Abstract

Moral and Affective Psychophysics:
Or, How (Not) to Study Subjective Magnitude

Vladimir Chituc

2024

People can feel more or less happy, hold opinions with more or less conviction, and find some criminal offenses to be more or less heinous. Though all of these examples concern something that is inherently subjective, it nonetheless seems as if each has a *magnitude*. Upon reflection, however, this raises two distinct puzzles, one practical and the other theoretical. Practically, how can we accurately measure these magnitudes, given their inherently subjective nature? And theoretically, magnitudes involve “more than” or “less than” judgments of quantity, but subjective stimuli provide no obvious answer to the question: “quantities of *what*?” So how can we say that subjective stimuli have a magnitude in the first place? This dissertation aims to answer these questions, and it does so by adapting methods and insights that were first developed to answer analogous questions about perceptual magnitudes. Across three case studies, I highlight and try to solve three salient problems involving labeled (e.g. Likert) scales: their *interpersonal relativity* (i.e. meaning different things to different people), their *intrapersonal elasticity* (i.e. meaning different things to the same person at different times), and their *nonlinearity* (i.e. compressing values in the upper range of the scale). In the first case study, I find that the interpersonal relativity of these scales can produce illusory differences in emotional intensity — suggesting, for example, that women experience anger more intensely than do men. In the second case study, I find that the intrapersonal elasticity of these scales produces absurd results in between-subjects experiments — showing, for example, that a particularly cruel prank seems just as immoral as (if not morally worse than) an internationally recognized war crime. And in my last case study, I find that the nonlinearity of these scales produces ratings that are a logarithmic compression of subjective moral magnitude, thus systematically underestimating moral judgment. In all three cases, I go on to demonstrate how each of these problems has a straightforward solution that can be easily adapted from work in sensory psychophysics. Collectively, these findings provide clear methodological lessons aimed toward improving psychological measurement, while also bearing on some of the most basic questions involving how the mind represents magnitude.

Moral and Affective Psychophysics:
Or, How (Not) to Study Subjective Magnitude

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

By
Vladimir Chituc

Dissertation Director: Brian J. Scholl

December 2024

For my mom and my dog...

...and for S.S. Stevens.

Contents

List of Figures

List of Tables

Acknowledgments

This is the dissertation that I dreamed of writing, and my ability to do so has fully depended on the kindness and generosity of an implausibly long list of people, each of whom was willing to make their life harder than it needed to be, solely for my benefit. To each and every one one of those people — and to the friends, mentors, and colleagues who have shared with me their time, wisdom, and support over the years — Thank you.

To Brian Scholl, who saw in this work the same promise that I did, oftentimes with more clarity than I did (for that alone I'll always be grateful). Because of you, this dissertation has sharper ideas, funnier jokes, and much more clever chapter titles. Thank you for your careful guidance and unwavering kindness, your reassuring sense of scale and perspective, and your willingness to spend so much of your sabbatical working late to see this through. I hope that you had even half as much fun working together as I did.

To Paul Bloom, for being my earliest cheerleader and most thoughtful critic. Since first introducing me to moral psychology, you've been there to moderate my excesses and polish my roughest drafts (my good talks all started as bad talks in your lab). But more than that, thank you for being an unfaltering model of the scholar I hope to be: lucid and compelling, supportive and intellectually generous, and provocative without being mean-spirited.

To Josh Knobe, who has done more to cultivate fledgling ideas than anyone else I know (and purely for the love of it, too). Thank you for your warmth and encouragement, and for studying the mind with such a contagious sense of earnestness and curiosity. To Sam McDougle, whose support and enthusiasm provided a needed boost of confidence during the early stages of this work. Thank you for suggesting the best papers on topics I've never heard of (and the killer James quote) — but more than that, thank you for being a friend.

And to Maria Gendron, for the inspiring and dedicated care with which you treat all of your students, even the fosters. Thank you for your sharp, attentive feedback (and for your patient willingness to help an interloper in affective science). You're an incredible researcher and a better person, and I can't thank you enough for setting all of this in motion.

To Yale Psychology — I feel impossibly lucky to have spent more than a decade of my life in such a vibrant, fascinating, and collaborative department. To Molly Crockett, for being a brilliant researcher and the reason I came back to Yale. I couldn't have asked for a more supportive advisor during the worst of a global pandemic, and Princeton is lucky to have you. I'm deeply grateful to Laurie Santos for being a phenomenal mentor to so many students (and for introducing me to research in psychology). Thank you for giving me a home in Silliman (and with Toad). To Mark Fedyk, Chaz Firestone, Ellen Furlong, Shelly Kagan, Eric Mandelbaum, Jonathan Phillips, Alex Shaw, Mark Sheskin, Chris Starmans, and everyone else I looked up to as an undergrad — I learned so much just by sharing a room with you. I'm grateful to Karin Gosselink and Carl Zimmer for showing me that good writing can be taught, and to Julia Istomina, Ryan Wepler, and my clients at the Graduate Writing Lab for keeping me sharp over the years. But my deepest debt of gratitude is owed to Melissa Ferguson — of everything you've done in your role as DGS, little (if any) was required or expected of you and most (if not all) was done for our benefit. This dissertation could not have happened without you, and that means more than I can say.

To Billy Brady, Megha Chawla, Brian Earp, Jim Everett, Judy Kim, Killian McLoughlin, Yeon Soon Shin, Jen Siegel, Hongbo Yu, Yoonseo Zoh, and the rest of the Crockett Lab, as well as the members of the Mind and Development Lab, the Perception and Cognition Lab, and the Affective Science and Culture Lab. To Margaret Clark for graciously sharing not only some useful references (and the occasional gossip session), but also an office in your lab space when I needed one. I'm grateful to Julian Jara-Ettinger for teaching me R, to Laurie Paul and David Rand for their helpful feedback as readers, and to the administrative staff who were there behind the scenes (especially those who reminded me to do what I'd reliably forgotten to do) — Lindsey Fuller, Fredericka Grant, Sam Haller, Christine James, Caleb Kealoha, Sophie Harrington, Kim Nicefaro, Akari Roudebush, Jenny Trujillo, and Jim Westgate.

To Dan Ariely, Felipe DeBrigard, Walter Sinnott-Armstrong, and Nina Strohminger for being such generous mentors during my time at Duke. To David Pizarro, for introducing me to supertasters (and being a friend when I needed it). To the undergraduate researchers who worked with me over the years, particularly Charlotte Austin, Keenan Miller, Breanna

Nguyen, and Amisha Vyas — watching you grow into such bright and independent scholars (who have gone on to do such cool work!) has easily been a highlight of graduate school.

To Alyssa Arre, May Conley, Maureen Gill, Kathryn Graves, Emily Gerdin, Srishti Goel, Michael Lopez-Brau, Aalap Shah, Kim Wong, Flora Zhang and my other friends at Yale (particularly Marlene Berke, Arin Korkmaz, Merve Erdogan, Joan Ongchoco, and Gracie Reinecke for tolerating my breathless rants about the earliest versions of these ideas). To Rosie Aboody, Dorsa Amir, April Bailey, Adam Bear, Clara Colombatto, Jill Jordan, Matt Jordan, Bud Lambert, Julia Marshall, Brynn Sherman, and Sami Yousif for doing such impressive work that I could look up to (and for your frequent, generous help). To Chance Adkins, Xanni Brown, and Ajua Duker for being the most supportive (and social) social cohort. To Lizzie Mundell-Perkins, Eli Neudstadter, and Julio Silva for making Toad Hall such a special place to call home. And to Ryan Carlson, who shared an office and lab with me for six years (in addition to *literally* all of the above). You're a caring friend, a deep and thorough thinker, and I'm awed by your capacity for warmth and earnestness. The best thing to happen to me in graduate school was getting to start it together with you.

This work is dedicated to my mom (you would've hated my tattoos) and my dog (you would've loved COVID) — I miss you both every day. To Chris, John, and Nadia who were there from the start (along with my twin brother Alex, though more literally). To my dad in Romania, and to Hannah, Jozie, Matt, Michael and the rest of my family in Durham. To Jesse for the folk songs (and callsigns), and to Paul, Dan, and The Jitter Bus for the unending stream of coffee. To Madi for getting me through the front half of this program (and the worst of the pandemic), and to Abby, Bijan, Isaiah, Kurt, and PJ for the hours we spent on Fortnite (it felt normal during a weird few years). To Robin and Marlow for Coqui (and your floor), and to Emily, Kibby, and Whitney for being there when I needed to talk. To Nick for the clambakes, Sarah for the dinner parties, and Tess for the nights out dancing. To Bun and Mie for the sushi, to Larissa for so often passing through, and to Rosa for bringing such kindness (and Poppy) into my life. To my brother Daniel for always being there (and for taking care of mom), and to Anabel, who kept me sane and fed over the manic few weeks I spent finishing this. I love you, and I'm so, so grateful.

And finally, this work was only possible because 31,359 people took part in one of the 94 online experiments I ran over the course of my PhD. Thank you for your help, especially the 7,060 of whom I've included in what's to follow — You taught me everything I know.

Psychology is passing into a less simple phase. Within a few years, what one may call a microscopic psychology has arisen in Germany, carried on by experimental methods, asking of course every moment for introspective data, but eliminating their uncertainty by operating on a large scale and taking statistical means. This method taxes patience to the utmost, and could hardly have arisen in a country whose natives could be *bored.* Such Germans as Weber, Fechner, and Wundt obviously cannot; and their success has brought into the field an array of younger experimental psychologists, bent on studying the elements of the mental life, dissecting them out from the gross results in which they are embedded, and as far as possible reducing them to quantitative scales. The simple and open method of attack having done what it can, the method of patience, starving out, and harassing to death is tried; the Mind must submit to a regular siege, in which minute advantages gained night and day by the forces that hem her in must sum themselves up at last into her overthrow. There is little of the grand style about these new prism, pendulum, and chronograph-philosophers. They mean business, not chivalry.

No general description of the methods of psychophysics would be instructive to one unfamiliar with their application, so we will waste no words upon the attempt.

— William James, *The Principles of Psychology, Vol. 1*

N.B. – Lightly edited for clarity.

1

Introduction

There is a deep methodological question posed in "Seasons of Love," perhaps the most widely known song from the Tony Award-winning musical *Rent*: how do you measure a year? Broadly, the song proposes two kinds of answers, one more plausible than the other. The first set of options are *metric* in nature, in that they have meaningful and countable units (e.g. "in sunsets" or "in 525,600 minutes"). The second set, however, are decidedly *nonmetric*, in that these options aren't really countable and contain no discernible unit at all (e.g. "in laughter" or "in strife"). Ultimately, the song opts for this second kind of answer, entreating the listener to "measure in love."

To the experimentalist, this is sure to be unsatisfying, least of all because it raises as many questions as it does answers — most immediately: "but how?" To state the obvious, the song is not making the literal claim that love can be used as a unit with which to measure time, and the word "measure" is clearly being used to mean something more evaluative, i.e. "we should judge the worth of a year by how much love it contains." But this hardly makes the question any easier to answer, since it just pushes the problem back one step further. Love is not a unit of time, sure, but there aren't exactly units of love, either. So we are again left asking: "but *how*?" If love has no countable unit with which it can be measured, then how can it be quantified in the measure of a year?

This dissertation is not about quantifying love, though it is about quantifying subjective magnitudes, of which love is just one of many possible examples. Two of the three chapters will center around moral judgment, but this is more or less arbitrary, stemming entirely from the accident of my having started graduate school doing work on moral judgment. Even so, I wanted to start with love (and this kind of question, in general), since it invites us to acknowledge something deeply puzzling about what is otherwise very mundane: hardly anyone could deny loving some things more than others, but where does that "more than" come from? We can surely quantify the *expression* of love — even young children can stretch out their arms and say "I love you *this* much" — but what is being quantified to *create* that feeling that the child is then matching with this gesture? How can we form magnitudes for nonmetric stimuli like love or morality, given that neither has a meaningful unit on which we can aggregate? — How is it possible to so casually make something, we might say, out of nothing?

1.1 *On psychophysical measurement*

The chapters in this dissertation aim to answer these questions, as well as others like it. Largely, they do so by adapting methods and insights from classic work in sensory psychophysics, which themselves were aimed at answering analogous questions in the domain of sensation and perception. Since a comprehensive historical overview is beyond the scope of this introduction (for such an overview, see Murray, 1993), what follows is a brief summary of the most relevant work.

Traditionally understood, psychophysics is most often associated with the work of its founder, Gustav Fechner, who aimed to bridge the psychological world with the physical one, mapping the objective magnitude of a stimulus (e.g. the number of photons emitted by some light, per square inch per second) onto the subjective magnitude of the resulting sensation (e.g. how bright that light seems; Fechner 1860). This approach, however, only produced indirect measures of sensation magnitude, since Fechner saw the idea of assigning numbers directly to sensations as wholly subjective if not incoherent.

To Fechner, one sensation can only be said to be greater than, less than, or equal to another sensation (e.g. one light can be said to be brighter than another, but there is no obvious basis on which to say just *how much* brighter). As such, Fechner approached psychophysical measurement via a process of so-called *indirect scaling*, using methods like thresholding (e.g. what increase in luminance is necessary to make one light seem brighter than another?), in which a scale can be created via summation of these so-called *just-noticeable differences* (JNDs), which serve as the unit. There was a crucial flaw with such procedures, however — equal changes in this unit did not create equal changes in apparent magnitude. Put another way, it would be as if the length of an inch shrank or expanded depending on how far one has extended a tape-measure. An inch should be an inch no matter what, but the same could not be said of Fechner's discriminability scales.

Decades later, S.S. Stevens (to whom this dissertation is partially dedicated) began developing the methods that would constitute his "New Psychophysics," including approaches like magnitude estimation and cross-modality matching (Stevens, 1960). In many ways, these so-called *direct scaling* procedures served as a return to what James may have considered a simpler and more open method of attack: use numbers. While discriminability served as the foundational operation in Fechner's psychophysics, Stevens instead relies on a process of matching, in which the intensity of one sensory magnitude is directly equated to the subjective intensity of another magnitude (of which number is included). This latter approach, broadly construed, serves as the basis for the work I will present in the remainder of this dissertation.

1.2 *On subjective magnitudes*

In perception research, the word "subjective" tends be used in two different ways. First, it can be used to refer to a fact about someone's subjective experience, meaning something more like "apparent" — what philosophers might call *qualia* (e.g. Jackson, 1986), or the *phenomenal character* (e.g. Lewis, 1990) of an experience. Crucially, this need not have anything at all to do with the external world; we can experience colors while we're dreaming or awake, and neither is more or less subjective than the other. Similarly, a light will seem subjectively brighter if you've spent the whole day in a darkened room, but this isn't because more photons are hitting your eye per square inch per second. This is the sense of subjective in which the methods of sensory psychophysics were developed to address, as contrasted with the objective, physical properties of the stimulus which produces this sensory magnitude.

In the second sense, subjective refers to a fact about the physical world, meaning something more like "lacking a physical correlate" — what I described as *nonmetric* at the start of this dissertation. Here, perceptual stimuli clearly *don't* involve this kind of subjectivity. While a photometer can objectively measure a light's brightness, just as a spectrometer can objectively measure that same light's wavelength, stimuli like love and morality have no such correspondingly objective measures (nor likely ever could they). In sensory psychophysics, the first kind of magnitude was described as *metric*, while the latter was described as *nonmetric*.

To avoid confusion, I'll be describing these four kinds of magnitudes using the following language throughout the remainder of this dissertation. I will describe magnitudes that are subjective in the first sense as being *psychological magnitudes*, in that the distinction is between the psychological properties of the stimulus (subjective in the first sense) and what I will describe as *objective magnitudes*, which involve the physical properties of that same stimulus. To avoid unnecessarily introducing unfamiliar terms (i.e. nonmetric and metric), I will instead be using the generic term "subjective magnitude" always and *only* in the second sense described above, in that it distinguishes between psychological magnitudes with no corresponding physical correlates (subjective in the second sense) and what I will describe as *perceptual magnitudes*, which involve psychological magnitudes that do concern physical correlates.

In sum, this dissertation ultimately proposes that methods that were initially developed to measure lower level *perceptual* magnitudes (e.g. brightness or loudness) can be successfully applied to the measurement of higher level *subjective* magnitudes (e.g. love or morality). I go on to propose that these methods work in both contexts, since subjective magnitudes rely on a domain-general magnitude representation that is shared by *all* psychological magnitudes, whether metric or nonmetric. In this way, the chapters in this dissertation demonstrate not only pragmatic and methodological lessons aimed toward improving psychological measurement, but they also bear on some of the most basic questions involved in the study of the mind.

1.3 *On a scale from* 1 *to* 10

Mental states cannot be observed, though they can be described. It should only be natural, then, that our field has made such liberal use of self-report (e.g. Baumeister et al., 2007). It is less natural, however, that our field so frequently does so by transforming those words and descriptions into numerical quantities like test statistics, parameters, and probability estimates, all via a process that William James may well have described as a perverse and brutal alchemy bereft of Human virtue.

The most common of such practices, by far, is to collect reports in the form of numerical responses — e.g. on a 7-point Likert scale (Likert, 1932) or a 100-point Visual Analog Scale (Aitken, 1969) usually aiming to capture some kind of judgment or evaluation as a point between two extremes. For Likert scales, this is reflected as a choice between a number of distinct categories, whereas the Visual Analog Scale is a continuous measure, allowing one to select any value between the two points by e.g. clicking along a slider on a computer or marking a line on a survey. Collectively, these and similar measures are often referred to as *labeled scales* (Bartoshuk et al., 2003).

For the moral psychologist, these labels might be phrases like "not at all immoral" and "very immoral," with the response quantifying the permissibility of some action or the nature of a person's character. For the political scientist, these labels might read "strongly agree" and "strongly disagree," with the response serving to quantify the strength of an attitude or support for one policy over another. And for the cognitive neuroscientist, these labels might say "0%" to "100%," with the response quantifying a subject's confidence in a decision. And so on.

According to Stevens — whom some may call an expert on the topic (1946) — the ancient Greek astronomer Hipparchus was the first person to use such a scale as a tool for scientific measurement (Stevens, 1975, p. 15). What Hipparchus developed was a method for cataloging the brightness of stars into a six-point scale, ranging from 1 (the brightest stars) to 6 (the faintest stars). This scale is remarkable for more than just its originality, however. It was also remarkable for its longevity and success — Ptolemy used this scale in his catalog of the stars, and some version of Hipparchus' scale was in steady use until the 19th Century, when it was replaced by the Pogson scale, named for the astronomer who developed it.

What's even *more* remarkable is that Pogson, having developed his scale by using photometers to objectively measure the brightness of stars, amounted to doing little more than mathematically formalizing and extending Hipparchus's scale. What Pogson found was an almost perfect logarithmic relationship between a star's category and the number of photons hitting a surface per square inch per second. Each successive level of Hipparchus's scale was about two and a half times brighter than the one preceding it (Burke-Gaffney, 1963). Even after 2,000 years and with the invention of photometry, all they could add was precision: Pogson's final ratio of 2.512.

I tell this story because I think it clearly illustrates something that can be too easily overlooked or taken for granted: labeled scales work astonishingly, almost *impossibly* well. Though much of what follows is undeniably critical of how labeled scales are used in our discipline, I can't help but find it genuinely awe-inspiring that Hipparchus's scale worked so well that it is still being used to measure the brightness of the stars even today. It may not seem like it, but over the course of doing the work that I present in this dissertation, my most common feeling was one of almost *reverence* towards labeled scales and how well they work. They have their problems, a handful of which I catalog here, but I've found that more often than not these problems arise from our attempts to contort these measures into doing what they were never meant to do.

1.4 *Three main problems*

In what remains of this introduction, I first provide an overview of the methodological limits and pitfalls of the labeled scales most commonly used in our discipline: their *interpersonal relativity*, their *intrapersonal elasticity*, and their *nonlinearity*. Briefly, scales can mean different things to different people (as many can attest after asking for a dish that is "medium" spicy at a Thai restaurant), they can change meaning on shorter time scales *within* a given person (we have no trouble understanding what someone means by saying that their house is small yet their dog is big), and they compress the most extreme values measured by the scale (as one might intuit by comparing the subjective distance between “good” and “great” with “pretty good” and “fine”). Then, I explain how methods from sensory psychophysics can help address each of these three problems, thus introducing the basic logic that underpins the research I present in the remainder of this dissertation.

1.4.1 *Interpersonal relativity*

Economists have long-appreciated how difficult it is to compare subjective magnitudes, and of chief concern was subjective utility. Though one person may describe their happiness as an 8 out of 10, and another may describe theirs as a 9 out of 10, it does not necessarily follow that the second person is happier than the first. Many economists and philosophers even go so far as to say that such *interpersonal utility comparisons* (Harsanyi, 1977; Jeffrey, 1974) are in principle impossible (Hausman, 1995). Lionel Robbins (1935), an early proponent of this view, articulates it nicely:

> *There is no means of testing the magnitude of A’s satisfaction as compared with B’s. If we tested the state of their blood-streams, that would be a test of blood, not satisfaction. Introspection does not enable A to measure what is going on in B’s mind, nor B to measure what is going on in A’s. There is no way of comparing the satisfactions of different people. (p. 140)*

The logic underpinning the problem of interpersonal utility comparisons applies to virtually any interpersonal comparison of subjective magnitude. As such, structurally similar issues have been considered under different names in different disciplines: psychophysicists have discussed interindividual comparisons (Borg, 1990) and across-group comparisons (Bartoshuk et al., 2003), while political scientists have considered the interpersonal comparability of responses (Brady, 1985) or the cross-cultural comparability of measurement (King et al., 2004).

This problem is most clearly illustrated by the existence of so-called “supertasters,” who are able to taste phenylthiocarbamide (PTC) and a related compound, 6-n-propylthiouracil (PROP). This ability is genetically inherited (Blakeslee & Fox, 1932), and it was initially understood through an explicit analogy to colorblindness, such that so-called “nontasters” were described as “taste blind,” being unable to taste some bitter compounds in the same way that some are unable to see red (Fox, 1932).

It was eventually discovered, however, that supertasters and nontasters differed also in the overall intensity of their taste experiences, though this difference was hidden for nearly a century. Though both nontasters and supertasters would likely describe soda as “very sweet” and rate it highly on a scale from 1 (not at all sweet) to 10 (very sweet; see Bartoshuk, 2014), the supertaster would nonetheless have a much more intense taste experience compared to the nontaster. How is this possible? The simple answer is that supertasters experience *all* tastes more intensely (even having a higher density of taste buds on their tongue; Bartoshuk et al., 1994), which means that they not only experience the soda as sweeter, but what they mean by the phrase “very sweet” *itself* is sweeter! Since the sweetness of the soda is magnified to the same extent as the label anchoring the scale used to measure sweetness, these effects effectively cancel out, and supertasters and nontasters provide virtually identical ratings (Bartoshuk, 2014; Bartoshuk et al., 2003).

In Chapter 2, I discuss this problem as it relates to purported group differences in emotional intensity.

