%), with the majority (56%) reporting between 5 and 15 minutes. Moreover, reports of the fraction of time spent in flow while playing Tetris is $\leq$ 50% in the vast majority of cases (74% both in lab and at home, 6-4 bottom-right), with the majority of reports suggesting $\leq$ 30% of the time in flow (55% over all conditions, 58% for Tetris 6-4 bottom-left). These self-reported transition times are similar for Tetris and for self-selected, practiced activity. When asked to indicate the dynamics of their flow states over time, participants communicate a diversity of temporal dynamics. Drawings of their mental states are surprising and diverse (Figure 6-5) and their self reports also indicate cycles, deep moments, and interruptions of flow (Table 6.4.2).

The results point to a complex, fragile, difficult to elicit, and time-varying experience of deep states of focus.

Implications for Short Duration Flow Tasks

Reports that flow states for Tetris take multiple minutes to achieve (after a short introductory play session and survey) undermine the idea that flow states are quick to elicit and

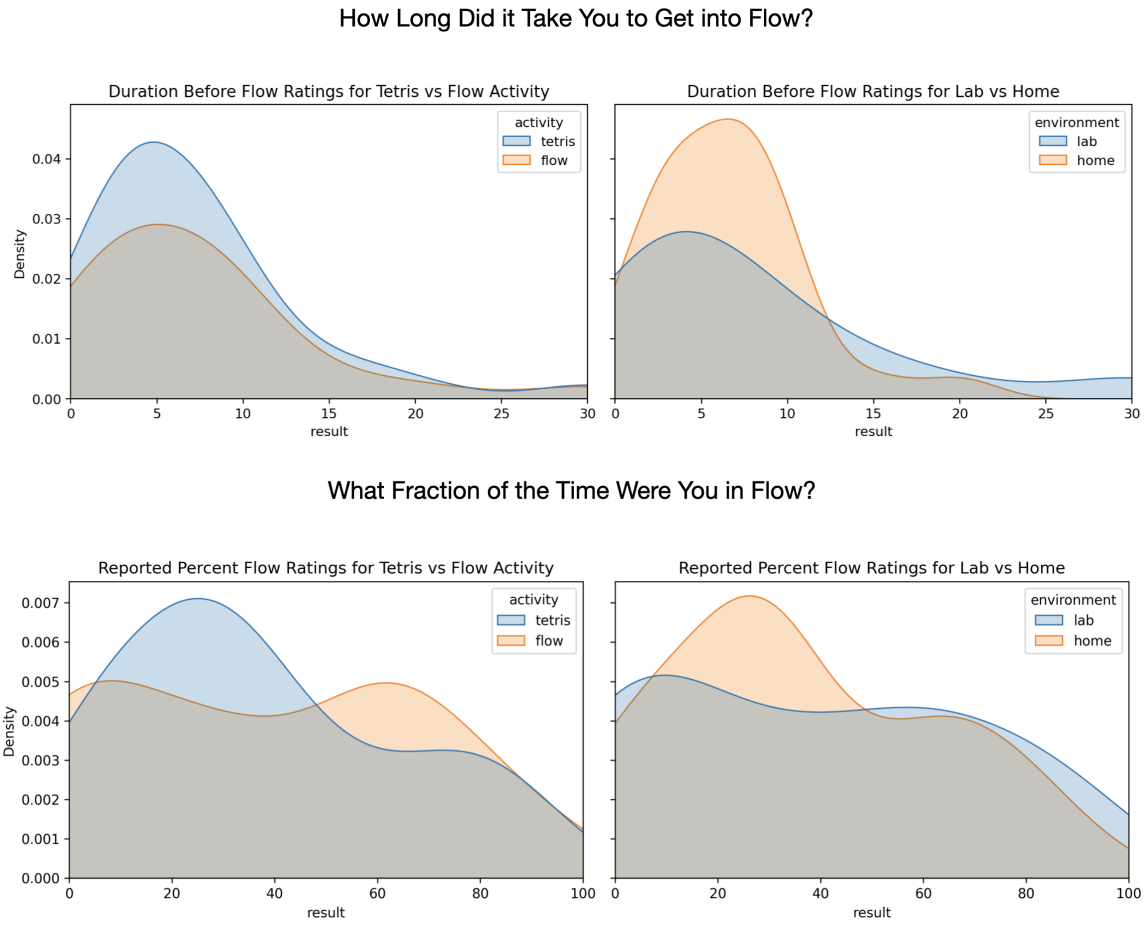


Figure 6-4: Self-reported duration to get into flow, and fraction of time spent in flow, by task and environment.

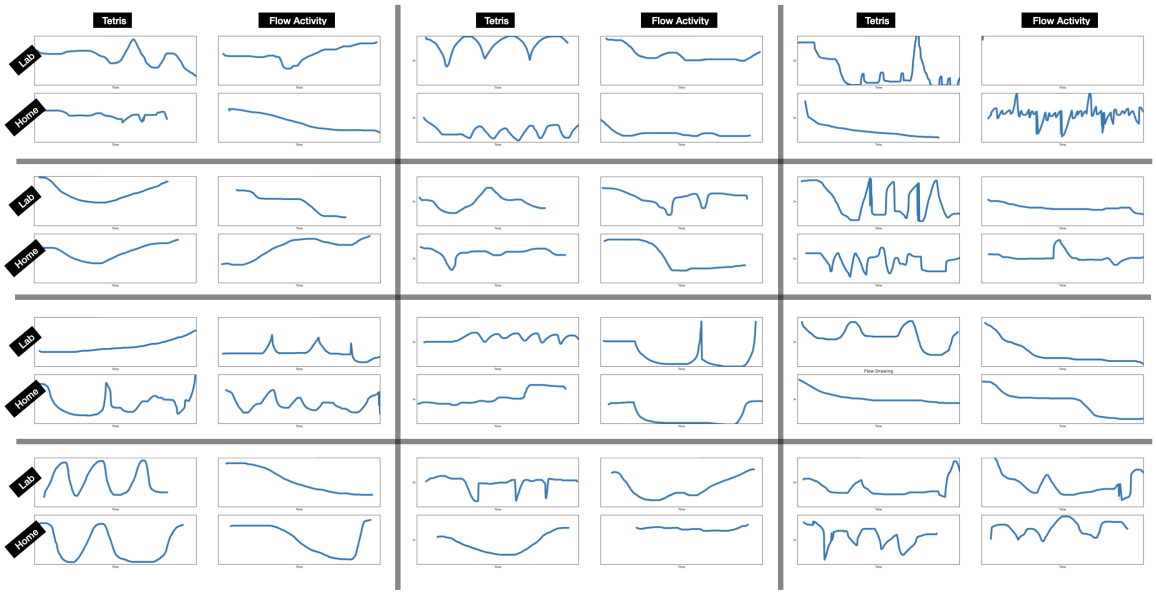


Figure 6-5: A subset of flow drawings for several participants across the four conditions (lab/home, Tetris/self-selected flow activity). Lower Y-axis ratings indicate deeper states of flow.

Quotes that Reference Variations over Time

I was very focused initially but then I got a bit distracted from my task and found it hard to get back.
There were sporadic moments of being very focused and I wasn't paying attention to the glasses at all. Then I'd remember and go through a couple of minutes of more actively paying attention to the lights. This cycle repeated several times.
Generally it was fairly good! I forgot about the experiment multiple times.
Near the beginning I became lost in the activity. Only infrequently did I snap back when I was switching between tasks and remembered to look at the light.
There were a few moments that were similar where I was busy doing a task I knew how to execute. When switching between tasks that's when I snapped out of it.
Distracted a little at the very beginning when doing reading and more focused when doing CAD design after reading.
Yes at the end of the second round I was in deep focus and cannot feel the outer world.
I pretty much got into flow and stayed there...
Several times I was aware I was distracted but also could focus a little.
I was really lost in crocheting for the first part. After that, I started struggling.
One moment where every piece I was laying down was fitting in perfectly and another right after I got a 4-line clear.
Honestly I just remember being highly focused and then making a dumb mistake and getting knocked out of it.
One of my last games — was focused on filling the rows didn't notice other things around me.

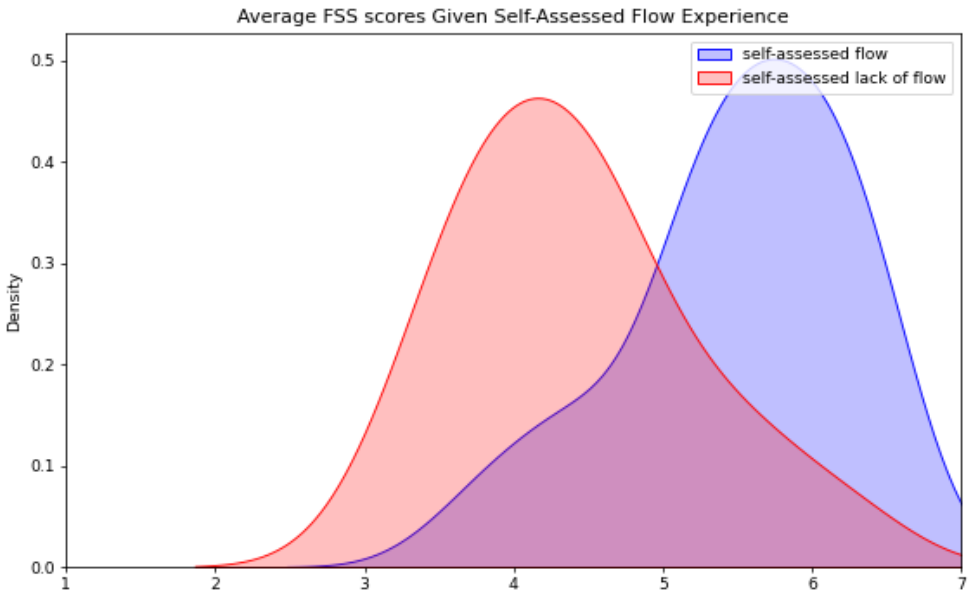


Figure 6-6: Kernel Density Estimate (base-rates removed) for FSS-2 results compared against self-reported flow experience.

easy to sustain. As stated, around 65% of this population report it took them more than five minutes to access flow– roughly the same duration as the typical lab flow task length. These reports remain consistent for the home Tetris session (62.5% report 5 minutes or longer before flow), in which the user is (1) unobserved in a comfortable environment, (2) has an hour of previous experience with experimental paradigm, and (3) has 25 minutes of previous experience with this version of Tetris and its controls. These reports also roughly match the distributions for naturalistic, well-practiced tasks that start half an hour into the session.

Moreover, the typical method to corroborate flow induction for lab-based Tetris is the FSS-2 value, which alone is not a strong signal of the underlying flow state. When we compare self-assessed flow induction with FSS-2 results, two overlapping skewed distributions emerge for which an FSS-2 score of 5.0/7.0 (or 3.6/5.0) gives roughly equal odds of representing a true flow state or not (Figure 6-6). Scores in this range should not be considered de facto evidence of flow induction.

These two points together– (1) that our data suggest flow takes a long time to achieve upon task initiation, and (2) the standard FSS-2 is a weak check for underlying flow states– imply more, careful work should be done to understand the value of short duration flow tasks. Strong FSS-2 scores of short duration Tetris may imply that the FSS-2 is failing to capture the core state as experienced by participants[1]; or it may imply the paradoxical conclusion that introducing consistent, relatively long survey interruptions to gameplay-as-designed *improves* flow induction. Future short-duration Tetris work should include survey improvements we introduce to interrogate the time dimensions of flow and its self-assessed experience.

Short vs Long Durations

We cannot definitively rule out the possibility that short bursts of gameplay punctuated by survey breaks may, in fact, quickly elicit a stressful state of deep attention.[2] However, if short duration task structures are found to elicit flow on these timescales, it is *because* they are psychologically (and likely physiologically) distinct from the typical flow dynamics reported herein (a more gradual process of focusing). We should thus move towards longer duration experiences– they more closely mimic real-world flow task structure and provide a suitable chance to get lost in an activity.

Regardless of task duration, we simply can't ignore time dynamics. If we assume a roughly fixed, minimum transition period to get 'in-the-zone', shorter and shorter task durations will be corrupted by a larger percentage of non-flow physiology. Longer-duration, naturalistic flow tasks force us to develop a strategy to identify the typically $\leq 50\%$ of the physiological data that actually corresponds to deep focus.

[1]This interpretation is in line with other criticism of the FSS-2, and could explain other unusual features of the relationship between Tetris and FSS-2 scores; i.e. trivially easy gameplay (which should be very boring on longer timescales) elicits equally strong FSS-2 results across multiple Tetris studies. It could also be the case that FSS-2 results suffer from context-related differences in their assessment, further undermining the implicit link between survey results and real phenomenology.

[2]In addition to our data showing that this is unlikely, this hypothesis is further complicated by the strong FSS-2 scores on trivially easy 3 minutes versions of the task.

6.4.2 H2: Generalization Across Contexts

Tetris and Lab-based Flow Induction

Location/Activity	Flow Reported?	Percentage
Home	31/40	77.50%
Lab	31/40	77.50%
Playing Tetris	30/40	75.00%
Doing Selected Flow Activity	32/40	80.00%

Table 6.1: Flow states for different locations and activities

How well does Tetris in the lab represent more natural experiences of flow activity? We start by evaluating whether reported flow experiences from Tetris or in the lab are different. Table 6.1 shows that, in our study, Tetris and Lab-based activities are (roughly) equally successful at inducing flow based on self-report and FSS-2 results. Moreover, our bio-behavioral indicators– data showing taking optional extra time, delay to notice the LED, time perception estimations– which seem to have worked as intended (Table 6.4.2) have similar trends for across all conditions (Figure 6-8). This data is preliminary, but there are no obvious 'smoking-gun' differences.

There is no evidence in the population statistics here that flow states as studied (Tetris in lab) are different from natural activities and environments based on survey data (the standard way we measure flow). While there are some minor differences that should be explored at higher power, none warrant a distinction between cases with this relatively small dataset– lab environments and Tetris activities look very similar to the real-world in survey results.

Psychological Similarity?

The overall results may obscure the picture at the individual level, though. Tetris seems to work well for flow induction generally, but many advanced players with prior experience mentioned differences in the interface from their expectations getting in the way of their flow. (These participants may otherwise be our best population to study, as we expect experts

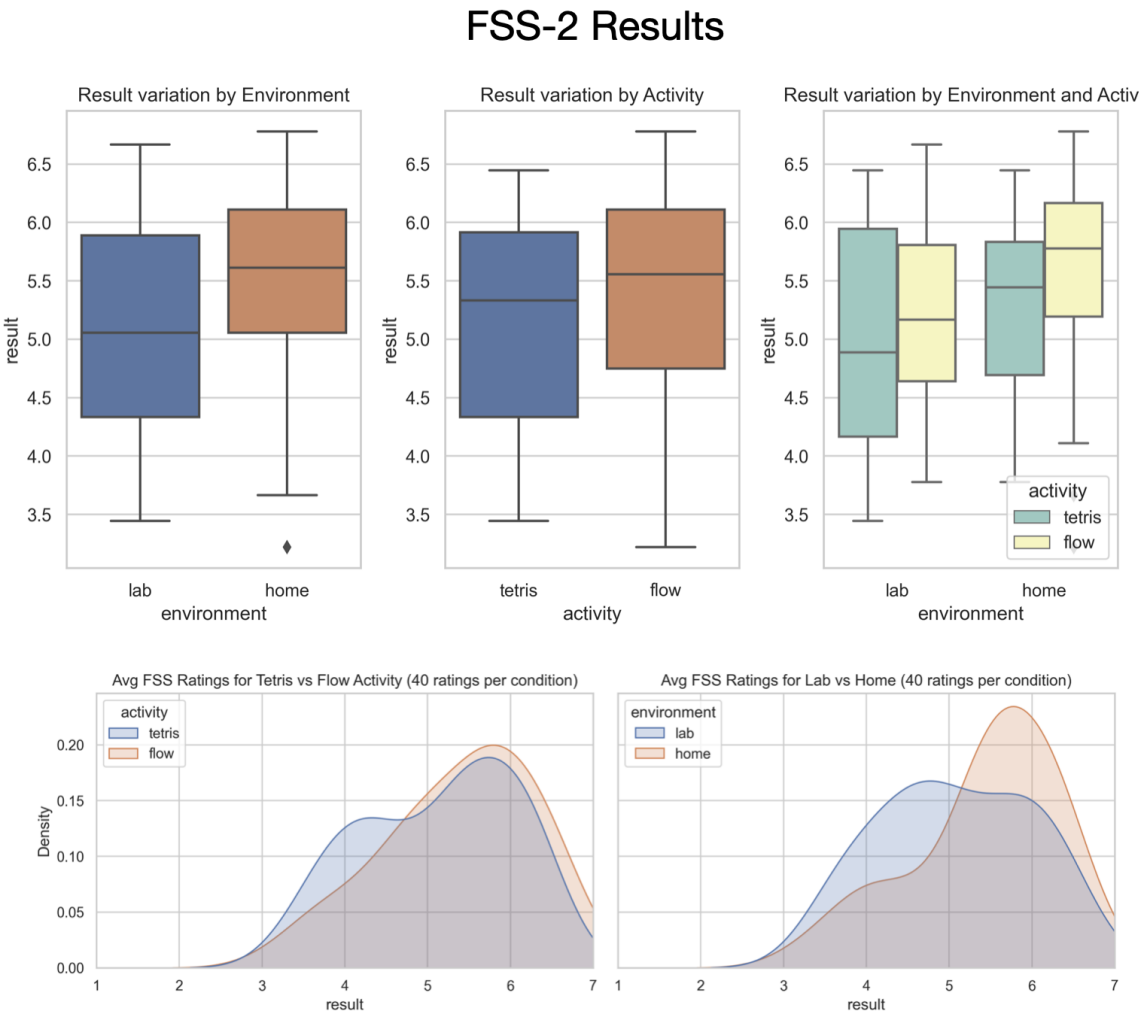


Figure 6-7: FSS-2 results by environment and activity.

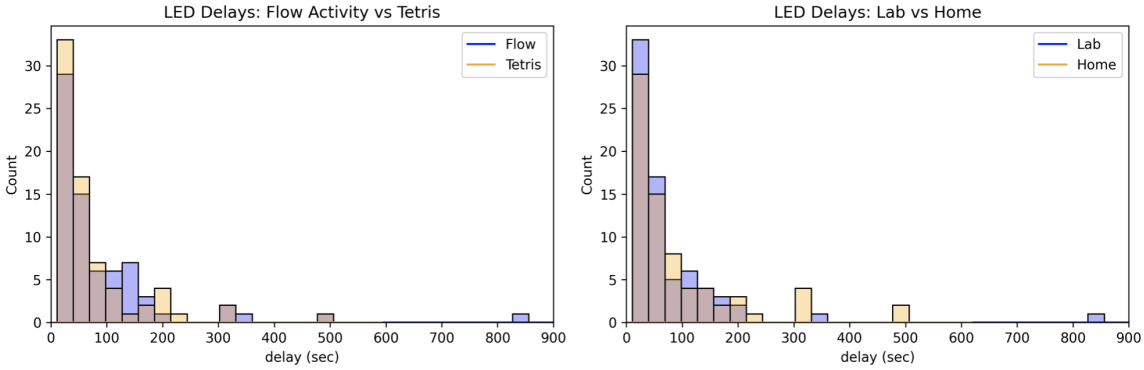


Figure 6-8: LED delays until noticing across conditions look very similar.

Quotes that Reference the LED

...during the second Tetris I was in deep focus and can't see the LED.
I started getting a hang of the strategy and I didn't realize the light had turned until I lost the game.
I also didn't notice the blue light and do not know how long it was there before I noticed.
I got deep into the game realized my light was on when I died violently.
Before the first LED transition I was pretty deep Tetris and then I eventually snapped out of it and noticed the LED was blue.
"There is nothing else in the world except me and 'Elia Maiden of Water'." "Oh sh*t did my light change during that song-PHEW" "Hmm that was pretty rough-OH THE LIGHT."
I was surprised by the first blue lights appearing since I didn't notice them change at all.
I think I was mostly in focus and there were some times where I saw the blue and snapped out.
During lulls in game play I thought to check the lights.
At the beginning I kept on checking the light but as I started to do better I became more focused.

to achieve flow more readily). Moreoever, it seems lab and home contexts were viewed by many participants as quite different and influential on their state of focus, but that they affect people in equal and opposite ways. General commentary suggest the lab is more conducive to work-like flow activities (coding, design) and the home is more conducive to creative, enjoyable activities (reading, crocheting, art, music) when the home environment is well-kept.

Physiology Across Contexts

Finally, we compare the physiological data across contexts to see if Tetris and lab studies capture representative flow physiology. We can see from a sampling of the individual raw data in Figures 6-9 and distributions of the data across conditions (Figure 6-10) that physiological data appears distinct across the four conditions. For each individual with a pair of activities that both successfully induced flow, we compare the distributions of the data using a Kolmogorov-Smirnov Test– a non-parametric test that estimates how likely two sets of data are drawn from the same underlying distribution. When we run this test

Quotes that Reference Variations over Time

Home is Better for Focus
I think it is easier/faster for me to get into a deep focus state at home in the position and environment that I always do this activity in.
Home- much more relaxed, easy to get into it. Lab- new environment, but less easily distracted.
I definitely focus easier at home but that's mostly because I have my setup (reclined chair, monitor close to face, AC, etc).
At home is better. Late at night is better, more peaceful and I can lose myself in what I am doing.
I think I was able to focus more strongly on the experiment at home than in the lab.
Periodically I find myself guessing the time or checking the LED ... I think that is because of the lab setting otherwise I wouldn't find breaks at all.
Either
I don't know if setting matters for my flow activity.
At home I am more comfortable but I have a lot more distractions and things to fidget with.
At home there is more comfort. At a lab there are many new things.
Lab is generally good if not too many loud noises. Home is great if I have a task I enjoy doing...
Lab is Better for Focus
Home is much more cozy and for me, it's harder to get into and stay focus at home.
I find more pressure to be focused in a new and different setting, especially an isolated one.
I was better focused in lab.
I am generally more focused in lab for work tasks, but at home I have about equal levels of focus for creative tasks.

