University of Groningen

Can positive emotions improve physical health?

Brown, Nick

DOI:
[10.33612/diss.99196913](https://doi.org/10.33612/diss.99196913)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Brown, N. (2019). *Can positive emotions improve physical health? an examination of some claims from positive psychology*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen.
<https://doi.org/10.33612/diss.99196913>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Can Positive Emotions Improve Physical Health?

Brown, N. J. L.

Can Positive Emotions Improve Physical Health? An Examination of Some Claims From
Positive Psychology

ISBN: 978-94-034-2009-7 (print), 978-94-034-2008-0 (e-book)

© Nicholas J. L. Brown, 2019

Cover design: Martin Brümmer

Printed by GVO Drukkers & Vormgevers, Ede.

This research was conducted within the Research Institute SHARE of the University Medical Center Groningen / University of Groningen. The printing of this thesis was financially supported by the Research Institute SHARE, the faculty of Medical Sciences (UMCG), and the University of Groningen.

university of
 groningen

Can Positive Emotions Improve Physical Health?

An Examination of Some Claims From Positive Psychology

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Monday 4 November 2019 at 11:00 hours

by

Nicholas John Laird Brown

born on 23 November 1960
in Birmingham, England

Supervisor

Prof. A. V. Ranchor

Co-supervisor

Prof. C. J. Albers

Assessment committee

Prof. P. de Jonge

Prof. M. Sprangers

Prof. C. Chambers

Contents

Chapter 1	Introduction	7
Chapter 2	A Critical Reanalysis of Fredrickson et al.'s Study of Genomics and Well-being <i>Proceedings of the National Academy of Sciences</i> , 111, 12705–12709.	23
Chapter 3	More Questions Than Answers: Continued Critical Reanalysis of Fredrickson et al.'s Studies of Genomics and Well-being <i>PLoS ONE</i> , 11(6), e0156415.	41
Chapter 4	The Elusory Upward Spiral: A Reanalysis of Kok et al. (2013) <i>Psychological Science</i> , 26, 1140–1143.	53
Chapter 5	Emodiversity: Robust Predictor of Outcomes or Statistical Artifact? <i>Journal of Experimental Psychology: General</i> , 146, 1372–1377.	63
Chapter 6	Does Twitter Language Reliably Predict Heart Disease? A Commentary on Eichstaedt et al. (2015a) <i>PeerJ</i> , 6, e5656.	91
Chapter 7	Easy as (Happiness) Pie? A Critical Evaluation of a Popular Model of the Determinants of Well-Being <i>Journal of Happiness Studies</i> .	121
Chapter 8	An Introduction to Criticality for Students of Positive Psychology <i>The Routledge Handbook of Critical Positive Psychology</i> .	157
Chapter 9	General Discussion	189
Chapter 10	Other Projects	209
Chapter 11	Samenvatting	221
Chapter 12	Acknowledgements	231

Notes

Chapters 2 through 8 are taken from material that has already been published in journals (Chapters 2 through 7) or a book (Chapter 8). A few typographical and grammatical errors have been corrected, and in the case of Chapters 2 and 3, the citations and references have been changed to APA style for consistency. Otherwise, the text of all of these chapters is identical to the published article or book, except where noted on an individual chapter's cover page.

The primary affiliation of each chapter co-author is given in a footnote on the chapter cover pages, where this is not the University of Groningen.

Chapter 1

Introduction

A very brief history of positive psychology

The positive psychology movement (Seligman & Csikszentmihalyi, 2000) has had a substantial impact on the public perception of psychology as a whole. One of positive psychology's core messages is that psychological interventions need no longer be reserved for people who "have something wrong with them." Numerous popular books (e.g., Duckworth, 2016; Fredrickson, 2009, 2013; Lyubomirsky, 2007, 2013; Seligman, 2002, 2011) have brought the endorsement of empirical science to the self-help section of bookstores, and considerable numbers of positive psychology coaches are plying their trade via web sites, seminars, and retreats.