1.4.2 *Intrapersonal elasticity*

To use the jargon of psychometrics, it is crucially important that our measures be invariant (see Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). Otherwise we would be unable to determine whether a given change was caused by the manipulation of the independent variable or merely how the subjects interpreted the meaning of the dependent variable. In his *Handbook of Experimental Psychology*, Steven’s straightforwardly describes this principle as follows: “the scientist seeks measures that will stay put while [their] back is turned" (1951, p. 21). Unfortunately, labeled scales are prone to do anything but.

In addition to the problem described above, in which there are individual differences in how people interpret a given scale, there can also be substantial *within-person* differences, such that same person using the same scale can mean different things at different times. The classic analogy is that labeled scales can act like elastic rulers or rubber bands (Volkmann, 1951). Postman and Miller (1945, p. 52), attributing this analogy to Volkmann, describe this problem (dryly) as follows:

> *The subjective scale is a flexible elastic scale. Introduction of an anchor outside the original range causes the entire subjective scale to be extended, in a manner analogous to a rubber band, beyond the original scale but remaining anchored to the stimulus of smallest magnitude. As a result the same number of judgment categories must cover a wider range.*

This problem, described in contemporary work as *response shift* (e.g. Ogden & Lo, 2012) or *scale recalibration* (e.g. Ubel et al., 2010) has received some attention, largely within the domain of happiness and quality of life research but especially in medical contexts. For example, someone in their 80’s does not use the phrase “perfect health” in the same way that they did in their 20’s (Ubel et al., 2005). Such effects can also operate at much shorter time scales, as illustrated in the classic finding by Birnbaum (1999): subjects rate the number 9 to be larger than the number 221 in a between-subjects design (see also Leong et al., 2019). The explanation for this effect is that the number 9 evokes a context of 1 or 2 digit numbers, within which it seems large (i.e. 9 out of 10), whereas 221 evokes a context of 3 or 4 digit numbers, within which it seems small (i.e. 221 out of 1000).

Other work, still, has illustrated this problem as it relates to stereotypes. For example, one study (Biernat et al., 1991) found what appears to be an inconsistent pattern of results involving judgments of financial success. When asked to estimate the yearly salary of men and women pictured in a photograph, subjects estimated higher salaries for the men than for the women. When asked to estimate financial success on a Likert scale, however, subjects showed the opposite pattern — women scored higher on this measure of financial success. It's clear, however, that this is merely reflecting the existence of a lower standard of "financial success" as applied to women compared to men.

I discuss this problem in Chapter 3 in the context of moral judgment.

1.4.3 *Nonlinearity*

I believe that this final problem is the least widely-appreciated. Through the simple act of averaging our data, we assume that there is a linear relationship between the scales we use and whatever it is that our scales are measuring. While there are some cases where labeled scales are in fact linearly related to the stimulus in question — what Stevens referred to as "metathetic" continua — they more often form a logarithmic relationship (similar to how Hipparchus's scale was logarithmically related to the objective luminosity).

A more detailed explanation is too technical for this brief overview, but I will discuss these issues in more detail in Chapter 4. On a more intuitive level, however, I've found it helpful to consider this idea in the context of pain. Most often, pain is measured using a 100-point Visual Analog Scale, which ranges from 0 (no pain) to 100 (the worst pain imaginable; e.g. Price et al., 1983). If this scale were linear, then a 50 should be half as painful as a 100, and a 25 should be half as painful as a 50, and so on. But suppose that this scale were *actually* linear, and suppose you were to rate your second and third most painful experiences. Even if the most painful experience was set to 100, I would likely give my second most painful experience a 30 (a value I would typically assign to something like rolling an ankle), and my third most painful experience would be around a 10 (a value I would typically assign to something like stubbing my toe).

This basic idea, more or less, is why so many scales are logarithmically compressed.

Were they *actually* linear, then they'd be useless for measuring anything that varied by more than one or two orders of magnitude. This is because meaningful differences on the lower end of the scale would be undetectable, and these differences can be important. For example, minor fluctuations in pain can reveal whether a conditioning is improving or worsening, but it would be impossible to show as much if this movement is captured within only a few percentage points out of 100. Thus, nonlinearity allows these scales to be more effective at their primary function — discriminating between stimuli — though at the expense of accurately estimating the true magnitude of a stimulus.

I discuss this problem in Chapter 4 (again in the context of moral judgment), and I propose a preliminary but straightforward solution to correct this type of nonlinearity.

1.5 *Overview of the current chapters*

Though the problems I've just described are often subtle and easy to miss, they have been extensively studied in the field of sensory psychophysics, particularly the so-called "New Psychophysics" (see: Stevens, 1960). The overall strategy from this work (which I will be adapting) is to collect subjective judgments in reference to a fixed benchmark. In Chapter 2, I evaluate work which alleges sex differences in how men and women experience the intensity of anger. Given the work on supertasters just described, I find this purported group difference *theoretically* possible, though unlikely given the measures used to find it. Thus, I adapt a method developed in taste psychophysics explicitly to detect these types of group differences that are hidden by labeled scales — the general Labeled Magnitude Scale (gLMS; Bartoshuk, Duffy, et al., 2004), which aims to minimize the problem of interpersonal relativity. The gLMS accomplishes this by anchoring judgments against a benchmark that is more widely fixed across people: the most intense sensation imaginable. When tested using this method, the purported difference reliably disappears.

In Chapter 3, I consider how the problem of intrapersonal elasticity can create implausible patterns of results in the context of moral judgment. I illustrate this in a vivid case study, in which subjects report (between-subjects) that a particularly cruel college prank is just as morally bad as (if not worse than) an internationally recognized war crime. Just as the meaning of the word "big" can change based on whether we are evaluating relatively small numbers like 9 or relatively big numbers like 221, so too can the meaning of "immoral" change based on context. Here, as well, the strategy involves making judgments in reference to a fixed benchmark, though in this context the benchmark is another stimulus: how immoral it is to steal a wallet. Since the immorality of stealing a wallet is stable (at least relative to vague descriptions like "very immoral," which can drastically change), it is less susceptible to these sorts of problems, and this paradoxical finding disappears.

In Chapter 4, I consider the problem of nonlinearity in concurrence with a related problem: how are subjective magnitudes encoded? This debate has been surprisingly difficult to resolve, since much of this work is almost necessarily confounded by the fact that subjective magnitudes are tightly correlated with physical magnitudes. Thus, it is sometimes unclear whether a subject's judgment reflects their experience of subjective magnitude, or merely a judgment of the relevant physical magnitude. In this context, a subjective stimulus like morality can actually provide a unique opportunity to address similar questions in the absence of a particularly stubborn confound. By applying straightforward psychophysical methods involving matching and bisection, I find compelling evidence first, that Likert scales are nonlinear, and second, that subjective moral magnitude has a linear representational format whose variability increases in proportion to the size of the stimulus.

These chapters were written as individual journal articles. The first two chapters are currently under review (Chituc & Scholl, under review; Chituc, Crockett & Scholl, under review), while the third chapter is in preparation for submission (Chituc, in prep). As a result, these chapters stand on their own and can be read independently and in any order.

2

The El Greco fallacy, this time with feeling: How (not) to measure group differences in emotional intensity

Abstract

Affective scientists often make claims about group differences in emotional intensity by comparing group averages on labeled (e.g. Likert) scales. Despite their ubiquity, these measures are susceptible to a subtle but notorious problem: *the El Greco fallacy*. (El Greco famously painted elongated figures, but this could not be because he perceived a stretched-out world due to astigmatism, since he would also see the canvas itself as stretched-out.) Here we consider the prominent claim that women experience anger more intensely than men. Across four experiments (n=4000, tested online), we replicate this finding with labeled scales, but show that it reliably disappears when tested with a general Labeled Magnitude Scale (gLMS) — a psychophysical measure designed to detect such differences, avoiding the El Greco fallacy. This demonstrates how insights from sensory psychophysics can be usefully employed in affective science, and supports skepticism about purported group differences in emotional intensity based on labeled scales.

2.1 *Introduction*

Do you feel what I feel? Maybe not: as we are all aware, there are substantial differences in emotional experiences across people (and even across the same people at different times). Some differences may be qualitative, but they may more often involve emotional intensity (e.g. Diener et al., 1985). What elates me may only tickle you, and what enrages you may only bother me. This much seems obvious from our everyday experience of feeling and sharing emotions with one another, but affective science has sometimes gone further, suggesting that some such individual differences are *systematic*, varying across groups. For example, some researchers have claimed that women feel emotions more strongly than men (Davis et al., 2012; Diener et al., 1985), or that people living in China experience emotions less intensely than people living in the West (Davis et al., 2012; Eid & Diener, 2001).

In the current short empirical report, we take no position on the truth or falsity of such claims. And we take no position on any of the ambient controversies surrounding the nature of emotion (basic and largely universal? conceptual and largely socially constructed?). The only claims about emotions that matter here are as follows: having an emotion *feels like something*, and this feeling can be stronger or weaker. Accordingly, the focus here will be only on how such claims about group differences could and should (and couldn't and shouldn't) be measured. In particular, we will suggest that certain (almost universal) ways of measuring such claims may be flawed, for a famously subtle reason.

2.1.1 *The* El Greco *fallacy*

Emotional experience is inherently private, making the measurement of emotion (or any sensation) a notoriously difficult (and historically controversial) enterprise. This isn't necessarily obvious, since it may seem at first blush like one could simply ask people to report details of their subjective experiences, including intensity. But report how? By far the most common answer involves labeled scales — e.g. reporting the intensity of felt anger on a Likert scale, anchored with labels from "not at all angry" to "very angry". Though pleasingly direct, this method masks a notorious problem: differences in underlying experiences may also create differences in the use of the scale.

The most infamous example of this may be a proposed explanation for the distinctive style of the Spanish Renaissance painter El Greco. El Greco famously painted figures that were exceptionally (and even oddly) *elongated*. Why? One famous proposal implicated El Greco's perception of the world: if he perceived the world as vertically stretched due to a type of astigmatism, then perhaps he just painted what he saw. On reflection, however, this explanation could not possibly work, since such an astigmatism would have caused El Greco to also experience a stretched-out canvas, such that the effects would cancel out (for a review, see Firestone, 2013). Thinking otherwise has come to be known as the El Greco fallacy, and this error in reasoning seems as stubborn as it is seductive. For example, modern research has argued that several alleged 'top-down' effects of cognition on perception could not possibly be true, because they inadvertently commit the fallacy (Firestone & Scholl, 2014). (And the fallacy may even persist in modern art appreciation, as when a team of otolaryngologists recently proposed that vestibular migraines could explain why Van Gogh painted with a tilt to the left; Dasgupta et al., 2022, cf. Huntley, 2022).

2.1.2 *The current studies*

In the present project, we consider the El Greco fallacy in the context of alleged group differences in emotional experience. While such claims could be true, it seems challenging to substantiate them with labeled scales — because such scales (like El Greco's canvas)

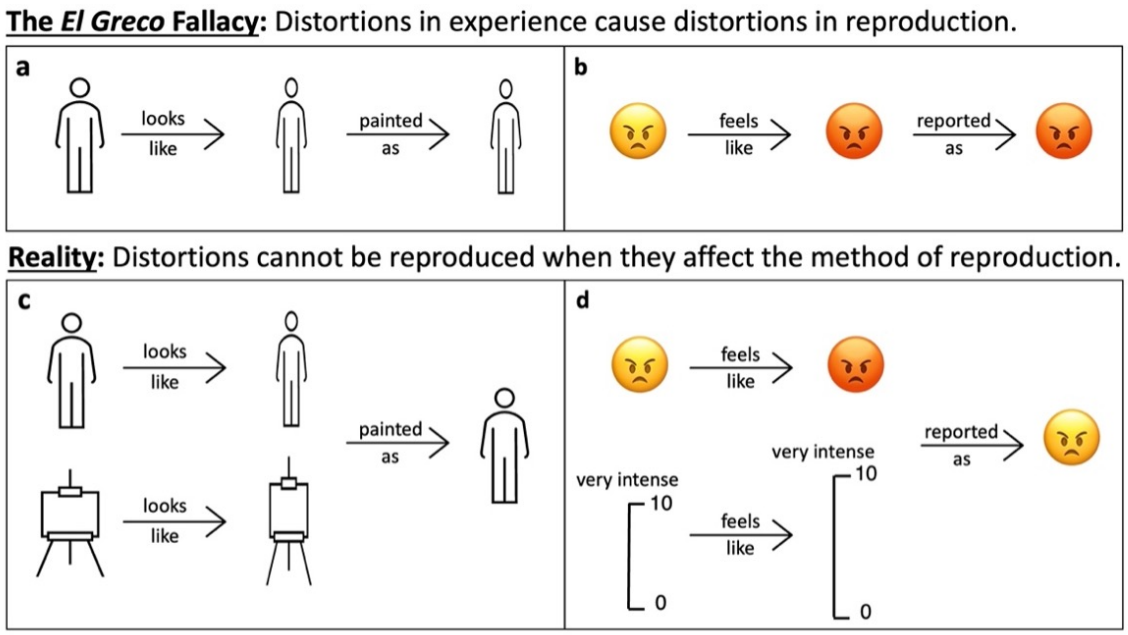
The El Greco Fallacy: Distortions in experience cause distortions in reproduction.
a
looks like
painted as
b
feels like
reported as
Reality: Distortions cannot be reproduced when they affect the method of reproduction.
c
looks like
looks like
painted as
d
feels like
very intense
10
0
feels like
very intense
10
0
reported as

Figure 2.1: Illustration of the *El Greco* fallacy. The distortion of a stimulus is assumed to be reflected in the reproduction of that stimulus, as depicted in the context of vertical elongation captured by a painting (a) and more intensely felt anger reported on a scale (b). In reality, a perceptual distortion should be undetectable if it also affects the method of reproducing that distortion. Vertical elongation of a figure cannot be captured by a painting of that figure, since the canvas would also be perceived as vertically elongated in kind (c). Just so, a magnified experience of anger cannot be captured on a labeled scale of intensity, since the meaning of "very intense" anger itself would be magnified in kind (d).

could be distorted in the same way as the emotional experience purportedly measured by those scales. Consider the claim that women experience *anger* more intensely than men (Simon & Nath, 2004). This could not be readily captured on a labeled scale, since the anchors themselves — e.g. "not at all intense" or "very intense" — would naturally be interpreted differently, due to the very same putative differences in emotional intensity (see Figure 2.1).

This point may be subtle, but it has been well understood in the study of sensory psychophysics. Consider, for example, *supertasters* — who experience tastes as much more intense than do most people (something true, incidentally, of both authors). Such differences are very real, but can often go undetected by labeled scales (for a review, see

Bartoshuk et al., 2005). Normal tasters and supertasters may both describe a soda as sweet, and may even provide numerically identical ratings on a scale ranging from “not at all sweet” to “very sweet”. But the actual anchoring experience of “very sweet” itself differs dramatically across such groups, such that (despite identical ratings!) supertasters may experience soda to be dramatically more sweet than would normal tasters (Bartoshuk, 2014; Bartoshuk et al., 2003).

Here, we directly apply this logic to the claim that women experience anger more intensely than men (Simon & Nath, 2004). We chose this target finding because it is widely cited (over 900 times as of this writing), it is replicable (e.g. Brebner, 2003; Fischer & Roseman, 2007), and it is bracingly counter-stereotypical (since anger is more often associated with men’s emotional responses; e.g. Kelly & Hutson-Comeaux, 1999).

We first aim, in Experiment 2.1a, to conceptually replicate the apparent finding that women experience more intense anger, using the same measure employed in previous work: a 10-point Likert scale, ranging from “not at all intense” to “very intense” — which is susceptible in principle to the El Greco fallacy. Then, in Experiment 2.1b, we aim to make the same comparison using a scale that was explicitly developed to capture true underlying differences in subjective sensory experience: the gLMS (general Labeled Magnitude Scale; Bartoshuk, Duffy, et al., 2004; Hayes et al., 2013). Importantly, the gLMS employs a benchmark systematically unrelated to the intensity of any one sensation, making it immune to the El Greco fallacy (at least in this context). More specifically, the gLMS asks subjects to rate their experience relative to “the strongest imaginable sensation” of any type, rather than to an extreme intensity of the target experience itself. In Experiments 2.2a and 2.2b, we then provide a further conceptual replication of both the Likert and gLMS experiments, using a subtly different variety of gLMS measurement.

2.2 *Experiments* 2.1*a*-2.1*b*

Based on the logic of the El Greco fallacy, we predicted that the apparent difference in emotional intensity — with women experiencing anger more intensely compared to men — would replicate with the typical Likert scale measurement (in Experiment 2.1a), but would disappear with the more careful gLMS measurement (in Experiment 2.1b).

2.2.1 *Methods*

All hypotheses, sample sizes, methods, and analyses were preregistered before data collection began (see https://aspredicted.org/MM4_KNV) — with the only deviation from preregistration being the exclusion of subjects who inadvertently participated in multiple experiments from this same project (though retaining these subjects produces qualitatively identical results).

Subjects

Via the Prolific survey platform (Palan & Schitter, 2018), we recruited a convenience sample of 1000 subjects for each experiment (500 men and 500 women, based on demographic information that they had previously reported to Prolific; *n*=4000). Though we preregistered no exclusions, there was a brief period when more than one experiment reported here was recruiting concurrently, inadvertently leading 525 subjects to complete more than one of them. In such cases, we retained only the first experiment completed by that subject, and we recruited additional subjects until we reached our preregistered sample size, which had 99% power to detect an effect as large as the one from the initial report of such effects (*d*=0.3, as calculated from the sample size, means, and *SD*s from Table B1 of Simon & Nath, 2004, p. 1172).

Stimuli and Procedure

All testing was conducted using the Qualtrics survey platform. Subjects were first instructed to think about the anger they felt the last time someone insulted them — describing what happened in a free-response box, per the instructions: "In only a few words, describe what was happening that made you feel this way."[1] They were then subsequently asked to report the intensity of the anger they felt using one of two measurement methods — a Likert scale (Experiment 2.1a) or the gLMS (Experiment 2.1b). For the Likert scale, following the initial report of such effects (Simon & Nath, 2004), subjects simply clicked on one of 10 visible numbers, ranging from 1 (anchored with the label "not at all intense") to 10 (anchored with the label "very intense"). For the gLMS, subjects clicked and dragged a slider along a vertical line, ranging from 0 (anchored with the label "no sensation") to 100 (anchored with the label "strongest imaginable sensation"), with other labels (e.g. "moderate" and "very strong") quasi-logarithmically spaced across the length of the scale (see Figure 2.2).

The instructions for this scale (adapted from Hayes et al., 2013) were as follows:

> *The scale you will use today captures the entire range of how intense experiences can possibly be. The top of the scale (100) is the strongest sensation you could even imagine, which should be the maximum amount of intensity possible for an experience. The bottom of the scale (0) is the absence of sensation, which should be the minimum amount of intensity possible for an experience.*
>
> *Between those two extremes, you should be able to rate the intensity of every other experience you have ever had or could even imagine having. If there's a sensation that would fall outside either end of the scale, then that sensation is what should be the endpoint, instead.*
>
> *This is very important: the endpoints of the scale represent the minimum and maximum level of intensity possible for an experience, so you should be able to assign any other experience a number between 0 and 100.*

[1] The original report of such differences (Simon & Nath, 2004) used data taken from the emotion module of the 1996 General Social Survey (GSS), which asked respondents multi-stage questions involving the frequency, intensity, and duration for 19 different emotions. Since we were only interested the intensity of anger, we elicited ratings in a more simplified way.

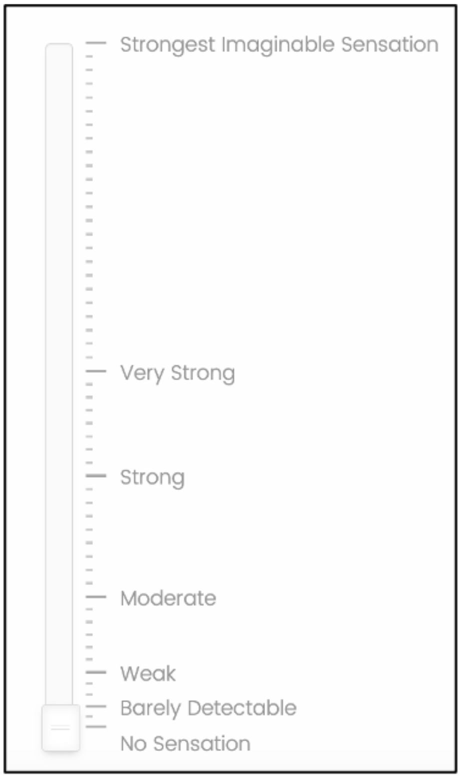

Figure 2.2: The general Labeled Magnitude Scale (gLMS) as displayed to subjects. Since group differences in the intensity of anger are unrelated to the strongest sensation one can imagine, the gLMS is not susceptible to the El Greco fallacy. It is for this reason that the gLMS has been widely used to detect (or rule out) group differences or changes in sensory experience, including for taste (Bartoshuk, Duffy, Green, et al., 2004), pain (Bartoshuk, Duffy, Chapo, et al., 2004; Čeko et al., 2022), and smell (Petrova et al., 2008).

2.2.2 *Results*

Per our preregistered analyses, we first tested for a difference between men and women on the reported intensity of anger using a Likert scale. Replicating the original result (Simon & Nath, 2004), women reported more intense anger than did men (7.0 [*SD* = 2.0] vs. 6.7 [*SD* = 2.1]; Welch two samples *t*(996.4)=2.60, *p*=.009, *d*=0.16 95% *CI* = [0.04, 0.29]; all tests 2-tailed). We next tested whether this difference would also appear using the gLMS. As predicted, it did not: women and men reported similarly intense anger (44.2 [*SD* = 25.0] vs. 42.5 [*SD* = 25.1]; *t*(998.0)=1.04, *p*=.301, *d*=0.07 [-0.06, 0.19]). We next tested whether the pattern of results produced by Likert ratings differed significantly from that produced by gLMS ratings. After normalizing each measure (*z*-scoring the raw Likert ratings and the logarithm of the gLMS ratings [incremented by one to avoid undefined values]), a 2x2 Between-subjects ANOVA (measure: Likert vs. gLMS; sex: male vs. female) revealed a significant interaction (*F*(3, 1996)=3.29, *p*=.020; η^2 =.005 [0.001, 0.012]).

2.3 *Experiments* 2.2*a*-2.2*b*

As typically used, the gLMS includes a standard battery of 15 practice stimuli to orient subjects to the use of the scale (e.g. Bartoshuk et al., 2003; Hayes et al., 2013). Because Experiment 2.1b did not employ these practice trials, we replicated the results of both experiments with the addition of this battery in Experiment 2.2b.