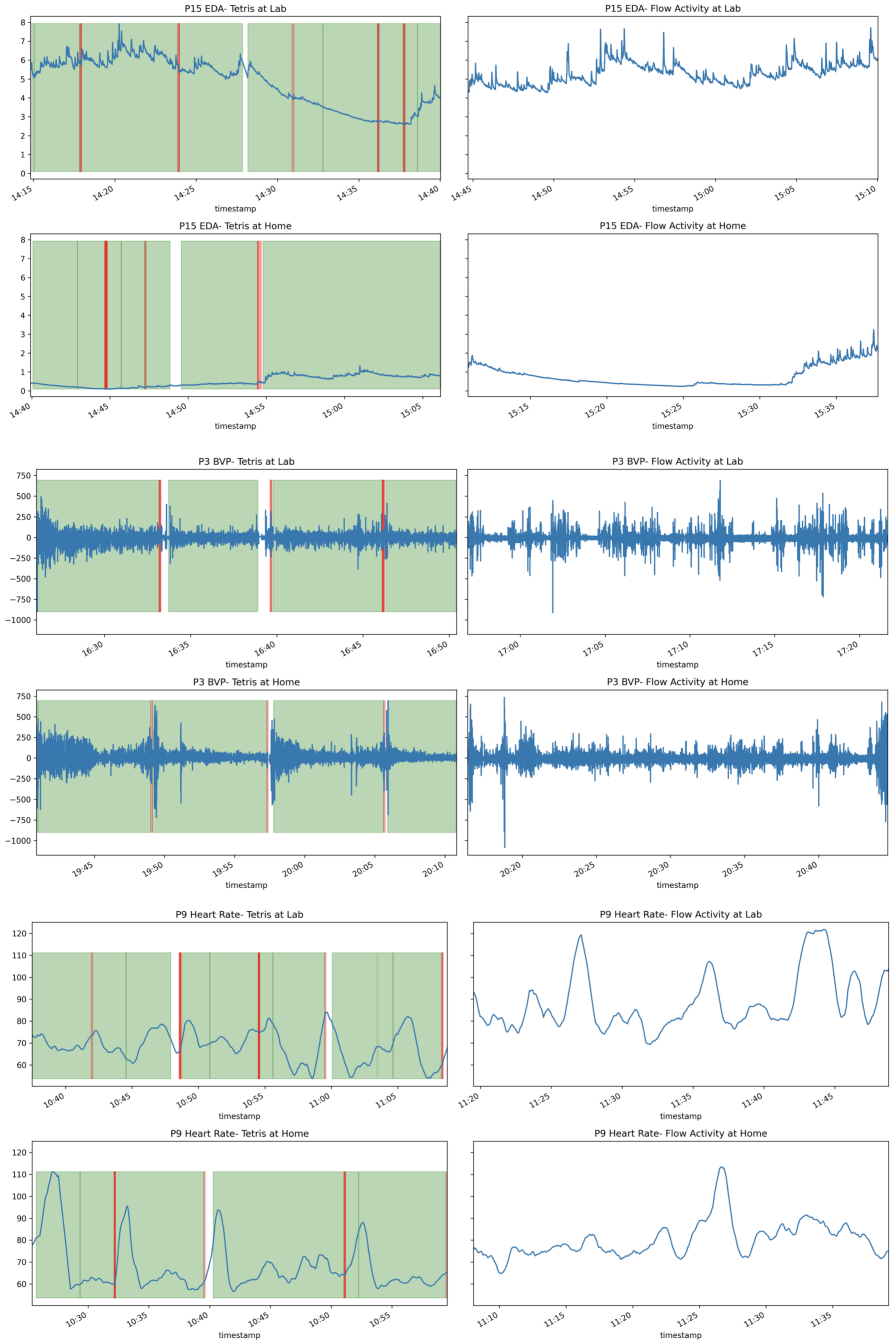


Figure 6-9: Example raw data from across the four conditions for random participants/data sources; for the left column (Tetris gameplay) green means a script designed to process the video of the Tetris footage identified the user is playing the game; red indicates a game-over event. Notice (1) the heterogeneity within an interval across time, and (2) the heterogenity across conditions. If we imagine condensing these data down to a single summary statistic– throwing away the time-relevant information– we might expect that the insights from any subset of the data will fail to generalize to any other subset when attempting to make a more specific prediction than the basic physiological variance we naturally experience.

on each individual (Table 6.2 compares the physiology of each individual, when in flow, between Tetris and their Self-selected Flow Activity; Table 6.3 is the same for Lab vs. Home environments), we see that the physiology across conditions almost never matches.

Kolmogorov-Smirnov Test - Tetris vs Flow Activity (when self-reported Flow = 1)				
Title	Tests with p<0.001	1st Quartile p-value	Median p-value	3rd Quartile p-value
HR	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
IBI	15 / 17 (88.24%)	<1e-5	<1e-5	<1e-5
RMSSD_1M	8 / 17 (47.06%)	0.00004	0.00151	0.04844
SKIN_TEMP	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
SCL	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
SCR_NUMPEAKS	9 / 17 (52.94%)	<1e-5	0.00052	0.12460
SCR_AMP	9 / 17 (52.94%)	<1e-5	0.00077	0.05000
iSCR	16 / 17 (94.12%)	<1e-5	<1e-5	<1e-5
WRIST_ACC	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
WRIST_ACC_MICRO	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
NOSE_DIFF	22 / 22 (100.00%)	<1e-5	<1e-5	<1e-5
HEAD_ACC	22 / 22 (100.00%)	<1e-5	<1e-5	<1e-5
HEAD_GYRO	22 / 22 (100.00%)	<1e-5	<1e-5	<1e-5
HEAD_ACC_MICRO	22 / 22 (100.00%)	<1e-5	<1e-5	<1e-5
HEAD_GYRO_MICRO	22 / 22 (100.00%)	<1e-5	<1e-5	<1e-5
NUMBLINK	20 / 22 (90.91%)	<1e-5	<1e-5	<1e-5
BLINK_VEL	16 / 22 (72.73%)	<1e-5	<1e-5	0.00709

Table 6.2: Kolmogorov-Smirnov test for Tetris vs Normal Flow Activity over all physiological indicators. Comparisons are performed where a participant reported that both activities performed in one environment successfully induced flow state. We see no strong indicators that the underlying data appears from a similar distribution, even for a very strict p value cutoff (p<0.001). The indicators that show rely on heavy averaging (RMSSD), have few samples drawn from a small range of possible values (SCR or NUMBLINK, which may have many intervals of zero), etc; still, the evidence is strong that even these seem to be from different distributions for 'normal' p-value thresholds.

Summary

It is unlikely that short-duration Tetris captures the same psychophysiological dynamics as typical long-form flow activities. We see that flow typically takes a long time to achieve and states of attention are dynamic. Our data suggests the typical methods may be too short, measures to corroborate flow induction too imprecise, and physiology too dissimilar to make strong claims. We suggest short-duration lab studies at a minimum should extend

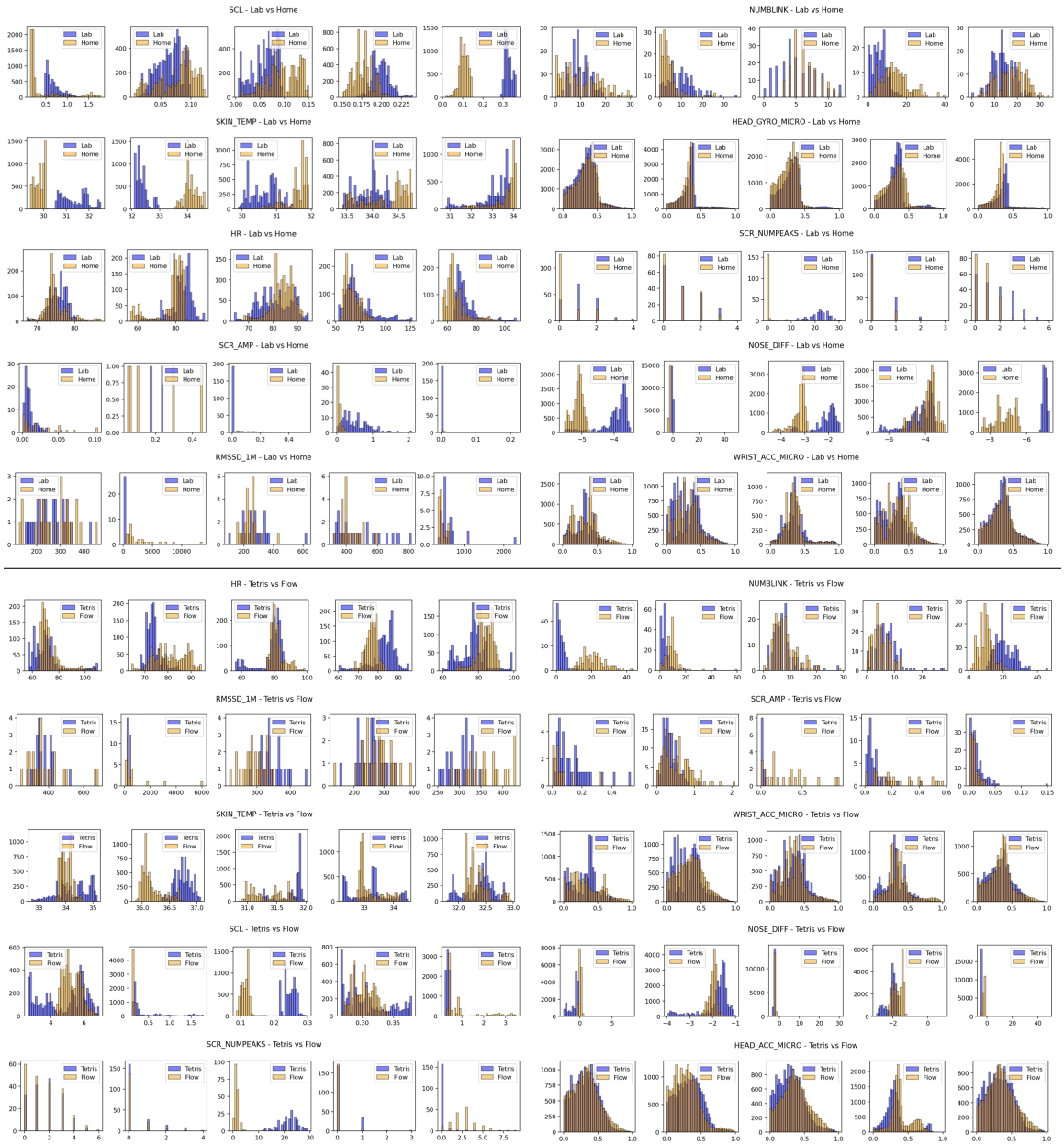


Figure 6-10: Randomly selected distributions of individual data from a few select physiological indicators. These distributions represent matched pairs of tasks in which both conditions were reported to have induced flow; we see they typically look quite distinct. The top compare lab vs home; the bottom compare Tetris vs selected flow activity.

Kolmogorov-Smirnov Test - Lab vs Home (when self-reported Flow = 1)				
Title	Tests with p<0.001	1st Quartile p-value	Median p-value	3rd Quartile p-value
HR	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
IBI	13 / 17 (76.47%)	<1e-5	<1e-5	0.00007
RMSSD_1M	8 / 17 (47.06%)	0.00007	0.06258	0.17684
SKIN_TEMP	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
SCL	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
SCR_NUMPEAKS	11 / 17 (64.71%)	<1e-5	<1e-5	0.04845
SCR_AMP	9 / 17 (52.94%)	<1e-5	0.00030	0.07204
iSCR	16 / 17 (94.12%)	<1e-5	<1e-5	<1e-5
WRIST_ACC	17 / 17 (100.00%)	<1e-5	<1e-5	<1e-5
WRIST_ACC_MICRO	16 / 17 (94.12%)	<1e-5	<1e-5	<1e-5
NOSE_DIFF	23 / 23 (100.00%)	<1e-5	<1e-5	<1e-5
HEAD_ACC	23 / 23 (100.00%)	<1e-5	<1e-5	<1e-5
HEAD_GYRO	23 / 23 (100.00%)	<1e-5	<1e-5	<1e-5
HEAD_ACC_MICRO	22 / 23 (95.65%)	<1e-5	<1e-5	<1e-5
HEAD_GYRO_MICRO	22 / 23 (95.65%)	<1e-5	<1e-5	<1e-5
NUMBLINK	21 / 23 (91.30%)	<1e-5	<1e-5	<1e-5
BLINK_VEL	23 / 23 (100.00%)	<1e-5	<1e-5	<1e-5

Table 6.3: Kolmogorov-Smirnov test for Lab vs Home environments over all physiological indicators. Comparisons are performed where a participant reported that the activity induced Flow state across both of these two contexts. We see no strong indicators that the underlying data appears from a similar distribution, even for a very strict p value cutoff (p<0.001). The indicators that show rely on heavy averaging (RMSSD), have few samples drawn from a small range of possible values (SCR or NUMBLINK, which may have many intervals of zero), etc; still, the evidence is strong that even these seem to be from different distributions for 'normal' p-value thresholds.

their survey techniques to probe the influence of time; however, a better approach favors uninterrupted, naturalistic task structures as we have captured here.

Within this new context, Tetris and lab environments do appear to elicit flow roughly as much as home/naturalistic activities based on survey data. However, individuals report psychological differences which may be obscured at the population level; additionally, basic tests to compare the statistics of physiology over the intervals of each task demonstrate distinctions at the individual level across all indicators.

These physiological comparisons, however, suffer from the same issues that we are using them to illustrate– when we ignore temporal dynamics, the results cannot generalize. While the physiology we capture is not trivially similar across conditions, it may be possible to identify temporal patterns or subsets of physiology that *will* generalize across conditions using the tools of probabilistic inference. Success demands that we preserve and consider uncertainty and temporal variation as first class features of our model.

6.5 Contribution: Open-Source Dataset

The biggest contribution of this effort is an open-source, one-of-a-kind psychophysiological dataset to explore the physiological dynamics of flow. Very few studies have collected this depth of data in flow-like scenarios, especially across locations, days, and activities. It's collection rests on a very robust, mobile testing platform that can be self-administered and run consistently over an hour and a half with no intervention. The two and a half hours of data per person should serve as a useful starting point for idiographic, bio-behavioral model design; it also represents the most diverse set of flow task data collected in this way.

The data from this study is organized into Participant folders and then by Session ('P1' $->$ 'P1sess1', 'P1sess2', 'P1sess3'). Each folder includes a video file of Tetris (timestamped at the end of the session); a folder of raw E4 data with descriptions from Empatica, and a set of CSV files which include all user interactions, glasses and watch data, and survey data timestamps. These sessions are timestamped and labeled in order (LAB_TETRIS,

TETRIS_CONTD, LAB_FLOW, FLOW_CONTD). An overall excel sheet is provided with introductory and final survey data; data processing scripts are provided to extract all of the data in an easy to use format in the 'data_processing' folder.

6.6 Limitations and Future Work

6.6.1 FSS-2 and 2-axes of Flow

We advance a 2-axis survey methodology for interrogating flow (time-depth) vs the current practice (depth). Future work includes a thorough interrogation of the relationship between single flow estimates and the time-varying nature of flow we capture, giving insight into the temporal dynamics and mental process behind these judgements.

6.6.2 Data and Feature Engineering

Besides the limitations outlined with the raw data itself in previous chapters, feature engineering is also an evolving science. The quality of these features rely on post-hoc signal processing of the raw signals in ways that cannot perfectly separate the signal from the noise.[3]

We are currently revising all of our feature engineering strategies; already under development are temperature models that try to combine the four different temperatures (ambient, skin, nose, temple) in a way that captures underlying sympathetic activity based on a model of blood diffusion; similarly, we are working on a probabilistic blink model that numerically optimizes a model of ambient lighting based on head Kalman-filtered head motion data. EDA processing could use motion artifact analysis and a Bayesian model of SCRs for iSCR deconvolution; movement features that capture fluid, repeated motions (or identify atypical

[3]Even for commercial devices with large teams and huge datasets this work is constantly evolving– Harvard's Onnela showed that, for the Apple Watch, exporting the same two years of heart rate data six months apart gave completely different heart rate data– a correlation coefficient of 0.67 (2021) [50]. For Apple, the algorithms are evolving so rapidly that the results over the same raw data look entirely different after just a few months of development. Apple does not provide access to the underlying raw data.

patterns of motion given a baseline corpus) are also necessary. HR data is provided based on a prioprietary Empatica algorithm; but we can also derive it from the raw PPG signal and from subtle head motion. We will build an ensemble model with confidence for HR and IBI data as well.

The features above are not population or inter-day normalized; this is an additional step that requires theory building, such that we can have a prior over all of the relevant data. In general, the goal is to move toward data generating feature models based on sympathetic and parasympathetic state where applicable, and a small subset of parameters otherwise.

Modeling

Armed with this data, we can begin modeling flow states. There are yet unanswered questions about optimal time-resolution for modeling work we might approach with predictive coding; we could also analyze basic patterns in the data at lower time-resolutions to make stronger claims about the likelihood of generalizability and overall physiological variation in the data. The most exciting work, however, will be probabilistic modeling development using this data as a way to evaluate theories. There are many ways to translate the data we've collected into priors of latent flow states, and many ways to evaluate how well these models fit the data or predict unseen data conditioned on other information. We believe this dataset will be most useful as an initial, exploratory test-bed for bio-behavioral flow ontologies.

Chapter 7

Interventions to Promote Flow

This chapter introduces a collection of interventions designed to influence flow state. While much of the work on *implicit* interventions have failed to replicate ('Primeworld'– interventions so small that they are unnoticed), automatic interventions are still among the most transformative (i.e. interventions that become quickly automatic *because they are so integral to our lives*). This distinction separates an intervention like introducing a phone into someone's life from nudging them; both require limited conscious consideration, but one is incredibly powerful and one is almost certainly marginal. I attempt to design interfaces that are automatic, but not implicit.

7.1 Huxley

7.1.1 Motivation

Multi-function devices like the smartphone suffer from function overloading– their interface evokes and imposes a state of constant decision-making on the user. Especially when a task requires effort, the device itself invites regular re-evaluation of the task at hand. This undermines focus; it's also possible that too many choices [384] and/or the ability to renege

Figure 7-1: The Huxley Smart Book, showing updating e-ink displays on the cover, spine, and inside. A USB port on top is available for charging, one button is available for turning pages. Huxley only updates when the user is away and the current book has been completed or ignored. Its selection is based on the measured interests and affective state of the user at work.

on our selections [174] lead to a paradoxical dissatisfaction, despite our instincts to the contrary.

While there are many open questions about the nature of the relationships between an artefact's affordances (what actions it physically invites– touching, grasping, throwing), its symbolic information (what the buttons, menus, labels and screens indicate you should do with it; including cultural assumptions you might bring to a new form factor), its conceptual model (the mental representation of the device and its state that you create in your mind) [315], and its cognitive effects (how it impacts the quality of your experience), it seems reasonable that we should avoid designs whose symbolic information has been rigorously paired with fractured attention over the last decade if our goal is to promote deep engagement. More research needs to be done to answer important questions:

- How tightly coupled are conceptual models and cognitive load? Is the relationship binary or continuous? (If your phone decreases your ability to focus, will a similar

phone? Any phone? Anything with a screen?)

- Have we learned a direct association between symbolic information and cognitive state? Do new conceptual models alter it? (Does a 'phone form factor' induce high cognitive load even without phone functionality? Is this reversible?)

- What aspects of these devices are most damaging to focus? (How much does the type and nature of a device's functionality versus addictive design choices like variable reward structures, badges, and infinite scroll contribute to cognitive load? What is the cost of convergence– platforms that run many diverse applications? How does that interact with each application's ability to notify you?)

In the spirit of 'Calm Technology' [453], Huxley is designed against the backdrop of these provocations, both as a tool for inquiry and as an example of attention-aware design. Other theories of design offer notions of psychological pairing between an object's sensory features and a user's emotional response to it [89]; we submit that this idea logically extends to the user's resting attentional demand when using the object as well. We believe the psychology supports a move toward essentialism [291] in design—environments and artefacts whose choice architecture promotes fewer, longer, deeply engaged experiences in line with a user's priorities for themselves.

There is a rich confluence of cause-and-effect to disentangle given the current state of screen-based, multifunction artefacts. We hope Huxley will spark fruitful conversation about the cultural and aesthetic implications of the Attention Economy's addictive design tradition and move us towards an empirical design language based on behavioural economics and cognitive models of deep engagement.

7.1.2 The System

Huxley (Fig. 7-1) is a book with three e-ink displays, a single button, and a charging USB port. It behaves like a book— its spine and cover display the title and cover of the book

contained within, and once opened the internal e-ink page can be turned with a button press.