An important part of the positive psychology worldview is that it is no longer sufficient to make a binary division of people into those who have a diagnosable mental health condition from the DSM (such as depression or anxiety)—the group who have historically been treated by clinical psychologists—and the rest. Indeed, if there is no way to distinguish between the members of the non-clinical population, the idea of such people "improving" their well-being—with or without help from psychologists—would seem to have little meaning from a scientific point of view, as such an improvement could not be empirically demonstrated. It is, therefore, perhaps not surprising that proponents of positive psychology have adopted the notion of different levels of healthy mental functioning. Thus, for example, Seligman (2002) wrote "Lying awake at night, you probably ponder, as I have, how to go from plus two to plus seven in your life" (p. xi), implying that well-being is that something that exists, for a person, on a numerical scale (say, from zero or minus 10 to plus 10), and that it is possible and desirable to improve one's "score." Keyes (e.g., 2002) introduced the ostensibly simple idea that people who are not depressed can either be *languishing* or *flourishing*, with the latter being the more desirable state of affairs:

The mental health continuum consists of complete and incomplete mental health. Adults with complete mental health are flourishing in life with high levels of well-being. To be flourishing, then, is to be filled with positive emotion and to be functioning well psychologically and socially. Adults with incomplete mental health are languishing in life with low well-being. Thus, languishing may be conceived of as emptiness and stagnation, constituting a life of quiet despair that parallels accounts of individuals who describe themselves and life as "hollow," "empty," "a shell," and "a void." (Keyes, 2002, p. 210)

Positive psychology quickly picked up on the idea that the binary (once depression has been ignored) states of “languishing” and “flourishing” might exist as observable psychological phenomena, and that people can be moved from the former to the latter with the right sort of help. Fredrickson and Losada (2005) even claimed to have discovered the exact mathematical formula for identifying the boundary between these states. Seligman (2011) gave the simple title *Flourish* to his book describing his revised (compared to his 2002 book, *Authentic Happiness*) model of well-being.

A common theme of this literature is that in order to flourish, one should ideally experience positive emotional states on a regular basis. Fredrickson (2009) expounded on her “broaden and build” theory, by which the experience of positive emotions expands thought–action repertoires and develops psychological resources such as resilience, and described an ideal “positivity ratio” of three positive emotions to each negative one. Seligman (2011) created an acronym—PERMA—for the five components of his model of a flourishing life, with the first of these letters representing “Positive emotions.”

By the end of the first decade of the 21st century, then, positive psychology had established its mission (to improve the psychological well-being of everyone, not just depressed and anxious people), devised the branding of this desirable outcome (flourishing), and identified the principal pathway by which this was to be achieved (positive emotions). Furthermore, because such populations would not be considered vulnerable, the people who were to help them could be less specialized. Hence the creation, in 2005, of the University of Pennsylvania’s Master’s in Applied Positive Psychology (MAPP) program (Seligman, 2011), followed by the introduction of numerous courses with similar titles at universities around the world. Many graduates of these programs have gone on to become trainers and coaches in positive psychology themselves. A number of alumni of the University of Pennsylvania MAPP program provide their services to the U.S. Army’s Comprehensive Soldier and Family Fitness initiative, under which positive psychology interventions are rolled out to hundreds of thousands of American service personnel every year (Brown, 2015).

Positive psychology and models of physical health

The World Health Organization (WHO) defines health as “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” (World Health Organization, 1946, p. 1), a definition that has not changed since its adoption more than 70 years ago. This definition, and in particular its somewhat disparaging final nine words, represented a radical shift away from the prevailing medical model. However, the

WHO definition has been criticized on a number of grounds, such as the absence of a precise definition of well-being (Boruchovitch & Mednick, 2002) or the unattainability of “complete” well-being (R. Smith, 2008). These limitations arguably make the WHO definition more political than practical (cf. Lewis, 1953), although it was defended as “correctly express[ing] the focus for health effort that is now attainable” by Breslow (1972, p. 348). Interestingly for what follows, Saracci (1997) argued that the complete state of well-being envisaged by the WHO definition appeared to be closer to a definition of happiness than of health.