2.3.1 *Method*

Experiment 2.2a was a direct replication of Experiment 2.1a. Experiment 2.2b was identical to Experiment 2.1b except that before rating the intensity of their anger, subjects first completed 15 practice trials, using the gLMS with sample items relating to other forms of experience (including “The strength of a firm handshake”, “The warmth of a summer breeze on your face”, and “The brightest light you have ever seen”; for full details, see Bartoshuk et al., 2003; Hayes et al., 2013). For each experiment, we recruited a new sample of 1000 subjects (500 men, and 500 women), post-exclusions (as described above), with this preregistered sample size chosen to match that of Experiments 2.1a and 2.1b.

2.3.2 *Results*

Replicating the original result (Simon & Nath, 2004), women again reported more intense anger than did men, when assessed with a Likert scale (7.2 [*SD* = 2.0] vs. 6.8 [*SD* = 2.0]; *t*(997.3)=2.57, *p*=.010, *d*=0.16 [0.04, 0.29]). But when tested with the gLMS, this

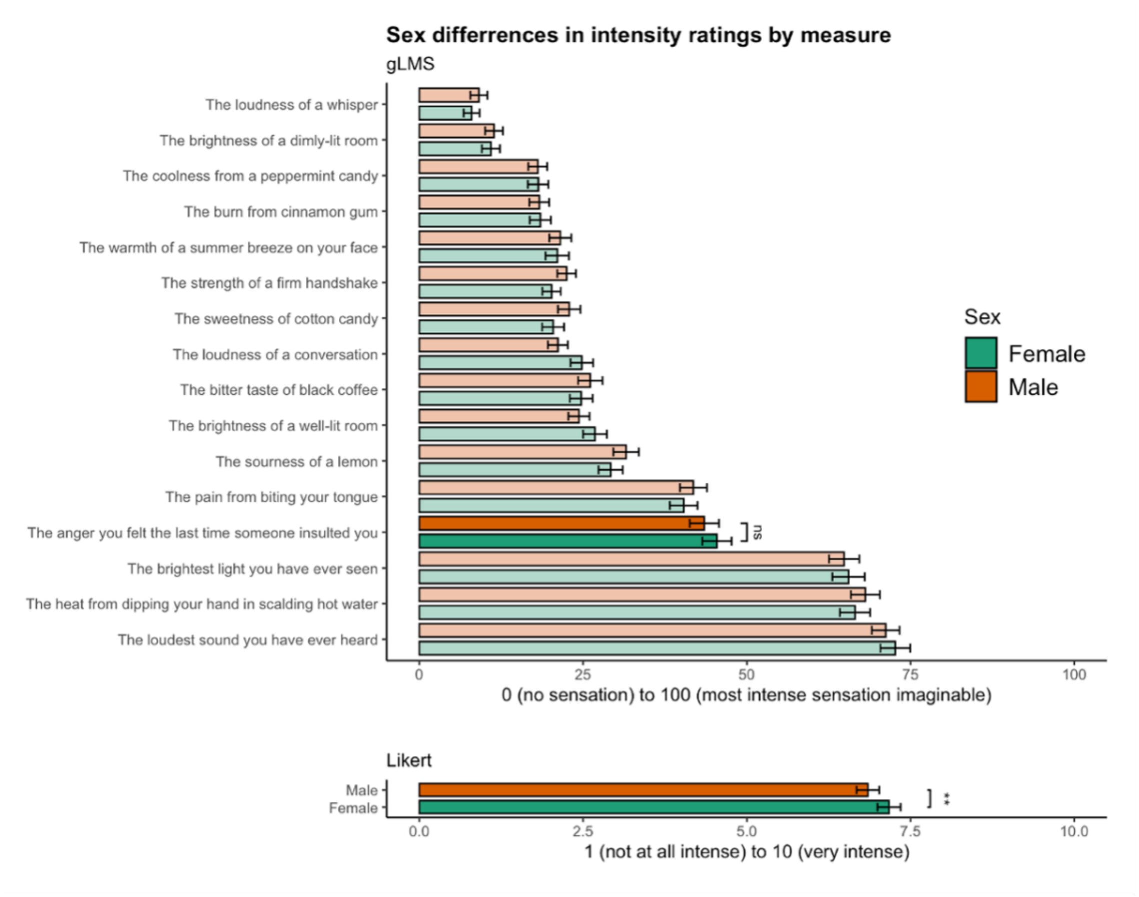

Figure 2.3: Mean ratings from Experiment 2.2a (bottom) and 2.2b (top), split by sex. Anger intensity is presented in the solid bars, and the battery of practice questions are presented in faded bars. Error bars represent 95% confidence intervals. ** indicates differences of p=.01.

difference again disappeared, with women and men reporting similarly intense anger (45.4 [*SD* = 25.5] vs. 43.5 [*SD* = 25.5]; *t*(998)=1.20, *p*=.23, *d*=0.08 [-0.06, 0.19]). After normalization, a 2x2 Between-subjects ANOVA again found a statistically significant interaction (*F*(3, 1996)=2.66, *p*=.047; η^2 =.004 [0.001, 0.010]).

Next, we preregistered an exploratory analysis to determine whether there were any overall sex differences in the use of the gLMS and, if so, to correct for them (see Figure 2.3). To determine whether there were such sex differences, we conducted a 2 (Between-subjects factor: sex) x 15 (Within-subjects factor: practice question) mixed-measures ANOVA. This revealed a significant interaction between sex and practice question (*F*(7.9, 7792)=2.94, *p*=.003; η^2=.003 [0.001, 0.005]). To correct for these differences, we then normalized all 16 gLMS responses (15 practice questions and the intensity of anger) separately for both men and women. First, we tested whether these corrected gLMS ratings would reveal a sex difference in anger, such that women might report higher ratings than did men, and they did not: the results were qualitatively identical (*t*(989.3)=1.55, *p*=.122, *d*=0.10 [-0.03, 0.22]). Next, we centered the gLMS means around 0, recombined it with the Likert data, and tested again for the interaction, which remained significant (*F*(3, 1996)=3.47, *p*=.016; η^2=.005 [0.001, 0.012]).

Finally, we report one additional exploratory analysis to support the overall robustness of our results. To achieve greater statistical power, we combined the datasets for Experiments 2.1b and 2.2b. Though this sample is sufficiently powered to detect an effect size of only *d*=0.14 with 99% power (which we note is even smaller than the one obtained by Likert ratings in either Experiment 2.1a [*d*=0.17] or Experiment 2.1b [*d*=0.16]) and less than half the size of the original finding [*d*=0.3], we again found no difference in anger between gLMS ratings provided by men and women; *t*(1998)=1.58, *p*=.114, *d*=0.07 [-0.02, 0.16].

2.4 *General Discussion*

Claims about group differences in emotional intensity are widespread in affective science, similar to the particular claim tested here — that women experience anger more intensely compared to men. And indeed, we ourselves replicated this very result twice in the current project (in Experiments 2.1a and 2.2a). We nevertheless remain skeptical about this and similar claims, both in principle and in practice, due to the potential (and unacknowledged) influence of the El Greco fallacy: El Greco's elongated paintings could not be explained by an elongated perception of the world, since he would then also see the canvas as elongated, and these distortions would cancel out in the paintings themselves. Analogously, we should be unable to detect that women experience anger more intensely than do men using labeled (e.g. Likert) scales, since such differences in intensity would affect the interpretations of the scale's extremities as well, such that these differences would cancel out.

As predicted by this interpretation, the putative sex difference in anger disappeared when tested with the more careful gLMS measure that was developed precisely to detect actual differences in experience. This pattern seemed especially robust. First, it was replicated multiple times (across Experiments 1 and 2). Second, the difference itself was also apparent in statistically reliable interactions that directly compared the two types of measures. And third, the gLMS failed to find an effect even when testing with extremely high power (combining data across Experiments 2.1b and 2.2b).

This pattern suggests that women do not in fact experience anger more intensely than men. But if so, why do such putative differences reliably appear when measured with labeled scales? An obvious limitation of the present study is that it cannot answer this question. According the logic of the El Greco fallacy, however, this question does not have to be answered in order for us to know that the effect cannot reflect *actual* differences in emotional experience, for the same reason that El Greco's elongated figures cannot reflect his elongated perception of the world (see Firestone & Scholl, 2014).

Nevertheless, these results are consistent with a number of possible explanations involving gender norms or stereotypes (e.g. Huntsinger & Raoul, 2022). More specifically, this difference may be somewhat ironically caused by the assumption that women experience anger *less* intensely than do men, not more. Consider the following paradoxical pattern of results from a study on stereotyping: when judging the financial success of different men and women, subjects reported that the men earned higher salaries than the women, yet they also reported that the women were more financially successful than the men (as rated on a 7-point scale; Biernat et al., 1991). The authors interpret these results as reflecting gendered stereotypes and expectations involving women's financial success. If women are held to a lower standard of financial success, then a phrase like “very financially successful” would correspond to a lower salary.

In this way, an apparent gender difference in the experience of anger may instead be reflecting gendered stereotypes involving the level of anger that is either expected or appropriate. More specifically, men may be interpreting the phrase “very angry” to mean “very angry *for a man*,” and women may be interpreting the phrase to mean “very angry *for a woman*.” Thus, it could be the case that women’s anger may be rated more highly (despite identical experience) in the same way that they may also be rated as more financially successful (despite lower salaries). Of course, this is only one of many such plausible accounts, and the explanation for this illusory difference in Likert rating remains an open question that future work may wish to explore.

2.4.1 *Theoretical and methodological lessons, beyond anger*

Though the empirical scope of this short report was intentionally narrow, the theoretical and methodological lessons are much broader. Theoretically, this work suggests that we should be skeptical of any purported group difference in emotional experience that is established using a labeled scale. And indeed, such skepticism may help make sense of confusing (if not contradictory) patterns of results, sometimes even obtained within a single study. For example, it has long been recognized that emotion ratings on labeled scales correlate only weakly (if at all) with physiological measures of emotion (Mauss et al., 2005). This has led some researchers to suggest abandoning certain physiological measures (Poláčková Šolcová & Lačev, 2017), but we suggest that the opposite conclusion may also be reasonable: subjective self-reports of emotion have been so poorly aligned with physiological measures because subjective self-reports of emotion have (so-far) been poorly aligned with emotional experience itself!

Methodologically, the central lesson of this work is that we can do better. Taking a cue from work on group differences in sensory psychophysics, work exploring group differences in emotional intensity should always (1) explicitly acknowledge and address the possibility of an El Greco fallacy, and (2) use methods to which this fallacy does not apply — such as the gLMS.[2] Such changes would be relatively simple to implement, as exemplified by the current project. They might nevertheless be highly consequential, if only because this insight from sensory psychophysics has almost never previously been adopted in this research community — which has instead relied almost universally on labeled scales. Indeed, we are aware of only a single instance of a different psychophysical method (magnitude estimation) being used in the context of emotion research (Hsee & Tang, 2007), and we are unaware of any instances of psychophysical scales being used to measure emotional intensity. We believe that a broader application of these measures may greatly aid affective scientists in more accurately measuring emotional experience — just as has been the case with the measurement of other types of sensations.

[2] As the gLMS may be somewhat awkward to implement in common platforms like Qualtrics (e.g. requiring custom code to space labels at unequal intervals), there are more streamlined alternatives such as the gVAS (general Visual Analog Scale; Hayes et al., 2013) which can also be used for this purpose.

3

How to show that a cruel prank is worse than a war crime: Shifting scales and missing benchmarks in the study of moral judgment

Abstract

Moral judgment is central to both everyday life and cognitive science, but how can it be studied with quantitative precision? By far the most direct and ubiquitous method is to simply ask people for their judgments, in the form of ratings on a labeled scale. Such scales are most often used without independent benchmarks, however, which raises subtle yet important problems: as has long been recognized in sensory psychophysics, such responses are meaningful only in a relative sense. (Is your dog "big"? Perhaps yes in the context of house pets, but perhaps not in the context of all mammals?) Here we illustrate the nature and extremity of this problem in a case study of moral judgment, by showing that it can readily lead people to seemingly rate a cruel prank (involving humiliation) to be just as immoral as (or even worse than) an internationally recognized war crime (involving murder). We show in a series of nine experiments that such effects arise in between-subjects designs (but not within-subjects designs) using what are perhaps the two most common measures in such work: Likert scales, and Visual Analog Scales. In contrast, such seemingly nonsensical results disappear when using magnitude estimation — a psychophysical method employing an explicit benchmark. These results were large and reliable (with a total *n*=3,000, and Cohen's *d* effect sizes between 1.19 and 1.67 for between-subjects comparisons), and were replicated using both overt and implicit context effects. This illustrates how insights from psychophysics can help improve measurement in moral psychology.

The LORD detests dishonest scales, but accurate weights find favor with him.
(Proverbs, 11:1)

3.1 *Introduction*

Morality is central not only to everyday life, but also to the study of how the mind works. Studying morality can reveal the nature of various underlying cognitive mechanisms (e.g. Cushman & Greene, 2012), and can serve as a hub which links together the central methods and concerns of many different parts of cognitive science and beyond — from philosophy and psychology, to economics and evolutionary biology (Heiphetz & Cushman, 2021). Moreover, morality influences many mental processes: it "predominates" social judgment (Goodwin et al., 2014; cf. Melnikoff & Bailey, 2018), is central to our sense of identity (Strohminger & Nichols, 2014), and affects judgments of beauty (Dion et al., 1972), causality (Kominsky et al., 2015), possibility (Acierno et al., 2022), intentionality (Knobe, 2003), agency (Khamitov et al., 2016), and emotion (Prinzing et al., 2023). And immorality seems particularly consequential for cognition: it stokes curiosity (Wylie & Gantman, 2023), grabs attention (Vanneste et al., 2007), distorts memory (Carlson et al., 2020; Pizarro et al., 2006), constrains imagination (Liao et al., 2014), and dilates subjective time (Jia et al., 2021) — and even pretending to act immorally can increase blood pressure (Cushman et al., 2012).

3.1.1 *Moral measurement*

One factor that may help to explain the current renaissance in explorations of moral psychology (Malle, 2021) is that moral intuitions seem relatively easy to study. If you want to know how motion is extracted from visual stimuli, you need to conduct subtle psychophysical experiments; but if you want to know whether people find something to be immoral, you can just ask them. And how can moral judgment be studied with quantitative precision? This also seems straightforward: you can still just ask them, but use numbers. As such, one simple experimental method predominates the study of moral judgment: subjects read a short vignette (about a person, an event, or a dilemma, etc.; for libraries of examples, see Clifford et al., 2015; Knutson et al., 2010), and then they record their judgment using a scale that is labeled with the relevant moral concept.

Perhaps the two most ubiquitous sorts of scales used for this purpose are the Likert scale (Likert, 1932) and the Visual Analog Scale (Aitken, 1969). Likert scales require subjects to choose among a limited number of discrete response categories — e.g. a 7-point scale that ranges from not at all wrong to very wrong (e.g. Gray & Keeney, 2015), extremely immoral to extremely moral (e.g. Minson & Monin, 2012), not at all blameworthy to very blameworthy (e.g. Young et al., 2010), forbidden to obligatory (e.g. Cushman et al., 2006), etc. Visual Analog Scales, in contrast, allow subjects to report a judgment that falls anywhere along a continuum that is anchored with the same sorts of labels (e.g. Siegel et al., 2017; Sosa et al., 2021) — e.g. clicking on a line with a left endpoint labeled not at all wrong, and a right endpoint labeled very wrong. These two measures are often collectively referred to as labeled scales (Bartoshuk et al., 2003). Labeled scales can differ dramatically from study to study in terms of their endpoints, number of labels, and wording, but in a general sense they are utterly ubiquitous in the study of moral psychology. (For example, out of the 50 most recent research articles published in *Cognition* that contained the word "moral" in the title or as a keyword, 47 used a labeled scale to test a key hypothesis: 31 used a Likert scale, 6 used a Visual Analog Scale, and 10 used both. Of the remaining articles, one was an infant study using dichotomous choice as a dependent measure, another used cheating behavior as a dependent measure, and the third was a meta-analysis. A complete list is included in the supplementary data file.)

This ubiquity is partially understandable, since of course such scales seem like the most direct possible ways of asking about moral judgment, and are easy to implement in online studies which have come to dominate social psychology over the past decade (C. A. Anderson et al., 2019). But this ubiquity is also surprising, since such scales have an especially deep problem: these scales are always used within a certain frame of reference (Parducci, 1965), and unless that frame of reference is in some way made explicit, then it is not possible to compare one judgment to another. Put differently, unless these scales are anchored to some shared baseline, then responses on them are inevitably relative. This problem is extremely familiar in the context of everyday adjectives. Consider "big", for example: 9 seems intuitively like a big number (since it implicitly evokes the range of 1-10), but 221 seems intuitively like a small number (since it implicitly evokes the range of 1-1000; Birnbaum, 1999; Leong et al., 2019). And similarly, you might naturally say that you have a big dog, but a small house — despite the obvious fact that the latter still dwarfs the former (but see Bridwell, 1963). Such observations seem obvious and pedestrian, but they can wreak havoc on experiments — especially those with between-subjects designs. Even in a uniform population, one group might give an average response of 5 out of 7 on a "How big is your dog?" Likert scale, and another group might give a 3 out of 7 on a "How big is your house?" Likert scale — but that doesn't provide experimental support for the notion that this population thinks they have houses that are smaller than their dogs. And to foreshadow, this problem — what has been called "one of the most difficult problems concerning intersubjectivity" (Borg, 2001) — is just as salient when the relevant adjective is "bad" instead of "big".

This challenge — both its conceptual foundation and the trouble it can cause for between-subjects experiments — has been widely appreciated in the study of psychophysics, but in an odd manner. The key insight has been recognized for a very long time. More than 65 years ago, for example, this same point was made by S. S. Stevens in his seminal studies of auditory psychophysics (Stevens, 1958), as when exploring how loud sounds appear to be (Stevens, 1956). Yet more than 40 years later, this same point had to be re-discovered in the context of taste psychophysics, e.g. to show how certain studies with labeled scales (say, of how strong some tastes are) missed large differences in actual sensory experience (e.g. Bartoshuk, 2000; Bartoshuk et al., 2003). And in general, this point seems to be strangely both intuitive yet subtle, such that it needs to keep being rediscovered, especially in different subfields (Borg, 2001). (For other examples, see Biernat & Manis, 1994; Biernat et al., 1991; Ubel et al., 2005)). As Linda Bartoshuk once noted, this insight "hasn't penetrated anywhere, because this mistake is being made all over the place" (as cited in Borg, 2001).

3.1.2 *The current studies*

Here we suggest that one of the places this mistake is (still) being made is the study of moral judgment. Even more, we think this issue may be particularly relevant in this domain, for two intertwined reasons. First, as noted above, the use of labeled scales has utterly dominated this field. Second, this field has been particularly sensitive to the dangers of within-subjects designs. Indeed, much of the work in this field employs different versions of the same scenario, carefully holding all factors constant except for one — e.g. whether harm was intentional or not (e.g. Cushman, 2008), whether personal force was used (Greene et al., 2009), or whether an outcome was framed in terms of the proportion that died or the proportion that lived (e.g. McDonald et al., 2021). In such cases, a within-subjects design might far too readily highlight the relevant factor (which might not otherwise be salient; see also Hsee, 1996) — and as a result, such studies have frequently relied on between-subjects comparisons.

These two ingredients (labeled scales, and between-subjects designs), when mixed together with morality, can be scientifically dangerous — but this problem can also be readily solved. We illustrate this here — both the problem and its solution — in a case study of moral judgment. We first developed a vignette describing an act that seemed highly immoral in the context of everyday behavior: a Prank scenario, in which someone is publicly humiliated for her weight. We then contrasted this with another vignette describing an act that also seemed highly immoral but in a far more extreme context: a War Crime scenario, in which a soldier kills a defenseless prisoner of war. (These vignettes were meant to evoke different moral contexts, just as one could evoke different numerical contexts when judging whether a number is "big"; Birnbaum, 1999; Leong et al., 2019.) Across several variations, we show that when participants evaluate such scenarios using labeled scales, between-subjects, they seemingly rate a cruel prank to be just as immoral as (or even worse than) an internationally recognized war crime.

How can one eliminate such seemingly nonsensical results? Here again, foundational work in psychophysics provides a solution: when making such measurements, use an explicit benchmark. This is what is done in the standard psychophysical method of magnitude estimation (e.g. Lodge, 1981; Stevens, 1956, 1966a; for a short primer, see Moskowitz, 1977). This method provides subjects with a benchmark stimulus with a certain pre-assigned value, and then asks raters to make all responses relative to that baseline, in a particular manner — e.g. assigning the brightness of a benchmark light as a 10, and then asking them to rate a light that seemed half as bright as a 5, a light that seemed twice as bright as a 20, etc. In the present context, we assigned a value of 10 to the immorality of stealing a wallet. Subjects then made judgments of immorality in reference to that benchmark, such that they would assign a value of 20 to something that seemed twice as immoral, and a 5 to something that seemed half as immoral, etc. (In contrast to the vast number of moral psychology studies that have used labeled scales, we are aware of only a handful of prior studies that have ever employed magnitude estimation in this domain; e.g. Cohen & Ahn, 2016; Sellin & Wolfgang, 1964).

We employed these manipulations and measures in a series of 9 experiments. We first demonstrated that the seemingly nonsensical pattern of results (where a prank is just as bad as a war crime) holds when tested between-subjects (but not within-subjects) using both a Likert scale (Experiment 3.1a) and a Visual Analog Scale (Experiment 3.1b), but not with magnitude estimation (Experiment 3.1c). We then replicated this pattern (for all three scale types) when the moral context was established by an explicit contrast stimulus, rather than being evoked implicitly by a scenario itself. This was implemented in both a relatively common way (using explicit contrasts of sending a prank email vs. multiple murder; Experiments 3.2a-3.2c), and in an even more extreme way (using explicit contrasts of jaywalking vs. terrorism targeting a preschool; Experiments 3.3a-3.3c).

3.2 *General Method*

Since all experiments shared the same basic structure, we first provide a general methodological overview. All hypotheses, analyses, and sample sizes were preregistered, and the Supplementary Data file provides all raw data and links to preregistrations.

3.2.1 *Subjects*

Online subjects (all from the US, who had completed at least 10 prior online studies with at least a 90% approval rate) were recruited using Prolific (Palan & Schitter, 2018). No subject participated in more than one experiment, and all testing was conducted using the Qualtrics survey platform.

3.2.2 *Measures*

The nine experiments collectively used three different measures (each of which is depicted in a screenshot in the Supplementary Data file). Subjects in Experiments 3.1a, 3.2a, and 3.3a made their judgments using a 7-point Likert scale (implemented using the "modern" Qualtrics layout and default font) that ranged from 1 ("not at all immoral") to 7 ("very immoral”) — with the numbers arrayed horizontally, and with the two labels presented just above the most extreme values. Subjects responded by clicking on one of the seven visible numbers.