There is no way for the user to change the book contained within. Instead, Huxley broadcasts an API over the local network— once loaded with PDF/EPUB files, it can be programmatically directed to become a random selection (favoring new books) or become a relevant book based on provided queries (using Doc2Vec embeddings as a similarity metric). It's designed to work invisibly with a custom chrome extension that selects a relevant book based off a user's browsing history once the current one is either complete or neglected (in this implementation, three ignored days triggers a new book).

To perform research in comparison to an iPad, a standard e-reader, and an old fashioned book, Huxley was created with a few motivating principles in mind. It minimizes the user's cognitive model– there are no menus, no battery or wifi indicators, no state to track, and no ability to summon any book in the universe or leave the current option behind. Moreover, Huxley only switches from one book to another when the user is away. The goal is to make Huxley feel like a permanent object. Despite its screen, it evokes the conceptual model of a book. It also takes physical space; an important and overlooked trait in the dematerialized era. It is still, though, a digital artifact with updating content.

Huxley reintroduces physicality, friction, and simplicity to digital reading devices; it evokes 'book' rather than 'screen'. It is a tool to test many of our most basic assumptions about the influence of UI on focused reading.

7.2 Guitarbot

7.2.1 Motivation

Friction has a big impact on our behavior– in my own life, I've noticed small amounts of friction drastically alter how I use social media and whether or not I put on music when I get home from work. The impact of small barriers is borne out in large data: milliseconds

Figure 7-2: Guitarbot is a guitar stand built on top of a iRobot Create 2, that can rove around the space and find the positions most likely to result in guitar use.

of loading delay can have a dramatic effect on click-through [405], for instance. Introducing friction can provide a moment to reconsider; over time, with consistent small decisions to stay on task, the urge to self-interrupt may diminish.

Huxley introduced friction into the e-reader experience to promote focus in this way; Guitarbot takes the opposite approach, eliminating friction from the decision to practice. Whenever you have a slight urge to play, the guitar should be within reach. In both cases, the goal is to feel as though someone who knows you well adjusts your space for you– picking out a book just for you and putting out your guitar just where you need it.

7.2.2 The System

Guitarbot (Figure 7-2) is a guitarstand that is built on top of a Roomba (the iRobot Create 2) and controlled using ROS running on a Raspberry Pi. The version pictured here is programmed to assume three separate locations in my apartment; by the front door, by the seating area, and against the wall. It moves only in the middle of the day, when no one is home.

As another live-with intervention, the goal of this device is to quantify how location and reachability of the guitar actually alters my engagement with it in a truly randomized way.

7.3 Stagehand

Figure 7-3: Stagehand marries a custom button which attaches to the guitarstrap (left) with an immersive living room system (right). Whenever the guitar is lifted and in use, the entire environment shifts to 'concert mode' (right bottom).

7.3.1 Motivation

After my undergraduate degree, I lived in Dublin for a year, where I'd go see live music every week. The music culture in Ireland is incredible, and I found myself getting chills reliably, nearly every night. At first I thought it was because of the quality of musicianship, but as I analyzed the experience I began to realize that the musician was a small part of the puzzle; more than the performer, the atmosphere was one of shared focus and attention. You could hear a pin drop. The acoustics also contributed– that silence allowed incredible dynamic range alongside just enough reverb to mask small mistakes; the system was both forgiving and enhancing.

Those experiences have stuck with me, and with the dozens of concerts I've been to and audio systems I've evaluated, I'm a firm believer in the power of gestalt ambiance (easily destroyed by one chatty concert goer), trust in the performer (easily destroyed by one obvious mistake), and the aural infrastructure (easily muffled, muddy, compressed, and one-dimensional). When they come together it can be transformative.

7.3.2 The System

To test some of these ideas with my own guitar playing, I created Stagehand– a BLE button armed with an accelerometer that snaps into a standard guitar strap and captures when the guitar is in use. This button controls the larger system shown in Figure 7-4.

In normal mode, the A/C is on, a fake fireplace is roaring, the lights are on and set to a normal color temperature, and subtle, synthetic cricket sounds are played in the space (when my girlfriend first moved in, she didn't notice them for 3 days, until I asked her if she did– from then on, she was very annoyed by them).

When the guitar is lifted, the crickets and air conditioning stop; the lights either turn off or turn to deep purple; the fireplace becomes a purple jellyfish; and a microphone in the room adds an extra layer of dense reverb to the space. The room stays in the mode as long as the guitar is in use.

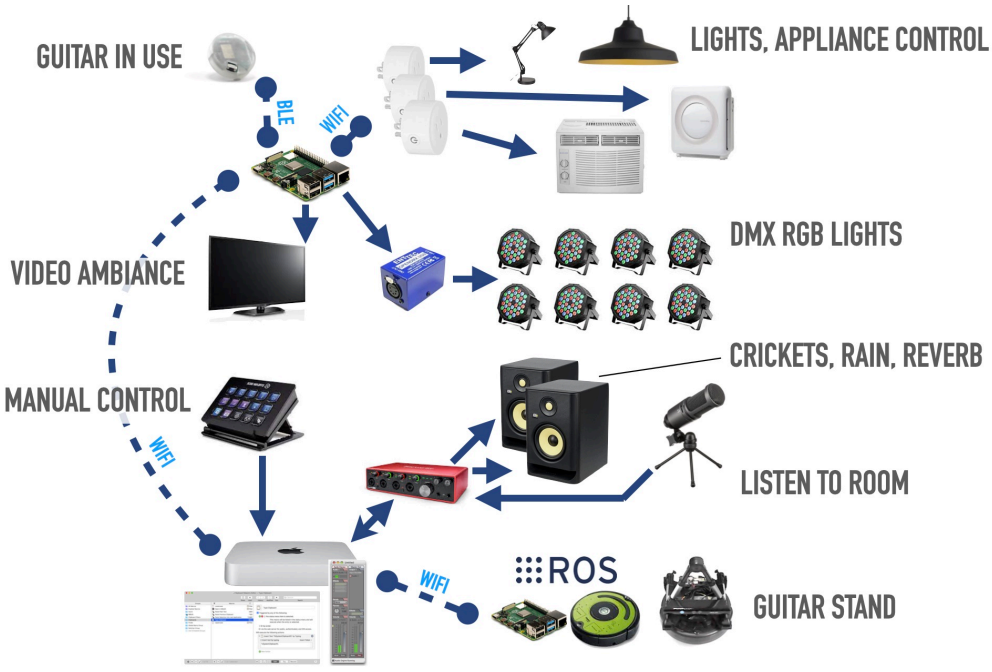


Figure 7-4: The system underlying Stagehand includes many aspects of environmental control.

As a musician, I know the power of a really nice, forgiving reverb on my ability to lose myself in playing. I wasn't expecting how visceral and exciting it felt with the other aspects

of the transition; the immediate stillness when background noises disappear is evocative and intense. From my anecdotal experience, this kind of intervention certainly changes the feeling of practice and I expect it to lead to longer, more frequent, and more rewarding sessions with my guitar.

7.4 Re-designing Asynchronous Communication

Figure 7-5: This email device is an e-ink screen VPNed to a small single board computer (latte panda); designed to evoke a typewriter and to feel different than a typical laptop.

7.4.1 Motivation

Second-Order Agency

In 'Brave New World', Aldous Huxley weaves a dystopian story– a story where every individual is matched in ability to their station in life, perfectly and completely happy, sexually gratified, and unburdened by existential threat. It is a fully optimized society; a hedonic paradise. What makes this society *so clearly* a dystopia, when many of its features are utopian?

Humans have the ability to cultivate tastes, to have preferences about their preferences. Our ability to control the influences on us that shape our preferences– our second-order agency– is fundamental to our freedom. We select friends, mentors, and colleagues; they shape us in turn. We select news curators whose epistemological frame matches our own; their curation has a tremendous effect on our beliefs and ideology.

Controlling this choice is fascism. In some very important sense, this is the most fundamental type of freedom; the freedom to shape and curate the forces in your lives that shape you.

The domain where I see the least practical second-order agency is our digital infrastructure. It's not for lack of trying– everyone has a technique to minimize their addiction or control their screentime– but we are sold digital ecosystems wholesale with little ability to adapt them. We do not have the same control over our digital environments that we have with our physical ones.[1]

One of my motivations with this project is to probe the limits of adversarial interoperability– making it seem to those in my social graph and service providers that I am still a normal part of the technology ecosystem and preserving my ability to function as a member of modern society, while redesigning from first principles the interaction paradigm.

Another of my goals is to make this philosophical point– much of the world of design vectors towards a Brave New World-like notion of utopia; maximizing convenience and hedonic reward rather than maximizing our ability to shape ourselves into the people we wish to become.

The Largest Plausible Intervention

If there is one kind of 'background' intervention that will have a measurable cognitive impact on the focus of a typical Western knowledge worker, I believe it's a redesign of their

[1] I don't want to sound too harsh, because I do believe that there is a robust conversation around these topics, robust market-forces keeping a lot of companies in check, and a clear tradeoff with innovation when legislation starts to become paternalistic and restrictive. This conversation deserves it's own book.

email and smartphone. A third goal with these interventions is to measure the impact of a *huge* UI change in someone's life; as large as I can instrument while keeping it reasonable for someone existing in the modern world. It can then serve as a useful benchmark to set expectations for other interventions related to ubiquitous technology.

These interventions are designed to remove all interrupt-based, asynchronous communication from a user's daily activities– email, text, and phone are only accessible from one communication desk. Any form of communication requires an explicit choice to visit the station; notifications and interruptions are completely removed.

7.4.2 Email System

The first device (Figure 7-5) is a small single-board-computer connected to an e-ink screen, designed to evoke a typewriter and feel distinct from a laptop. This device is used just to answer emails; it is automatically synced with a dropbox folder so that attachments can be easily interfaced from other computers. The screen refresh is slow; there is friction. As a result, actions feel more deliberate. The threshold for opening an email is slightly higher.

This device is designed to separate email as a task from the user's main computer. All email interaction requires a deliberate decision to get up and move to a separate physical space. It was originally designed to run with a modified version of the open-source Mailspring email client to track email habits; its latest incarnation runs with a simple chrome extension to track email habits.

7.4.3 Phone Replacement System

With the goal of preserving phone functionality while removing the smartphone itself, the first and most simple part of this interface is to handle phone calls. Using Cell2Jack (Figure 7-6), it's easy to move phone calls back to a landline phone. This device is incredibly reliable, and combined with a paper address book, easily stands in for call functionality.

Figure 7-6: Cell2Jack is a solution for your grandparents to allow their cellphones to operate like the familiar phones of yester-year. It's been very reliable as my bluetooth-to-landline bridge.

To text, I designed a system that takes advantage of Apple's iMessage functionality, a receipt printer, and the remarkable tablet, an e-ink note-taking device. My main iMac runs a script that monitors incoming texts; upon reception, it sends the text to the receipt printer as shown in Figure 7-7.

Also upon receipt of a text, the system generates a PDF of that text with lines to respond. This PDF gets pushed automatically to the remarkable tablet, where it appears in chronological order, labeled by names in my address book. Tapping a name brings up a screen with their last message at the top, and several lines available to write a response. Upon writing the response and exiting the document, the iMac automatically pulls the edited document, OCRs it using Google's API, and sends out a text version as well as a snapshot of the handwritten message (Figure 7-8). On successful send, the text appears in the iMessages database, which triggers the receipt printer to display the outgoing text; thus the success of an outgoing message is easily verifiable.

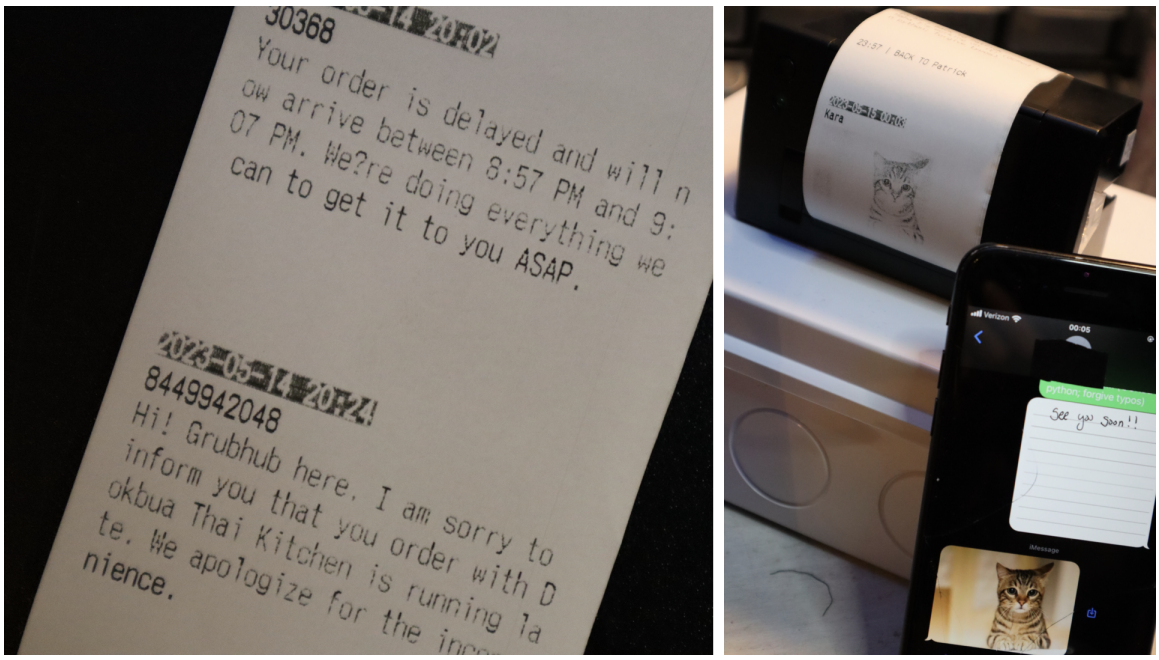


Figure 7-7: Incoming texts get printed, in chronological order, on a receipt printer. Images are also printed.

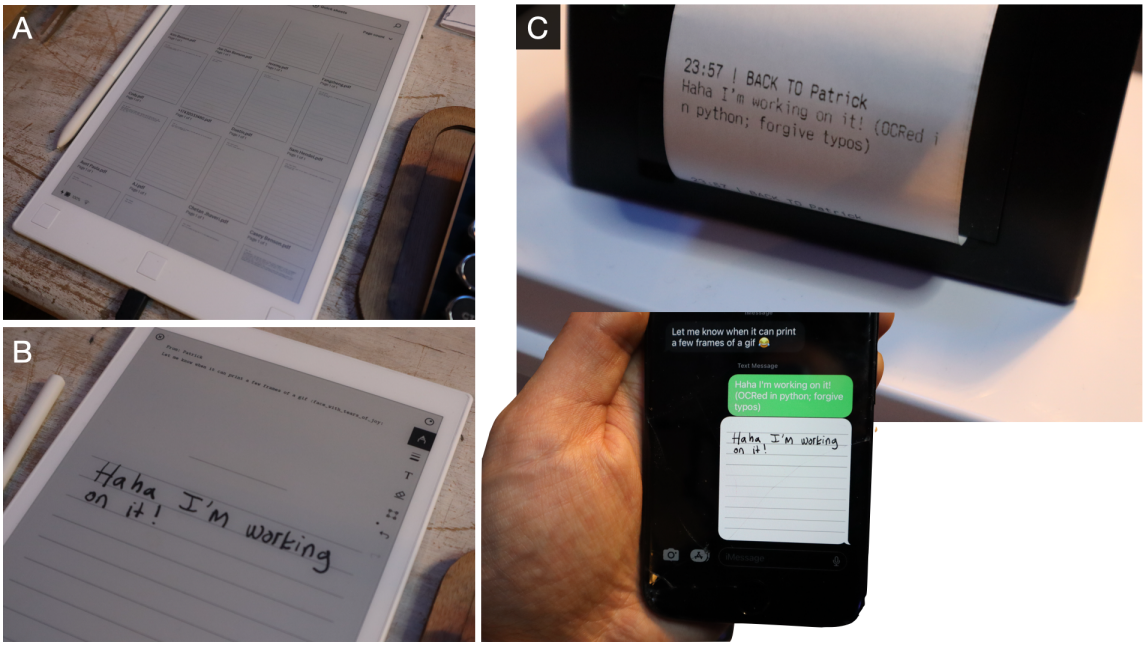


Figure 7-8: To send texts, the user selects one of the auto-generated documents that correspond to every incoming texts, sorted chronologically (and seeded with everyone in the address book) [A]. After clicking on a document name (labeled by sender) the user handwrites a reply [B], which automatically gets OCRed and sent in both text and image form. Upon success, it is also printed to the receipt printer [C].

Chapter 8

The Smartphone Free Living Experiment

8.1 A Brief Justification of N=1

In 2015, one Stanford professor spent two mornings per week over 18 months in an fMRI machine studying himself. Russ Poldrack published his N=1 self-experimentation results that year [341], demonstrating the powerful influence of a cup of coffee on his imaging results.

Russ is not a typical Stanford professor; he is also the Associate Director of Stanford Data Science, and the Director of the SDS Center for Open and Reproducible Science. He happens to be a leading thinker on statistics and methodology.

Self-experimentation and N=1 experiments are not highly regarded by the mainstream academy; we expect results that generalize over a population. Unfortunately, the constraints of nomothetic approaches obscure many of the interaction effects between a person's state (emotion, mood, and personality), their environment (social and physical), and a new intervention.

Idiographic techniques are on the rise. Outside of Poldrack, Aaron Fisher runs the UC Berkley Idiographic Dynamics laboratory where they "believe that the concept of personalization extends beyond treatment delivery and should encompass study design, data collection, and statistical analysis." His most cited paper 'Lack of group-to-individual generalizability is a threat to human subjects research', [147] quantifies the cost of aggregation across subjects when basic assumption of ergodicity are violated (as we would expect). Peter Molenaar, at Penn State, has highly cited works on the topic ('A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever' [302], 'The new person-specific paradigm in psychology' [303]) and has worked to integrate nomothetic and indiographic approaches using statistical tools for clinical mental health work [51].

In medicine, several articles are starting to appear on the topic of N-of-1 trials for drug effectiveness. [419]. Lillie et al. wrote in a 2011 article [262]:

> Coordinated n-of-1 trials have the potential to radically change the way in which evidence-based and individualized medicine is pursued. The availability of relevant wireless clinical monitoring devices that are largely invisible to the user will enhance their value. These enhancements may involve the collection of data, such as continuous time heart rate or blood pressure variability that have never been considered in population-based trials. Not only are the results of n-of-1 trials of immediate benefit to the patient and the treating physician, but if enough of them are pursued, patient characteristics that ultimately differentiate those that benefit from a particular intervention from those that do not can be explored, allowing for stratification of future patient groups in a way that would further benefit patient care.

This statement largely sums up my own view of how research should proceed– identify interventions that have a large impact for an individual first. Once we've established that we can make a difference for one person, the challenge of finding similar <individual, environment>

constellations where we might predict similar impact becomes much more tractable. People are too diverse and interaction effects are too strong for average effects to be meaningful.

8.2 The N=1 Experiment

Figure 8-1: All of my asynchronous communication is designed to happen at this desk– texts are handwritten on the tablet (and automatically sent once written), incoming texts are printed, phone calls are routed through the rotary phone, and email happens on the main typewriter-like interface with an e-ink screen.

My belief in in this approach led me to design the 'Smartphone Free Living Experiment'. For two months, I measured everything about myself while I went about my daily life. For the second of those two months I locked my smartphone in a box and removed email from my personal and work computers.

To communicate with the outside world, I was forced to sit down at the 'asynchronous communication desk' pictured in Figure 8-1 and described in Chapter 7. At this desk, I could hand-write outgoing texts, read printed incoming texts, make/answer phone calls on my rotary phone, and slowly answer email using a type-writer like, e-ink interface set up

with gmail. Away from this desk I received no notifications and no communications; I was impossible to reach on demand.

I picked this intervention to serve as a strong line-in-the-sand– it is very disruptive, something few are willing to consider for themselves. The size of its effect will serve as a useful benchmark for other interventions of more and less severity. If it fails to produce a measurable result– washed out by the joy and the stress of daily living– that would also be a very telling result about the challenges of generalization to real-world contexts, and the relative importance of this style of intervention in daily life.

This is my attempt to design the most impactful, realistic 'background' intervention for focus that I can– removing all of my notifications, my email, and my smartphone from my normal course of living. This also represents a unique, large scale, in-the-wild dataset that fully contextualizes the bio-behavioral and physiological data we collect with our new wearable tools.

8.3 Contributions

This is a unique, large-scale, idiographic collection of bio-behavioral data as introduced in earlier chapters. This data represents two full months of naturalistic data collection for one person– one month of normal living and one month with a major UI intervention. It contextualizes as much as possible about that physiological data through secondary measures– videos, structured interviews, surveys, open-ended notes, and passive metadata monitoring.

This data is open-source. Email and text data is sorted in folders by device (main laptop, home desktop, and 'typewriter' interface); each timestamped entry includes interaction type ('view', 'compose', 'send', 'delete', 'poll'), hashed to/from fields, and summarized text content (emotion analysis using Vader and Flair) [213, 22]. Text data is sorted by day, identities are hashed and texts are summarized as with the email; call logs are included with hashed phone numbers.

Beiwe metadata (see Figure 8-2) [320], E4 data, Captivates and Equinox data (including

surveys), video diary footage (alongside a daily json file that timestamps each structured section of the diary footage), and personal notes are all freely available in raw format. Beiwe data formats are discussed here: [320], Empatica formats are described in the folder with each record.