Since the WHO declaration of 1946, a number of definitions, models, or frameworks of health have been proposed that attempt to circumvent these limitations while incorporating the psychological and social dimensions envisaged by the WHO definition of health. These more recent models can be divided into two broad categories, corresponding to what Larson (1999) defined as “Wellness” and “Environmental” models. Wellness models (“Health promotion and progress toward higher functioning, energy, comfort, and integration of mind, body, and spirit”; Larson, 1999, p. 125) assume that people’s mental states influence most of the biological processes that take place in their bodies, so that physical and mental functioning go closely hand in hand on a tightly coupled basis. Empirical evidence to support such models appears to be lacking, perhaps not least because of the difficulty of measuring (necessarily subjective) perceptions of wellness in an environment where healthcare policymakers and medical practitioners tend to demand objective outcomes. Proponents of well-being models also need to be aware of the danger of sailing too close to the rocks of “New Age” thinking, “alternative” medicine, and other forms of pseudoscience that can arise as a result of these models’ romantic appeal (e.g., “Human bodies become healthier when repeatedly nourished by positivity resonance with others, with the result that human communities become more harmonious and loving”; Fredrickson, 2013, p. 58).

In contrast, environmental models describe health in terms of “[a]daptation to physical and social surroundings—a balance free from undue pain, discomfort, or disability¹” (Larson, 1999, p. 125), a description that lends itself particularly well to healthcare funding models that integrate physical, mental, and social care into one overall system with close

¹ It is interesting to compare this definition with a quote from Lewis (1953, p. 116): “According to the most widely used of current textbooks, *psychiatry* [emphasis added] is concerned with ‘the study of the individual as a psycho-biological organism perpetually called upon to adapt to a social environment’.”

cooperation between agencies, such as can be found in many Western countries where health and social care are funded or guaranteed by taxpayers. These models attempt to situate health in a more pragmatic context in which infectious diseases and industrial-scale war have been replaced by the chronic conditions and diseases of old age, and the need for tradeoffs—for example, between purely numerical life expectancy and quality of life—is acknowledged. In such models, improved psychological well-being is typically considered as a desirable outcome in itself, although people with higher levels of psychological well-being may also be more likely to have healthy lifestyles in other respects, such as avoiding smoking or eating a balanced diet (Trudel-Fitzgerald, Boehm, Tworoger, & Kubzansky, 2018). However, the pathways by which well-being is associated with physical health outcomes are normally presumed to be complex and indirect, to the extent that they are causal at all (as opposed to being the results of a common cause, such as when higher household income or better education lead both to higher psychological well-being and—through factors such as easier access to healthcare—better physical health).

A recent example of such a model is that put forward by Huber and colleagues (Huber et al., 2011, 2016), which has been adopted by The Netherlands Organisation for Health Research and Development (ZonMw) under the title of Positive Health. This model defines health as “the ability to adapt and to self manage, in the face of social, physical and emotional challenges.” It has six dimensions: bodily (physical) functions, mental functions and (psychological) perceptions, a spiritual/existential dimension, quality of life, social and societal participation, and daily functioning. These are further divided into 32 aspects that can typically each be measured with a specific instrument (Huber et al., 2016). The Positive Health model was developed in collaboration with a large variety of stakeholders in the Netherlands, and hence fits with the idea, hinted at in the previous paragraph, that models of health should match the social systems of the country in which they are to be used. Although Huber and colleagues positioned their model as a response to the limitations of the WHO definition, it should be mentioned that the WHO itself has developed models of health that are considerably more usable in practice, such as Health-related Functioning (HrF), which also includes dimensions such as quality of life and activities of daily living (Salvador-Carulla & Garcia-Gutierrez, 2011).