Subjects in Experiments 3.1b, 3.2b, and 3.3b made their judgments using a 100-point Visual Analog Scale that ranged from 0 to 100 — with these values placed just above the endpoints of a visible horizontal line, and with the same two labels presented just above their respective values. Subjects responded by clicking and/or dragging a slider (implemented in Qualtrics as a small blue disc) arrayed along the visible line (and initially placed at the line’s center).

Subjects in Experiment 3.1c, 3.2c, and 3.3c made their judgments using magnitude estimation, instructed as follows:

> *Please use a 0 to mean “neither moral nor immoral.” (Imagine something like playing with a pen. It’s neither morally bad nor morally good, it just is.)*
>
> *As a 10, we want you to think about the morality of the following event: stealing a wallet. This event is called your benchmark.*
>
> *In this study, we will ask you to rate other events relative to this benchmark. If the event describes something just as immoral as the benchmark, you should also rate it a 10. If it describes something half as immoral as the benchmark, you should rate it a 5. If it is twice as immoral as the benchmark, rate it a 20, and so on. You can use decimals (e.g. “7.5”) if you feel like you need to.*
>
> *Please try to keep the benchmark and these instructions carefully in mind when you give your responses. If you read about something that seems extremely immoral for example, please don’t respond with an arbitrarily high number. Instead, really think about just how much more immoral it is relative to the benchmark, and rate it accordingly.*
>
> *There’s no right or wrong answer to this question, so just do your best and try to be as careful and thoughtful as possible.*

Subjects entered their rating in a free-response text box which accepted only a non-negative number. (Since subjects were free to enter a number of any size, we preregistered 10,000 as an arbitrary cutoff point for exclusions, since our results might otherwise be skewed by implausibly large responses. In practice, this resulted in the exclusion of 3 subjects in Experiment 3.1c, 2 subjects in Experiment 3.2c, and 7 subjects in Experiment

3.3c (or 1%, 0.7%, and 4.7% of subjects for each study, respectively). Following standard practices for magnitude estimation (Butler et al., 1987), data were log-transformed prior to analysis (with raw ratings all first incremented by one, to avoid undefined values).

3.2.3 *Stimuli*

The primary goal of each experiment was to compare how subjects rated the morality of the event described in each of two scenarios. The first scenario (Prank) was inspired by a story the first author read years ago online:

> *A college freshman has a crush on a boy in her class. He invites her to a party, but it turns out to be a prank hosted by a fraternity to see which pledge can invite the heaviest date. The boy she has a crush on oinks at her until she cries and leaves.*

The second scenario (War Crime) was inspired by the real story of Eddie Gallagher, a navy SEAL who was accused and partially convicted of war crimes:

> *A soldier in an enemy army is receiving medical treatment after surrendering. A marine sees this and kills the prisoner of war with a knife, an action which is internationally recognized as a war crime.*

Design and Analyses

Each experiment tested the scenarios described in the previous section in two conditions: a between-subjects condition (in which independent groups of subjects completed the Prank and War Crime scenarios) and a within-subjects condition (in which the same group of subjects completed both the Prank and the War Crime scenarios). For Experiments 3.1a-3.1c, subjects in the within-subjects condition saw both scenarios concurrently on the same page, presented in a random order. Since Experiments 3.2a-3.2c and 3.3a-3.3c paired each scenario with its own specific contrast stimulus (presented first), subjects in the within-subjects condition always saw the Prank scenario presented above the War Crime scenario (still all on the same page). During recruitment for each experiment, subjects were assigned sequentially to each of these three possibilities (Prank only, War Crime only, Both).

Per our preregistration, all data were primarily analyzed using *t*-tests for the Prank vs. War Crime scenarios (paired samples for within-subjects data, independent samples for between-subjects data), with additional interactions as described below. Per our preregistered hypotheses, we predicted that the War Crime scenario would be rated as more immoral than the Prank scenario for all three measures in the Within-Subjects condition, but only for magnitude estimation (and not for Likert and Visual Analog Scales) in the Between-subjects condition.

3.3 *Experiments* 3.1*a*-3.1*c*: *Raw scenarios*

We first tested whether the key nonsensical result (where a war crime is seemingly rated as no worse than a prank, in between-subjects conditions) would occur with the raw scenarios presented without any additional contrast stimuli, with a Likert scale (Experiment 3.1a), a Visual Analog Scale (Experiment 3.1b), and magnitude estimation (Experiment 3.1c).

3.3.1 *Method*

Separate groups of 300 subjects completed each of the three experiments (with a single subject excluded with replacement from Experiment 3.1c) — 100 subjects each for the Prank scenario, the War Crime scenario, and both. This preregistered sample size was chosen to be roughly in line with past experiments in this domain (e.g. Johnson & Ahn, 2021; Kneer & Machery, 2019).

3.3.2 *Results and Discussion*

The mean immorality ratings for each scenario and condition are depicted in Figures 3.1a-3.1c (for the three measures respectively). The key comparisons in this study always involved a prank (green bar) vs. a war crime (orange bar) — with both common sense and ethical considerations suggesting that the latter should be rated as more immoral than the former (i.e. that the orange bar should be higher than the green bar). Inspection of these figures suggests three primary patterns: First, this sensible result (with orange higher than green) was obtained for all three measures in the within-subjects condition. Second, this sensible result was also obtained for the between-subjects condition with magnitude estimation. But third, this sensible result was not obtained for the between-subjects conditions with either the Likert scale or the Visual Analog Scale.

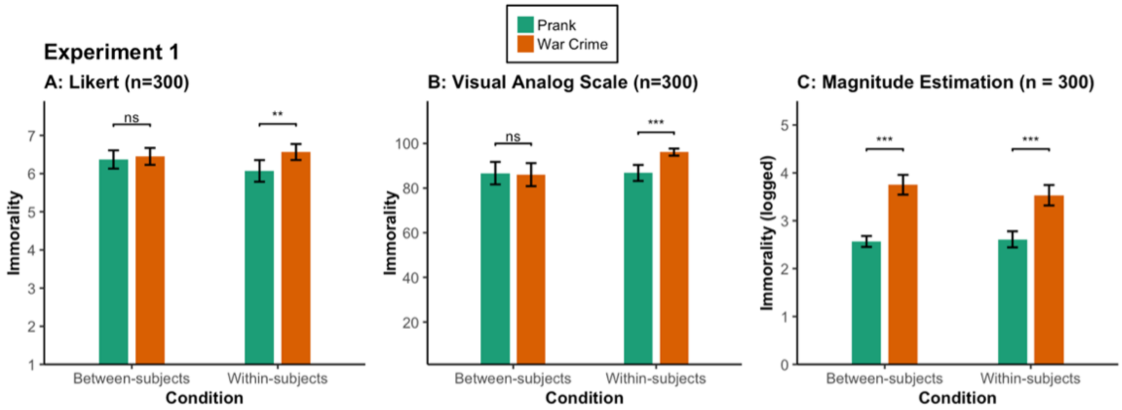

Figure 3.1: Mean ratings of the Prank and War Crime scenarios for (a) the Likert scale, (b) Visual Analog Scale, and (c) magnitude estimation. Error bars represent 95% confidence intervals. *** indicates differences of p < .001. ** indicates differences of p < .01. ns indicates non-significant differences.

These impressions were all verified by the *t*-tests reported in Table 3.1 — which for the between-subjects condition indicate a reliable difference for magnitude estimation, but null effects for the Likert scale and Visual Analog Scale (and of course with reliable differences for all three measures for the within-subjects comparisons). That these measures

Table 3.1: Means and t-test results for Experiments 3.1a–3.1c

Experiment	Measure	Design	Prank	War Crime	t	p	d
3.1a	Likert	Between	6.37	6.45	0.49	.63	
		Within	6.03	6.57	3.17	.002	0.32
3.1b	VAS	Between	86.68	86.02	0.18	.86	
		Within	86.79	96.11	5.20	<.001	0.52
3.1c	ME	Between	13.12	42.11	9.68	<.001	1.37
		Within	13.44	34.01	9.70	<.001	0.97

Note: We report the mean immorality ratings for the Prank and War Crime scenarios under their respective columns for the Likert scale, Visual Analog Scale (VAS), and magnitude estimation (ME) — with raw ratings for Experiments 3.1a and 3.1b, and geometric means for Experiment 3.1c (to make the log-transformed values comparable). Paired t-tests were conducted for within-subjects comparisons, while Welch's Two Sample t-tests were conducted for between-subjects comparisons.

actually differed from each other is also clear from Figures 3.1a-3.1c, but to verify this we also conducted a supplementary analysis in which we first z-scored the between-subjects ratings in each experiment (to make them directly comparable) and then computed two mixed ANOVA interactions between scenario (Prank vs. War Crime) and experiment (3.1a vs. 3.1c, and 3.1b vs. 3.1c); this confirmed that the between-subjects difference with magnitude estimation differed from the null effect for both the Likert scale (F(1, 396)=33.44, p<.001, η^2=.08) and the Visual Analog Scale (F(1, 396)=39.66, p<.001, η^2=.09). (And for completeness, we also computed the mixed ANOVA interactions between scenario [Prank vs. War Crime] and condition [Within-subjects vs. Between-subjects] for each measure. This interaction was reliable for the Visual Analog Scale [F(1, 396)=5.82, p=.016, η^2=.01], but not for either the Likert scale [F(1, 395)=2.95, p=.09] or magnitude estimation [F(1, 396)=1.69, p=.20].) These results provide an initial suggestion of the problematic nature of the labeled scales in this domain, along with the advantages of magnitude estimation.

3.4 *Experiments* 3.2*a*-3.2*c*: *Explicit Contrast Stimuli*

In Experiments 3.1a-3.1c, the relative nature of the ratings obtained with labeled scales was made clear by using scenarios that themselves implicitly triggered different contexts (everyday pranks vs. war crimes), but it is also possible to make such contexts even more explicit, by having subjects rate additional vignettes which "set the stage." Here, we replicated Experiment 3.1 while using an especially mild explicit contrast for the Prank scenario (signing your boss up for spam mail), and an especially extreme explicit contrast for the War Crime scenario (multiple murder).

3.4.1 *Method*

These experiments were identical to Experiments 3.1a-3.1c (testing independent groups of subjects), with the only difference being an addition of two new explicit contrast stimuli. We paired the Prank scenario with the following mild contrast ('Spam'), and the War Crime scenario with the following extreme contrast ('Arson'):

> *Spam*: After receiving a bad performance review at work, an employee signs their boss up to receive more junk mail.
>
> *Arson*: After receiving a bad performance review at work, an employee sets the boss's house on fire late at night, killing the boss along with the boss's spouse and three young children.

3.4.2 *Results and Discussion*

The mean immorality ratings for each scenario and condition are depicted in Figures 3.2a-3.2c (for the three measures respectively) — again with the key comparisons involving a prank (green bar) vs. a war crime (orange bar), and now with the added ratings of the two explicit contrast stimuli also included as the faded bars for their respective scenarios. Inspection of these figures suggests that the key results fully replicated the pattern observed in Experiments 3.1a-3.1c — with the War Crime being rated as more immoral than the Prank in all Within-subjects conditions, but only for magnitude estimation when tested Between-subjects.

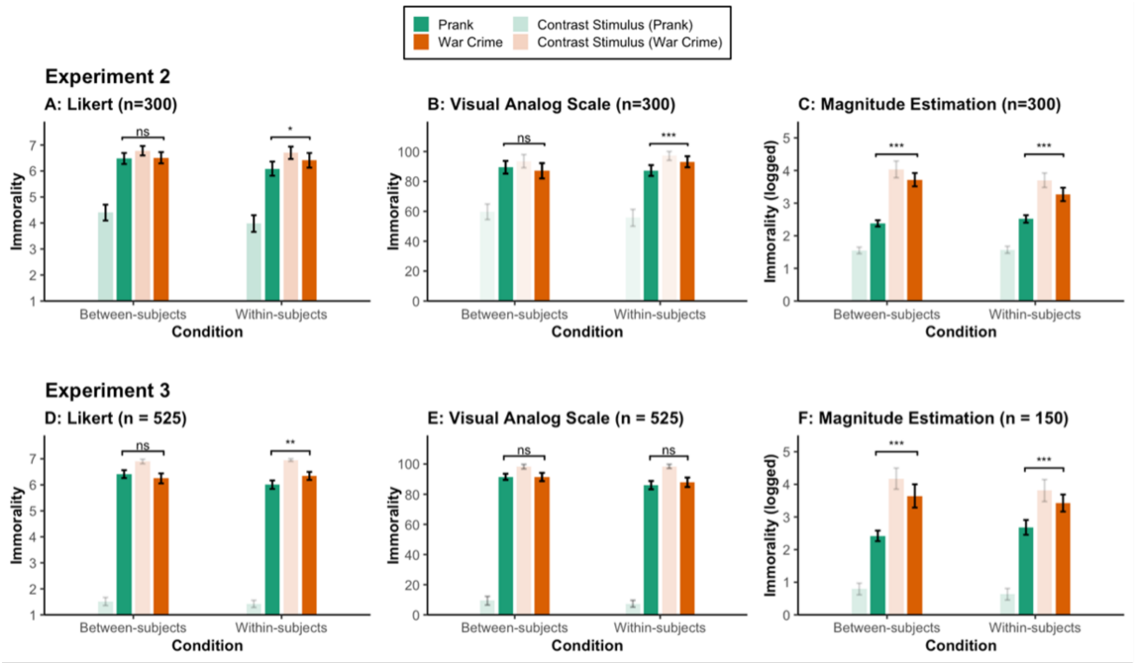

Figure 3.2: Mean ratings of the Prank and War Crime scenarios (solid bars) and their respective explicit contrast scenarios (faded bars) for each of the three measures, for both Experiment 3.2 (a-c) and Experiment 3.3 (d-f). Error bars represent 95% confidence intervals. *** indicates differences of $p < .001$. ** indicates differences of $p < .01$. * indicates differences of $p < .05$. ns indicates non-significant differences.

These impressions were all verified by the *t*-tests reported in Table 3.2 — These results thus provide a conceptual replication of Experiment 3.1 (further reinforcing the problematic nature of the labeled scales in this domain), along with even stronger evidence for the advantages of magnitude estimation (which still produced sensible results even with the additional explicit contrast stimuli).

Table 3.2: Means and *t*-test results for Experiments 3.2a–3.2c

Experiment	Measure	Design	Prank	War Crime	*t*	*p*	*d*
3.2a	Likert	Between	6.48	6.51	0.20	.84	
		Within	6.09	6.41	2.44	.016	0.24
3.2b	VAS	Between	89.42	87.07	0.70	.48	
		Within	87.24	93.11	3.78	<.001	0.38
3.2c	ME	Between	10.80	41.21	11.81	<.001	1.67
		Within	12.38	26.26	7.69	<.001	0.77

Note: We report the mean immorality ratings for the Prank and War Crime scenarios under their respective columns for the Likert scale, Visual Analog Scale (VAS), and magnitude estimation (ME) — with raw ratings for Experiments 3.2a and 3.2b, and geometric means for Experiment 3.2c (to make the log-transformed values comparable). Paired *t*-tests were conducted for Within-subjects comparisons, while Welch's Two Sample *t*-tests were conducted for Between-subjects comparisons.

3.5 *Experiments* 3.3*a*-3.3*c*: *Extreme Contrast Stimuli*

Our final three studies replicated Experiments 3.2a-3.2c in an even larger sample, while also using explicit contrast stimuli that were even more extreme (to our knowledge providing a wider range of immorality than used in almost any past study) — jaywalking in a residential neighborhood (as an explicit contrast for the Prank scenario), and terrorism targeting a preschool (as an explicit contrast for the War Crime scenario).

3.5.1 *Method*

These experiments were identical to Experiments 3.2a-3.2c, with the only differences being (a) larger sample sizes, (b) a tweak in the wording of the War Crime scenario, and (c) the addition of two even more extreme explicit contrast stimuli. We recruited a preregistered 525 subjects for each of Experiments 3.3a and 3.3b, as a power analysis conducted using g*Power software (Faul et al., 2007) found that this was sufficient to detect the smallest within-subjects effect obtained in Experiments 3.1a or 3.1b (*d*=.32) with 95% power and an alpha of .05. We recruited a preregistered 150 subjects for Experiment 3.3c, as a power analysis conducted using g*Power software found that this was sufficient to detect the within-subjects effect obtained for magnitude estimation in Experiment 3.1c (*d*=.97) with 95% power and an alpha of .05. We also tweaked the wording of the War Crime scenario to address possible ceiling effects, and we paired the Prank scenario with an even milder contrast ('Jaywalk'), and the War Crime scenario with an even more extreme contrast ('Terrorism'):

> *War Crime (Experiment 3.3c)*: A soldier in an enemy army is receiving medical treatment after surrendering. A marine sees this and kills the enemy soldier with a knife.
>
> *Jaywalk*: A teenager jaywalks across a quiet, residential street.
>
> *Terrorism*: A member of a radical paramilitary group detonates a bomb at a preschool, targeting the two young children of a political opponent. All fifteen children are killed in the explosion, as well as their 23-year-old teacher.

3.5.2 *Results and Discussion*

The mean immorality ratings for each scenario and condition are depicted in Figures 3.2d-3.2f (for the three measures respectively) — again with the key comparisons involving a prank (green bar) vs. a war crime (orange bar), and with the added ratings of the two explicit contrast stimuli also included as the faded bars for their respective scenarios. Inspection of these figures suggests that for the Between-subjects condition, the key results fully replicated the pattern observed in Experiments 3.1a-3.1c and 3.2a-3.2c — with the War Crime being rated as more immoral than the Prank only for magnitude estimation. For the within-subjects condition, however, these figures suggest that the sensible pattern (with the War Crime rated as more immoral than the Prank) was still obtained for the Likert scale and magnitude estimation, but now (in contrast to Experiments 3.1b and 3.2b) not for the Visual Analog Scale.

These impressions were all verified by the *t*-tests reported in Table 3.3. That these measures actually differed from each other is also again apparent from Figures 3.2d-3.2f,

Table 3.3: Means and t-test results for Experiments 3.3a–3.3c

Experiment	Measure	Design	Prank	War Crime	t	p	d
3.3a	Likert	Between	6.41	6.25	1.34	.18	
		Within	6.01	6.34	3.17	.002	0.24
3.3b	VAS	Between	91.55	91.44	0.06	.95	
		Within	86.03	87.92	0.97	.33	
3.3c	ME	Between	10.33	35.57	5.96	<.001	1.19
		Within	14.08	30.58	4.36	<.001	0.62

Note: We report the mean immorality ratings for the Prank and War Crime scenarios under their respective columns for the Likert scale, Visual Analog Scale (VAS), and magnitude estimation (ME) — with raw ratings for Experiments 3.3a and 3.3b, and geometric means for Experiment 3.3c (to make the log-transformed values comparable). Paired t-tests were conducted for Within-subjects comparisons, while Welch's Two Sample t-tests were conducted for Between-subjects comparisons.

but to verify this we again conducted a supplementary analysis in which we first z-scored the between-subjects ratings in each experiment and then computed two mixed ANOVA interactions between scenario (Prank vs. War Crime) and experiment (3.3a vs. 3.3c, and 3.3b vs. 3.3c); this confirmed that the between-subjects difference with magnitude estimation differed from the null effect for both the Likert scale ($F(1, 446)=28.29$, $p<.001$, $\eta^2=.06$) and the Visual Analog Scale ($F(1, 446)=21.98$, $p<.001$, $\eta^2=.05$) . (And for completeness, we also computed the mixed ANOVA interactions between scenario [Prank vs. War Crime] and condition [Within-subjects vs. Between-subjects] for each measure. This interaction was reliable for the Likert scale [$F(1, 696)=9.00$, $p=.003$, $\eta^2=.01$] but not for the Visual Analog Scale [$F(1, 695)=0.53$, $p=.47$] or magnitude estimation [$F(1, 196)=2.72$, $p=.10$]). When looking at the p-values in Table 3.3, one might be initially worried that the difference between the two scenarios for the Likert scale is also trending to some degree, but note that these values are in the opposite direction from the sensible pattern — with the Prank now being rated numerically as even worse than the War Crime. (6 of these 350 subjects gave nonsensical responses even for the explicit contrast stimuli vs. the relevant condition: 3 of 175 subjects rated jaywalking as more immoral than the Prank, and 3 of 175 subjects rated the preschool bombing as less immoral than the War Crime. We note in passing that if these 6 subjects are excluded, then the remaining 344 subjects' Likert scale data actually demonstrate a reliable effect in the opposite direction, with the Prank being rated worse than the War Crime in a Welch Two Sample *t*-test; $t(286.5)=2.04$, $p=.042$, $d=.22$).

These results, while providing another conceptual replication of the primary effects, show that the problematic nature of labeled scales — and the advantages of magnitude estimation — hold even in the face of the explicit moral contrasts that are especially extreme.

3.6 *General Discussion*

This project contrasted the immorality of a prank (resulting in humiliation) with the immorality of a war crime (resulting in murder), with the assumption that any reasonable measure of moral judgment should capture the fact that the latter is morally worse than the former. When tested with the psychophysically-inspired method of magnitude estimation (involving an explicit benchmark stimulus, as in Experiments 3.1c, 3.2c, and 3.3c), the results of the current project were entirely pedestrian, and even boring: the war crime was indeed rated as more immoral than the prank, just as everyone would expect — and this was true regardless of several sources of experimental variation (e.g. when tested both with and without an explicit contrast stimulus, and when tested both within-subjects and between-subjects).

The central result of this project, however, was that everything went haywire (with the trolley going off the rails, so to speak) when this same simple contrast was tested using the most common method of moral psychology: labeled scales, tested between-subjects. In six separate experiments (with both Likert scales as in Experiments 3.1a, 3.2a, and 3.3a;

and with Visual Analog Scales as in Experiments 3.1b, 3.2b, and 3.3b) we consistently obtained results in which the prank was rated as just as immoral as (or even worse than) the war crime. This nonsensical result was especially striking in the current project, since the difference that these measures failed to capture was anything but subtle: when tested between-subjects with magnitude estimation, the war crime wasn't just rated as more immoral, but as definitively more immoral — with Cohen's *d* effect sizes of 1.37, 1.67, and 1.19, for Experiments 3.1c, 3.2c, and 3.3c, respectively. Effects of this size are on par with the strongest effects in psychology (e.g. Table 3.1 from Lovakov & Agadullina, 2021), making it especially concerning when common methods cannot capture them.

(And while these common methods did, unsurprisingly, capture the central intuitive result when tested within-subjects, it is worth noting that even here such effects were much weaker with labeled scales, in at least three ways. First, the average effect sizes for the within-subjects effects when tested with labeled scales [Likert Scale: d = .27; Visual Analog Scale: d = .32] were notably smaller than those with magnitude estimation [d = .79]. Second, this contrast was not statistically reliable for the Visual Analog Scale in Experiment 3.3b. And third, as depicted in the nonparametric within-subjects data from Figure 3.3: while the vast majority of individual subjects [on average 84%] rated the war crime as worse than the prank when tested with magnitude estimation, far fewer did when tested with the Likert scale [37%] or Visual Analog Scale [47%].)