While raw video data is not publicly posted, examples of the footage are available and requests to run scripts on the raw video feeds for computer screens and video captures will be honored. In addition, extracted pyfeat facial features are provided for every 15s of face footage, as well as toolbar analysis and basic motion data from each screen.

The complete video collection represents 16.5 TB of footage; including 57 continuous days of room footage, 30 days of streaming of all cameras in my living space, 156 hours at my workspace in the Media Lab, and 61 video diaries totally 21.5 hours of structured vlog-style self-reflection (see Figure 8-3). The physiological data includes 48 and 41 days of E4 data on the Right and Left hand respectively with over 25 days of overlap alongside 42.5 Days worth of Captivates/Equinox data. As far as behavioral indicators go, the data includes 92 duration estimates, 223 LED notice events, 117 time estimates, 815 survey responses, and 956 notes timestamped throughout the day. Finally, communication data includes 220 phone calls, 1,513 texts, 10k email interactions, and 2.6 GB of passive phone metadata.

This dataset represents a one-of-a-kind, contextualized, idiographic dataset featuring a large-scale intervention. It should serve as a useful benchmark for individualized, bio-behavioral modeling of deep attentional states in realistic contexts.

8.4 Tools

To enable this data collection effort, I installed state-of-the-art tools, and created new tools where off-the-shelf solutions were not available. I used Beiwe [320]– the top digital phenotyping platform from Harvard– on my iPhone alongside a personal back-end server to collect data. I downloaded call logs from my phone provider and translated these files into CSVs; built a chrome extension to capture my email usage statistics; built a text scraping

Physiological Measures	
Empatica E4	*EDA, Wrist Temperature, PPG, Wrist Acceleration*
Captivates	*Nose and Temple Temperature, Blink Signal, Head Rotation and Acceleration*
Equinox	*Ambient Temperature, Humidity, and Light Levels*
Behavioral Data	
Captivates	*Duration to Notice LED Color Change*
Equinox	*Time Estimates*
Spotify	*Now Playing*
Communication Device Statistics	
Beiwe Phenotyping Data (iPhone)	*Power Status, WIFI/Cellular Connection Status, GPS data, Magnetic Field, Phone Rotation, Phone Acceleration*
Email Statistics	*Activity Timestamps, Unread Message Statistics, Delete, Send, and Open Events with Subject, Body, To, and From Fields*
Text Records	*Timestamps, To, From, Body*
Call Logs	*Timestamps, To, From, Duration*
Screen-time Screenshots	*Screentime for Each Day*
Self-report and Twitch User Data	
Equinox Surveys	*Likert Scales for Cognitive State, Focus, and Task Duration*
Structured Daily Video Diary	*Discussion of Events of the Day, Overall Feelings, Social Interactions, Caffeine, Food, Exercise, Sleep, Productivity and Focus, Screen-time, Experiment Reflections, Gratitude*
Note App	*Free-text application on laptop and computer to enter relevant notes/thoughts*
Twitch Chat	*Chat logs, Twitch User Labels via Chatbot (stress, focus, emotional state)*
Video and Screen Recordings	
At Work	*Recording of Laptop screen, iPhone screen, and Face Camera*
At Home	*24/7 Room Camera, Couch Camera, Face Camera, Communication Desk Camera, Laptop Screen, iMac Screen, iPhone Screen*

Table 8.1: A summary of the types of data collected as part of the smartphone free living experiment. All data is freely available after appropriate anonymization except full video feeds, which can be analyzed on request.

and anonymization tool to pull my text messages from the Apple Chat application.

My physiological data was logged on-device and downloaded to the cloud for the E4 and to a small base-station I carried with me for the Captivates and Equinox data. This base-station hosted a dashboard, which would automatically stream data to the twitch interface when in wifi range.

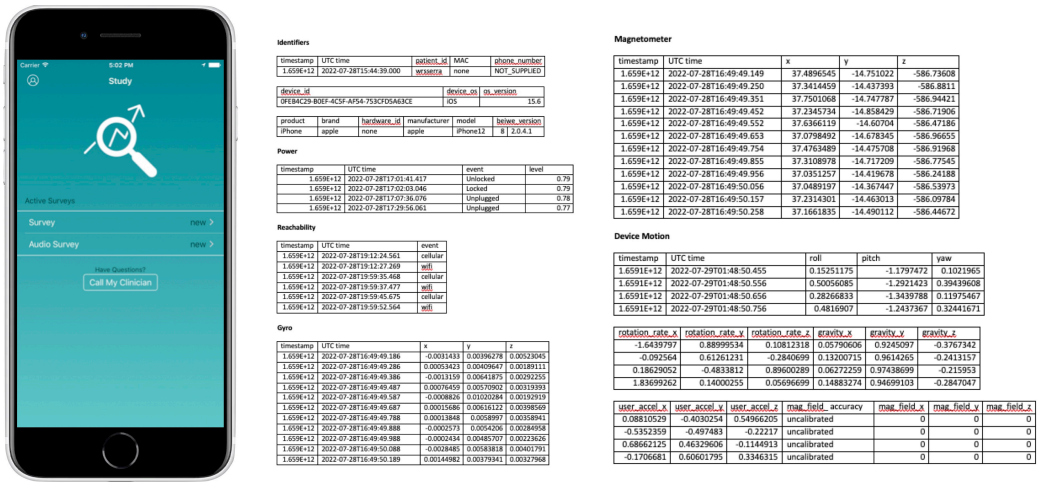


Figure 8-2: A sample of Beiwe data; this open-source application is run by the Onella lab at Harvard.

I had access to three E4 bands at various points during this experiment; when possible I wore E4 bands on both wrists simultaneously, including overnight. The physiological data is also available online.

Throughout the day, I wore the Equinox watch and Captivates glasses. The watch was set up so I could indicate when I noticed the LED in my peripheral vision change (once every 20 minutes to 2 hours). It also allowed time perception estimates and survey data collection. I attempted to cover all of my normal clocks with post-it notes, and rely on the Equinox watch to check the time as much as possible (and guess the time frequently). The watch one survey with the following questions for time estimates: *What time do you think it is? How certain are you?, stress?, alertness?, % time in flow?, max flow?* and one survey intended to capture duration estimates as I switched tasks: *How long were you engaged in the previous activity? How certain are you? Emotional state?, mental effort?, next activity?* My goal was to answer these questions about the time when I needed to check the time or when I noticed a light change (i.e randomly throughout the day); and to catalog my current activities and duration estimates as I went through various tasks in my daily life. Compliance on these surveys fell over the two months of the experiment; not only were they onerous and self-initiated, task boundaries were more difficult to identify than I initially anticipated.

After starting the experiment, I quickly realized that there was a lot of missing context about my environment and mental state that would be more easily captured with a simple text box on my screen. I created a simple one line note-taking app that timestamped and saved any entry, and set it up to run on my main laptop and my home computer (see 8-6). I used this app to make quick notes when something distracted me, or when I felt I had really lost track of the time; it became a less structured (and thus more frequent) supplement to the Equinox watch survey interface.

For video capture, I created two setups– one at my desk at the Media Lab and one at home– using OBS to record many streams at once. An example of the OBS videos I captured at my desk at the Media Lab can be seen in Figure 8-4 top; when I got to work every day, I would plug in my laptop and my iPhone in (during month 1) such that their screens were mirrored to HDMI inputs on my video recorder; I then captured my phone and laptop screen and video footage of my face. At home, I ran a 24/7 room camera that looks out over my apartment (Figure 8-4, bottom two dashboards, small inset). When at home, I would record my activity. In addition to the 24/7 room camera, I would capture my phone screen, my laptop screen, and my main desktop screen; I would capture camera footage of the couch, and camera footage of my desk area (face if working on the desktop computer). During the phone-free month, I set up an additional fish-eye lens and captured footage of the desk I used for communication.

Every day I recorded a video diary about the day, usually lasting around 20 minutes (see Figure 8-3). I created a custom dashboard that I could advance, with structured topics, such that there is a json file for each day noting the timestamps at which each topic starts and ends (the topics I discussed each day are: my activities, my general feelings about the day, my social interactions, my stressors, caffeine intake, food, exercise, sleep, productivity and focus, screen-time, experiment reflections, and gratitude). Not only is this data interesting for transcription and affect analysis, I also wore my full suite of physiological sensors during every video diary.

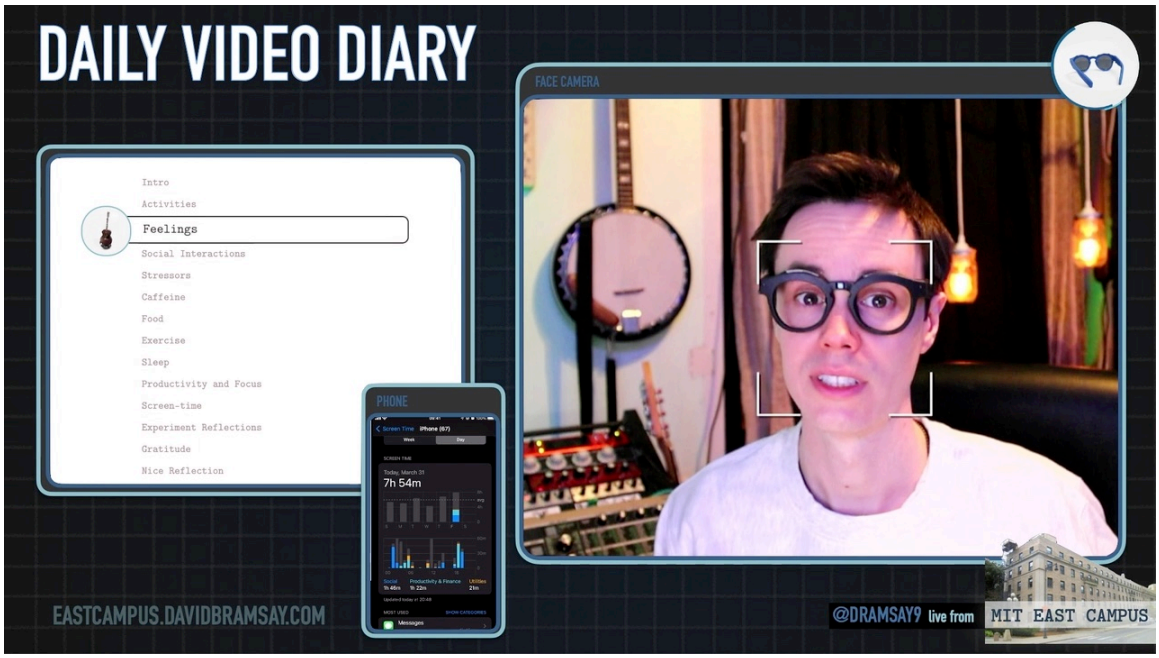


Figure 8-3: Every day I would stream a video diary using this web interface. On the left, you can see a list of topics I discussed every day; For me, this interface had a button to advance the topic, which recorded timestamps so I could easily slice together videos of all 60 days by topic. It's also easy to do emotion and speech analysis on these daily diaries. They also included screen shots of my screen-time app for the day when I had my phone. Typically these sessions would last 20 minutes.

8.4.1 Twitch Streaming

As part of this experiment, I streamed quite a bit of the video footage to Twitch, an ephemeral platform (videos last two weeks) for this type of content. My goal was (1) accountability, (2) to incentivize viewers to label my data in real-time, and (3) to curate an open-source dataset of the video footage (anything that is streamed should be something that can be posted/hosted publicly without having to scrub it for privacy concerns).

To build an interesting streaming platform, I had my physiological data streaming to a live dashboard hosted on my portable raspberry pi base-station. This data interface is shown in Figure 8-5 with all of the data described. Summary videos streamed in real-time are shown in Figure 8-4; at the Media Lab (office), E4 data was collected offline (I only had one BLE Streaming Server implementation, and kept it at home)– Captivates and Equinox data was sent to the dashboard. At home (East Campus), the E4 data was also live-streamed,

Figure 8-4: Examples of the video stream I had on Twitch 4-8 hours a day. At the top we see the dashboard I used at the media lab; at the middle and bottom we see the dashboard at home, which features more configurations depending on my activity, location, and screen usage.

with different configurations of the overlay depending on whether I was (1) working at my desktop, (2) working on my laptop on the couch, or (3) using my phone.

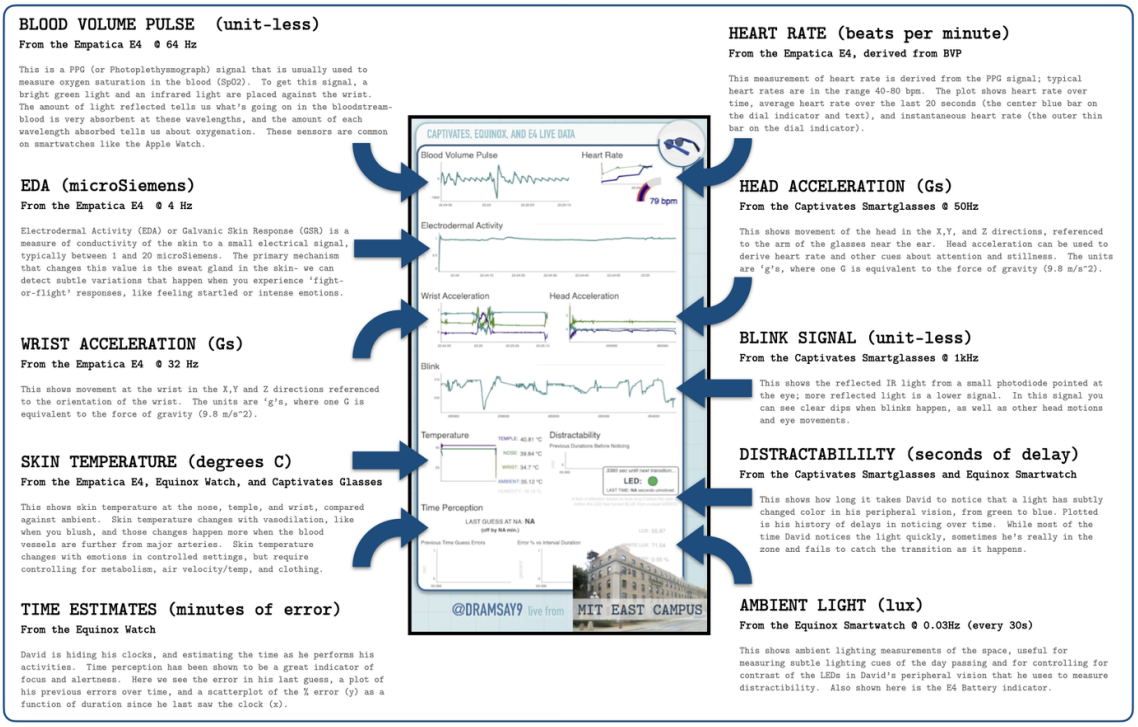


Figure 8-5: The dashboard of data I actively collected at all times over two months. This dashboard was hosted on a raspberry pi that I carried around with me in a backpack (since I gave up my phone for one of the two months).

For both media lab and home overlays, I included a real-time deepface face classifier to display real-time emotion and gaze predictions. I also implemented a Spotify monitoring script which logs and displays the current audio track (bottom of the dashboard). At the office, I streamed my audio headphone feed to Twitch, so I listened to free-to-stream playlists by Streambeats. The dashboard also displays the last reported activity from the Equinox watch (options included 'Reading', 'Writing', 'Viewing', 'Browsing', 'Video Games', 'Playing Music', 'Recording Music', 'Socializing', 'Coding', 'Building', and 'Misc. Work').

Finally, I implemented a Twitch chatbot; it gives instructions to viewers such that they can use certain commands to contribute stress, altertness, focus, and emotion labels to the dataset (i.e. 'Does David look tired? type "!tired $< 0-5 >$" to label his energy state, with 0=very tired and 5=very energized!') Should users follow this command structure, their

assessment will be added in real-time to the Twitch overlay and saved for analysis.

To stream effectively, I designed tools to obscure my phone and video screens– either covering them with a solid block of color, or blurring them so any incoming communication, personal information, or password would be hidden.

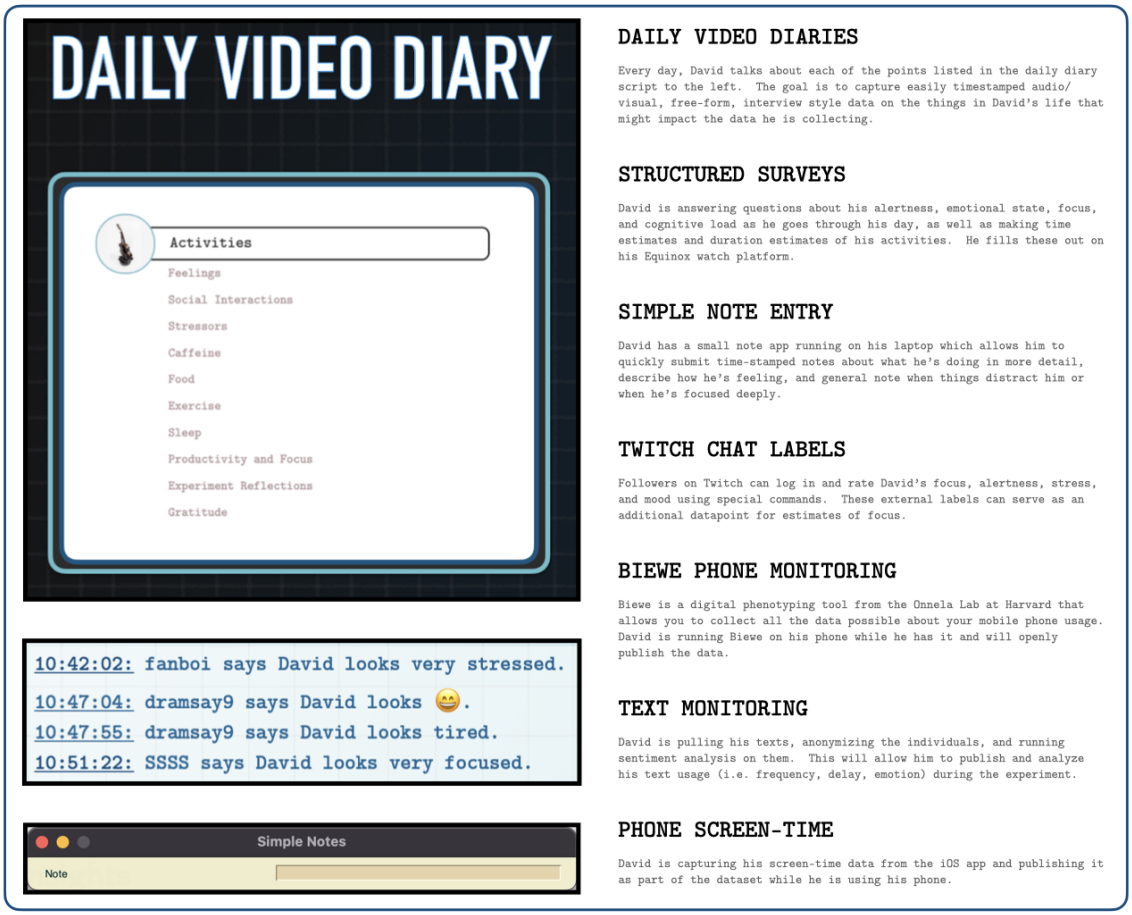


Figure 8-6: A summary of the types of reporting data. On the left you can see the video diary sections (top); part of the twitch dashboard driven by a chatbot that monitored user labels of my focus, stress, altertness, and emotional state (middle), and a small, always-open note-taking app (bottom).

8.4.2 Takeaways

I felt this intervention had a big impact on my cognition with a minimal cost (outside of traveling away from the home), and I plan to enhance this design so I can integrate it

permanently at home and work. I felt less stress, less pressure to always be updating and communicating others, less frenetic energy when I had multiple tasks to get done, and a greater comfort with missing trivial information (turns out I don't need to need to know the weather forecast or the name of that actor).[1] Behavioral indicator corroborate a large change. My email discipline improved dramatically– my time on email was cut in half, with no hanging emails. My texting response times dropped from 2 to 7 hours, but my words per text quadrupled; I sent half the number of texts but twice as many words. My number of phone calls was similarly cut in half, while my time on the phone doubled.

Future analysis will reveal whether and how the physiological data bears out the phenomenological and behavioral impact. Regardless of the specific findings in the data, they should provide valuable insights for probabilistic models of cognition, and serve as a case study to contextualize the cognitive impact of ubiquitous interface design.

[1] Some of these changes took almost three weeks to become noticeable.

Chapter 9

Conclusion

In this thesis, we examine and contextualize the lessons of the replication crisis for the study of flow cognition. The lessons of the crisis re-contextualize what kinds of interventions we expect to matter; instead of implicit interventions that are too small to be consciously perceived, we design automatic interventions that are large and obvious, but integrate intuitively into a user's life. We introduce four such interventions that were built to be tested in naturalistic settings.

In order to make strong claims about the influence of these interventions on flow states, we again turn to lessons from the broader methodological challenges facing psychology and the broader statistical challenges facing science. We distill these lessons into an approach that prioritizes phenomenological, gestalt representations of the concepts under study, as demonstrated in work on auditory perception. We also lay the groundwork for probabilistic, data-generating ontological theories of flow that can be easily compared based on the naturalistic physiological, behaviorial, and self-report data. This approach requires novel tools to capture these signals in the course of normal daily living, of which we introduce three, and modified survey techniques that provide insight into the time-varying nature of flow over a task interval, which we introduce as well.