Positive psychology clearly has a role to play in environmental models, which emphasize the ability of people to adapt to changes in their physical health or their social and environmental conditions, and to deal adequately with a range of stressors. Higher levels of positive affect and lower levels of negative affect are associated with better adherence to

medical treatments (Nsamenang & Hirsch, 2015). Furthermore, positive traits such as optimism (Scheier & Carver, 1985) or hope (Snyder, Irving, & Anderson, 1991) can lead to enhanced coping skills, while others, such as gratitude, can lead to the creation or maintenance of social support networks (Wood, Maltby, Gillett, Linley, & Joseph, 2008). For example, Kubzansky et al. (2018) noted that people who are higher in optimism may be better at both devising strategies to manage controllable stressors, and finding other coping mechanisms when faced with uncontrollable stressors, both of which would clearly be propitious when managing a diagnosis of, say, cardiovascular disease. These relations are typically biopsychosocial in nature, with the relation between psychological and physical health outcome variables being explained by explicit mediation processes, such as cognitive decisions or social factors. Huber (2014, pp. 218–219) noted the congruence between positive psychology and several dimensions of the Positive Health model—indeed, she indicated that this was at least part of the reason for the choice of its name—as well as the utility of familiarizing health professionals with some knowledge of positive psychology.

Matters become a little more complex when positive psychologists propose topics of study that fit under the umbrella of wellness models. For example, Seligman (2011, ch. 9) devoted an entire chapter entitled “Positive Physical Health: The Biology of Optimism” to the idea that there could be a direct (causal) relation behind some observed correlations between optimism and other positive psychological traits and states on the one hand, and health outcomes on the other. He proposed to start a research program to investigate this, under the title of Positive Health², using longitudinal studies of existing public data sets. A later working paper on the same topic (Seligman et al., 2013) defined the concept of “health assets,” which can be biological (e.g., favorable levels of HDL cholesterol), functional (e.g., social support), or subjective (e.g., psychological states and traits), and proposed to investigate the effects of these assets—with no a priori differentiation among their types in terms of their immediacy or directness—on health. This paper clearly stated the authors’ belief in the possibility of identifying one or more direct causal relations between psychological factors and health: “Further investigation of these relationships is essential, in

² As long ago as 1972, Breslow used the term “Positive Health” to describe “a state of being not just disease-free but of wellness beyond the norm” (p. 348). It is slightly unfortunate, but perhaps unavoidable, that so many researchers have chosen to use this pair of words to mean somewhat different things. Perhaps we should refer to “Seligman’s Positive Health,” “Huber’s Positive Health,” etc.

particular to establish whether the relationship between psychological characteristics and physical health is a direct physiological one or whether it is mediated through health- and illness-related behaviors” (Seligman et al., 2013, p. 16). However, it appears (e.g., from the very small number of citations of this working paper) that little rigorous research has subsequently been published in this area. Whether or not such research has in fact been conducted, but the results remain unpublished because no relation was found, is of course difficult to ascertain.

The idea that positive mental functioning might be in some lawful way associated with better physical health has an obvious intuitive appeal, and builds naturally on the uncontroversial and widely demonstrated negative relation between diagnoses of mental illness and life expectancy (e.g., Chesney, Goodwin, & Fazel, 2014). However, although there are a number of established physiological mechanisms that link psychological problems and poor health, such as the well-documented effects of chronic stress on the hypothalamic-pituitary-adrenal (HPA) axis (e.g., Levine, 2000), the idea that *positive* mental states might somehow directly and measurably *improve* physical health in the general population, perhaps mediated by some biological process internal to the person (such as gene expression or vagal nerve tone; see Chapters 2, 3, and 4), represents a radical claim for which there currently appears to be little good evidence. It is not clear why an equivalent of the relation between negative psychological states and poor health outcomes should necessarily hold for their opposites; indeed, psychologists do not even agree among themselves as to whether a simple positive–negative dichotomy is an appropriate way to classify emotional states at all (e.g., Lazarus, 2003; Wood & Tarrier, 2010). Reviews of positive psychology interventions and the effects of positive mental states on cancer (Coyne, Tennen, & Ranchor, 2010) or other chronic conditions (Ghosh & Deb, 2017) have found no strong evidence for any effect. Furthermore, there seems to be no obvious mechanism by which such an effect might work. HPA-related health problems are believed to arise because, for example, stress leads to the release of undesirable levels of hormones such as cortisol. But there seems to be no reason, another than a fondness for symmetry, to believe that positive emotional states should correspondingly cause the release of “good” hormones, or activate some other physiological mechanism, which would then have salutary effects. It does not seem unreasonable to assume that humans have evolved so that normally functioning individuals already have appropriate levels of these hormones, just as people with a balanced diet do not need to consume vitamin supplements.