The central lesson of this project is thus that the most commonly used measures in moral psychology are deeply problematic, insofar as they may mischaracterize or even entirely miss real effects — and that researchers should instead use magnitude estimation to measure moral judgments. In particular, magnitude estimation provides at least four different advantages in this context, with no appreciable disadvantages:

First, and most directly, magnitude estimation makes moral ratings meaningful on an absolute scale — since the different measurements (even when tested in different groups

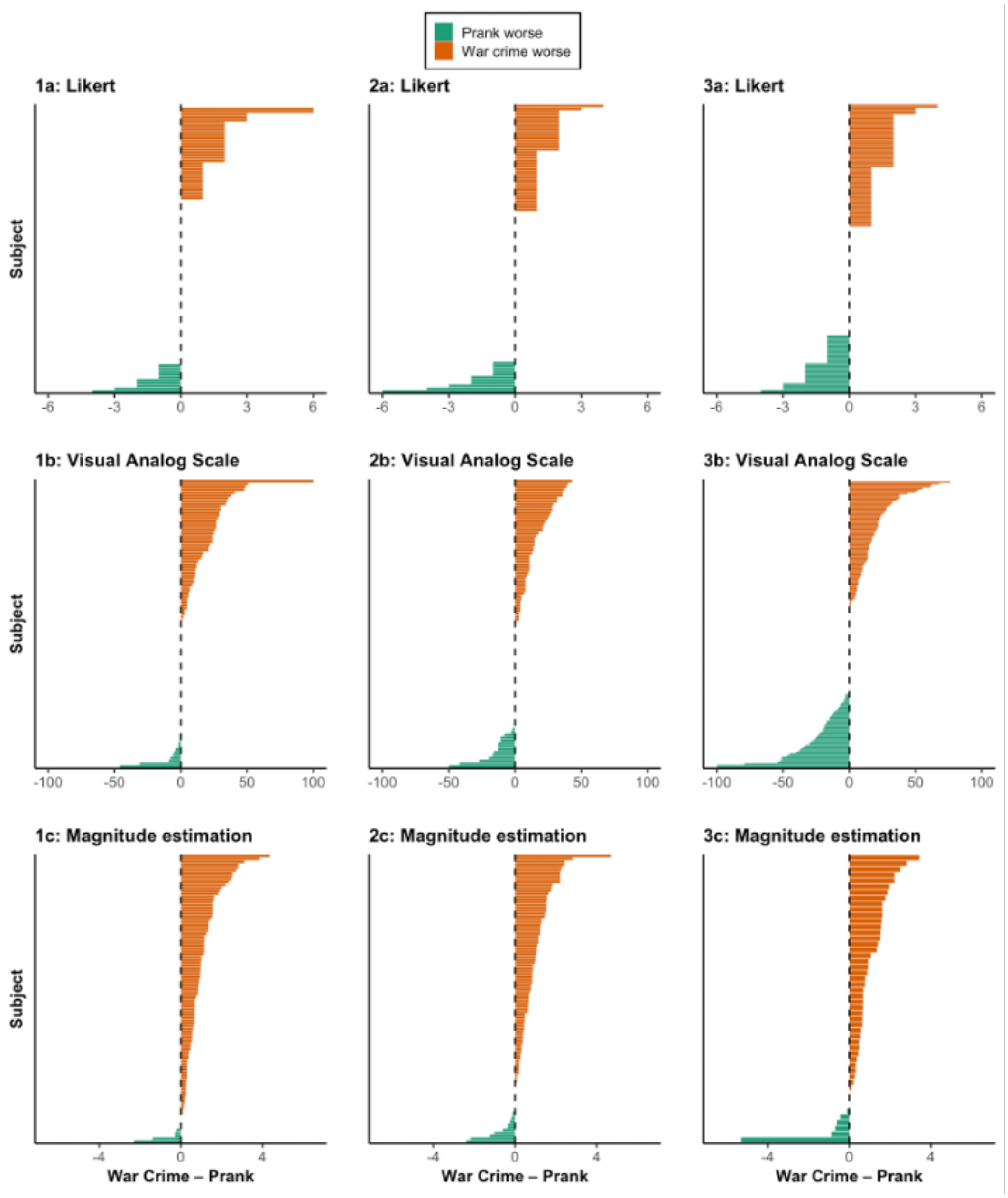

Figure 3.3: Within-subject difference between War Crime and Prank scenarios. Each bar represents an individual subject. Positive bars (in orange) represent subjects who rated the War Crime to be more immoral than the Prank. Negative bars (in green) represent subjects who rated the Prank to be more immoral than the War Crime.

of subjects) can be meaningfully related to each other. In particular, as long as there is no systematic deviation in the ratings of the (ideally non-controversial) benchmark stimulus across the different groups, one could meaningfully compare their relative moral judgments

— whereas this is simply not possible with labeled scales. Indeed, the entire purpose of magnitude estimation as applied to judgment is to make such comparisons meaningful (e.g. Stevens, 1966).

Second, magnitude estimation also allows for meaningful comparison across different experiments (even by different research groups), as long as the studies employ comparable benchmark stimuli. This seems of potentially great value in the study of moral psychology — which currently features hundreds of individual studies, few of which can be directly compared with each other in terms of their absolute ratings. If this research community instead pivoted to the use of magnitude estimation with a set of common shared benchmark stimuli, our science might become more cumulative, such that the results of different studies could be directly combined and compared. (And for discussion of the sometimes nuanced factors involved in choosing benchmark stimuli, we refer readers to standard magnitude estimation primers such as Lodge, 1981 and Moskowitz, 1977).

Third, magnitude estimation also has direct pragmatic advantages for researchers in this domain, due to its ability to more efficiently reveal underlying patterns. This was especially apparent in the current project in the comparison of the effect sizes across the different measures. And in practice, this means that magnitude estimation can reveal effects using far fewer subjects than are required by studies using labeled scales — in a way that can be quantified. For example, if we aimed to conduct a replication of the Within-subjects conditions of Experiments 3.1a-3.1c with 95% power and an alpha of .05, we would need a total of 129 subjects using a Likert scale, 51 using Visual Analog Scale, but only 16 using magnitude estimation.

Finally, the use of magnitude estimation may help to methodologically integrate the study of moral psychology with those of other subfields of cognitive science. As noted above, the problems with labeled scales (and the corresponding advantages of magnitude estimation) are quite familiar in other areas, especially sensory psychophysics. In recent years, there have been several attempts to synthesize research on morality and perception (for a review, see Schein et al., 2016), though it has been suggested that some such efforts go too far (e.g. Firestone & Scholl, 2016). The current project suggests that this useful crosstalk may also extend to how the methods of different fields can enrich each other.

4

Representing moral magnitude:

Linear encoding, scalar variability, and logarithmic compression onto Likert scales

Abstract

While it may seem obvious that subjective stimuli like feelings and judgments can differ in magnitude, it is not clear how such magnitudes could be psychologically realized or encoded. Unlike perceptual magnitudes like space, number, and time, subjective magnitudes have no objective unit on which to aggregate. To address this puzzle, this paper builds on the idea of a domain general magnitude system which is theorized to apply to all *prothetic* (or extensive) stimuli. Through reanalyses of a large-scale data set on judgments of crime severity, as well as conceptual replications in online samples, I first demonstrate that a subjective stimulus like morality nonetheless bears the two hallmarks of a prothetic continuum — variability that increases proportionally with the size of the stimulus and a (near perfect) logarithmic relationship (R^2 = .99) between category (e.g. labeled or Likert) scales and magnitude estimates (which have increasing scalar variability). Crucially, however, this pattern of results is equally consistent with either scale being the "true" measure of subjective magnitude, with the other scale undergoing some transformation (i.e. compression or expansion). Using psychophysical paradigms like magnitude production and bisection, I attempt to disentangle these possibilities. Ultimately, I find that magnitude estimates track psychological magnitudes more accurately than Likert scales, and they more effectively predict real world outcomes like prison sentence length. I conclude by considering practical ramifications of these results given the widespread use of Likert scales in psychological science (and the virtually nonexistent use of magnitude estimation), and I propose a preliminary but simple correction that can be applied to remove the nonlinearity in a scale, requiring just one additional rating in a bisection task and a few lines of code.

4.1 *Introduction*

Of the myriad features that the mind extracts from the world, magnitude is one of the most foundational. It's hard to imagine traveling without "how far?"; seeing without "how big?"; chewing without "how hard?"; and planning without "how long?" But not all magnitudes concern the external world — we may find some behaviors morally appalling while others merely evoke minor disapproval, and we may find some people exceptionally kind while others just mildly nice. Importantly, however, such judgments of magnitude involve subjective stimuli which have no objective physical correlate or meaningful unit from which the strength of these judgments can be objectively quantified. This is in contrast with perceptual magnitudes like space, time, and number (e.g. Dehaene & Brannon, 2010), which do have a clear physical correlate and a meaningful unit.

In the study of perception, "subjective magnitude" is often used to simply mean the sensation magnitude as contrasted with the stimulus magnitude (e.g. the perceptual experience of brightness corresponding to a certain number of photons hitting the eye per square inch per second; see Stevens, 1957). Here, I use subjective magnitude in a narrower sense in primary reference to the stimulus itself, not whether the stimulus is being quantified from an objective or subjective standpoint. More specifically, the distinction is between internal and not countable (i.e. nonmetric) stimuli, as contrasted with external and countable (i.e. metric) stimuli.

Of course, there may be rare cases where such subjective stimuli may involve objective magnitudes — as when making moral judgments involving number of lives lost or saved (Shenhav & Greene, 2010) — but it does not seem immediately obvious how such magnitudes could be psychologically realized or encoded. Magnitudes involve greater than or less than judgments of quantity, but subjective magnitudes provide no obvious answer to the question: "quantities of *what?*" A parsimonious explanation would be that subjective magnitudes are quantifiable via a mapping onto other magnitudes that *are* objective, perhaps as part of a domain-general magnitude system (e.g. Bueti & Walsh, 2009; Gallistel, 1989; Walsh, 2003). Accounts of such a domain-general systems theoretically apply to *all* representations of magnitude (e.g. Walsh, 2003, see Box 1), though subjective stimuli have thus far largely gone unstudied in contemporary work on this topic (with few exceptions; e.g. Powell & Horne, 2017).

While there is growing consensus about the existence of such a system, there is ongoing debate about how those magnitudes are represented. Under one account, psychological magnitudes are represented logarithmically with uniform variability, as has been proposed for numerical representations that are logarithmically compressed onto a mental number line (e.g. Dehaene et al., 2008). Under another account, however, these magnitudes are represented linearly and with increasing scalar variability, in which magnitudes are encoded as noisy signals that become noisier in proportion to the size of the stimulus (e.g. Cantlon et al., 2009; Gallistel & Gelman, 2000). It has been difficult to garner evidence in favor one representational format over the other, since both accounts make virtually identical empirical predictions (e.g. Cantlon et al., 2009; Wearden & Jones, 2007).

This mirrors a long-standing and little-discussed controversy in subjective measurement and psychophysical scaling, which also centers around logarithmic representations and scalar variability. Briefly, one of the most consistent findings in psychophysical scaling is that scales whose fundamental operation is discrimination (i.e. category scales, or what we would call today *labeled* or Likert scales) are logarithmically related to scales whose fundamental operation is matching (i.e. magnitude estimation, where judgments are made in reference to a benchmark stimulus). Category scales tend to have variability that is uniform across the range of the scale, but magnitude scales have variability that increases linearly with the stimulus. Here, both accounts *also* seem empirically equivalent, with some favoring magnitude scales (Marks, 1978b), while others maintain that the category scales "constitute the true measure of sensation" (N. H. Anderson, 1972, p. 137) and represent "the way value judgments are expressed in everyday life" (Parducci, 1972, p. 89).

Though they have thus far been largely overlooked in these contexts, subjective magnitudes nonetheless provide a unique opportunity to address both of these puzzles (supposing they are actually distinct — which, for what it's worth, I doubt). Proponents of both views will often respond to unfavorable empirical results (Dehaene, 2001, cf. Brannon et al., 2001) by noting that subjects may not be making judgments about psychological magnitude, *per se*, but rather the *physical* magnitude of the stimulus, itself (as may happen when one reflexively says that winning $1,000 would feel five times as good as winning $200, despite the subjective difference being significantly smaller; Kahneman & Tversky, 1979). Such an effect can even be obtained by merely *associating* subjective magnitudes with corresponding objective magnitudes; the following quote explains this somewhat counter-intuitive view in the context of time perception:

> *Surely, if timed behaviour varies as a linear function of real time, then subjective time must also appreciate linearly? This conjecture, although reasonable, is completely false. All that is needed to perform as participants do in the two experiments quoted is that they possess a set of subjective internal states $s_1, s_2, \ldots s_n$,each member of which corresponds to a real time $T_1, T_2, \ldots T_n$ and that the subjective internal states can be reliably distinguished. To perform, the participant needs only to learn to respond in the presence of some state s_x when the real-time requirement (or real-time stimulus duration) will be Tx. (Wearden & Jones, 2007, p. 1290)*

Though most often treated as a liability, the subjectivity of moral judgment is in this context a valuable asset, in that it serves as a straightforward bulwark against these so-called "physical correlate theories" (e.g. Lockhead, 1992; Warren, 1969). Namely: such accounts require (for obvious reasons) *an actual physical correlate*, which subjective magnitudes like morality necessarily do not have. By removing the possibility for physical magnitude to confound judgments of psychological magnitude, one can straightforwardly apply relatively simple psychophysical paradigms like bisection and completion to make progress on problems that may otherwise remain intractable.

4.1.1 *The current studies*

There are two clear factors that differentiate so-called "prothetic" continua (derived from the Greek *prosthesis*, though slightly modified since medicine got there first; Stevens & Galanter, 1957, Footnote 2), from "methathetic" continua (derived from the Greek *metamorphōsis;* Stevens, 1957, 1960). The first factor involves variability. Since prothetic continua change as the relevant property is added or removed (e.g. loudness or heaviness), variability changes in direct proportion to the size of the prothetic stimulus. In contrast, methathetic stimuli change as the relevant property is replaced with one of a different kind (e.g. angle or hue), creating variability that is uniform across the range of the continuum. The second factor that differentiates these continua involves the relationship between category (e.g. Likert) scale ratings and magnitude estimates. For prothetic stimuli, there is a logarithmic (or power-law) relationship between category scales and magnitude estimates, while this relationship is linear in the case of metathetic stimuli (Stevens & Galanter, 1957).

In the context of moral judgment, these factors are most clearly illustrated by judgments of crime severity collected by Sellin and Wolfgang (1964). This exhaustive project aimed to apply psychophysical methods in the service of impartially quantifying the seriousness of different offenses based on relevant factors (e.g. the number of victims, the use of physical or verbal intimidation, the value of lost property, etc). Since judgments were collected using both magnitude estimation and an 11-point scale, it is possible to directly compare these methods. Such an analysis reveals precisely the two factors just described.

While these data clearly illustrate a prothetic continuum, they cannot determine whether that magnitude is represented in a logarithmic or linear format (in the same way that they cannot determine whether the category scale or magnitude estimate is the more valid measure). Though these two representational formats generate empirically indistinguishable predictions about the relationship between category scales and magnitude estimates, they do not, however, make empirically indistinguishable predictions, *simpliciter*. So long as judgments of crime severity influence more than just how subjects use scales to rate the severity of crime, we should naturally expect any such outcome to be more accurately predicted by the more valid measure. Here, I use the severity of institutional punishment as a test case.

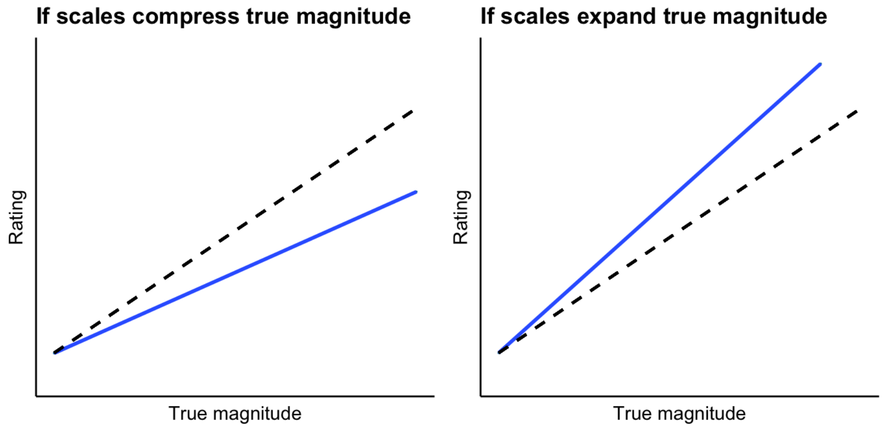

Figure 4.1: Possible patterns of nonlinearity for Likert ratings or magnitude estimates. The dotted line represents a “true” measure of psychological magnitude.

Of course the severity of institutional punishment is highly multidimensional, and reflects many different factors including the past record of the offender, the biases of a judge or jury, or the intent to deter milder crimes whose perpetrators may less often be caught. Despite this, more severe violations reliably elicit harsher punishments, so it stands to reason that one of these multidimensional factors is likely to include judgments of crime severity on at least *some* level, whether on the part of lawmakers, the voting public, or whoever else. Fortunately, Sellin and Wolfgang also report the maximum prison sentence lengths for the 21 primary offenses, allowing the predictive power of both magnitude estimation and category ratings to be directly compared.

Finally, I build on these analyses in a straightforward experiment making use of two very simple psychophysical paradigms which I have adapted for use in moral judgment: magnitude production or completion (in which subjects generate a stimulus of certain size; see Stevens, 1975, pp. 30-31) and bisection (in which subjects generate a stimulus that lies halfway between two other stimuli; see Stevens, 1975 pp. 154-155). In either case, I am able to compare the actual ratings obtained for each generated stimulus with the rating we should expect given a "true" scale of psychological magnitude (as calculated from the ratings provided to the other relevant stimuli; see Figure 4.1). In doing so, I am able to make use of the most foundational operation in measurement: concatenation (which some have explicitly suggested would be necessary to differentiate between logarithmic encoding and linear encoding with scalar variability; Anderson 1978). And since these judgments have no obvious connection to a potentially confounding physical magnitudes, one can safely interpret that these judgments as reflecting subjective magnitude.

4.2 *Study* 4.1: *Reanalyzing judgments of crime severity*

One of the most consistent findings in psychophysical scaling is that prothetic stimuli — which is to say stimuli with magnitude — produce a logarithmic (or power-law) relationship between category scale ratings and magnitude estimates. In this first study, I demonstrate exactly such a relationship in an analysis of data that was originally presented in *The Measurement of Delinquency* by Sellin and Wolfgang (1964), a work that was frequently cited by Stevens as a key example of how psychophysical methods could be applied to problems in social measurement (e.g. Stevens, 1966b, 1975, Chapter 8).

This data proves particularly useful for several reasons: first, it tested a wide range of stimuli using both magnitude estimation and an 11-point scale, allowing direct comparison of the two methods; second, it tested a wide sample (including students, judges, and police officers); and third, it contains other useful data like the maximum prison sentence length (at the time it was written) for the 21 primary offenses my analyses will focus on.

4.2.1 *Methods*

Stimuli

Over the course of three years, two criminologists at the University of Pennsylvania undertook a comprehensive effort to quantify the seriousness of different crimes by using psychophysical methods that had recently been developed by S.S. Stevens (Sellin & Wolfgang, 1964). They began by randomly sampling 10% of the universe of criminal events recorded in Philadelphia in 1960, yielding 1,313 offenses taken from arrest records, remedial files of juvenile offenders, and "The Moral Squads file" (further details available on pp. 132 - 136). These offenses were categorized based on a classification scheme developed by the Federal Bureau of Investigations and adjusted according to existing police data to ensure accuracy. These offenses were eventually condensed and refined to a list of 141 offenses, which are listed in Appendix D (pp. 381-386).

All offenses were pretested on a sample of 17 subjects using a 7-point scale of crime seriousness (see p. 247), and three offenses were selected for each scale value, based on which items had the least variability and medians closest to that value's midpoint. The primary focus of my analyses will be on the 21 offenses that were selected in this way, which constituted the "Primary Index Scale" that was used to validate and determine the weights of the other 120 offenses. These 120 offenses were the main focus of Sellin and Wolfgang's project, and they ultimately made up the final items in the Wolfgang-Sellin Index of Crime Seriousness (see Appendix F, pp. 401-412).

Procedure

The seriousness of these offenses were evaluated using two different methods, either an 11-point scale ranging from 1 (Least Serious) to 11 (Most Serious) or via magnitude estimation, in which judgments are made in reference to the seriousness of stealing a bicycle, which was assigned the value 10.

The instructions for the category ratings were slightly more involved than what is typical in contemporary research, and a few differences are worth noting. First, Sellin and Wolfgang provide two anchor stimuli representing the least and most serious crimes, thus making concrete the range of stimuli considered.[1] Second, they explicitly tell participants how to interpret the different levels of the scale, writing that "[e]ach of the eleven categories is an equal step on the scale of seriousness so that 6 is one step more serious than 5 and 10 is one step more serious than 9, and so forth."

For magnitude estimation, the most relevant instructions were as follows:

> *Use [stealing a bicycle] as a standard. Every other violation should be scored in relation to this standard violation. For example, if any violation seems ten times as serious at the standard violation, write in a score of 100. If a violation seems half as serious as the standard, write in a score of 5. If a violation seems only a twentieth as serious as the standard, write in a score of 1/2 or .50. You may use any whole or fractional numbers that are greater than zero, no matter how small or large they are just so long as they represent how serious the violation is compared to the standard violation (pp. 254-255).*

The scenarios were written on cards, shuffled, and administered in booklets containing the primary index of 21 offenses (which were rated by all subjects) and a further 30 that were randomly sampled from the larger index of 120 offenses. The offender in each case was described as a male of an unidentified age (with a few exceptions later noted).

[1] Unfortunately, Sellin and Wolfgang do not report the violations used as anchor stimuli, though I have reached out to the University of Pennsylvania archives in an effort to obtain the original study materials.

Sample Characteristics

This data comprises of 10 samples taken from 251 college students, 286 police officers, and 38 juvenile court judges. The college students were men enrolled in sociology courses at Temple University (*Temple,* of which 85 made category ratings and 78 made magnitude estimates) and Penn State Ogontz Center (*Ogontz*, of which 45 made category ratings and 43 made magnitude estimates), while the police officers were recruited at the district level, with inspectors selecting a representative portion of their division to take part (*Police*, of which 144 made category ratings and 142 made magnitude estimates), while juvenile judges were mailed test booklets to complete, of which 38 did (*Judges*, of which 20 made category ratings and 18 made magnitude estimates).

Of note, some subjects rated the seriousness of offenses committed by males aged 13 years (all *Judges*, 40 *Temple*, and 62 *Police*), males aged 17 years (40 *Temple* and 63 *Police*), and males aged 27 years (42 *Temple* and 59 *Police*). Since Sellin and Wolfgang found consistent judgments across all offender ages, I'll be omitting this variable from my analyses.