To test and compare bio-behavioral definitions of flow, naturalistic data is required. We contribute two unique open source datasets– one in which 20 users contribute 2.5 hours of

focused time across multiple days, multiple environments, and multiple activities. This one-of-a-kind dataset relies on a portable test setup that can be self-administered and robust hardware monitoring. The second dataset features 60 days of invasive monitoring of one user (me), with half of that time spent with a major intervention on my daily user interfaces. This dataset fully contextualizes the bio-behavioral and physiological data of interest with as many secondary measures as possible, enabling deep idiographic techniques across the contexts in which I spend my time (work and home) and serving as a case study for major UI interventions. Both datasets feature a rich collection of bio-behavioral and physiological data recorded by robust, product-like hardware we developed ourselves.

The integration of low-level psychological ontology, hardware development, and advanced modeling come together in this thesis to chart a novel path forward for theory building related to cognition. We apply these principles in how we reason about flow states; we hope they are illustrative of broader approaches in psychology. We believe the combination of these tools, the naturalistic datasets they've enabled, and the probabilistic representations that motivated them can serve to push the study of flow beyond the time-naive, lab-based results that dominate current state-of-the-art psychophysiological flow research.

Bibliography

[1] [n. d.]. https://www.npr.org/programs/ted-radio-hour/545024014/hardwired

[2] [n. d.]. https://radiolab.org/podcast/stereothreat

[3] [n. d.]. Frequently Asked Questions - Research — The Gottman Institute - gottman.com. https://www.gottman.com/about/research/faq/. [Accessed 08-Jul-2023].

[4] 2004. Relationships between quality of experience and participation in diverse performance settings. (2004).

[5] 2020 (accessed November 3, 2020). *Base Station.* https://www.vive.com/us/accessory/base-station/

[6] 2020 (accessed November 3, 2020). *Focals by North.* https://www.bynorth.com/

[7] 2020 (accessed November 3, 2020). *FocusEDU.* https://www.brainco.tech/

[8] 2020 (accessed November 3, 2020). Meditation Made Easy. https://choosemuse.com/

[9] 2020 (accessed November 3, 2020). *Open Source Tools for Neuroscience.* https://openbci.com/

[10] 2020 (accessed November 3, 2020). *OpenThread.* https://openthread.io/

[11] 2020 (accessed November 3, 2020). *Quality RTOS Embedded Software.* https://www.freertos.org/

[12] 2020 (accessed November 3, 2020). *Thread*. https://www.threadgroup.org/

[13] 2020 (accessed November 3, 2020). *Your Everyday Smart Glasses*. https://vueglasses.com/

[14] 2021 (accessed April 2, 2021). *The world's first wearable eyewear that lets you see yourself*. https://jins-meme.com/en/

[15] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. *IMWUT* 1 (2017), 33:1–33:20.

[16] Yomna Abdelrahman, Eduardo Velloso, Tilman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive heat: exploring the usage of thermal imaging to unobtrusively estimate cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–20.

[17] Sami Abuhamdeh. 2020. Investigating the "flow" experience: Key conceptual and operational issues. *Frontiers in psychology* 11 (2020), 158.

[18] Felix Acker. 2008. New findings on unconscious versus conscious thought in decision making: additional empirical data and meta-analysis. *Judgment and Decision making* 3, 4 (2008), 292–303.

[19] Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. 2021. Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems* 34 (2021), 29304–29320.

[20] Salvatore M Aglioti and Mariella Pazzaglia. 2010. Representing actions through their sound. *Experimental brain research* 206, 2 (2010), 141–151.

[21] Eugene Aidman, Carolyn Chadunow, Kayla Johnson, and John Reece. 2015. Real-time driver drowsiness feedback improves driver alertness and self-reported driving performance. *Accident Analysis & Prevention* 81 (2015), 8–13.

[22] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[23] Claude Alain, David L Woods, and Robert T Knight. 1998. A distributed cortical network for auditory sensory memory in humans. *Brain research* 812, 1-2 (1998), 23–37.

[24] Alan Allport. 1993. Attention and control: Have we been asking the wrong questions? A critical review of twenty-five years. (1993).

[25] Teresa Amabile and Steven Kramer. 2011. *The progress principle: Using small wins to ignite joy, engagement, and creativity at work*. Harvard Business Press.

[26] Amphenol 2018. *ZTP-135SR Thermometrics Thermopile IR Sensor*. Amphenol. Rev. H.

[27] Valentin Amrhein, Sander Greenland, and Blake McShane. 2019. Scientists rise up against statistical significance. *Nature* 567, 7748 (2019), 305–307.

[28] Eran Amsalem and Alon Zoizner. 2022. Real, but limited: A meta-analytic assessment of framing effects in the political domain. *British Journal of Political Science* 52, 1 (2022), 221–237.

[29] Ishwarya Ananthabhotla and Joseph A Paradiso. 2018. SoundSignaling: Realtime, Stylistic Modification of a Personal Music Corpus for Information Delivery. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 2. 1–23.

[30] Ishwarya Ananthabhotla and Joseph A Paradiso. 2018. SoundSignaling: Realtime, stylistic modification of a personal music corpus for information delivery. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.

[31] Ishwarya Ananthabhotla, David Ramsay, and Joe Paradiso. 2018. HCU400: An Annotated Dataset for Exploring Aural Phenomenology through Causal Uncertainty. *International Conference on Acoustics, Speech, and Signal Processing* (2018). under review; http://arxiv.org/abs/1811.06439.

[32] Ishwarya Ananthabhotla, David B Ramsay, and Joseph A Paradiso. 2019. HCU400: An Annotated Dataset for Exploring Aural Phenomenology Through Causal Uncertainty. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 920–924.

[33] Celia Andreu-Sánchez, Miguel Ángel Martín-Pascual, Agnès Gruart, and José María Delgado-García. 2017. Looking at reality versus watching screens: Media professionalization effects on the spontaneous eyeblink rate. *PLoS One* 12, 5 (2017), e0176030.

[34] Sinan Aral. 2021. What Digital Advertising gets wrong. https://hbr.org/2021/02/what-digital-advertising-gets-wrong

[35] Dan Ariely. [n. d.]. http://danariely.com/2008/05/05/3-main-lessons-of-psychology/

[36] Dominik R Bach and Karl J Friston. 2013. Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology* 50, 1 (2013), 15–22.

[37] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. 2013. The Intrinsic Memorability of Face Photographs. *Journal of Experimental Psychology: General* 142, 4 (2013), 1323.

[38] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. 2013. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* 142, 4 (2013), 1323.

[39] Monya Baker. 2015. Irreproducible biology research costs put at $28 billion per year. *Nature* 533 (2015).

[40] Guha Balakrishnan, Fredo Durand, and John Guttag. 2013. Detecting pulse from head motions in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3430–3437.

[41] James A Ballas. 1993. Common factors in the identification of an assortment of brief everyday sounds. *Journal of experimental psychology: human perception and performance* 19, 2 (1993), 250.

[42] Tadas Baltrusaitis, Amirali Bagher Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (2018), 59–66.

[43] Mahzarin R Banaji and Anthony G Greenwald. 2013. Implicit stereotyping and prejudice. In *The psychology of prejudice*. Psychology Press, 55–76.

[44] John A Bargh. 2011. Unconscious thought theory and its discontents: A critique of the critiques. *Social Cognition* 29, 6 (2011), 629–647.

[45] James Craig Bartlett. 1977. Remembering environmental sounds: The role of verbalization at input. *Memory & Cognition* 5, 4 (1977), 404–414.

[46] Tom Bartlett. [n. d.]. The Magic Ratio that Wasn't. https://www.chronicle.com/blogs/percolator/the-magic-ratio-that-wasnt

[47] František Bartoš and Ulrich Schimmack. 2022. Z-curve 2.0: Estimating replication rates and discovery rates. *Meta-Psychology* 6 (2022).

[48] Hamit Basgol, Inci Ayhan, and Emre Ugur. 2021. Time perception: A review on psychological, computational and robotic models. *IEEE Transactions on Cognitive and Developmental Systems* (2021).

[49] CG Begley and LM Ellis. 2012. Drug development: Raise standards for preclinical cancer research. Nature.[Online]. 483 (7391).

[50] Beiwe. 2021. Exporting the same data from a wearable twice doesn't give you the same data. https://www.beiwe.org/exporting-the-same-data-from-a-wearable-twice-doesnt-give-you-the-same-data/

[51] Adriene M Beltz, Aidan GC Wright, Briana N Sprague, and Peter CM Molenaar. 2016. Bridging the nomothetic and idiographic approaches to the analysis of clinical data. *Assessment* 23, 4 (2016), 447–458.

[52] Daryl Bem, Patrizio Tressoldi, Thomas Rabeyron, and Michael Duggan. 2015. Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research* 4 (2015).

[53] Daryl J Bem. 2011. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology* 100, 3 (2011), 407.

[54] Craig M Bennett, Michael B Miller, and George L Wolford. 2009. Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. *Neuroimage* 47, Suppl 1 (2009), S125.

[55] Alex Bertrams, Roy F Baumeister, Chris Englert, and Philip Furley. 2015. Ego depletion in color priming research: Self-control strength moderates the detrimental effect of red on cognitive test performance. *Personality and Social Psychology Bulletin* 41, 3 (2015), 311–322.

[56] Thomas Blake, Chris Nosko, and Steven Tadelis. 2015. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* 83, 1 (2015), 155–174.

[57] Arthur C Bohart and Thomas Greening. 2001. Humanistic psychology and positive psychology. (2001).

[58] Oliver Bones, Trevor J Cox, and William J Davies. 2018. Distinct categorization strategies for different types of environmental sounds. (2018).

[59] Stefanos Bonovas and Daniele Piovani. 2023. Simpson's Paradox in Clinical Research: A Cautionary Tale. , 1633 pages.

[60] Han Bossier, Sanne P Roels, Ruth Seurinck, Tobias Banaschewski, Gareth J Barker, Arun LW Bokde, Erin Burke Quinlan, Sylvane Desrivieres, Herta Flor, Antoine Grigis, et al. 2020. The empirical replicability of task-based fMRI as a function of sample size. *NeuroImage* 212 (2020), 116601.

[61] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Adcock, et al. 2020. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582, 7810 (2020), 84–88.

[62] Margaret M Bradley and Peter J Lang. 2007. The International Affective Digitized Sounds (; IADS-2): Affective ratings of sounds and instruction manual. *University of Florida, Gainesville, FL, Tech. Rep. B-3* (2007).

[63] Jason J Braithwaite, Derrick G Watson, Robert Jones, and Mickey Rowe. 2013. A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49, 1 (2013), 1017–1034.

[64] Philip Brickman, Dan Coates, and Ronnie Janoff-Bulman. 1978. Lottery winners and accident victims: Is happiness relative? *Journal of personality and social psychology* 36, 8 (1978), 917.

[65] David Brooks. 2013. Beware stubby glasses. https://www.nytimes.com/2013/01/11/opinion/brooks-beware-stubby-glasses.html

[66] Nicholas JL Brown and James AJ Heathers. 2017. The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science* 8, 4 (2017), 363–369.

[67] Nicholas JL Brown and Julia M Rohrer. 2020. Easy as (happiness) pie? A critical evaluation of a popular model of the determinants of well-being. *Journal of Happiness Studies* 21 (2020), 1285–1301.

[68] Nicholas JL Brown, Alan D Sokal, and Harris L Friedman. 2013. The complex dynamics of wishful thinking: the critical positivity ratio. (2013).

[69] Nicholas JL Brown, Alan D Sokal, and Harris L Friedman. 2014. The persistence of wishful thinking. (2014).

[70] Scott W Brown. 1985. Time perception and attention: The effects of prospective versus retrospective paradigms and task demands on perceived duration. *Perception & psychophysics* 38, 2 (1985), 115–124.

[71] Scott W Brown and D Alan Stubbs. 1988. The psychophysics of retrospective and prospective timing. *Perception* 17, 3 (1988), 297–310.

[72] Brian Bruya. 2010. Effortless attention. *A Bradford Book, the MIT Press, Cambridge, Massachusetts Institute of Technology* (2010), 75–92.

[73] Brian Bruya and Yi-Yuan Tang. 2018. Is attention really effort? Revisiting Daniel Kahneman's influential 1973 book attention and effort. *Frontiers in psychology* 9 (2018), 1133.

[74] Bradley R Buchsbaum, Rosanna K Olsen, Paul Koch, and Karen Faith Berman. 2005. Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron* 48, 4 (2005), 687–697.

[75] Bradley R Buchsbaum, Rosanna K Olsen, Paul Koch, and Karen Faith Berman. 2005. Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron* 48, 4 (2005), 687–697.

[76] Jens Bühren, Evdoxia Terzi, Michael Bach, Wolfgang Wesemann, and Thomas Kohnen. 2006. Measuring contrast sensitivity under different lighting conditions: comparison of three tests. *Optometry and Vision Science* 83, 5 (2006), 290–298.

[77] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature human behaviour* 2, 9 (2018), 637–644.

[78] Cristina Candal-Pedreira, Alberto Ruano-Ravina, Esteve Fernández, Jorge Ramos, Isabel Campos-Varela, and Mónica Pérez-Ríos. 2020. Does retraction after misconduct have an impact on citations? A pre–post study. *BMJ Global Health* 5, 11 (2020), e003719.

[79] Eugene M Caruso, Oren Shapira, and Justin F Landy. 2017. Show me the money: A systematic exploration of manipulations, moderators, and mechanisms of priming effects. *Psychological Science* 28, 8 (2017), 1148–1159.

[80] Juan A. Castro-García, Alberto J. Molina-Cantero, Manuel Merino-Monge, and Isabel M. Gómez-González. 2019. An Open-Source Hardware Acquisition Platform for Physiological Measurements. *IEEE Sensors Journal* 19 (2019), 11526–11534.

[81] Christopher F Chabris, Patrick R Heck, Jaclyn Mandart, Daniel J Benjamin, and Daniel J Simons. 2018. No evidence that experiencing physical warmth promotes interpersonal warmth. *Social Psychology* (2018).

[82] Matthew Charlton, Sophie A Stanley, Zoë Whitman, Victoria Wenn, Timothy J Coats, Mark Sims, and Jonathan P Thompson. 2020. The effect of constitutive pigmentation on the measured emissivity of human skin. *Plos one* 15, 11 (2020), e0241843.

[83] D Paice Chris et al. 1990. Another stemmer. In *ACM SIGIR Forum*, Vol. 24. 56–61.

[84] Patrick Chwalek, David Ramsay, and Joseph A Paradiso. 2021. Captivates: A Smart Eyeglass Platform for Across-Context Physiological Measurement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 3 (2021), 1–32.

[85] Robert Cialdini. 2016. *Pre-suasion: A revolutionary way to influence and persuade.* Simon and Schuster.

[86] RP Clark, BJ Mullan, and LG Pugh. 1977. Skin temperature during running–a study using infra-red colour thermography. *The Journal of physiology* 267, 1 (1977), 53–62.

[87] John Cochrane. [n. d.]. https://johnhcochrane.blogspot.com/2015/05/homo-economicus-or-homo-paleas.html

[88] Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science* 349, 6251 (2015), aac4716.

[89] Sara Colombo. 2016. *Dynamic products: shaping information to engage and persuade.* Springer.

[90] Alper Cömert and Jari Hyttinen. 2015. Investigating the possible effect of electrode support structure on motion artifact in wearable bioelectric signal monitoring. *BioMedical Engineering OnLine* 14 (2015).

[91] R Conrad. 1972. The developmental role of vocalizing in short-term memory. *Journal of Memory and Language* 11, 4 (1972), 521.

[92] Enrico Costanza, Samuel Inverso, Elan Pavlov, Rebecca Allen, and Pattie Maes. 2006. eye-q: eyeglass peripheral display for subtle intimate notifications.

[93] Nelson Cowan. 1984. On short and long auditory stores. *Psychological bulletin* 96, 2 (1984), 341.

[94] David R Cregg and Jennifer S Cheavens. 2021. Gratitude interventions: Effective self-help? A meta-analysis of the impact on symptoms of depression and anxiety. *Journal of Happiness Studies* 22 (2021), 413–445.

[95] Rodney J. Croft and Roger J. Barry. 2000. Removal of ocular artifact from the EEG: a review. *Neurophysiologie Clinique/Clinical Neurophysiology* 30 (2000), 5–19.

[96] Mihaly Csikszentmihalyi. 1975. *Beyond boredom and anxiety.* Jossey-bass.

[97] Mihaly Csikszentmihalyi. 1999. If we are so rich, why aren't we happy? *American psychologist* 54, 10 (1999), 821.

[98] Mihaly Csikszentmihalyi and Judith LeFevre. 1989. Optimal experience in work and leisure. *Journal of personality and social psychology* 56, 5 (1989), 815.

[99] Mihály Csíkszentmihályi. 2008. Flow: The Psychology of Optimal Experience.

[100] Erin Cvejic, Jeesun Kim, and Chris Davis. 2010. Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Communication* 52, 6 (2010), 555–564.

[101] Laura Dabbish, Gloria Mark, and Víctor M González. 2011. Why do I keep interrupting myself? Environment, habit and self-interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3127–3130.

[102] Ian Daly, Nicoletta Nicolaou, Slawomir Nasuto, and Kevin Warwick. 2013. Automated Artifact Removal From the Electroencephalogram: A Comparative Study. *Clinical EEG and neuroscience : official journal of the EEG and Clinical Neuroscience Society (ENCS)* 44 (05 2013). https://doi.org/10.1177/1550059413476485

[103] Linh C Dang, Gregory R Samanez-Larkin, Jaime J Castrellon, Scott F Perkins, Ronald L Cowan, Paul A Newhouse, and David H Zald. 2017. Spontaneous eye blink rate (EBR) is uncorrelated with dopamine D2 receptor availability and unmodulated by dopamine agonism in healthy adults. *Eneuro* 4, 5 (2017).

[104] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108, 17 (2011), 6889–6892.

[105] Debraj De, Pratool Bharti, Sajal K Das, and Sriram Chellappan. 2015. Multimodal wearable sensing for fine-grained activity recognition in healthcare. *IEEE Internet Computing* 19, 5 (2015), 26–35.

[106] Örjan De Manzano, Töres Theorell, László Harmat, and Fredrik Ullén. 2010. The psychophysiology of flow during piano playing. *Emotion* 10, 3 (2010), 301.

[107] Marcelo Felipe de Sampaio Barros, Fernando M Araújo-Moreira, Luis Carlos Trevelin, and Rémi Radel. 2018. Flow experience and the mobilization of attentional resources. *Cognitive, Affective, & Behavioral Neuroscience* 18 (2018), 810–823.

[108] Angus Deaton and Nancy Cartwright. 2018. Understanding and misunderstanding randomized controlled trials. *Social science & medicine* 210 (2018), 2–21.

[109] Daniel Defraia. 2020. The unknown legacy of Military Mental Health Programs. https://taskandpurpose.com/military-life/the-unknown-legacy-of-military-mental-health-programs/

[110] Stefano DellaVigna and Elizabeth Linos. 2022. RCTs to scale: Comprehensive evidence from two nudge units. *Econometrica* 90, 1 (2022), 81–116.

[111] Varinthira Duangudom Delmotte. 2012. *Computational auditory saliency.* Ph. D. Dissertation. Georgia Institute of Technology.

[112] Artem Dementyev and Christian Holz. 2017. DualBlink: A Wearable Device to Continuously Detect, Track, and Actuate Blinking For Alleviating Dry Eyes and Computer Vision Syndrome. *IMWUT* 1 (2017), 1:1–1:19.

[113] Lydia Denworth. 2019. Debate arises over teaching "growth mindsets" to motivate students. *Scientific American* (2019).

[114] Carolina Diaz-Piedra, Emilo Gomez-Milan, and Leandro L Di Stasi. 2019. Nasal skin temperature reveals changes in arousal levels due to time on task: An experimental thermal infrared imaging study. *Applied Ergonomics* 81 (2019), 102870.

[115] Ed Diener, Richard E Lucas, and Christie Napa Scollon. 2009. Beyond the hedonic treadmill: Revising the adaptation theory of well-being. *The science of well-being: The collected works of Ed Diener* (2009), 103–118.

[116] Arne Dietrich and Oliver Stoll. 2010. Effortless attention, hypofrontality, and perfectionism. *Effortless attention: A new perspective in the cognitive science of attention and action* (2010), 159–178.

[117] Ap Dijksterhuis. 2004. Think different: the merits of unconscious thought in preference development and decision making. *Journal of personality and social psychology* 87, 5 (2004), 586.

[118] Eugen Dimant, Gerben A Van Kleef, and Shaul Shalvi. 2020. Requiem for a nudge: Framing effects in nudging honesty. *Journal of Economic Behavior & Organization* 172 (2020), 247–266.

[119] Amy M Do, Alexander V Rupert, and George Wolford. 2008. Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic bulletin & review* 15, 1 (2008), 96–98.

[120] Georgios Doganis, P Iosifidou, and S Vlachopoulos. 2000. Factor structure and internal consistency of the Greek version of the Flow State Scale. *Perceptual and Motor Skills* 91, 3_suppl (2000), 1231–1240.

[121] Stéphane Doyen, Olivier Klein, Cora-Lise Pichon, and Axel Cleeremans. 2012. Behavioral priming: it's all in the mind, but whose mind? *PloS one* 7, 1 (2012), e29081.

[122] Charlotte L Doyle. 2017. Creative flow as a unique cognitive process. *Frontiers in psychology* 8 (2017), 1348.

[123] Sylvie Droit-Volet, Sophie L Fayolle, and Sandrine Gil. 2011. Emotion and time perception: effects of film-induced mood. *Frontiers in integrative neuroscience* 5 (2011), 33.