The core of the present thesis is an examination of a number of articles, published in the past few years by prominent positive psychology researchers, that attempt to establish a positive monotonic relation, in the general adult population, between higher levels of psychological well-being—such as having a more meaningful life, or reporting that one experiences high levels of positive emotions—and objectively better physical health, as instantiated by a variety of surrogate or self-report measures. In several cases, it is implied—sometimes with only limited hedging—that this relation is probably causal in nature. My principal aim here, then, is to establish whether these claims stand up to critical scrutiny, and to identify the extent to which these articles provide support for the well-being models that their authors appear to favor.

I should note here that the articles that are the subject of Chapters 2 through 7 were not chosen in any kind of systematic way. This thesis does not claim to be a comprehensive review of the literature relating positive psychology to physical health, and I have not attempted to make any type of quantitative estimate of the reliability of that literature. In fact, Chapters 2 through 4 were already published, and Chapter 6 was a work in progress (although it took more than three years to complete), before I entered the PhD program at the UMCG; indeed, my work on these chapters was the principal motivation for Jim Coyne to suggest that I should consider joining that program. Some of the articles appeared on my radar as part of my general reading, while others were brought to my attention by colleagues who thought that they might be interesting for me to look at. A chapter resulted in those cases where strong claims and (usually) media coverage were combined with apparently weak evidence, and when I found the subject matter interesting.

Structure of the following chapters

Chapters 2 and 3 are closely related, in that they examine the evolution of the same line of research over several articles and rebuttals. We discuss the claims by Fredrickson et al. (2013, 2015) concerning the purportedly distinct effects on health—represented by the differential expression of a number of genes known to be phenotypically associated with the immune system—of “hedonic” versus “eudaimonic” well-being. Fredrickson and colleagues claimed that people with proportionally higher levels of eudaimonic well-being (“a meaningful life”)—versus hedonic well-being (“a pleasant life”)—have relatively higher expression of anti-viral genes, because, in evolutionary terms, their “more fulfilling” lives will include more social contacts, leading to more incidences of contagious (typically viral) infections. Meanwhile, people in the opposite situation (i.e., with a preponderance of hedonic

well-being) exhibit relatively higher expression of pro-inflammatory, antibacterial genes, because (again, in evolutionary terms) their “self-centered” lives will lead to them getting into more fights and suffering more bacterial infections as a result of their wounds (E. E. Smith, 2013). Chapter 2 was a direct reaction to the appearance of the first of these studies (Fredrickson et al., 2013). The publication of this chapter in *PNAS* was followed by a second article by the same authors (Fredrickson et al., 2015) in *PLoS ONE*, which appeared to have been written at least partly in response to our critique of their first article. In their 2015 article, Fredrickson and colleagues claimed to have reproduced and indeed extended their original results in two new samples. However, the later article made some subtle, yet crucial, changes to their hypotheses and analytical methods, which necessitated the writing of a new reply.

In Chapter 4, we critique an article (Kok et al., 2013) that claimed that a specific meditation practice (“loving-kindness meditation”) produced beneficial changes in the physical health of participants, as revealed by improvements in vagal nerve tone (VNT). Of all the articles examined in the chapters of this thesis, Kok et al.’s longitudinal study is perhaps the most ambitious in terms of its direct claims of a causal link between an intervention (the meditation program), leading to enhanced positive emotional states among participants, which in turn purportedly led to improved indices of physical health. Again, this article was widely covered in the popular media, with the *Daily Mail*’s headline claiming that a “[p]ositive outlook on life and making friends could be as good for you as diet and exercise” (“Don’t bother with the gym today,” 2013).