The data

Means and prison sentence length data are printed directly in tables found in Appendices E1-E8 of the manuscript (pp. 387-400).[2] There are a few important details to note. First, the central tendency of magnitude estimates was calculated by taking the geometric mean, rather than the familiar arithmetic mean one typically associates with averaging. Second, standard deviations were only listed for magnitude estimates, so variance cannot be directly compared across category ratings and magnitude estimates. And third, the mean magnitude estimates that make up the primary index of 21 offenses was calculated by taking

[2] Data entry for this study was performed by Breanna Nguyen, to whom I am grateful.

the average of the 10 geometric means calculated for each offense in each sample, with minor corrections based on standardization in one subsample (*Ogontz*, see pp. 278-279). I similarly averaged the category ratings across the 10 samples to construct a comparable mean for the 21 offenses. Finally, Appendix E-8 contains data on sentence length in 30 day units for the 21 primary offenses, as established in the Pennsylvania Penal Code (with an upper limit set at 50 years, a value which was assigned to death sentences).

Portions of this data were sometimes lightly analyzed or plotted throughout the original manuscript, though never fully aggregated and typically presenting results from individual samples. This is partly because the main focuses of this data were to 1) justify the use of magnitude estimates over category ratings for the final index (largely based on the presence of ceiling effects in the latter; see Chapter 16) and 2) use the 21 items to determine the weights and scoring scheme of the final Wolfgang-Sellin index (see Chapter 17). The data involving prison sentences lengths were correlated only for magnitude estimation in the *Ogontz* and *Police* samples, though no explanation was given as to why only these samples were chosen (pp. 327-328).[3] Given this (and given that modern statistical software straightforwardly allows computations that would have been prohibitively time-consuming at the time these data were collected) none of the following analyses were presented in the original manuscript.

4.2.2 *Results and Discussion*

The central finding of this study should be readily apparent upon visual inspection of Figure 4.2, which plots the means of the category ratings and magnitude estimates for the

[3] Note: there is also a discrepancy between the information in Appendix E-8 and the analysis described on p. 327. The table reports that the death penalty was assigned a maximum value of 50 years, whereas p. 327 says the value chosen was 43 years (a number they say approximates the life expectancy of the median age revealed in studies of criminal homicide). Regardless, they recognize that the choice of this value is arbitrary, and show an attenuated (though still clear) correlation on p. 328 when dropping capital offenses from the analysis.

21 offenses in Sellin and Wolfgang's primary index. No statistical tests are needed to see that the relationship between these measures is clearly (and nearly perfectly) logarithmic, indicating that judgments of crime severity are represented as having a magnitude. Were crime severity a metathetic continua — which is to say one involving qualitative, rather than quantitative changes — this plot would instead demonstrate a linear relationship between the two scales.

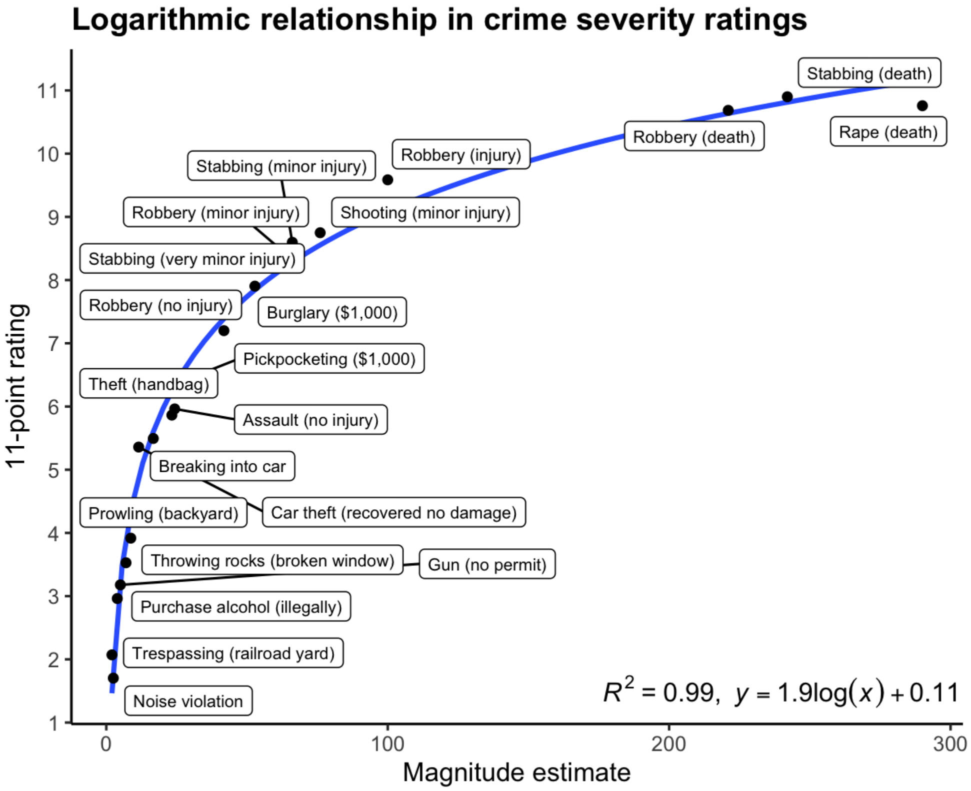

Figure 4.2: Mean ratings of crime severity (from Sellin & Wolfgang, 1964) reveal the expected logarithmic relationship classically found in prothetic continua. The y-axis plots arithmetic means calculated from ratings on an 11-point scale, while the x-axis plots geometric means calculated from magnitude estimates in reference to the violation "stealing a bike," which was assigned the value 10. Each point is labeled with a shortened description of the offense, and the blue line corresponds to the equation printed in the lower right corner of the graph.

I ran several more analyses for thoroughness. First, I obtained the R^2 value for the uncorrected primary index ratings (which I calculated by simply taking the arithmetic mean across the 10 samples for each of the 21 offenses). This yields virtually identical results as those presented in Figure 4.2: $R^2 = .99$, $y = 2.0\log(x) + .12$. Next, I calculated the R^2 values individually for each of these 10 samples. The lowest value was obtained in the *Police* sample that rated offenders who were thirteen years old, $R^2 = .92$.

Next, I tested whether variability remained uniform or increased in proportion to crime severity. Appendix E-5 reports the z-scored means and standard deviations for the magnitude estimates (p. 395), though no such data were available for the category scale ratings. Nonetheless, this data can be used to calculate the relative variability of magnitude estimates by dividing the mean for each offense by its standard deviation (see e.g. John, 1971). This reveals an exceptionally clear pattern, which is illustrated below in Figure 4.3. Variability increases as a direct function of stimulus magnitude.

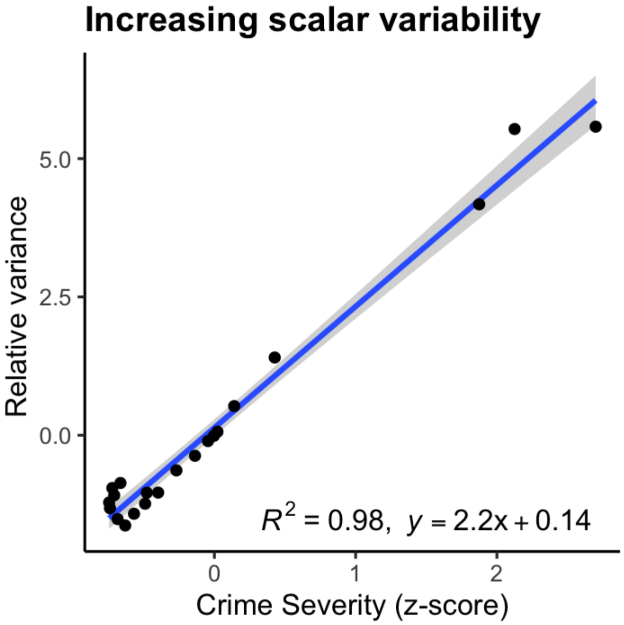

Figure 4.3: Relative variability increases as a function of stimulus magnitude.

It is important to note, however, that there are two possible interpretations for this overall pattern of results: the magnitude estimate may be capturing the "true" judgment of crime severity (which is then compressed onto the category scale), but so too may the category scale be capturing the "true" judgment (which is then expanded onto the magnitude scale).

As a first step toward answering this question, I directly compared how well each measure predicted the maximum prison sentence recorded by Sellin and Wolfgang as taken from the Pennsylvania Penal Code.

To start, I ran two simple regression models to confirm that both magnitude estimates and category ratings (both z-scored) predicted the maximum prison sentence length for different offenses. Results were significant for both category ratings (β=162.7, p<0.001, *adjusted* R^2=.59) and magnitude estimates (β=194.5, p<0.001, *adjusted* R^2=.87), suggesting that a one standard deviation increase in crime severity corresponds to about 16 additional years in prison as measured using magnitude estimation, and 14 additional years as measured using the 11-point scale. Next, I used a Fisher's z-test to compare the correlation coefficients between each rating and the maximum prison sentence length. This test, however, was not significant, z=1.94, p=0.053, though it is worth nothing that these correlations were based on only 21 ratings for each measure, each representing the average across 10 different samples. Thus, I conducted the same analysis using all 10 means per offense per scale, rather than 1. This analysis did reveal a significant difference, z=2.05, p=.04, such that magnitude estimates correlated with prison sentence lengths more strongly than did category scales.

Taken together, these analyses provide strong evidence that moral judgments, though subjective, are nonetheless represented as having a magnitude. These data also provide some initial evidence in favor of two related views. First, that subjective moral magnitudes are encoded linearly and with increasing scalar variability; and second, magnitude scaling provides a more valid measure of this subjective magnitude. The final study aims to provide more conclusive evidence in favor of these two views.

4.3 *Study* 4.2: *Moral bisection and completion tasks*

Since perceptual magnitudes correlate with physical magnitudes, judgments of the former may be confounded by judgments of the latter. Because moral magnitudes are inherently subjective, however, it becomes possible to circumvent this problem. Thus, one can straightforwardly measure subjective magnitudes by applying similar procedures that have been used (inconclusively) in the study of perceptual magnitudes — e.g. time (Wearden & Jones, 2007), length (Brannon et al., 2001; Dehaene, 2001), and brightness (Stevens, 1961).

The basic logic for both procedures is as follows. In a completion task, subjects produce a stimulus to match a given psychological magnitude — in this context, a crime that would make a second list of offenses just as bad as the first. If a scale either compresses or expands that psychological magnitude, then the rating given for the chosen stimulus task would either be smaller than expected (if the scale compresses, which would be the expected outcome given that magnitude estimates reflect the "true" psychological magnitude) or larger than expected (if the scale expands, which would be the expected outcome given that the Likert scale reflects the "true" psychological magnitude). A similar logic holds for the bisection task, except that subjects are choosing a stimulus whose psychological magnitude is halfway between two stimuli of a given magnitude.

4.3.1 *Methods*

Subjects and procedure

A convenience sample of 60 subjects was recruited via the Prolific survey platform (Palan & Schitter, 2018) as part of a short study on estimating the wrongness of different crimes. The study took approximately six minutes to complete, and subjects were paid a dollar each for participating.

There were three main components to this study, which subjects completed in the following order: a completion task (loosely adapted from studies involving magnitude production; Stevens, 1966a, Stevens, 1975, pp. 31-32), a bisection task (Stevens, 1975, pp. 154-155), and stimulus rating (using both a Likert scale and magnitude estimation in counterbalanced order).

Stimuli and materials

Completion: For this task, subjects were presented with two lists of stimuli (List A and List B), each containing violations that were chosen to be at approximately the same level of seriousness in the Sellin-Wolfgang index (1964): one high severity, one moderate severity, and one low severity. However, while List A contained all three stimuli, List B contained only two. Thus, the subject was prompted to note the overall wrongness of List A, then add a stimulus of their choosing to make List B and List A equally bad.

The items for each list were as follows:

List A:
Stabbing resulting in death (premeditated)
Burglary (>$10,000)
Owning a gun with no permit

List B:
Attempted murder resulting in minor injury
Stealing a car and abandoning it (no damages)
____________________?

The logic of this task is as follows. Subjects were instructed to choose a stimulus whose subjective magnitude, when added to List B, produces two lists of (approximately) equal subjective magnitude. If one scale is the “true” measure of psychological magnitude, and if the other scale merely transforms that true magnitude (whether via compression or expansion), then we should expect only the "true" scale to produce ratings such that the sum of List A will be (approximately) equal to the sum of List B.

Bisection: For this task, subjects were prompted to consider the wrongness of two crimes, one from List A (stabbing resulting in death) and one from List B (stealing a car). In a free response box, subjects were prompted to generate a stimulus that fell halfway between the other two stimuli in terms of how wrong it was.

This procedure has effectively the same logic as was just described for the completion task, with the only difference being that the halfway point should be distorted on the transformed scale. Thus, we should expect the more valid scale to produce a rating for the bisected stimulus that is (approximately) equal to the average of the two stimuli that were bisected.

Ratings: All subjects rated each stimulus twice, including the completion and bisection stimulus they entered as a free response. They made these ratings using either a 10-point Likert scale ranging from 1 (Not at all bad) to 10 (Extremely bad), or magnitude estimation (in which ratings were made in reference to a benchmark stimulus: stealing a wallet).

The instructions for magnitude estimation were presented as follows:

> *For this kind of rating, we'll ask you to describe the badness of a crime by directly comparing it to a different crime:* ***stealing a wallet****.*
>
> *First, we want you to give this crime any number that feels right to you: people tend to overthink this, but it really doesn't matter what you choose. What does matter is that you assign the other crimes a number in relationship to stealing a wallet. Suppose you gave it a "12."*
>
> *If another crime is just as bad as stealing a wallet, then you should also give it a 12. If it is half as bad, you should give it a 6. If it's ten times as bad, give it a 120, and so on. Please use decimals if you feel like you need to, and try to remember that there's no right or wrong answer to these questions, so just do your best.*

Subjects were then prompted to enter a value for the benchmark stimulus. Before analysis, each subject's ratings were standardized by dividing all ratings by the initial value assigned to this benchmark stimulus — in effect transforming each magnitude estimate into a ratio judgment in relation to how wrong it is to steal a wallet (e.g. a standardized rating of 5 would then be interpreted as something that is five times as wrong as stealing a wallet). Then, in line with standard procedure for data obtained through magnitude estimation (Butler et al., 1987), all magnitude estimates were log-transformed after having been incremented by one to avoid undefined values.

In sum, subjects rated a total of 15 stimuli: the 7 rated using both methods (3 from List A, 2 from List B, 1 from the completion task, and 1 from the bisection task), as well as the rating given to the benchmark stimulus during magnitude estimation.

4.3.2 *Results*

The central finding of this study should also be readily apparent upon visual inspection of Figure 4.4, which simply plots the summed ratings of List A on one axis and the summed ratings of List B on the other. It is immediately clear that the magnitude estimates form nearly a straight line, and almost just as clear that the Likert ratings do not.

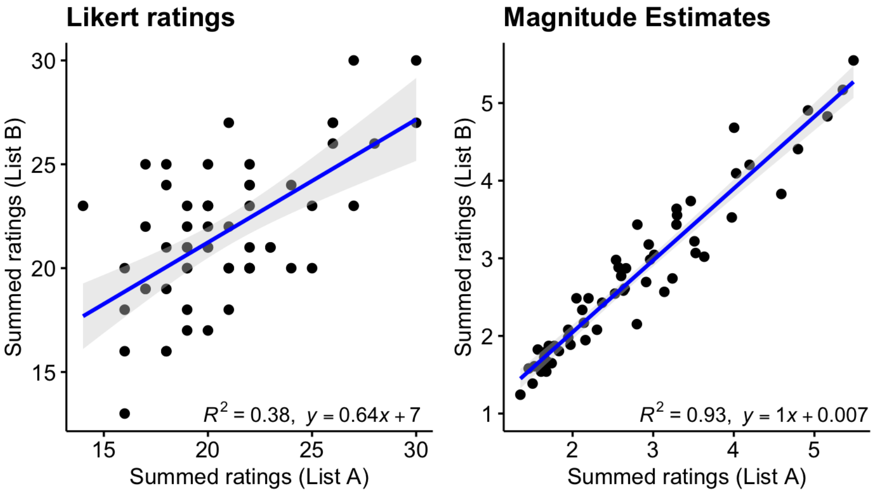

Figure 4.4: Summed Likert ratings (left) and magnitude estimates (right) for Study 4.2's completion task. Subjects were instructed to match the subjective magnitude of List B to List A. The blue line corresponds to the equation printed in the bottom right corner, and shaded areas represents 95% confidence intervals.

A Fisher's *z*-test confirms what is obvious in this figure: the correlation between the two lists is significantly stronger when measured using magnitude estimation compared to when measured using Likert ratings, *z*=6.60, *p*<.001. Descriptively, it is also striking to compare the slopes of the two measures. The obtained slope for magnitude estimation (*m*=1) is exactly what we should hope to obtain in a task involving the one-to-one match of one stimulus to another. The obtained slope for Likert ratings (*m*=.64) is more troubling, and what one should in fact predict if Likert ratings are compressing the "true" subjective magnitude, as discussed above.

This same pattern also obtains when plotting the bisection ratings against the average of the two bisected ratings, which should ideally form a straight line. However, a visual inspection of Figure 4.5 below makes clear that the only approximately linear pattern of responses was obtained using magnitude estimation. Again, a Fisher's *z*-test confirms what is obvious in this figure: the correlation between the predicted and actual bisection ratings is significantly stronger when measured using magnitude estimation compared to when measured using Likert ratings, *z*=3.60, *p*<.001.

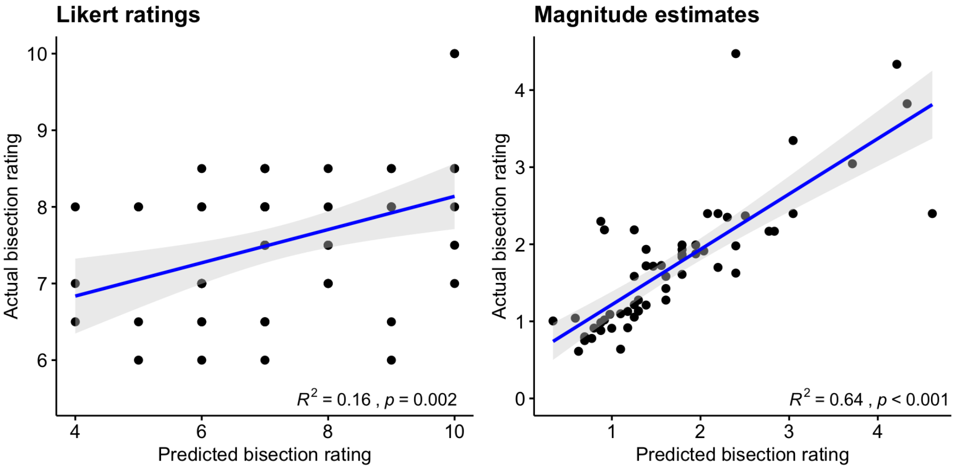

Figure 4.5: Predicted vs. actual ratings in Study 4.2's bisection task, in which subjects were instructed to generate a stimulus that would be halfway between two other stimuli. Ratings for the generated stimuli are on the y-axis, while the average of the two bisected stimuli is presented on the x-axis. Shaded areas represents 95% confidence intervals.

In both the bisection task and the completion tasks, subjects matched one subjective moral magnitude with another. This would only be evident, however, by looking at the responses from one measure: magnitude estimation. From this, we can straightforwardly infer that magnitude estimates are more closely aligned with *actual* subjective magnitudes.

Correcting the nonlinearity of Likert scores

It is clear from these data that Likert ratings are not linearly related to subjective magnitude, and I believe that the clear lesson given the above is that researchers should use magnitude estimation, rather than the ubiquitous Likert scale. Even so, I recognize that Likert scales are deeply entrenched, and methodological changes come slowly, if at all. Thus, it'd be pragmatically and methodologically valuable to have a straightforward procedure to correct the nonlinearity present in our most commonly used measures. Below, I describe just such a procedure in brief. Full details are available in Appendix A.[4]

Given the relationship classically obtained in psychophysical scaling (e.g. Stevens, 1966a), we can represent the relationship between Likert ratings R_n and the subjective magnitude M_n to be as follows:

$$M_n = kR_n^x + c \tag{4.1}$$

where k and c are arbitrary constants related to the unit and values unique to the Likert scale based on the number of items, etc, and x is the exponent creating the nonlinearity. However, if x is known for each subject, then it would be possible to undo that nonlinearity by obtaining a corrected value R'_n that *is* linearly related to M_n in the form:

$$M_n = kR'_n + c \tag{4.2}$$

[4] I'm grateful to Marlene Berke for sanity-checking the basic math and helping to clarify parts of what follows.

There is some precedence for a procedure like this, though all cases I'm aware of require an objective physical magnitude that can be known precisely (e.g. Marks, 1968). However, the rating obtained via the bisection task R_b can serve the same function, since it not only tell us what the rating for the bisected stimuli $\frac{1}{2}(M_1 + M_2)$ actually *is*, but more importantly, it tells us what $\frac{1}{2}(R_1 + R_2)$ *would have been* if the rating scale were, in fact, linear — which is to say, it provides $\frac{1}{2}(R'_1 + R'_2)$. What this means, concretely, is that it allows us to remove constants k and c from the equation. And since, by stipulation $R'_b = R_b$, we can represent the relationship between R_b and the bisected ratings R_1 and R_2 as follows:

$$R_b = \frac{1}{2}(R_1^x + R_2^x) \tag{4.3}$$

The above equation makes it possible calculate from each subject's ratings — R_b, R_1, and R_2 — an exponent x that is unique for each subject. Then, any rating by any subject can be corrected by raising it to their exponent x. Applying such a procedure to the Likert data from Study 4.2 reveals the following:

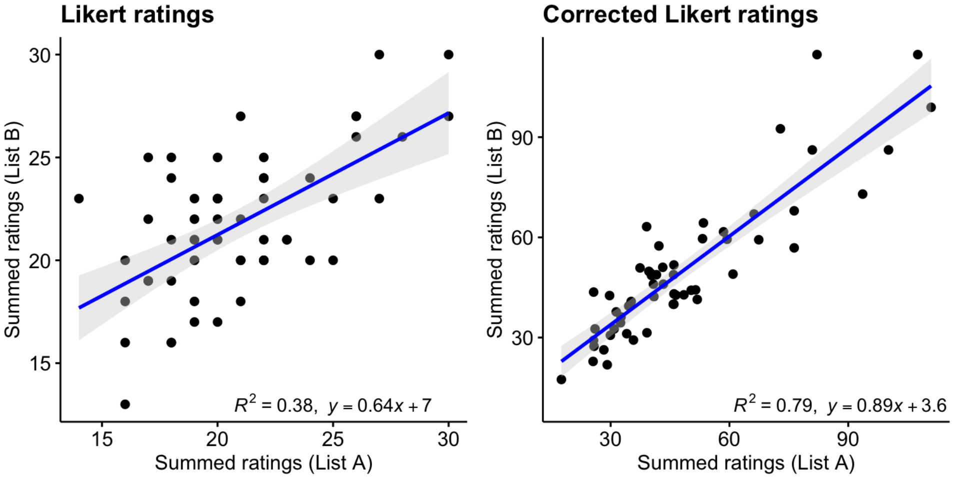

Figure 4.6: Raw (left) and corrected (right) Likert ratings for the completion task from Experiment 2. Shaded areas represents 95% confidence intervals.