[124] Sylvie Droit-Volet and Warren H Meck. 2007. How emotions colour our perception of time. *Trends in cognitive sciences* 11, 12 (2007), 504–513.

[125] Gershon Dublon. 2018. *Sensor(y) Landscapes: Technologies for New Perceptual Sensibilities*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[126] Danièle Dubois, Catherine Guastavino, and Manon Raimbault. 2006. A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories. *Acta acustica united with acustica* 92, 6 (2006), 865–874.

[127] Clement Duhart, Gershon Dublon, Brian Mayton, and Joseph Paradiso. 2018. Deep Learning Locally Trained Wildlife Sensing in Real Acoustic Wetland Environment. In *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 3–14.

[128] Adam Dunkels, Bjorn Gronvall, and Thiemo Voigt. 2004. Contiki-a lightweight and flexible operating system for tiny networked sensors. In *29th annual IEEE international conference on local computer networks*. IEEE, 455–462.

[129] Carol Dweck. 2022. Growth mindset interventions yield impressive results. https://theconversation.com/growth-mindset-interventions-yield-impressive-results-97423

[130] Lisa Valentina Eberhardt, Christoph Strauch, Tim Samuel Hartmann, and Anke Huckauf. 2022. Increasing pupil size is associated with improved detection performance in the periphery. *Attention, Perception, & Psychophysics* 84, 1 (2022), 138–149.

[131] Editorial. 2019. It's time to talk about ditching statistical significance. *Nature* 567, 7748 (2019), 283.

[132] Roy Eidelson, Marc Pilisuk, and Stephen Soldz. 2011. The dark side of comprehensive soldier fitness. (2011).

[133] Anders Eklund, Thomas E Nichols, and Hans Knutsson. 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences* 113, 28 (2016), 7900–7905.

[134] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. 2013. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PloS one* 8, 10 (2013), e76585.

[135] Maxwell L Elliott, Annchen R Knodt, David Ireland, Meriwether L Morris, Richie Poulton, Sandhya Ramrakha, Maria L Sison, Terrie E Moffitt, Avshalom Caspi, and Ahmad R Hariri. 2020. What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological science* 31, 7 (2020), 792–806.

[136] Christoph Ellmer. 2017. OpenThread vs. Contiki IPv6: An Experimental Evaluation.

[137] Stefan Engeser and Falko Rheinberg. 2008. Flow, moderators of challenge-skill-balance and performance. *Motivation and Emotion* 32, 3 (2008), 158–172.

[138] Stefan Engeser, Anja Schiepe-Tiska, and Corinna Peifer. 2021. Historical lines and an overview of current research on flow. *Advances in flow research* (2021), 1–29.

[139] Stefan Ed Engeser. 2012. *Advances in flow research.* Springer Science+ Business Media.

[140] Timothy M Errington, Alexandria Denis, Anne B Allison, Renee Araiza, Pedro Aza-Blanc, Lynette R Bower, Jessica Campos, Heidi Chu, Sarah Denson, Cristine Donham, et al. 2021. Experiments from unfinished registered reports in the reproducibility project: cancer biology. *Elife* 10 (2021), e73430.

[141] Susann Eschrich, Thomas F Münte, and Eckart O Altenmüller. 2008. Unforgettable film music: the role of emotion in episodic long-term memory for music. *BMC neuroscience* 9, 1 (2008), 48.

[142] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 1459–1462.

[143] Ayman Farahat and Michael C Bailey. 2012. How effective is targeted advertising?. In *Proceedings of the 21st international conference on World Wide Web*. 111–120.

[144] Miguel Farias and Catherine Wikholm. 2016. Has the science of mindfulness lost its mind? *BJPsych bulletin* 40, 6 (2016), 329–332.

[145] Szymon Fedor, Peggy Chau, Nicolina Bruno, Rosalind W Picard, Joan Camprodon, et al. 2016. Can we predict depression from the asymmetry of electrodermal activity? *Iproceedings* 2, 1 (2016), e6117.

[146] Melissa J Ferguson, Travis J Carter, and Ran R Hassin. 2014. Commentary on the attempt to replicate the effect of the American flag on increased Republican attitudes. (2014).

[147] Aaron J Fisher, John D Medaglia, and Bertus F Jeronimus. 2018. Lack of group-to-individual generalizability is a threat to human subjects research. *Proceedings of the National Academy of Sciences* 115, 27 (2018), E6106–E6115.

[148] Ronald A Fisher. 1956. Statistical methods and scientific inference. (1956).

[149] Gianluca Di Flumeri, Pietro Aricò, Gianluca Borghini, Nicolina Sciaraffa, Antonello Di Florio, and Fabio Babiloni. 2019. The Dry Revolution: Evaluation of Three Different EEG Dry Electrode Types in Terms of Signal Spectral Features, Mental States Classification and Usability. *Sensors (Basel, Switzerland)* 19 (2019).

[150] Frederic Font, Tim Brookes, George Fazekas, Martin Guerber, Amaury La Burthe, David Plans, Mark D Plumbley, Meir Shaashua, Wenwu Wang, and Xavier Serra. 2016. Audio Commons: bringing Creative Commons audio content to the creative industries. In *Audio Engineering Society Conference: 61st International Conference: Audio for Games*. Audio Engineering Society.

[151] Patrick S Forscher, Calvin K Lai, Jordan R Axt, Charles R Ebersole, Michelle Herman, Patricia G Devine, and Brian A Nosek. 2019. A meta-analysis of procedures to change implicit measures. *Journal of personality and social psychology* 117, 3 (2019), 522.

[152] Barbara L Fredrickson. 2013. Updated thinking on positivity ratios. (2013).

[153] Barbara L Fredrickson and Marcial F Losada. 2005. Positive affect and the complex dynamics of human flourishing. *American psychologist* 60, 7 (2005), 678.

[154] Leonard P Freedman, Iain M Cockburn, and Timothy S Simcoe. 2015. The economics of reproducibility in preclinical research. *PLoS biology* 13, 6 (2015), e1002165.

[155] Harris L Friedman and Nicholas JL Brown. 2018. Implications of debunking the "critical positivity ratio" for humanistic psychology: Introduction to special issue. *Journal of humanistic psychology* 58, 3 (2018), 239–261.

[156] Shelly L Gable and Jonathan Haidt. 2005. What (and why) is positive psychology? *Review of general psychology* 9, 2 (2005), 103–110.

[157] Suzanne H Gage and Harry R Sumnall. 2019. Rat Park: How a rat paradise changed the narrative of addiction. *Addiction* 114, 5 (2019), 917–922.

[158] Inc. Gallup. 2023. State of the global workplace report 2023. https://www.gallup.com/workplace/349484/state-of-the-global-workplace.aspx

[159] William W Gaver. 1993. How do we Hear in the World? Explorations in Ecological Acoustics. *Ecological Psychology* 5, 4 (1993), 285–313.

[160] William W Gaver. 1993. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology* 5, 1 (1993), 1–29.

[161] William W Gaver. 1993. What in the World do we Hear?: An Ecological Approach to Auditory Event Perception. *Ecological psychology* 5, 1 (1993), 1–29.

[162] Andrew Gelman. [n. d.]. Stereotype threat! https://statmodeling.stat.columbia.edu/2013/07/07/stereotype-threat/

[163] Andrew Gelman. 2018. You need 16 times the sample size to estimate an interaction than to estimate a main effect. https://statmodeling.stat.columbia.edu/2018/03/15/need16/

[164] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[165] Xiaowei Geng, Ziguang Chen, Wing Lam, and Quanquan Zheng. 2013. Hedonic evaluation over short and long retention intervals: The mechanism of the peak–end rule. *Journal of Behavioral Decision Making* 26, 3 (2013), 225–236.

[166] Peter Gerjets, Carina Walter, Wolfgang Rosenstiel, Martin Bogdan, and Thorsten O. Zander. 2014. Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience* 8 (2014).

[167] Eric Ghelfi, Cody D Christopherson, Heather L Urry, Richie L Lenne, Nicole Legate, Mary Ann Fischer, Fieke MA Wagemans, Brady Wiggins, Tamara Barrett, Michelle Bornstein, et al. 2020. Reexamining the effect of gustatory disgust on moral judgment: A multilab direct replication of Eskine, Kacinik, and Prinz (2011). *Advances in Methods and Practices in Psychological Science* 3, 1 (2020), 3–23.

[168] Theodoros Giannakopoulos. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PloS one* 10, 12 (2015).

[169] James J Gibson and Eleanor J Gibson. 1955. Perceptual learning: Differentiation or enrichment? *Psychological review* 62, 1 (1955), 32.

[170] Gerd Gigerenzer. 1991. How to make cognitive illusions disappear: Beyond "heuristics and biases". *European review of social psychology* 2, 1 (1991), 83–115.

[171] Gerd Gigerenzer. 1996. On narrow norms and vague heuristics: A reply to Kahneman and Tversky. (1996).

[172] Gerd Gigerenzer. 2004. Mindless statistics. *The Journal of Socio-Economics* 33, 5 (2004), 587–606.

[173] Gerd Gigerenzer. 2018. Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science* 1, 2 (2018), 198–218.

[174] Daniel Gilbert. 2009. *Stumbling on happiness*. Vintage Canada.

[175] Bruno L Giordano, John McDonnell, and Stephen McAdams. 2010. Hearing living symbols and nonliving icons: category specificities in the cognitive processing of environmental sounds. *Brain and cognition* 73, 1 (2010), 7–19.

[176] Stefan Glasauer and Zhuanghua Shi. 2021. The origin of Vierordt's law: the experimental protocol matters. *PsyCh Journal* 10, 5 (2021), 732–741.

[177] Andreas Glöckner. 2016. The irrational hungry judge effect revisited: Simulations reveal that the magnitude of the effect is overestimated. *Judgment and Decision making* 11, 6 (2016), 601–610.

[178] Claudia Goldin and Cecilia Rouse. 2000. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American economic review* 90, 4 (2000), 715–741.

[179] Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. A comparison of approaches to advertising measurement: Evidence from big field experiments at Facebook. *Marketing Science* 38, 2 (2019), 193–225.

[180] John M Gottman, James D Murray, Catherine C Swanson, Rebecca Tyson, and Kristin R Swanson. 2005. *The mathematics of marriage: Dynamic nonlinear models*. mit press.

[181] Gabriele Gratton. 1998. Dealing with artifacts: The EOG contamination of the event-related brain potential. *Behavior Research Methods, Instruments, Computers* 30 (1998), 44–53.

[182] Anthony G Greenwald, Mahzarin R Banaji, and Brian A Nosek. 2015. Statistically small effects of the Implicit Association Test can have societally large effects. (2015).

[183] Richard A Griggs. 2015. The disappearance of independence in textbook coverage of Asch's social pressure experiments. *Teaching of Psychology* 42, 2 (2015), 137–142.

[184] Vladas Griskevicius, Noah J Goldstein, Chad R Mortensen, Jill M Sundie, Robert B Cialdini, and Douglas T Kenrick. 2009. Fear and loving in Las Vegas: Evolution, emotion, and persuasion. *Journal of Marketing Research* 46, 3 (2009), 384–395.

[185] Simon Grondin. 2010. Timing and time perception: a review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception, & Psychophysics* 72, 3 (2010), 561–582.

[186] Jongseong Gwak, Motoki Shino, Kazutaka Ueda, and Minoru Kamata. 2019. An investigation of the effects of changes in the indoor ambient temperature on arousal level, thermal comfort, and physiological indices. *Applied Sciences* 9, 5 (2019), 899.

[187] Brian Gygi and Valeriy Shafiro. 2007. General functions and specific applications of environmental sound research. *Frontiers in bioscience: a journal and virtual library* 12 (2007), 3152–3166.

[188] Carlton Gyles. 2015. Skeptical of medical science reports? *The Canadian Veterinary Journal* 56, 10 (2015), 1011.

[189] Wayne Hall and Megan Weier. 2017. Lee Robins' studies of heroin use among US Vietnam veterans. *Addiction* 112, 1 (2017), 176–180.

[190] Claudia Hammond. 2012. *Time warped: Unlocking the mysteries of time perception.* House of Anansi.

[191] PA Hancock, AD Kaplan, JK Cruit, GM Hancock, KR MacArthur, and JL Szalma. 2019. A meta-analysis of flow effects and the perception of time. *Acta psychologica* 198 (2019), 102836.

[192] László Harmat, Örjan de Manzano, Töres Theorell, Lennart Högman, Håkan Fischer, and Fredrik Ullén. 2015. Physiological correlates of the flow experience during computer game playing. *International Journal of Psychophysiology* 97, 1 (2015), 1–7.

[193] S Alexander Haslam, Stephen D Reicher, and Jay J Van Bavel. 2019. Rethinking the nature of cruelty: The role of identity leadership in the Stanford Prison Experiment. *American Psychologist* 74, 7 (2019), 809.

[194] James A Heathers, Jordan Anaya, Tim van der Zee, and Nicholas JL Brown. 2018. *Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (SPRITE).* Technical Report. PeerJ Preprints.

[195] Kiran B. Hebbar, J. Dennis Fortenberry, Kristine Rogers, Robert Merritt, and Kirk Easley. 2005. Comparison of temporal artery thermometer to standard temperature measurements in pediatric intensive care unit patients. *Pediatric critical care medicine : a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies* 6 5 (2005), 557–61.

[196] Barbara S Held. 2018. Positive psychology's a priori problem. *Journal of Humanistic Psychology* 58, 3 (2018), 313–342.

[197] Javier Hernandez, Zicheng Liu, Geoff Hulten, Dave DeBarr, Kyle Krum, and Zhengyou Zhang. 2013. Measuring the engagement level of TV viewers. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG).* IEEE, 1–7.

[198] Javier Hernandez, Daniel McDuff, Christian Infante, Pattie Maes, Karen Quigley, and Rosalind Picard. 2016. Wearable ESM: differences in the experience sampling method across wearable devices. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*. 195–205.

[199] Francisco Hernando-Gallego, David Luengo, and Antonio Artés-Rodríguez. 2017. Feature extraction of galvanic skin responses by nonnegative sparse deconvolution. *IEEE journal of biomedical and health informatics* 22, 5 (2017), 1385–1394.

[200] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN Architectures for Large-scale Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 131–135.

[201] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://arxiv.org/abs/1609.09430

[202] Robert E Hicks, George W Miller, and Marcel Kinsbourne. 1976. Prospective and retrospective judgments of time as a function of amount of information processed. *The American journal of psychology* (1976), 719–730.

[203] Michael J Hiscox, Tara Oliver, Michael Ridgway, Lilia Arcos-Holzinger, Alastair Warren, and Andrew Willis. 2017. Going blind to see more clearly: Unconscious bias in Australian Public Service shortlisting processes. *Behavioural Economics Team of the Australian Government* (2017).

[204] Bernhard Hommel, Craig S Chapman, Paul Cisek, Heather F Neyedli, Joo-Hyun Song, and Timothy N Welsh. 2019. No one knows what attention is. *Attention, Perception, & Psychophysics* 81 (2019), 2288–2303.

[205] Yong-Wook Hong, Yejong Yoo, Jihoon Han, Tor D Wager, and Choong-Wan Woo.

2019. False-positive neuroimaging: Undisclosed flexibility in testing spatial hypotheses allows presenting anything as a replicated finding. *NeuroImage* 195 (2019), 384–395.

[206] David Hoppe, Stefan Helfmann, and Constantin A Rothkopf. 2018. Humans quickly learn to blink strategically in response to environmental task demands. *Proceedings of the National Academy of Sciences* 115, 9 (2018), 2246–2251.

[207] Daniel Horowitz. 2017. *Happier?: the history of a cultural movement that aspired to transform America*. Oxford University Press.

[208] Yu-Yu Hsiao, Davood Tofighi, Eric S Kruger, M Lee Van Horn, David P MacKinnon, and Katie Witkiewitz. 2019. The (lack of) replication of self-reported mindfulness as a mechanism of change in mindfulness-based relapse prevention for substance use disorders. *Mindfulness* 10 (2019), 724–736.

[209] Chia-Fen Hsu, Lee Propp, Larissa Panetta, Shane Martin, Stella Dentakos, Maggie E Toplak, and John D Eastwood. 2018. Mental effort and discomfort: Testing the peak-end effect during a cognitively demanding task. *PloS one* 13, 2 (2018), e0191479.

[210] Lifu Huang, Heng Ji, et al. 2017. Learning Phrase Embeddings from Paraphrases with GRUs. In *Proceedings of the First Workshop on Curation and Applications of Parallel and Comparable Corpora*. 16–23.

[211] Hilde M Huizenga, Ruud Wetzels, Don van Ravenzwaaij, and Eric-Jan Wagenmakers. 2012. Four empirical tests of unconscious thought theory. *Organizational Behavior and Human Decision Processes* 117, 2 (2012), 332–340.

[212] Louise Hume, Christopher A Dodd, and Nigel P Grigg. 2003. In-Store Selection of Wine—No Evidence for the Mediation of Music? *Perceptual and motor skills* 96, 3_suppl (2003), 1252–1254.

[213] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.

[214] Koji Inui, Tomokazu Urakawa, Koya Yamashiro, Naofumi Otsuru, Makoto Nishihara, Yasuyuki Takeshima, Sumru Keceli, and Ryusuke Kakigi. 2010. Non-linear laws of echoic memory and auditory change detection in humans. *BMC neuroscience* 11, 1 (2010), 80.

[215] Stephanos Ioannou, Vittorio Gallese, and Arcangelo Merla. 2014. Thermal infrared imaging in psychophysiology: Potentialities and limits. In *Psychophysiology*.

[216] Tomoya Ishige and Jun-ichi Imai. 2020. Peripheral Notification That Does Not Interfere with User's Concentration by Interactive Projection. In *2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS)*. IEEE, 1–5.

[217] Shoya Ishimaru, Kai Kunze, Koichi Kise, Jens Weppner, Andreas Dengel, Paul Lukowicz, and Andreas Bulling. 2014. In the Blink of an Eye - Combining Head Motion and Eye Blink Frequency for Activity Recognition with Google Glass. *ACM International Conference Proceeding Series*, 150–153. https://doi.org/10.1145/2582051.2582066

[218] Shoya Ishimaru, Kai Kunze, Koichi Kise, Jens Weppner, Andreas Dengel, Paul Lukowicz, and Andreas Bulling. 2014. In the blink of an eye: combining head motion and eye blink frequency for activity recognition with Google Glass. In *AH '14*.

[219] Sheena S Iyengar and Mark R Lepper. 2000. When choice is demotivating: Can one desire too much of a good thing? *Journal of personality and social psychology* 79, 6 (2000), 995.

[220] Susan A Jackson and Robert C Eklund. 2002. Assessing flow in physical activity: The flow state scale–2 and dispositional flow scale–2. *Journal of sport and exercise psychology* 24, 2 (2002), 133–150.

[221] Susan A Jackson and Herbert W Marsh. 1996. Development and validation of a scale to measure optimal experience: The Flow State Scale. *Journal of sport and exercise psychology* 18, 1 (1996), 17–35.

[222] Susan A Jackson, Andrew J Martin, and Robert C Eklund. 2008. Long and short measures of flow: The construct validity of the FSS-2, DFS-2, and new brief counterparts. *Journal of Sport and Exercise Psychology* 30, 5 (2008), 561–587.

[223] William James, Frederick Burkhardt, Fredson Bowers, and Ignas K Skrupskelis. 1890. *The principles of psychology*. Vol. 1. Macmillan London.

[224] Lutz Jäncke. 2008. Music, memory and emotion. *Journal of biology* 7, 6 (2008), 21.

[225] Melinda S Jensen, Richard Yao, Whitney N Street, and Daniel J Simons. 2011. Change blindness and inattentional blindness. *Wiley Interdisciplinary Reviews: Cognitive Science* 2, 5 (2011), 529–546.

[226] MW Johns et al. 2003. The amplitude-velocity ratio of blinks: a new method for monitoring drowsiness. *Sleep* 26, SUPPL. (2003).

[227] Eric J Johnson and Daniel Goldstein. 2003. Do defaults save lives? , 1338–1339 pages.

[228] Jillian A Johnson, Matthew J Zawadzki, Dusti R Jones, Julia Reichenberger, and Joshua M Smyth. 2022. Intra-individual associations of perceived stress, affective valence, and affective arousal with momentary cortisol in a sample of working adults. *Annals of behavioral medicine* 56, 3 (2022), 305–310.

[229] Lee Jussim. 2015. Is stereotype threat overcooked, overstated, and oversold. *Heterodox Academy* (2015).

[230] Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.

[231] Daniel Kahneman and Amos Tversky. 1996. On the reality of cognitive illusions. (1996).

[232] Ozlem Kalinli, Shiva Sundaram, and Shrikanth Narayanan. 2009. Saliency-driven unstructured acoustic scene classification using latent perceptual indexing. In *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*. IEEE, 1–6.

[233] Habib Karaki and Vasco Polyzoev. 2020 (accessed May 3, 2020). Demystifying Thermopile IR Temp Sensors. https://www.fierceelectronics.com/components/demystifying-thermopile-ir-temp-sensors-0

[234] Christoph Kayser, Christopher I Petkov, Michael Lippert, and Nikos K Logothetis. 2005. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology* 15, 21 (2005), 1943–1947.

[235] Saskia M Kelders and Hanneke Kip. 2019. Development and initial validation of a scale to measure engagement with eHealth technologies. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[236] Johannes Keller and Herbert Bless. 2008. Flow and regulatory compatibility: An experimental approach to the flow model of intrinsic motivation. *Personality and social psychology bulletin* 34, 2 (2008), 196–209.