Chapter 5 examines the claims of an article by Quoidbach et al. (2014) concerning a novel construct that these authors called *emodiversity*. The idea here is that, as well as having a preponderance of experiences of (positive) emotions, it is important for an individual to experience a wide range of distinct emotions (both positive and negative) in his or her life. The authors claimed that, in two large samples, higher self-reported emodiversity predicted better mental (Study 1) and physical (Study 2) health outcomes, after controlling for the expected correlations with the extent to which participants experienced overall positive or negative emotions. Quoidbach et al. claimed that their results showed emodiversity to be a better predictor of health than measures of diet, exercise, and smoking.

Chapter 6 is a critique of the claims of Eichstaedt et al. (2015), who asserted that the mortality rate from atherosclerotic heart disease (AHD) per U.S. county can be predicted—better than by demographic, SES, and health behavior indicators—by analyzing the degree of positivity or negativity expressed in the Twitter timelines of the people who live in each

county. Eichstaedt et al. suggested that the language choices of Twitter users in any given county were a proxy for the “combined psychological character of the community” (p. 166). This counter-intuitive claim received extensive media coverage (summarized at <https://sage.altmetric.com/details/3084830/news>).

Chapter 7 is an examination of Lyubomirsky, Sheldon, and Schkade’s (2005) “happiness pie” model of the determinants of well-being. Lyubomirsky et al. claimed that as much as 40% of the variance in individual well-being is due to volitional activities, with the remainder being accounted for by genetic factors (50%) and “life circumstances” (10%). While Lyubomirsky et al.’s article did not make direct claims about physical health—although its theoretical arguments formed the basis of Lyubomirsky’s (2007) popular book in which she claimed that “happy people . . . have stronger immune systems, and are physically healthier . . . [and] even live longer” (p. 25)—this chapter has a place here because it illustrates a number of other points about what might be termed the “business model” of positive psychology, such as the idea that people can make substantial improvements to their own happiness, which can outweigh the potentially negative effects of the circumstances of one’s life. Lyubomirsky et al.’s article did not contain any original empirical evidence; however, by following their chain of references, we were able to identify the two studies conducted in the 1970s on which these authors’ key claim—namely, that only 10% of the variance in well-being is accounted for by life circumstances, leaving four times as much available for individuals to change themselves—was based. Our commentary includes a complete reanalysis of the original data sets from which Lyubomirsky et al.’s claims ultimately derive their empirical support.

Chapter 8 is my contribution to a recent handbook (Brown, Lomas, & Eiroa-Orosa, 2017) of which I was also the lead editor. This book represents an attempt to move the focus of discussion in positive psychology away from its tendency (as we, the editors, saw it) to focus mostly on improving happiness and well-being among affluent Western populations (cf. Coyne, 2013) and towards its ostensible original goal of developing a psychology of optimal human functioning (Rathunde, 2001). In particular, this chapter presents a number of suggestions for ways in which students of positive psychology, such as those enrolled on MAPP programs, can raise a critical voice to question some of the more ambitious claims of positive psychology research. The chapter was developed from a number of talks that I have been invited to give at various universities over the past five years, in which the material of Chapters 2 through 7 plays a prominent role; Chapters 4, 5, and 6 are referenced directly, and

some of the ideas behind Chapter 7 (which was written after Chapter 8) are also present in this chapter.

In Chapter 9 I discuss the findings from the preceding chapters, and draw some general conclusions. Finally, in Chapter 10, I present some other projects that I have undertaken during the preparation of this thesis, which have at least some degree of relevance to my overall conclusions.