Though as much should be obvious from Figure 4.6 above, Fisher's z-test reveals that the correlation between List A and List B was higher for the corrected Likert values compared to the uncorrected values(z=3.81, p<.001).

4.4 *General Discussion*

It seems intuitively obvious that only quantifiable stimuli could possibly have magnitude. Given this, it's perhaps unsurprising that moral judgments and other subjective stimuli have largely been absent from the broader literature exploring psychological magnitudes and how they're represented. Skepticism about subjective magnitudes is surely understandable, but the result of these studies has been decisive. First, I found that morality shows all the hallmarks of a stimulus with a psychological magnitude — specifically, increasing scalar variability and a logarithmic relationship between magnitude estimates and category ratings. Second, I found that morality is more accurately measured by magnitude estimation and best described by a representational format that is linear with increasing scalar variability. And third, I illustrated how a simple psychophysical task can be used to correct the nonlinearity found in one of the most commonly used scales.

This project has aimed, in part, to gently invite researchers in this area to reconsider their stance on the utility and measurability of subjective magnitudes more generally and moral magnitudes in particular. While it is true that morality is inherently subjective, this is precisely what makes it so useful as a tool to study the basic features of how the mind encodes magnitude. Indeed, one of the most stubborn confounds in this work is the correlation between subjective magnitudes and physical magnitudes, and this problem disappears entirely in the context of something like moral judgment.

One skeptical response worth considering, however, is something like the following: perhaps moral magnitudes really *do* have a physical correlate that could be confounding these results. Though this certainly has an *ad hoc* flavor to it, it very well could be the case that people may build *some* numerical associations with certain moral magnitudes over time. This objection is worth taking seriously, but it's first worth noting how radical it would be to accept even this deflationary account. If true, this would serve to entirely dissolve the distinction between perceptual and subjective magnitudes, full stop, at least as it relates to how the mind represents magnitude. If even *morality* doesn't qualify as subjective, then it's not clear what else possibly could. More substantively, however, it's true that there are a host of physical correlates one could point to as relating in some way to subjective moral magnitude — such as time spent in prison, the pain observed in a victim, the amount of hateful comments received online in response to some transgression, and so on. It doesn't take much imagination to come up with more examples, so perhaps some combination of these sorts of physical correlates could be enough to produce a confounding effect.

Though I'd agree that it's possible that these prospective physical correlates may have *some* kind of systematic and predictable relationship to judgments of subjective moral magnitude, I think there's one crucial reason why such an account couldn't serve as a plausible confound for the results presented above: it wouldn't just be the case that these physical representations of magnitude are *correlated* with subjective moral magnitude, they would in fact be *caused* by it — in other words, this wouldn't be a confound, but an *effect*. The number of photons hitting a surface per square inch per second does not increase or decrease as a function of apparent magnitude, nor does the duration of 9,192,631,770 periods of the radiation corresponding to a caesium-133 atom transitioning between the two hyperfine levels of the ground state become longer or shorter based on one's subjective experience of time. However, this is exactly how these candidate physical correlates *must* work in the context of moral judgment — we punish the worst offenders so severely because our moral sentiments are the most severely offended in precisely those cases; it's not the other way around.

It seems clear, then, that there must be *some* subjective moral magnitude which preceded any of these physical correlates, so there's a very real sense in which — even if the "confound" is producing the given pattern of results, it's ability to do so depends fully on the phenomenon that it is being marshalled in service of explaining away. It seems as if the only way to avoid this would be to appeal to some kind of incidental confound relating to *another* physical correlate, as we might see when asking subjects to compare how wrong it is to steal this or that amount of money, or kill this many or that many people, and so on. As such, I believe that the most obvious explanation for these results is that they are reflecting the true representational format of moral magnitude: it is linearly encoded, it has scalar variability, and it is logarithmically compressed onto Likert scales.

4.4.1 *Implications for domain general magnitudes*

Finally, it is worth considering how these findings fit into the broader theoretical work concerning domain-general magnitudes and how they're represented. To start: these findings have essentially no direct bearing on this debate, whatsoever. Though I certainly believe it would be parsimonious for there to be just one domain general representation of magnitude shared across all prothetic continua — as well as a compelling explanation for the question I started with: how exactly *do* subjective magnitudes get their sense of quantity in the absence of something to quantify? — it is also very plausible that subjective stimuli and perceptual stimuli are just represented in entirely different ways. Though some recent work has started to look at cross-modal interference between apparent moral magnitudes and perceptual magnitudes (Powell & Horne, 2017), I believer further research in this area is the most straightforward way to move forward on this problem.

Theoretical implications aside, I believe the practical implications of these findings are by far the most salient. There are an enormous number of researchers in our discipline who make use of Likert scales within *some* aspect of their work. The findings I've reported here, however, suggest that data from these scales could be systematically distorting whatever it is that they're measuring. Though I believe the most promising solution would be to use magnitude estimation (at least for the measurement of prothetic continua), I believe that there is another clear benefit to the methods I've adapted in this project: the inclusion of a bisection task can correct for nonlinearity in subjects' use of category scales. A brief summary of this procedure appears in the appendix on the following page.

4.5 *Appendix*: *Correcting Likert ratings*

Via 4.2, we can relate all the collected ratings to their corresponding magnitudes as follows:

$$\begin{aligned} M_1 &= kR_1^x + c \\ M_2 &= kR_2^x + c \end{aligned} \tag{4.4}$$

Next, recall that we are trying to obtain a set of corrected values, of which R_b is something like a bridge, telling us what $\frac{1}{2}(R_1 + R_2)$ *would have been* if the scale were actually linear. Let R'_n represent those corrected (linear) scale values, such that $R'_n = R_n^x$.

Since M_n and R'_n are, by stipulation, linearly related, we can express the relationship between M_b and R_b without any exponent x

$$\begin{aligned} M_b &= kR'_b + c \\ R'_b &= R_b \\ M_b &= kR_b + c \end{aligned} \tag{4.5}$$

Since $M_b = \frac{1}{2}(M_1 + M_2)$, we can relate R_1 and R_2 to M_b as follows:

$$M_b = \frac{1}{2}(kR_1^x + kR_2^x + 2c) \tag{4.6}$$

Substitute for R_b and multiply both sides by 2:

$$\begin{aligned} kR_b + c &= \frac{1}{2}(kR_1 + kR_2 + 2c) \\ 2kR_b + 2c &= kR_1^x + kR_2^x + 2c \end{aligned} \tag{4.7}$$

Subtract $2c$ from both sides and divide by k:

$$2R_b = R_1^x + R_2^x \tag{4.8}$$

Then simplifying:

$$R_b = \frac{1}{2}(R_1^x + R_2^x) \tag{4.9}$$

Thus, we can express R_b in terms of R_1 and R_2 without having to solve for k or c. By substituting each subject's ratings for R_b, R_1, and R_2, it is possible to calculate a unique x for each subject. Each subjects ratings can then be corrected by raising them to the power x.

5

Conclusion

In one sense, this dissertation began with the idea that methods of sensory psychophysics could be successfully applied to more accurately measure higher-level phenomenon like morality, love, and other subjective magnitudes. In another sense, this dissertation began with an extended quote by William James, who said that the methods of psychophysics were boring.

I did this, in part, because it's easy to see where he was coming from. The heaviness of a weight or the brightness of a light involves only a narrow slice of what there is to learn about the inner workings of our psychological lives, and these topics rarely have any more than limited contact with the issues about which we care most deeply. Even so, I believe that James was wrong to blame this on the psychophysical methods, themselves. Accordingly, this dissertation has been my best effort to provide a more sympathetic perspective on the work of the chronograph philosophers whom James so unfairly maligned.

On the topic of emotion, I have illustrated how a ubiquitous measure can produce an illusory difference in emotional experience — specifically, that women (appear to) experience anger more intensely than do men. This work was inspired by a scale developed by taste psychophysicists to detect group differences in oral sensation, since existing measures had hidden such differences for the better part of a century (Bartoshuk, Duffy, et al., 2004).

On the topic of morality, I have illustrated how that same ubiquitous measure can produce implausible findings when used between-subjects — specifically, that killing a prisoner of war is no worse than a committing a fraternity prank, however cruel. This work was inspired by a method that S.S. Stevens developed in his efforts to create a ratio scale of loudness (the exact kind of enterprise that James had disparaged; Stevens, 1956).

And on the topic of magnitude, one of the most basic and essential features of the world that our minds represent, I have not only illustrated that a subjective stimulus like moral judgment exhibits all the classic hallmarks of having a magnitude, but I have also revealed the format in which that magnitude is mentally encoded, as well as a procedure for straightforwardly correcting the nonlinear use of labeled scales. This work was inspired by studies that meticulously tested the effect of subtle changes to the methods used to measure physical magnitudes like brightness (e.g. Marks, 1968).

In all three cases, I have aimed to demonstrate how metaphorical prisms, pendulums, and chronographs can be applied to answer important and theoretically interesting questions about the mind. I hope that the reader has found the result of these efforts to be anything but boring.

This final chapter will proceed as follows. I first consider a broader question that I alluded to both at the start of my dissertation and in the preceding chapter: many of the psychophysical methods that I have adapted within this work were initially developed to precisely quantify the intensity of a given perceptual sensation after tightly controlling for and manipulating the properties of an objective physical magnitude. How could these methods possibly work for something as subjective as *morality*? I answer this question in two ways. In the following section, I provide a theoretical basis for the application of psychophysical methods to the measurement of subjective stimuli. Then, I put this approach into a broader context — the largest figures in the field considered subjective stimuli to be well within the purview of psychophysical measurement.

5.1 *Theories of subjective magnitude*

It's not obvious how a subjective stimulus like morality can be represented as having a magnitude, and it's less obvious how the magnitude of that stimulus can be successfully measured using psychophysical methods. Nonetheless, there is a coherent theoretical picture that has been emerging from several disparate lines of work, and I believe this account provides a straightforward answer to these questions: subjective stimuli make use of a well-documented domain-general system of magnitude which applies to all prothetic continua. Given this, subjective magnitudes can be quantified because they share a representational format with perceptual magnitudes that *do* correspond to real, objective, and countable units in the external world. Such an account would be attractively parsimonious from both a theoretical and methodological perspective: since subjective magnitudes would be quantifiable by dint of their connection to perceptual magnitudes, it would thus be entirely natural to apply psychophysical methods to a subjective stimulus like morality, since these measures would still be doing what they were (in a sense) designed to do.

In the following section, I aim to bring together several lines of work to support this general theoretical account. It proceeds as follows. First, I provide some initial but compelling evidence to suggest that subjective and perceptual stimuli share a domain-general magnitude representation, which is grounded in a well-documented pattern of cross-modal correspondences between perceptual and subjective stimuli. Then, I summarize a body of work involving stimulus generalization that has rarely (if ever) been discussed in the context of the relevant work I've been considering involving representations of magnitude. Taken together, this works provides a straightforward explanation for the similarity, convergence, and mutual correspondence of different representations of magnitude, broadly construed. Though admittedly speculative, I believe this nonetheless provides a theoretically plausible basis for the application of psychophysical measures to the study of subjective magnitude.

5.1.1 *Matching experiments*

As discussed in Chapter 4, nearly all of the contemporary work on domain-general representations of magnitude concern one of the so-called “big three”— space, time, and number (e.g. Dehaene & Brannon, 2010). Having reviewed some of that more contemporary literature already, I will instead focus on another compelling (and often overlooked) source of evidence which concerns a psychophysical method called *cross-modality matching*. Cross-modality matching is when intensity in one sensory modality is equated or matched with an intensity in another sensory modality (Stevens, 1959, 1975 Chapter 4), like adjusting the loudness of a sound to match the brightness of a light, and vice versa. Such work has been conducted on an extensive number of sense modalities — Stevens, for example, published one paper matching 10 such sensory continua (Stevens, 1966a) — and results tend to reveal remarkable consistency across the different modalities (even, for example, with something like stage fright: Latané & Harkins, 1976). And while the operation can sometimes seem counter-intuitive at first, even five-year-olds eventually do so reliably (Bond & Stevens, 1969), as do preschoolers (to some extent; Blank & Bridger, 1964).

Perhaps most surprisingly, there are also a number of studies that used cross-modality matching to scale the magnitude of subjective stimuli, like occupational prestige (Cross, 1982; Kuennapas & Wikstroem, 1963), political opinion (Lodge et al., 1975), and though this final study was never published, Stevens has described work conducted by a thesis student that involved matching brightness and loudness to the seriousness of offenses (Stevens, 1975, p. 254). As Stevens described it, each crime had both a brightness associated with its severity and a loudness associated with its severity. Plotted against one another, the brightness and loudness intensities for each crime had a slope of .9, suggesting a high level of agreement between the modalities. Should data like this replicate more broadly, this seems as if it would be the clearest evidence that subjective and perceptual magnitudes share a representational format. It's not obvious how such data could be otherwise explained, since virtually no one would have any past experience correlating moral violations with sounds or lights of different intensities.

This raises another subtle but remarkable aspect of the results I presented in Chapter 4. On reflection, it seems deeply surprising that representations of moral magnitude should in any way resemble representations of perceptual magnitude. Something like sensory transduction can easily explain the kind of logarithmic relationship reflected by Hipparchus's scale of stellar magnitude (indeed, this was broadly how Fechner interpreted this finding), but magnitude estimates of crime severity provide nothing for sensory receptors to transduce. Why then, does an objective stimulus like brightness and a subjective stimulus like morality each display the exact same relationship to ratings collected on a category scale? I address this question in the section that follows.

5.1.2 *The Universal Law of Generalization*

In some ways, the remarkable consistency across magnitude representations mirrors a line of work that has largely been conducted in the field of animal behavior. Briefly, this work concerns how we are able to form generalizations about (and judge the similarity of) different kinds of stimuli (for an overview, see Ghirlanda & Enquist, 2003). No two stimuli are perfectly alike, yet it is nonetheless crucially important that we be able to generalize from a given stimulus to the overall category to which it belongs.

This can be a difficult needle to thread. Suppose you lived in area with abundant wild mushrooms. Given this, it would be very useful to be able to differentiate between the poisonous ivory funnel mushroom and the similar-appearing (and delicious) oyster mushroom. If eating a given mushroom goes on to make you sick, and if you fail to sufficiently generalize based on the features of that mushroom, you are thus prone to continue ingesting poisonous mushrooms. It would obviously be preferable to avoid this if at all possible, but there is also the danger of moving too far in the other direction. If you generalize too broadly from those same features, you may instead start avoiding mushrooms altogether, thus forgoing an otherwise valuable and plentiful source of food.

To be useful, the ability to recognize relevant similarities (or irrelevant dissimilarities) must be both general and universal — general because it is crucially important to recognize such similarities and differences for nearly any kind of stimulus that we might encounter (whether subjective or perceptual), and universal because any given stimulus might differ across a number of different sensory modalities (in the same way that larger mushrooms tend to both look bigger and feel heavier, etc). To satisfy these requirements, Shepard (1987) proposed the following principle, which he called "The Universal Law of Generalization" — the perceived similarity of two stimuli is a negative exponential function of the distance between those stimuli in a hypothesized psychological space.

Though we are of course unable to directly measure distance in psychological space, Shepard developed a method that allows us to calculate an estimate for these distances via a process called "nonmetric multidimensional scaling" (Shepard, 1962). As an *extremely* simplified illustration (adapted from Guttman & Kalish, 1956), suppose that a pigeon was trained to peck a certain key when shown a red light and a different key when shown a yellow light. Once trained, we can then add a third color, like green. While the pigeon will sometimes press the key corresponding to the red light when shown a green light, the pigeon will be more likely to press the key that corresponds to the yellow light (since green is more similar to yellow than it is to red). We can then determine the distance between these colors in psychological space based on how often the pigeon presses one key over the other, with higher proportions reflecting stimuli that are closer together. We can then add a fourth color, like orange. Here, pigeons should press both keys about equally often (since orange is just as similar to yellow as it is red). By iterating a much more complex version of this basic procedure, it is possible to calculate precise psychological distances between stimuli that differ in numerous ways and even along a number of different "spatial" dimensions; all based on whether a pigeon presses one key or another.

Though similar experiments as the one just described nearly always produce the exponential function proposed by Shepard, there is a class of stimulus in which perceived similarity appears to instead be a Gaussian function of psychological distance. The commonality shared by all these stimuli is that they vary along so-called "rearrangement dimensions" like pitch and angle (see Ghirlanda & Enquist, 2003, Table 2). To my knowledge, however, the clear parallel between these Guassian functions and methathetic continua (e.g. continua that change via a substitutive process), as well as between exponential functions and prothetic continua (e.g. continua that change via an additive process) has thus far gone largely (if not entirely) unnoticed. Indeed, I have been unable to find any instances in which either literature has made more then a perfunctory reference to the other.[1]

5.1.3 *Psychological distance and equivalent information*

The work I've just described lends broad support to a number of ideas that are complementary to the overall theoretical account that I've presented thus far. First, the requirement for generalization provides a straightforward explanation for why there is such remarkable consistency across magnitude representations, even for drastically different kinds of stimuli. If a prothetic stimulus is to be amenable to generalization, then it must be possible to represent that stimulus as capturing some distance in psychological space. Thus, insofar as the exponential decay function applies to all prothetic stimuli which are represented in this way (which almost universally appears to be the case), we should naturally expect to see a broad uniformity in the way that psychological magnitudes are represented.

Second, the requirement for universality provides a similarly straightforward explana-

[1] I am grateful to Sam McDougle who first noticed this parallel and kindly brought it to my attention.

tion for why there should be cross-modal correspondences in the first place. This reflects the basic fact that we rarely assess the properties of a stimulus along only one sense modality. By way of explanation, consider different ways that one might differentiate coins: one might use touch (as when rummaging through a pocket in search of exact change), or by sight (as when looking into a change jar on laundry day), or perhaps even by sound (as when deciding whether to pick up a coin that has fallen out of one's pocket and onto the floor). Each case uses a different sense modality to evaluate the similarity of coins, and since similarity is a function of psychological distance, we should naturally hope that each sense modality converges to the same position in psychological space.

This resembles a similar idea proposed by Marks in relation to commonalities across sense modalities, generally speaking (1978a, Chapter 2). Marks called this principle "The Doctrine of Equivalent Information," based on the idea that different sense modalities provide converging information about magnitude for the simple reason that magnitude can be perceived by more than one sense modality. Marks credits this basic idea to E.J. Gibson (1969), who wrote:

> *There must be [a] simple type of perceptual development, the registering of concurrent covariation from different organs... Insofar as this linkage is invariant, the information is the same in all of them, that is, the systems are equivalent (p. 289, as cited in Marks, 1978, p. 28)*

Put another way, we don't want to extract information about e.g. how an object looks, or how it feels, or how it sounds; we want to extract information *about the object*.

Importantly, all of the above holds in light of recent attempts to reformulate or reinterpret Shepard's universal law (e.g. Frank, 2018; Ghirlanda & Enquist, 2003; Sims, 2018). This is because each of these accounts continues to make reference to similarity in psychological space. Given this, I believe that the above presents a plausible theoretical account for the similarities between how subjective and perceptual magnitudes are represented.

5.2 *The purview of psychophysics*

In the final section of my dissertation, I consider the scope of my overall project by addressing an obvious skeptical response: is any of this *actually* psychophysical measurement? Wouldn't it more appropriate to call it something like *subjective measurement* (derogatory)? In response, I'll argue that yes, this project broadly construed is one of *subjective measurement* (respectful), but this dichotomy is a false one, and defining psychophysics purely in terms of how objective stimulus features map onto perceptual experience is entirely at odds with how the largest figures in the field saw their project.

5.2.1 *On the psychophysical law*

Traditionally, psychophysics is most often understood as the branch of psychology whose aim is to map the objective features of some stimulus (e.g. the number of photons emitted by some light, per square inch per second) to the subjective experience of that stimulus (e.g. how bright that light seems). Indeed, one of the largest sources of controversy in the field, over which dozens if not hundreds of articles have been written, is on the exact nature of the so-called "psychophysical law," or the mathematical relationship between external stimuli and sensory experience.

For Fechner, the psychophysical law was a logarithmic one, where the magnitude of a sensation ψ is the logarithm of the ratio between stimulus magnitude ϕ and smallest detectable magnitude of that stimulus b, multiplied by a proportionality constant k.

$$\psi = k \log(\phi \,/b)$$

For Stevens, the psychophysical law was a power law, where the magnitude of a sensation ψ is determined by the magnitude of a stimulus ϕ raised to a power n that depends on the given modality, and the constant k that, in Steven's words, "depends on the units of measurement and is not very interesting" (p. 13).

$$\psi = k\phi^n$$

In both cases, scores of experiments were conducted to see whether one law or the other was better supported by different sensory modalities such as sound, touch, heat, brightness, roughness of sand paper, and so on. In all cases, the ability to know and carefully control the precise stimulus magnitude ϕ allowed for a precise measurement of sensory magnitude ψ, and in Steven's case, the determination of the specific exponent n unique to each kind of stimulus.

In either interpretation of the psychophysical law, it seems clear that calculating a sensory magnitude ψ requires a corresponding stimulus magnitude ϕ on which to operate. Given this, it may seem radical (if not incoherent) to apply psychophysical methods to subjective stimuli like emotional experience or moral judgment that, by definition, have no objective physical magnitude on which to calculate a corresponding sensory magnitude. It is no surprise then, that psychophysics is often explicitly defined as the mapping between sensation, specifically, and the physical properties of a stimulus (e.g. Gescheider, 1997).

My approach is not, however, as radical or incoherent as it may first seem. Though Fechner seemed optimistic that a sufficiently mature psychological science might someday reveal an objective stimulus magnitude that could apply to *any* kind of psychological phenomenon, I don't find this particular plausible. But more importantly, I also don't find it in any way necessary (nor, for what it's worth, did Fechner). Rather, my contention is that having such a limited focus on objective, external stimuli is at odds with the history of psychophysics as a discipline more broadly — psychophysics is (and always has been) more than just painstakingly dull and iterative experiments supporting this or that psychophysical law.

Since I have discussed Stevens and his work extensively by now (and since he wrote several papers explicitly on the topic of "social psychophysics" (e.g. Stevens 1966), I'll instead focus on two earlier figures: Gustav Fechner and L.L. Thurstone.