[237] Johannes Keller, Herbert Bless, Frederik Blomann, and Dieter Kleinböhl. 2011. Physiological aspects of flow experiences: Skills-demand-compatibility effects on heart rate variability and salivary cortisol. *Journal of Experimental Social Psychology* 47, 4 (2011), 849–852.

[238] Johannes Keller and Frederik Blomann. 2008. Locus of control and the flow experience: An experimental analysis. *European Journal of Personality* 22, 7 (2008), 589–607.

[239] Johannes Keller and Anne Landhäußer. 2012. The flow model revisited. *Advances in flow research* (2012), 51–64.

[240] Shiva Khoshnoud, Federico Alvarez Igarzábal, and Marc Wittmann. 2020. Peripheral-physiological and neural correlates of the flow experience while playing video games: A comprehensive review. *PeerJ* 8 (2020), e10520.

[241] Matthew A Killingsworth and Daniel T Gilbert. 2010. A wandering mind is an unhappy mind. *Science* 330, 6006 (2010), 932–932.

[242] J Matias Kivikangas et al. 2006. Psychophysiology of flow experience: An explorative study. (2006).

[243] David Klahr. 2012. Beyond Piaget: a perspective from studies of children's problem solving abilities. (2012).

[244] Michaela Klauck, Yusuke Sugano, and Andreas Bulling. 2017. Noticeable or Distractive? A Design Space for Gaze-Contingent User Interface Notifications. In *Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems*. 1779–1786.

[245] Rebecca Kleinberger, Gershon Dublon, Joseph A Paradiso, and Tod Machover. 2015. PhoxEars: a Parabolic, Head-mounted, Orientable, Extrasensory Listening Device.. In *NIME*. 30–31.

[246] Wolfgang Klimesch. 1999. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews* 29 (1999), 169–195.

[247] Nataliya Kosmyna, Caitlin Morris, Utkarsh Sarawgi, and Pattie Maes. 2019. AttentivU: A Biofeedback System for Real-time Monitoring and Improvement of Engagement. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[248] Max Kozlov. 2022. NIH issues a seismic mandate: share data publicly. *Nature* 602, 7898 (2022), 558–559.

[249] Ariella S Kristal, Ashley V Whillans, Max H Bazerman, Francesca Gino, Lisa L Shu, Nina Mazar, and Dan Ariely. 2020. Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences* 117, 13 (2020), 7103–7107.

[250] Daniel Lakens. 2017. Impossibly hungry judges. http://daniellakens.blogspot.com/2017/07/impossibly-hungry-judges.html

[251] Saroj Lal and Ashley Craig. 2005. Reproducibility of the spectral components of the electroencephalogram during driver fatigue. *International journal of psychophysiology*

: official journal of the International Organization of Psychophysiology 55 2 (2005), 137–43.

[252] Anne Landhäußer and Johannes Keller. 2012. Flow and its affective, cognitive, and performance-related consequences. *Advances in flow research* (2012), 65–85.

[253] Ellen J Langer. 2009. *Counterclockwise: Mindful health and the power of possibility.* Ballantine Books.

[254] G Daniel Lassiter, Matthew J Lindberg, Claudia González-Vallejo, Francis S Bellezza, and Nathaniel D Phillips. 2009. The deliberation-without-attention effect: Evidence for an artifactual interpretation. *Psychological Science* 20, 6 (2009), 671–675.

[255] Raymond Lavoie, Kelley Main, and Anastasia Stuart-Edwards. 2022. Flow theory: Advancing the two-dimensional conceptualization. *Motivation and Emotion* 46, 1 (2022), 38–58.

[256] Hayley Ledger. 2013. The effect cognitive load has on eye blinking.

[257] Joseph E LeDoux. 1994. Emotion, memory and the brain. *Scientific American* 270, 6 (1994), 50–57.

[258] Joy Lee-Shi and Robert G Ley. 2022. A critique of the Dispositional Flow Scale-2 (DFS-2) and Flow State Scale-2 (FSS-2). *Frontiers in Psychology* 13 (2022), 992813.

[259] Guillaume Lemaitre, Olivier Houix, Nicolas Misdariis, and Patrick Susini. 2010. Listener expertise and sound identification influence the categorization of environmental sounds. *Journal of Experimental Psychology: Applied* 16, 1 (2010), 16.

[260] Dyani Lewis. 2021. Autocorrect errors in Excel still creating genomics headache. *Nature* (2021).

[261] James W Lewis, Frederic L Wightman, Julie A Brefczynski, Raymond E Phinney, Jeffrey R Binder, and Edgar A DeYoe. 2004. Human brain regions involved in recognizing environmental sounds. *Cerebral cortex* 14, 9 (2004), 1008–1021.

[262] Elizabeth O Lillie, Bradley Patay, Joel Diamant, Brian Issell, Eric J Topol, and Nicholas J Schork. 2011. The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalized medicine* 8, 2 (2011), 161–173.

[263] Siân Lindley, Robert Corish, Elsa Kosmack Vaara, Pedro Ferreira, and Vygandas Simbelis. 2013. Changing perspectives of time in HCI. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 3211–3214.

[264] Brian Little. 2014. *Me, myself, and us: The science of personality and the art of well-being*. Public Affairs.

[265] Vanessa M Loaiza, Kayla A Duperreault, Matthew G Rhodes, and David P McCabe. 2015. Long-term semantic representations moderate the effect of attentional refreshing on episodic memory. *Psychonomic bulletin & review* 22, 1 (2015), 274–280.

[266] German Lopez. 2017. For years, this popular test measured anyone's racial bias. but it might not work afternbsp;all. https://www.vox.com/identities/2017/3/7/14637626/implicit-association-test-racism

[267] Diana Sofía Lora Ariza, Antonio A Sánchez-Ruiz, and Pedro A González-Calero. 2019. Towards finding flow in Tetris. In *Case-Based Reasoning Research and Development: 27th International Conference, ICCBR 2019, Otzenhausen, Germany, September 8–12, 2019, Proceedings 27*. Springer, 266–280.

[268] Diana Sofía Lora-Ariza, Antonio A Sánchez-Ruiz, Pedro Antonio González-Calero, and Irene Camps-Ortueta. 2022. Measuring Control to Dynamically Induce Flow in Tetris. *IEEE Transactions on Games* 14, 4 (2022), 579–588.

[269] ZL Lu, SJ Williamson, and L Kaufman. 1992. Behavioral lifetime of human auditory sensory memory predicted by physiological measures. *Science* 258, 5088 (1992), 1668–1670.

[270] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2018. Are gans created equal? a large-scale study. *Advances in neural information processing systems* 31 (2018).

[271] Sonja Lyubomirsky, Kennon M Sheldon, and David Schkade. 2005. Pursuing happiness: The architecture of sustainable change. *Review of general psychology* 9, 2 (2005), 111–131.

[272] S Helen Ma and John D Teasdale. 2004. Mindfulness-based cognitive therapy for depression: replication and exploration of differential relapse prevention effects. *Journal of consulting and clinical psychology* 72, 1 (2004), 31.

[273] Wei Ji Ma, Masud Husain, and Paul M Bays. 2014. Changing concepts of working memory. *Nature neuroscience* 17, 3 (2014), 347.

[274] Alexis Madrigal. 2009. Scanning dead salmon in fmri machine highlights risk of Red Herrings. https://www.wired.com/2009/09/fmrisalmon/

[275] Maximilian Maier, František Bartoš, TD Stanley, David R Shanks, Adam JL Harris, and Eric-Jan Wagenmakers. 2022. No evidence for nudging after adjusting for publication bias. *Proceedings of the National Academy of Sciences* 119, 31 (2022), e2200300119.

[276] Robert Makepeace and Julien Epps. 2015. Automatic task analysis based on head movement. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015), 5167–5170.

[277] Naomi Mandel and Eric J Johnson. 2002. When web pages influence choice: Effects of visual primes on experts and novices. *Journal of consumer research* 29, 2 (2002), 235–245.

[278] Michael M Marcell, Diane Borella, Michael Greene, Elizabeth Kerr, and Summer Rogers. 2000. Confrontation naming of environmental sounds. *Journal of clinical and experimental neuropsychology* 22, 6 (2000), 830–864.

[279] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 107–110.

[280] Gloria Mark, Shamsi T Iqbal, Mary Czerwinski, Paul Johns, Akane Sano, and Yuliya Lutchyn. 2016. Email duration, batching and self-interruption: Patterns of email use on productivity and stress. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 1717–1728.

[281] Jeremy Marty-Dugas and Daniel Smilek. 2019. Deep, effortless concentration: Re-examining the flow concept and exploring relations with inattention, absorption, and personality. *Psychological research* 83, 8 (2019), 1760–1777.

[282] Sebastiaan Mathôt and Yavor Ivanov. 2019. The effect of pupil size and peripheral brightness on detection and discrimination performance. *PeerJ* 7 (2019), e8220.

[283] Maurizio Mauri, Pietro Cipresso, Anna Balgera, Marco Villamira, and Giuseppe Riva. 2011. Why is Facebook so successful? Psychophysiological measures describe a core flow state while using Facebook. *Cyberpsychology, Behavior, and Social Networking* 14, 12 (2011), 723–731.

[284] Brian Mayton, Gershon Dublon, Spencer Russell, Evan F Lynch, Don Derek Haddad, Vasant Ramasubramanian, Clement Duhart, Glorianna Davenport, and Joseph A Paradiso. 2017. The Networked Sensory Landscape: Capturing and Experiencing Ecological Change Across Scales. *Presence: Teleoperators and Virtual Environments* 26, 2, 182–209.

[285] Stephen Ed McAdams and Emmanuel Ed Bigand. 1993. Thinking in sound: The cognitive psychology of human audition.. In *Based on the fourth workshop in the Tutorial Workshop series organized by the Hearing Group of the French Acoustical Society.* Clarendon Press/Oxford University Press.

[286] Randy J McCarthy, John J Skowronski, Bruno Verschuere, Ewout H Meijer, Ariane Jim, Katherine Hoogesteyn, Robin Orthey, Oguz A Acar, Balazs Aczel, Bence E Bakos, et al. 2018. Registered replication report on Srull and Wyer (1979). *Advances in Methods and Practices in Psychological Science* 1, 3 (2018), 321–336.

[287] Author Alison McCook. 2017. "I placed too much faith in underpowered studies:"

nobel prize winner admits mistakes. https://retractionwatch.com/2017/02/20/placed-much-faith-underpowered-studies-nobel-prize-winner-admits-mistakes/

[288] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. 2013. Summary statistics in auditory perception. *Nature neuroscience* 16, 4 (2013), 493.

[289] Richard McElreath. 2018. *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

[290] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*. 18–25.

[291] Greg McKeown. 2014. *Essentialism: The disciplined pursuit of less*. Currency.

[292] Douglas L Medin. 2011. The case of the invisible experimenter (s). *APS Observer* 24 (2011).

[293] David E Melnikoff, Ryan W Carlson, and Paul E Stillman. 2022. A computational theory of the subjective experience of flow. *Nature communications* 13, 1 (2022), 2252.

[294] Michelle N. Meyer and Christopher Chabris. 2014. Why Psychologists' Food Fight Matters. https://slate.com/technology/2014/07/replication-controversy-in-psychology-bullying-file-drawer-effect-blog-posts-repligate
html

[295] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[296] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[297] Talya Miron-Shatz. 2009. Evaluating multiepisode events: Boundary conditions for the peak-end rule. *Emotion* 9, 2 (2009), 206.

[298] Anjali Mishra. 2009. Positivity, by Barbara Fredrickson.

[299] C Thomas Mitchell and Roy Davis. 1987. The perception of time in scale model environments. *Perception* 16, 1 (1987), 5–16.

[300] Helen F Mitchell and Raymond AR MacDonald. 2011. Remembering, Recognizing and Describing Singers' Sound Identities. *Journal of New Music Research* 40, 1 (2011), 75–80.

[301] Tota Mizuno and Yuichiro Kume. 2015. Development of a Glasses-Like Wearable Device to Measure Nasal Skin Temperature. In *HCI*.

[302] Peter CM Molenaar. 2004. A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement* 2, 4 (2004), 201–218.

[303] Peter CM Molenaar and Cynthia G Campbell. 2009. The new person-specific paradigm in psychology. *Current directions in psychological science* 18, 2 (2009), 112–117.

[304] AC Moller, M Csikszentmihalyi, J Nakamura, and EL Deci. 2007. Developing an experimental induction of flow. In *Poster presented at the Society for Personality and Social Psychology conference, Memphis, TN, USA*.

[305] Arlen C Moller, Brian P Meier, and Robert D Wall. 2010. Developing an experimental induction of flow: Effortless action in the lab. *Effortless attention: A new perspective in the cognitive science of attention and action* (2010), 191–204.

[306] Giovanni B Moneta. 2010. Flow in work as a function of trait intrinsic motivation, opportunity for creativity in the job, and work engagement. *The Handbook of employee engagement: Perspectives, issues, research and practice* (2010), 262–269.

[307] Giovanni B Moneta. 2012. On the measurement and conceptualization of flow. *Advances in flow research* (2012), 23–50.

[308] Giovanni B Moneta and Mihaly Csikszentmihalyi. 1996. The effect of perceived challenges and skills on the quality of subjective experience. *Journal of personality* 64, 2 (1996), 275–310.

[309] Stephen Monsell. 2003. Task Switching. *Trends in Cognitive Sciences* 7, 3 (2003), 134–140.

[310] Christine Moorman, Nader Tavassoli, and Megan Ryan. 2022. Why marketers are returning to traditional advertising. https://hbr.org/2022/04/why-marketers-are-returning-to-traditional-advertising

[311] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 681–699.

[312] Jeanne Nakamura, Mihaly Csikszentmihalyi, CR Snyder, and Shane J Lopez. 2009. Oxford handbook of positive psychology. *Flow Theory an Research* (2009), 195–206.

[313] Makoto Nishihara, Koji Inui, Tomoyo Morita, Minori Kodaira, Hideki Mochizuki, Naofumi Otsuru, Eishi Motomura, Takahiro Ushida, and Ryusuke Kakigi. 2014. Echoic memory: investigation of its temporal resolution by auditory offset cortical responses. *PloS one* 9, 8 (2014), e106553.

[314] A Imran Nordin, Jaron Ali, Aishat Animashaun, Josh Asch, Josh Adams, and Paul Cairns. 2013. Attention, time perception and immersion in games. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 1089–1094.

[315] Donald A Norman. 1999. Affordance, conventions, and design. *interactions* 6, 3 (1999), 38–43.

[316] Nargess Nourbakhsh, Yuhuai Wang, and Fang Chen. 2013. GSR and Blink Features for Cognitive Load Classification. In *INTERACT*.

[317] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[318] William Odom, Siân Lindley, Larissa Pschetz, Vasiliki Tsaknaki, Anna Vallgårda, Mikael Wiberg, and Daisy Yoo. 2018. Time, temporality, and slowness: Future directions for design research. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*. 383–386.

[319] Jukka-Pekka Onnela. 2021. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology* 46, 1 (2021), 45–54.

[320] Jukka-Pekka Onnela, Caleb Dixon, Keary Griffin, Tucker Jaenicke, Leila Minowada, Sean Esterkin, Alvin Siu, Josh Zagorsky, and Eli Jones. 2021. Beiwe: A data collection platform for high-throughput digital phenotyping. *Journal of Open Source Software* 6, 68 (2021), 3417.

[321] Andreas Ortmann. 2021. On the foundations of behavioral and experimental economics. *a modern guide to philosophy of economics* (2021), 157–181.

[322] Magda Osman, Scott McLachlan, Norman Fenton, Martin Neil, Ragnar Löfstedt, and Björn Meder. 2020. Learning from behavioural changes that fail. *Trends in Cognitive Sciences* 24, 12 (2020), 969–980.

[323] Amani Yousef Owda, Neil Salmon, Nacer Ddine Rezgui, and Sergiy Shylo. 2017. Millimetre wave radiometers for medical diagnostics of human skin. In *2017 IEEE SENSORS*. IEEE, 1–3.

[324] Michael O'donnell, Leif D Nelson, Evi Ackermann, Balazs Aczel, Athfah Akhtar, Silvio Aldrovandi, Nasseem Alshaif, Ronald Andringa, Mark Aveyard, Peter Babincak, et al. 2018. Registered replication report: Dijksterhuis and van Knippenberg (1998). *Perspectives on Psychological Science* 13, 2 (2018), 268–294.

[325] Pragya Pandey and Supriya Ray. 2021. Influence of the Location of a Decision Cue on the Dynamics of Pupillary Light Response. *Frontiers in human neuroscience* 15 (2021).

[326] Rafal Paprocki and Artem Lenskiy. 2017. What does eye-blink rate variability dy-

namics tell us about cognitive performance? *Frontiers in human neuroscience* 11 (2017), 620.

[327] Viral Parekh, Pin Sym Foong, Shengdong Zhao, and Ramanathan Subramanian. 2018. Aveid: automatic video system for measuring engagement in dementia. In *23rd International Conference on Intelligent User Interfaces*. 409–413.

[328] AC Parks and HA Victor. 2006. Are positive and negative moods causesor correlates of flow. In *Poster presented at the 18th Annual Convention of the American Psychological Society, New York, NY, USA*.

[329] Harold Pashler, Noriko Coburn, and Christine R Harris. 2012. Priming of social distance? Failure to replicate effects on social and food judgments. (2012).

[330] Harold Pashler, Doug Rohrer, and Christine R Harris. 2013. Can the goal of honesty be primed? *Journal of Experimental Social Psychology* 49, 6 (2013), 959–964.

[331] B Keith Payne, Jon A Krosnick, Josh Pasek, Yphtach Lelkes, Omair Akhtar, and Trevor Tompson. 2010. Implicit and explicit prejudice in the 2008 American presidential election. *Journal of Experimental Social Psychology* 46, 2 (2010), 367–374.

[332] John W Payne, Adriana Samper, James R Bettman, and Mary Frances Luce. 2008. Boundary conditions on unconscious thought in complex decision making. *Psychological Science* 19, 11 (2008), 1118–1123.

[333] Corinna Peifer. 2012. Psychophysiological correlates of flow-experience. *Advances in flow research* (2012), 139–164.

[334] Maquita Peters. 2019. 'Power Poses' Co-Author: 'I Do Not Believe The Effects Are Real'. https://www.npr.org/2016/10/01/496093672/power-poses-co-author-i-do-not-believe-the-effects-are-real

[335] BF Petrie. 1996. Environment is not the most important variable in determining oral morphine consumption in Wistar rats. *Psychological reports* 78, 2 (1996), 391–400.

[336] Rosalind W Picard, Szymon Fedor, and Yadid Ayzenberg. 2016. Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion review* 8, 1 (2016), 62–75.

[337] Martin Pielot and Rodrigo de Oliveira. 2013. Peripheral vibro-tactile displays. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*. 1–10.

[338] Steven Pinker. 2003. *The blank slate: The modern denial of human nature*. Penguin.

[339] Kelsey Piper. 2019. The conversation about diversity in Tech is getting hijacked by Bad Research. https://www.vox.com/2019/2/20/18232762/gender-diversity-tech-bad-research-recruiting-new-york-times

[340] L Pizzamiglio, T Aprile, G Spitoni, S Pitzalis, E Bates, S D'amico, and F Di Russo. 2005. Separate neural systems for processing action-or non-action-related sounds. *Neuroimage* 24, 3 (2005), 852–861.

[341] Russell A Poldrack, Timothy O Laumann, Oluwasanmi Koyejo, Brenda Gregory, Ashleigh Hover, Mei-Yen Chen, Krzysztof J Gorgolewski, Jeffrey Luci, Sung Jun Joo, Ryan L Boyd, et al. 2015. Long-term neural and physiological phenotyping of a single human. *Nature communications* 6, 1 (2015), 8885.

[342] Francesco Pompei. [n. d.]. Temporal artery temperature detector.

[343] Michael I Posner. 2016. Orienting of attention: Then and now. *Quarterly journal of experimental psychology* 69, 10 (2016), 1864–1875.

[344] Florian Prinz, Thomas Schlange, and Khusru Asadullah. 2011. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery* 10, 9 (2011), 712–712.

[345] Han Qiao, Yazhe Li, Jingyu Zhang, E Xiaotian, Xiangying Zou, You Jiang, Lin Xiong, and Xianghong Sun. 2018. The peak-end effects in controllers' mental workload evaluation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 62. SAGE Publications Sage CA: Los Angeles, CA, 87–91.

[346] Markus Quirin, Miguel Kazén, and Julius Kuhl. 2009. When nonsense sounds happy or helpless: the implicit positive and negative affect test (IPANAT). *Journal of personality and social psychology* 97, 3 (2009), 500.

[347] Mirko Raca, Lukasz Kidzinski, and Pierre Dillenbourg. 2015. Translating head motion into attention-towards processing of student's body-language. In *Proceedings of the 8th international conference on educational data mining*.

[348] David Ramsay, Ishwarya Ananthabhotla, and Joseph Paradiso. 2019. The Intrinsic Memorability of Everyday Sounds. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society.

[349] David Ramsay and Joe Paradiso. 2022. Peripheral Light Cues as a Naturalistic Measure of Focus. In *ACM International Conference on Interactive Media Experiences*. 375–380.

[350] David Ramsay and Joseph A Paradiso. 2019. YourAd: A User Aligned, Personal Advertising System. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[351] David B Ramsay, Ishwarya Ananthabhotla, and Joseph A Paradiso. 2019. The Intrinsic Memorability of Everyday Sounds. In *AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society.