5.2.2 *Fechner and the internal stimulus*

If anyone has the right to define the scope of psychophysics, it would be the founder of the discipline (and, arguably, of experimental psychology as a field): Gustav Fechner. In one of the first chapters of his *Elements of Psychophysics* (1860), Fechner considers a point that he takes to concern "the future of the whole of psychophysics" (p. 12). He writes:

> *If we classify thinking, willing, and the finer esthetic feelings as higher mental activities, and sensations and drives as lower mental activities, then... the higher mental activities can go on no less than the lower without involving physical processes or being tied to psychophysical processes.*

It is these higher mental activities that Fechner spent much of his later career studying. Most notably, he published a dozen papers and one book in the area of empirical aesthetics, in which he often used precisely the same methods as he applied in his psychophysical laboratory, involving thresholds, forced-choice, and production (see Höge, 1995). This aspect of Fechner's work has largely been ignored, likely owing to the fact that they were never translated to English, but one (and only one) article of his, "The Aesthetic Association Principle" has been recently translated (Ortlieb et al., 2020). The translators of this article comment that, owing to his reliance on psychophysical methods in his work on aesthetics, this line of work "is still widely mistaken for an application of psychophysics, rather than a full-fledged research programme in its own right" (p. 2). I'd interpret this confusion differently: this dichotomy is a false one, and Fechner was doing both psychophysics *and* experimental aesthetics.

As such, Fechner clearly did not envision a psychophysics restricted to the study of low-level perception and sensation. He also did not envision a psychophysics which was restricted to considering solely external stimuli, as was still arguably the case in his study of empirical aesthetics. Early in his *Elements*, Fechner draws a distinction between *objective* sensations — like "light and sound" which involve "a source external to the sensory organ"— and *common* sensations — like "pain, pleasure, hunger, and thirst" which are "felt only as conditions of our own bodies" (p. 15).

Fechner would go on to explicitly consider how psychophysics ought to engage with internal stimuli, arguing that internal sources of sensation should be treated "under the same concepts, standpoints, and formulas as the external sources." He writes:

> *At this time internal stimuli are an unknown* x *as to their location and quality, although they enter despite this limitation into the phenomenal sphere with a quantitative effect that is comparable to that of an external stimulus. The internal stimulus derives its name and value from this effect (pp. 16-17).*

Here, we can see that the crucial feature is the *quantitative effect* that an internal stimulus has on our subjective experience. It should follow, then, that subjective stimuli fall under the purview of psychophysics, as conceived by its founder, insofar as they share such a quantitative effect with the external stimuli classically associated with the field. I demonstrated precisely this kind of relationship in Chapter 4.

5.2.3 *Thurstone and scaling attitudes*

A year before Thurstone (1928) proposed that attitudes can, in fact, be measured, he provided a theoretical account of psychophysical measurement explicitly based on Fechner's work (and which didn't require any knowledge of the objective magnitude of a stimulus; Thurstone, 1927). For Thurstone, psychophysics was not concerned primarily with sensation magnitudes, *per se*, but rather with the process of discriminability *writ large*. Looking back later in his life, Thurstone wrote as follows:

> *Psychophysical Analysis was my first paper in the field of psychophysics, or psychological measurement proper. In my judgment, this is my best contribution to psychology. It probably has more implications for psychological science than any other paper that I have written (1959, p. 15).*

Thurstone explicitly mentions his work in psychophysics as motivating his later work on values and attitudes, noting that "new psychophysical concepts are applicable to the measurement of affective values, such as in aesthetics, and to many other forms of cognitive and affective discrimination." And though Thurstone considered his "Law of Comparative Judgment" to be a third psychophysical law, it is worth mentioning, however, that Thurstone was not necessarily thrilled by his association with the field, preferring to describe his project as "subjective measurement." Just a few sentences later, he writes:

> *The title 'psychophysics' has been unfortunate because it denotes the dead subject of lifted weights and limen determinations and arguments about the best way to compute somebody's limen for lifted weights to many decibels. Graduate students are right when they agree with William James that psychophysics of his day was, and still is, the dullest part of psychology.*

Given this, it is perhaps no surprise that Thurstone described his initial approach to psychophysical measurement as follows — something that is qualitatively similar to the one I've been defending, here:

> *Instead of asking students to decide which of two weights seemed to be the heavier, it was more interesting to ask, for example, which of two nationalities they would generally prefer to associate with, or which they would prefer to have their sister marry, or which of two offenses seemed to them to be the more serious (1959, p. 16).*

5.3 *The future of subjective measurement*

Taken together, I believe this brief history makes clear that the approach that I've taken in this dissertation is not nearly as strange as it might initially seen. Subjective stimuli have *always* been included under the banner of psychophysics, and they have been tested in psychophysics labs for as long as there have *been* psychophysics labs.

For now, however, I think that the approach that I’ve outlined here extends well beyond moral judgment and emotion. Future work can take similar approaches to topics like political attitudes, aesthetics, physical or emotional pain, subjective utility, or even the mapping between something like love and time. These methods should apply to just about any prothetic continua that one might imagine. Thankfully, it seems as though we’ve only started to uncover just how many there may actually be.

Bibliography

Acierno, J., Mischel, S., & Phillips, J. (2022). Moral judgements reflect default representations of possibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1866), 20210341.

Aitken, R. C. B. (1969). A growing edge of measurement of feelings [abridged]: Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine*, *62*(10), 989–993.

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850.

Anderson, N. H. (1972). Cognitive algebra and social psychophysics. In *Social attitudes and psychophysical measurement* (pp. 123–148). Psychology Press.

Bartoshuk, L. M. (2014). The measurement of pleasure and pain. *Perspectives on Psychological Science*, *9*(1), 91–93.

Bartoshuk, L. M., Duffy, V. B., Fast, K., Green, B. G., Prutkin, J., & Snyder, D. J. (2003). Labeled scales (e.g., category, Likert, VAS) and invalid across-group comparisons: What we have learned from genetic variation in taste. *Food Quality and Preference*, 14.

Bartoshuk, L. M., Duffy, V. B., Chapo, A. K., Fast, K., Yiee, J. H., Hoffman, H. J., Ko, C.-W., & Snyder, D. J. (2004). From psychophysics to the clinic: Missteps and advances. *Food Quality and Preference*, *15*(7), 617–632.

Bartoshuk, L. M., Duffy, V., Green, B., Hoffman, H., Ko, C.-W., Lucchina, L., Marks, L., Snyder, D., & Weiffenbach, J. (2004). Valid across-group comparisons with labeled scales: The gLMS versus magnitude matching. *Physiology & Behavior*, *82*(1), 109–114.

Bartoshuk, L. M., Fast, K., & Snyder, D. J. (2005). Differences in our sensory worlds: Invalid comparisons with labeled scales. *Current Directions in Psychological Science*, *14*(3), 122–125.

Bartoshuk, L. M., Duffy, V. B., & Miller, I. J. (1994). PTC/PROP tasting: Anatomy, psychophysics, and sex effects. *Physiology & Behavior*, *56*(6), 1165–1171.

Baumeister, R. F., Vohs, K. D., & Funder, D. C. (2007). Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior? *Perspectives on Psychological Science*, *2*(4), 396–403.

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66*(1), 5–20.

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60*(4), 485–499.

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods*, *4*(3), 243.

Blakeslee, A. F., & Fox, A. L. (1932). Our different taste worlds. *Journal of Heredity*, *23*(3), 97–107.

Blank, M., & Bridger, W. H. (1964). Cross-modal transfer in nursery-school children. *Journal of Comparative and Physiological Psychology*, *58*(2), 277–282.

Bond, B., & Stevens, S. S. (1969). Cross-modality matching of brightness to loudness by 5-year-olds. *Perception & Psychophysics*, *6*(6), 337–339.

Borg, G. (1990). A general model for interindividual comparison. In W. J. Baker, M. E. Hyland, R. van Hezewijk, & S. Terwee (Eds.), *Recent trends in theoretical psychology* (pp. 439–444). Springer US.

Borg, G. (2001). Are we subjected to a "long-standing measurement oversight?" In R. Kompass (Ed.), *Fechner day 2001. proceedings of the seventeenth annual meeting of the international society for psychophysics* (pp. 1–44).

Brady, H. E. (1985). The perils of survey research: Inter-personally incomparable responses. *Political Methodology*, *11*(3), 269–291.

Brannon, E. M., Wusthoff, C. J., Gallistel, C., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science*, *12*(3), 238–243.

Brebner, J. (2003). Gender and emotions. *Personality and Individual Differences*, *34*(3), 387–394.

Bridwell, N. (1963). *Clifford the big red dog*. Scholastic US.

Bueti, D., & Walsh, V. (2009). The parietal cortex and the representation of time, space, number and other magnitudes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1525), 1831–1840.

Burke-Gaffney, M. W. (1963). Pogson's scale and Fechner's law. *Journal of the Royal Astronomical Society of Canada*, *57*(1), 3–8.

Butler, G., Poste, L., Wolynetz, M., Agar, V., & Larmond, E. (1987). Alternative analyses of magnitude estimation data. *Journal of Sensory Studies*, *2*(4), 243–257.

Cantlon, J. F., Cordes, S., Libertus, M. E., & Brannon, E. M. (2009). Comment on "log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures". *Science*, *323*(5910), 38–38.

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, *11*(1), 2100.

Čeko, M., Kragel, P. A., Woo, C.-W., López-Solà, M., & Wager, T. D. (2022). Common and stimulus-type-specific brain representations of negative affect. *Nature Neuroscience*, *25*(6), 760–770.

Clifford, S., Iyengar, V., Cabeza, R., & Sinnott-Armstrong, W. (2015). Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior Research Methods*, *47*(4), 1178–1198.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General*, *145*(10), 1359–1381.

Cross, D. V. (1982). On judgments of magnitude. In *Social attitudes and psychophysical measurement* (pp. 73–88). Psychology Press.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: The aversion to harmful action. *Emotion (Washington, D.C.)*, *12*(1), 2–7.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

Dasgupta, S., Vanspauwen, R., Guneri, E. A., & Mandala, M. (2022). Vincent Van Gogh and the elusive diagnosis of vestibular migraine. *Medical Hypotheses*, *159*, 110747.

Davis, E., Greenberger, E., Charles, S., Chen, C., Zhao, L., & Dong, Q. (2012). Emotion experience and regulation in China and the United States: How do culture and gender shape emotion responding? *International Journal of Psychology*, *47*(3), 230–239.

Dehaene, S. (2001). Subtracting pigeons: Logarithmic or linear? *Psychological science*, *12*(3), 244–246.

Dehaene, S., & Brannon, E. M. (2010). Space, time, and number: A Kantian research program. *Trends in Cognitive Sciences*, *14*(12), 517–519.

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian indigene cultures. *Science*, *320*(5880), 1217–1220.

Diener, E., Larsen, R. J., Levine, S., & Emmons, R. A. (1985). Intensity and frequency: Dimensions underlying positive and negative affect. *Journal of personality and social psychology*, *48*(5), 1253–1265.

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*(3), 285–290.

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology*, *81*(5), 869–885.

Fechner, G. (1860). *Elements of psychophysics* (D. Howes & E. Boring, Eds.; H. Adler, Trans.). Holt, Rinehart; Winston.

Firestone, C., & Scholl, B. J. (2014). "Top-down" effects where none should be found: The El Greco fallacy in perception research. *Psychological Science*, *25*(1), 38–46.

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *The Behavioral and Brain Sciences*, *39*, e229.

Fischer, A. H., & Roseman, I. J. (2007). Beat them or ban them: The characteristics and social functions of anger and contempt. *Journal of Personality and Social Psychology*, *93*(1), 103–115.

Fox, A. L. (1932). The relationship between chemical constitution and taste. *Proceedings of the National Academy of Sciences*, *18*(1), 115–120.

Frank, S. A. (2018). Measurement invariance explains the universal law of generalization for psychological perception. *Proceedings of the National Academy of Sciences*, *115*(39), 9803–9806.

Gallistel, C. R. (1989). Animal cognition: The representation of space, time and number. *Annual Review of Psychology*, *40*(1), 155–189.

Gallistel, C., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, *4*(2), 59–65.

Gescheider, G. A. (1997). *Psychophysics: The fundamentals*. Psychology Press.

Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*(1), 15–36.

Gibson, E. J. (1969). *Principles of perceptual learning and development.* Appleton-Century-Crofts.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168.

Gray, K., & Keeney, J. E. (2015). Impure or just weird? Scenario sampling bias raises questions about the foundation of morality. *Social Psychological and Personality Science*, *6*(8), 859–868.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*(1), 79–88.

Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social Research*, 623–656.

Hausman, D. M. (1995). The impossibility of interpersonal utility comparisons. *Mind*, *104*(415), 473–490.

Hayes, J. E., Allen, A. L., & Bennett, S. M. (2013). Direct comparison of the generalized visual analog scale (gVAS) and general labeled magnitude scale (gLMS). *Food Quality and Preference*, *28*(1), 36–44.

Heiphetz, L., & Cushman, F. (2021). Introduction to morality as a hub: Connections within and beyond social cognition. *Social Cognition*, *39*(1), 1–3.

Hsee, C. K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes*, *67*(3), 247–257.

Hsee, C. K., & Tang, J. N. (2007). Sun and water: On a modulus-based measurement of happiness. *Emotion*, *7*(1), 213–218.

Huntley, J. S. (2022). Van Gogh, lateral tilt, and the El Greco fallacy. *Medical Hypotheses*, *163*, 110861.

Huntsinger, J. R., & Raoul, A. (2022). Only as a last resort: Sociocultural differences between women and men explain women's heightened reaction to threat, not evolutionary principles. *Behavioral and Brain Sciences*, *45*, e140.

Jackson, F. (1986). What Mary didn't know. *The Journal of Philosophy*, *83*(5), 291.

Jeffrey, R. C. (1974). Remarks on interpersonal utility theory. In S. Steunlund (Ed.), *Logical theory and semantic analysis* (pp. 35–44). D. Riedel.

Jia, L., Shao, B., Wang, X., & Shi, Z. (2021). Phrase depicting immoral behavior dilates its subjective time judgment. *Frontiers in Psychology*, *12*.

John, I. D. (1971). The properties of distributions of magnitude estimates of loudness and softness. *Scandinavian Journal of Psychology*, *12*(1), 261–270.

Kelly, J. R., & Hutson-Comeaux, S. L. (1999). Gender-emotion stereotypes are context specific. *Sex Roles*, *40*(1), 107–120.

Khamitov, M., Rotman, J. D., & Piazza, J. (2016). Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition*, *146*, 33–47.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, *98*(1), 191–207.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190–194.

Knutson, K. M., Krueger, F., Koenigs, M., Hawley, A., Escobedo, J. R., Vasudeva, V., Adolphs, R., & Grafman, J. (2010). Behavioral norms for condensed moral vignettes. *Social Cognitive and Affective Neuroscience*, *5*(4), 378–384.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Kuennapas, T., & Wikstroem, I. (1963). Measurement of occupational preferences: A comparison of scaling methods. *Perceptual and Motor Skills*, *17*(2), 611–624.

Latané, B., & Harkins, S. (1976). Cross-modality matches suggest anticipated stage fright a multiplicative power function of audience size and status. *Perception & Psychophysics*, *20*(6), 482–488.

Leong, L. M., McKenzie, C. R. M., Sher, S., & Müller-Trede, J. (2019). Illusory inconsistencies in judgment: Stimulus-evoked reference sets and between-subjects designs. *Psychonomic Bulletin & Review*, *26*(2), 647–653.

Lewis, D. (1990). What experience teaches. In *Mind and cognition* (pp. 29–57). Blackwell.

Liao, S., Strohminger, N., & Sripada, C. S. (2014). Empirically investigating imaginative resistance. *The British Journal of Aesthetics*, *54*(3), 339–355.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22 140*, 55–55.

Lockhead, G. R. (1992). Psychophysical scaling: Judgments of attributes or objects? *Behavioral and Brain Sciences*, *15*(3), 543–558.

Lodge, M. (1981). *Magnitude scaling: Quantitative measurement of opinions*. SAGE Publications, Incorporated.

Lodge, M., Cross, D. V., Tursky, B., & Tanenhaus, J. (1975). The psychophysical scaling and validation of a political support scale. *American Journal of Political Science*, *19*(4), 611.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*(1), 293–318.

Marks, L. E. (1968). Stimulus-range, number of categories, and form of the category-scale. *The American Journal of Psychology*, *81*(4), 467.

Marks, L. E. (1978a). *The unity of the senses: Interrelations among the modalities*. Academic Press.

Marks, L. E. (1978b). Binaural summation of the loudness of pure tones. *The Journal of the Acoustical Society of America*, *64*(1), 107–113.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175–190.

McDonald, K., Graves, R., Yin, S., Weese, T., & Sinnott-Armstrong, W. (2021). Valence framing effects on moral judgments: A meta-analysis. *Cognition*, *212*, 104703.

Melnikoff, D. E., & Bailey, A. H. (2018). Preferences for moral vs. immoral traits in others are conditional. *Proceedings of the National Academy of Sciences*, *115*(4), E592–E600.

Minson, J. A., & Monin, B. (2012). Do-gooder derogation: Disparaging morally motivated minorities to defuse anticipated reproach. *Social Psychological and Personality Science*, *3*(2), 200–207.

Moskowitz, H. R. (1977). Magnitude estimation: Notes on what, how, when, and why to use it. *Journal of Food Quality*, *1*(3), 195–227.

Murray, D. J. (1993). A perspective for viewing the history of psychophysics. *Behavioral and Brain Sciences*, *16*(1), 115–137.

Ogden, J., & Lo, J. (2012). How meaningful are data from likert scales? An evaluation of how ratings are made and the role of the response shift in the socially disadvantaged. *Journal of Health Psychology*, *17*(3), 350–361.

Ortlieb, S. A., Kügel, W. A., & Carbon, C.-C. (2020). Fechner (1866): The aesthetic association principle—a commented translation. *i-Perception*, *11*(3), 204166952092030.

Palan, S., & Schitter, C. (2018). Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, *72*(6), 407–418.

Parducci, A. (1972). Category ratings: Still more contextual effects! In *Social attitudes and psychophysical measurement* (pp. 89–105). Psychology Press.

Petrova, M., Diamond, J., Schuster, B., & Dalton, P. (2008). Evaluation of trigeminal sensitivity to ammonia in asthmatics and healthy human volunteers. *Inhalation toxicology*, *20*(12), 1085–1092.

Pizarro, D. A., Laney, C., Morris, E. K., & Loftus, E. F. (2006). Ripple effects in memory: Judgments of moral blame can distort memory for events. *Memory & Cognition*, *34*(3), 550–555.

Poláčková Šolcová, I., & Lačev, A. (2017). Differences in male and female subjective experience and physiological reactions to emotional stimuli. *International Journal of Psychophysiology*, *117*, 75–82.

Postman, L., & Miller, G. A. (1945). Anchoring of temporal judgments. *The American Journal of Psychology*, *58*(1), 43.

Powell, D., & Horne, Z. (2017). Moral severity is represented as a domain-general magnitude. *Experimental Psychology*, *64*(2), 142–147.

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain: *Pain*, *17*(1), 45–56.

Prinzing, M., Earp, B. D., & Knobe, J. (2023). Why do evaluative judgments affect emotion attributions? The roles of judgments about fittingness and the true self. *Cognition*, *239*, 105579.

Robbins, L. (1935). *An essay on the nature and significance of economic science* (1st. ed. 1932). Macmillan.

Schein, C., Hester, N., & Gray, K. (2016). The visual guide to morality: Vision as an integrative analogy for moral experience, variability and mechanism. *Social and Personality Psychology Compass*, *10*(4), 231–251.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222.

Sellin, J. T., & Wolfgang, M. E. (1964). *The measurement of delinquency*. Patterson Smith.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, *67*(4), 667–677.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Siegel, J. Z., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211.

Simon, R. W., & Nath, L. E. (2004). Gender and emotion in the United States: Do men and women differ in self-reports of feelings and expressive behavior? *American Journal of Sociology*, *109*(5), 1137–1176.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.

Sosa, F. A., Ullman, T., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*, *217*, 104890.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stevens, S. S. (1951). *Handbook of experimental psychology*. Wiley.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American Journal of Psychology*, *69*(1), 1.

Stevens, S. S. (1957). On the psychophysical law. *The Psychological Review*, *64*(3), 153–181.

Stevens, S. S. (1958). Adaptation-level vs. the relativity of judgment. *The American Journal of Psychology*, *71*(4), 633–646.

Stevens, S. S. (1959). Cross-modality validation of subjective scales for loudness, vibration, and electric shock. *Journal of Experimental Psychology*, *57*(4), 201–209.

Stevens, S. S. (1960). On the new psychophysics. *Scandinavian Journal of Psychology*, *1*(1), 27–35.

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science*, *133*(3446), 80–86.

Stevens, S. S. (1966a). Matching functions between loudness and ten other continua. *Perception & Psychophysics*, *1*(1), 5–8.

Stevens, S. S. (1966b). A metric for the social consensus: Methods of sensory psychophysics have been used to gauge the intensity of opinions and attitudes. *Science*, *151*(3710), 530–541.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural, and social prospects*. Wiley.

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, *54*(6), 377–411.

Strohminger, N., & Nichols, S. (2014). The essential moral self. *Cognition*, *131*(1), 159–171.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*(4), 529–554.

Thurstone, L. L. (1927). Psychophysical analysis. *The American journal of psychology*, *38*(3), 368–389.

Thurstone, L. L. (1959). *The measurement of values.* The University of Chicago.

Ubel, P. A., Jankovic, A., Smith, D., Langa, K. M., & Fagerlin, A. (2005). What is perfect health to an 85-year-old?: Evidence for scale recalibration in subjective health ratings. *Medical Care*, *43*(10), 1054–1057.

Ubel, P. A., Peeters, Y., & Smith, D. (2010). Abandoning the language of "response shift": A plea for conceptual clarity in distinguishing scale recalibration from true changes in quality of life. *Quality of Life Research*, *19*(4), 465–471.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Vanneste, S., Verplaetse, J., Van Hiel, A., & Braeckman, J. (2007). Attention bias toward noncooperative people. a dot probe classification study in cheating detection. *Evolution and Human Behavior*, *28*(4), 272–276.

Volkmann, J. (1951). Scales of judgment and their implications for social psychology. In *Social psychology at the crossroads; the university of oklahoma lectures in social psychology.* (pp. 273–298). Harper.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. *Trends in Cognitive Sciences*, *7*(11), 483–488.

Warren, R. M. (1969). Visual intensity judgments: An empirical rule and a theory. *Psychological Review*, *76*(1), 16–30.

Wearden, J., & Jones, L. A. (2007). Is the growth of subjective time in humans a linear or nonlinear function of real time? *Quarterly Journal of Experimental Psychology*, *60*(9), 1289–1302.

Wylie, J., & Gantman, A. (2023). People are curious about immoral and morally ambiguous others. *Scientific Reports*, *13*(1), 7355.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, *1*(3), 333–349.