[352] Fabian Ramseyer and Wolfgang Tschacher. 2014. Nonverbal synchrony of head-and body-movement in psychotherapy: different signals have different associations with outcome. *Frontiers in psychology* 5 (2014), 979.

[353] Eva Ranehill, Anna Dreber, Magnus Johannesson, Susanne Leiberg, Sunhae Sul, and Roberto A Weber. 2015. Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological science* 26, 5 (2015), 653–656.

[354] Isabelle Régner, Catherine Thinus-Blanc, Agnès Netter, Toni Schmader, and Pascal Huguet. 2019. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature human behaviour* 3, 11 (2019), 1171–1179.

[355] Umair Rehman and Shi Cao. 2017. Augmented-Reality-Based Indoor Navigation: A Comparative Analysis of Handheld Devices Versus Google Glass. *IEEE Transactions on Human-Machine Systems* 47 (2017), 140–151.

[356] Stephen D Reicher, S Alexander Haslam, and Arthur G Miller. 2014. What makes a person a perpetrator? The intellectual, moral, and methodological arguments for revisiting Milgram's research on the influence of authority. *Journal of Social Issues* 70, 3 (2014), 393–408.

[357] F Rheinberg and R Vollmeyer. 2003. Flow experience in a computer game under experimentally controlled conditions. *ZEITSCHRIFT FUR PSYCHOLOGIE-JOURNAL OF PSYCHOLOGY* 211, 4 (2003), 161–170.

[358] Falko Rheinberg, Regina Vollmeyer, and Stefan Engeser. 2003. Die erfassung des flow-erlebens. (2003).

[359] Grant J Rich. 2018. Positive psychology and humanistic psychology: Evil twins, sibling rivals, distant cousins, or something else? *Journal of Humanistic Psychology* 58, 3 (2018), 262–283.

[360] F Dan Richard, Charles F Bond Jr, and Juli J Stokes-Zoota. 2003. One hundred years of social psychology quantitatively described. *Review of general psychology* 7, 4 (2003), 331–363.

[361] Gaël Richard, Shiva Sundaram, and Shrikanth Narayanan. 2013. An overview on perceptually motivated audio indexing and classification. *Proc. IEEE* 101, 9 (2013), 1939–1954.

[362] Amber Rithalia, Catriona McDaid, Sara Suekarran, Lindsey Myers, and Amanda Sowden. 2009. Impact of presumed consent for organ donation on donation rates: a systematic review. *Bmj* 338 (2009).

[363] Katja Rogers, Maximilian Milo, Michael Weber, and Lennart E Nacke. 2020. The Potential Disconnect between Time Perception and Immersion: Effects of Music on VR Player Experience. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. 414–426.

[364] Sergio Romero, Miguel Angel Mañanas, and Manel J. Barbanoj. 2008. A comparative study of automatic techniques for ocular artifact reduction in spontaneous EEG signals based on clinical target variables: A simulation case. *Computers in biology and medicine* 38 3 (2008), 348–60.

[365] Rachel L Roper. 2019. Does gender bias still affect women in science? *Microbiology and Molecular Biology Reviews* 83, 3 (2019), e00018–19.

[366] Soha Rostaminia, A. Mayberry, D. Ganesan, Benjamin M Marlin, and Jeremy Gummeson. 2017. iLid: Low-power Sensing of Fatigue and Drowsiness Measures on a Computational Eyeglass. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 1 2 (2017).

[367] Thijs Roumen, Simon T Perrault, and Shengdong Zhao. 2015. Notiring: A comparative study of notification channels for wearable interactive rings. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2497–2500.

[368] Chris J Rudge. 2018. Organ donation: opting in or opting out? , 62–63 pages.

[369] Spencer Franklin Russell. 2020. *Resynthesizing Volumetric Soundscapes: Low-rank Subspace Methods for Soundfield Estimation and Reconstruction*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[370] Timothy Sanders and Paul Cairns. 2010. Time perception, immersion and music in videogames. *Proceedings of HCI 2010 24* (2010), 160–167.

[371] Lawrence J Sanna, Edward C Chang, Paul M Miceli, and Kristjen B Lundberg. 2011. RETRACTED: Rising up to higher virtues: Experiencing elevated physical height uplifts prosocial actions.

[372] Akane Sano, Rosalind W Picard, and Robert Stickgold. 2014. Quantitative analysis of wrist electrodermal activity during sleep. *International Journal of Psychophysiology* 94, 3 (2014), 382–389.

[373] Phattarapong Sawangjai, Supanida Hompoonsup, Pitshaporn Leelaarporn, Supavit Kongwudhikunakorn, and Theerawit Wilaiprasitporn. 2020. Consumer Grade EEG Measuring Sensors as Research Tools: A Review. *IEEE Sensors Journal* 20 (2020), 3996–4024.

[374] Benjamin Scheibehenne, Rainer Greifeneder, and Peter M Todd. 2010. Can there ever be too many options? A meta-analytic review of choice overload. *Journal of consumer research* 37, 3 (2010), 409–425.

[375] Ulrich Schimmack. 2017. Hidden Figures: Replication Failures in the Stereotype Threat Literature. https://replicationindex.com/2017/04/07/hidden-figures-replication-failures-in-the-stereotype-threat-literature/

[376] Ulrich Schimmack. 2020. A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie Canadienne* 61, 4 (2020), 364.

[377] U Schimmack. 2020. Personality science: The science of human diversity. TopHat.

[378] Ulrich Schimmack. 2021. The Implicit Association Test: A method in search of a construct. *Perspectives on Psychological Science* 16, 2 (2021), 396–414.

[379] Ulrich Schimmack. 2023. Why the journal of personality and social psychology should retract article doi: 10.1037/A0021524 "feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect" by Daryl J. Bem. https://replicationindex.com/2018/01/05/bem-retraction/

[380] Annett Schirmer, Yong Hao Soh, Trevor B Penney, and Lonce Wyse. 2011. Perceptual and conceptual priming of environmental sounds. *Journal of cognitive neuroscience* 23, 11 (2011), 3241–3253.

[381] Julia Schüler. 2012. The dark side of the moon. *Advances in flow research* (2012), 123–137.

[382] Richard Schulz and Barbara H Hanusa. 1978. Long-term effects of control and predictability-enhancing interventions: Findings and ethical issues. *Journal of personality and social psychology* 36, 11 (1978), 1194.

[383] Charles R Schuster and Travis Thompson. 1969. Self administration of and behavioral dependence on drugs. *Annual review of pharmacology* 9, 1 (1969), 483–502.

[384] Barry Schwartz. 2004. The paradox of choice: Why more is less. *New York* (2004).

[385] Barry Schwartz. 2014. Is the famous "paradox of choice" a myth? https://www.pbs.org/newshour/economy/is-the-famous-paradox-of-choic

[386] Martin EP Seligman. 2002. *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. Simon and Schuster.

[387] Marta Serra-Garcia and Uri Gneezy. 2021. Nonreplicable publications are cited more than replicable ones. *Science advances* 7, 21 (2021), eabd1705.

[388] Fred Shaffer and Jay P Ginsberg. 2017. An overview of heart rate variability metrics and norms. *Frontiers in public health* (2017), 258.

[389] Alexander Shtuchkin. 2020. DIY Position Tracking using HTC Vive's Lighthouse. https://github.com/ashtuchkin/vive-diy-position-sensor.

[390] Mark Simmonds. 2015. Quantifying the risk of error when interpreting funnel plots. *Systematic reviews* 4 (2015), 1–7.

[391] Daniel J Simons, Steven L Franconeri, and Rebecca L Reimer. 2000. Change blindness in the absence of a visual disruption. *Perception* 29, 10 (2000), 1143–1154.

[392] Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. 2014. P-curve: a key to the file-drawer. *Journal of experimental psychology: General* 143, 2 (2014), 534.

[393] Jesse Singal. [n. d.]. Positive psychology goes to war. https://www.chronicle.com/article/positive-psychology-goes-to-war

[394] Jesse Singal. 2017. Psychology's favorite tool for measuring racism isn't up to the job. https://www.thecut.com/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html

[395] Jesse Singal. 2021. *The quick fix: Why fad psychology can't cure our social ills*. Farrar, Straus and Giroux.

[396] Victoria F Sisk, Alexander P Burgoyne, Jingze Sun, Jennifer L Butler, and Brooke N Macnamara. 2018. To what extent and under which circumstances are growth mindsets important to academic achievement? Two meta-analyses. *Psychological science* 29, 4 (2018), 549–571.

[397] Rose L Siuta and Mindy E Bergman. 2019. Sexual harassment in the workplace. In *Oxford research encyclopedia of business and management*.

[398] Gary Smith. 2016. Hurricane names: A bunch of hot air? *Weather and Climate Extremes* 12 (2016), 80–84.

[399] Richard W Smith. 2009. In search of an optimal peer review system. *Journal of Participatory Medicine* 1, 1 (2009), e13.

[400] Vernon L Smith. 2003. Constructivist and ecological rationality in economics. *American economic review* 93, 3 (2003), 465–508.

[401] Joel S Snyder and Mounya Elhilali. 2017. Recent advances in exploring the neural underpinnings of auditory scene perception. *Annals of the New York Academy of Sciences* 1396, 1 (2017), 39–55.

[402] Robert Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge.. In *AAAI*. 4444–4451.

[403] Branka Spehar and Caleb Owens. 2012. When do luminance changes capture attention? *Attention, Perception, & Psychophysics* 74, 4 (2012), 674–690.

[404] Richard St Jean and Carrie MacLeod. 1983. Hypnosis, absorption, and time perception. *Journal of Abnormal Psychology* 92, 1 (1983), 81.

[405] Wiktor Stadnik and Ziemowit Nowak. 2018. The impact of web pages' load time on the conversion rate of an e-commerce platform. In *Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology–ISAT 2017: Part I*. Springer, 336–345.

[406] Claude M Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology* 69, 5 (1995), 797.

[407] Jacob Stern. 2023. An unsettling hint at how much fraud could exist in science. https://www.theatlantic.com/science/archive/2023/08/gino-ariely-data-fraud-allegations/674891/

[408] John A. Stern, Larry Walrath, and Rebecca Newberger Goldstein. 1984. The endogenous eyeblink. *Psychophysiology* 21 1 (1984), 22–33.

[409] Victoria Stern. 2015. A short history of the rise, fall and rise of subliminal messaging. https://www.scientificamerican.com/article/a-short-history-of-the-rise-fall-and-rise-of-subliminal-messaging/

[410] Jonathan AC Sterne, Betsy Jane Becker, and Matthias Egger. 2005. The funnel plot. *Publication bias in meta-analysis: Prevention, assessment and adjustments* (2005), 73–98.

[411] Chess Stetson, Matthew P Fiesta, and David M Eagleman. 2007. Does time really slow down during a frightening event? *PloS one* 2, 12 (2007), e1295.

[412] Rainer Stiefelhagen. 2002. Tracking focus of attention in meetings. *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces* (2002), 273–280.

[413] Maja Stikic, Chris Berka, Daniel J. Levendowski, Roberto F. Rubio, Veasna Tan, Stephanie Korszen, Douglas Barba, and David Wurzer. 2014. Modeling temporal sequences of cognitive state changes based on a combination of EEG-engagement, EEG-workload, and heart rate metrics. *Frontiers in Neuroscience* 8 (2014).

[414] Kristina Stojmenova and Jaka Sodnik. 2018. Detection-response task—uses and limitations. *Sensors* 18, 2 (2018), 594.

[415] Arthur A Stone, Joan E Broderick, Alan T Kaell, Philippe AEG DelesPaul, and Laura E Porter. 2000. Does the peak-end phenomenon observed in laboratory pain studies apply to real-world pain in rheumatoid arthritics? *The Journal of Pain* 1, 3 (2000), 212–217.

[416] Fritz Strack, Leonard L Martin, and Sabine Stepper. 1988. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of personality and social psychology* 54, 5 (1988), 768.

[417] Tim Claudius Stratmann, Andreas Löcken, Uwe Gruenefeld, Wilko Heuten, and Susanne Boll. 2018. Exploring vibrotactile and peripheral cues for spatial attention guidance. In *Proceedings of the 7th ACM International Symposium on Pervasive Displays*. 1–8.

[418] J Ridley Stroop. 1992. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General* 121, 1 (1992), 15.

[419] Bas C Stunnenberg, Jaap Deinum, Tom Nijenhuis, Frans Huysmans, Gert Jan van der Wilt, Baziel GM van Engelen, and Frans van Agt. 2020. N-of-1 trials: evidence-based clinical care or medical research that requires IRB approval? A practical flowchart based on an ethical framework. In *Healthcare*, Vol. 8. MDPI, 49.

[420] Kate Sweeny, Kyla Rankin, Xiaorong Cheng, Lulu Hou, Fangfang Long, Yao Meng, Lilian Azer, Renlai Zhou, and Weiwei Zhang. 2020. Flow in the time of COVID-19: Findings from China. *PloS one* 15, 11 (2020), e0242043.

[421] B. Tag, Andrew W. Vargo, Aman Gupta, G. Chernyshov, K. Kunze, and Tilman Dingler. 2019. Continuous Alertness Assessments: Using EOG Glasses to Unobtrusively Monitor Fatigue Levels In-The-Wild. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[422] Leena Tähkämö, Timo Partonen, and Anu-Katriina Pesonen. 2019. Systematic review of light exposure impact on human circadian rhythm. *Chronobiology international* 36, 2 (2019), 151–170.

[423] Sundeep Teki, Manon Grube, and Timothy D Griffiths. 2012. A unified model of time perception accounts for duration-based and beat-based timing mechanisms. *Frontiers in integrative neuroscience* 5 (2012), 90.

[424] Alexander Tekles, Katrin Auspurg, and Lutz Bornmann. 2022. Same-gender citations do not indicate a substantial gender homophily bias. *Plos one* 17, 9 (2022), e0274810.

[425] Auke Tellegen and Gilbert Atkinson. 1974. Openness to absorbing and self-altering experiences (" absorption"), a trait related to hypnotic susceptibility. *Journal of abnormal psychology* 83, 3 (1974), 268.

[426] Texas Instruments 2011. *TMP006/B Infrared Thermopile Sensor in Chip-Scale Package*. Texas Instruments. Rev. C.

[427] Richard H Thaler. 2015. *Misbehaving: The making of behavioral economics*. WW Norton & Company.

[428] Chinchu Thomas and Dinesh Babu Jayagopi. 2017. Predicting student engagement in classrooms using facial behavioral cues. In *Proceedings of the 1st ACM SIGCHI international workshop on multimodal interaction for education*. 33–40.

[429] Paul Thomas, Heather O'Brien, and Tom Rowlands. 2016. Measuring engagement with online forms. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 325–328.

[430] Sven Thönes and Daniel Oberfeld. 2015. Time perception in depression: A meta-analysis. *Journal of affective disorders* 175 (2015), 359–372.

[431] Robert L Thorndike. 1968. Reviews: Rosenthal, Robert, and Jacobson, Lenore. Pygmalion in the Classroom. New York: Holt, Rinehart and Winston, 1968. 240+ xi pp. 3.95. *American Educational Research Journal*5, 4(1968), 708 − −711.

[432] Jason Tipples. 2008. Negative emotionality influences the effects of emotion on time perception. *Emotion* 8, 1 (2008), 127.

[433] Simon Tobin, Nicolas Bisson, and Simon Grondin. 2010. An ecological approach to prospective and retrospective timing of long durations: a study involving gamers. *PloS one* 5, 2 (2010), e9271.

[434] Elizabeth B Torres, Joe Vero, and Richa Rai. 2018. Statistical platform for individualized behavioral analyses using biophysical micro-movement spikes. *Sensors* 18, 4 (2018), 1025.

[435] Michel Treisman. 1963. Temporal discrimination and the indifference interval: Implications for a model of the" internal clock". *Psychological Monographs: General and Applied* 77, 13 (1963), 1.

[436] Civil Rights Division United States Department of Justice. 2015. Investigation of the Ferguson Police Department.

[437] Chandan J Vaidya, Margaret Zhao, John E Desmond, and John DE Gabrieli. 2002. Evidence for cortical encoding specificity in episodic memory: memory-induced re-activation of picture processing areas. *Neuropsychologia* 40, 12 (2002), 2136–2143.

[438] Nicholas T Van Dam, Marieke K Van Vugt, David R Vago, Laura Schmalzl, Clifford D Saron, Andrew Olendzki, Ted Meissner, Sara W Lazar, Catherine E Kerr, Jolie Gorchov, et al. 2018. Mind the hype: A critical evaluation and prescriptive agenda for research on mindfulness and meditation. *Perspectives on psychological science* 13, 1 (2018), 36–61.

[439] Vincent T Van Hees, Lukas Gorzelniak, Emmanuel Carlos Dean León, Martin Eder, Marcelo Pias, Salman Taherian, Ulf Ekelund, Frida Renström, Paul W Franks, Alexander Horsch, et al. 2013. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PloS one* 8, 4 (2013), e61691.

[440] Thijs Verwijmeren, Johan C Karremans, Wolfgang Stroebe, and Daniël HJ Wigboldus. 2011. The workings and limits of subliminal advertising: The role of habits. *Journal of consumer psychology* 21, 2 (2011), 206–213.

[441] Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. 2009. Voodoo correlations in social neuroscience. *Perspectives on psychological Science* 4, 3 (2009), 274–290.

[442] E-J Wagenmakers, Titia Beek, Laura Dijkhoff, Quentin F Gronau, A Acosta, RB Adams Jr, DN Albohn, ES Allard, Stephen D Benning, E-M Blouin-Hudon, et al. 2016. Registered replication report: strack, martin, & stepper (1988). *Perspectives on Psychological Science* 11, 6 (2016), 917–928.

[443] Garrick L. Wallstrom, Robert E. Kass, Anita Miller, Jeffrey F. Cohn, and Nathan A. Fox. 2004. Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods. *International journal of psychophysiology : official journal of the International Organization of Psychophysiology* 53 2 (2004), 105–19.

[444] R Warne. 2020. The one variable that makes growth mindset interventions work.

[445] Russell T. Warne. 2020. The $67.5 million wasted on Stereotype Threat Research. https://russellwarne.com/2020/08/10/the-67-5-million-wasted-on-stereotype-threat-research/

[446] Laurent Waroquier, David Marchiori, Olivier Klein, and Axel Cleeremans. 2009. Methodological pitfalls of the unconscious thought paradigm. *Judgment and Decision Making* 4, 7 (2009), 601–610.

[447] SV Wass, K de Barbaro, and K Clackson. 2015. Tonic and phasic co-variation of peripheral arousal indices in infants. *Biological Psychology* 111 (2015), 26–39.

[448] Ronald L Wasserstein and Nicole A Lazar. 2016. The ASA statement on p-values: context, process, and purpose. , 129–133 pages.

[449] Ronald L Wasserstein, Allen L Schirm, and Nicole A Lazar. 2019. Moving to a world beyond "p¡ 0.05". , 19 pages.

[450] Alan S Waterman. 2013. The humanistic psychology–positive psychology divide: Contrasts in philosophical foundations. *American Psychologist* 68, 3 (2013), 124.

[451] Tyler W Watts, Greg J Duncan, and Haonan Quan. 2018. Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological science* 29, 7 (2018), 1159–1177.

[452] Nancy J Wei, Bryn Dougherty, Aundria Myers, and Sherif M Badawy. 2018. Using Google Glass in Surgical Settings: Systematic Review. *JMIR mHealth and uHealth* 6 (2018).

[453] Mark Weiser and John Seely Brown. 1996. Designing calm technology. *PowerGrid Journal* 1, 1 (1996), 75–85.

[454] Carmela A White, Bob Uttl, and Mark D Holder. 2019. Meta-analyses of positive psychology interventions: The effects are much smaller than previously reported. *PloS one* 14, 5 (2019), e0216588.

[455] István Winkler, Susan L Denham, and Israel Nelken. 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends in cognitive sciences* 13, 12 (2009), 532–540.

[456] Christoph Witzel and KR Gegenfurtner. 2015. Chromatic contrast sensitivity. *Encyclopedia of color science and technology* (2015), 1–7.

[457] Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41 (2019), 135–163.

[458] Kevin JP Woods, Max H Siegel, James Traer, and Josh H McDermott. 2017. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics* 79, 7 (2017), 2064–2072.

[459] David S Yeager, Paul Hanselman, Gregory M Walton, Jared S Murray, Robert Crosnoe, Chandra Muller, Elizabeth Tipton, Barbara Schneider, Chris S Hulleman, Cintia P Hinojosa, et al. 2019. A national experiment reveals where a growth mindset improves achievement. *Nature* 573, 7774 (2019), 364–369.

[460] Mert Yildiz and Aykut Coşkun. 2020. Time Perceptions as a Material for Designing New Representations of Time. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.

[461] Ed Yong. 2012. Nobel laureate challenges psychologists to clean up their act. *Nature* 490 (2012), 7418.

[462] Richard A Young, Li Hsieh, and Sean Seaman. 2013. The tactile detection response task: preliminary validation for measuring the attentional effects of cognitive load. (2013).

[463] Fadilla Zennifa, Junko Ide, Yukio Noguchi, and Keiji Iramina. 2015. Monitoring of cognitive state on mental retardation child using EEG, ECG and NIRS in four years study. *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015), 6610–6613.

[464] Shoshana Zuboff. 2020. You are now remotely controlled. https://www.nytimes.com/2020/01/24/opinion/sunday/surveillance-capitalism.html