References

- Breslow, L. (1972). A quantitative approach to the World Health Organization definition of health: Physical, mental and social well-being. *International Journal of Epidemiology*, 1, 347–355. <https://doi.org/10.1093/ije/1.4.347>
- Boruchovitch, E., & Mednick, B. R. (2002). The meaning of health and illness: Some considerations for health psychology. *Psico-USF*, 7, 175–183. <https://doi.org/10.1590/S1413-82712002000200006>
- Brown, N. J. L. (2015). A critical examination of the U.S. Army's Comprehensive Soldier Fitness program. *The Winnower*, 3, e143751.17496. <https://doi.org/10.15200/winn.143751.17496>
- Brown, N. J. L., Lomas, T., & Eiroa-Orosa, F. J. (2018). *The Routledge international handbook of critical positive psychology*. London, England: Routledge.
- Chesney, E., Goodwin, G. M., & Fazel, S. (2014). Risks of all-cause and suicide mortality in mental disorders: A meta-review. *World Psychiatry*, 13, 153–160. <https://doi.org/10.1002/wps.20128>
- Coyne, J. C. (2013, August 21). *Positive psychology is mainly for rich white people* [Weblog post]. Retrieved from <https://mindthebrain.blog/2013/08/21/positive-psychology-is-mainly-for-rich-white-people>
- Coyne, J. C., Tennen, H., & Ranchor, A. V. (2010). Positive psychology in cancer care: A story line resistant to evidence. *Annals of Behavioral Medicine*, 39, 35–42. <https://doi.org/10.1007/s12160-010-9157-9>.
- Don't bother with the gym today: Positive outlook on life and making friends could be as good for you as diet and exercise (2013, May 11). *Daily Mail*. Retrieved from <http://www.dailymail.co.uk/health/article-2323054/Dont-bother-gym-today-Positive-outlook-life-making-friends-good-diet-exercise.html>
- Duckworth, A. (2016). *Grit: The power of passion and perseverance*. New York, NY: Scribner.
- Fredrickson, B. L. (2009). *Positivity: Groundbreaking research reveals how to embrace the hidden strength of positive emotions, overcome negativity, and thrive*. New York, NY: Crown.
- Fredrickson, B. L. (2013). *Love 2.0: Finding happiness and health in moments of connection*. New York, NY: Hudson Street Press.
- Fredrickson, B. L., Grewen, K. M., Coffey, K. A., Algoe, S. B., Firestone, A. M., Arevalo, J. M. G., . . . Cole, S. W. (2013). A functional genomic perspective on human well-being. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 13684–13689. <https://doi.org/10.1073/pnas.1316887110>
- Fredrickson, B. L., Grewen, K. M., Algoe, S. B., Firestone, A. M., Arevalo, J. M. G., Ma, J., & Cole, S. W. (2015). Psychological well-being and the human conserved transcriptional response to adversity. *PLoS ONE*, 10(3), e0121839. <https://doi.org/10.1371/journal.pone.0121839>
- Fredrickson, B. L., & Losada, M. F. (2005). Positive affect and the complex dynamics of human flourishing. *American Psychologist*, 60, 678–686. <https://doi.org/10.1037/0003-066X.60.7.678>
- Ghosh, A., & Deb, A. (2017). Positive psychology interventions for chronic physical illnesses: A systematic review. *Psychological Studies*, 62, 213–232. <https://doi.org/10.1007/s12646-017-0421-y>

- Heathers, J. A. J., Brown, N. J. L., Friedman, H. L., & Coyne, J. C. (2015). The elusory upward spiral: A reanalysis of Kok et al. (2013). *Psychological Science*, 26, 1140–1143. <https://doi.org/10.1177/0956797615572908>
- Huber, M., Knottnerus, J. A., Green, L., van der Horst, H., Jadad, A. R., Kromhout, D., . . . Smid, H. (2011). How should we define health? *British Medical Journal*, 343, d4163. <https://doi.org/10.1136/bmj.d4163>
- Huber, M. (2014). *Towards a new, dynamic concept of health: Its operationalisation and use in public health and healthcare, and in evaluating health effects of food* [Doctoral dissertation]. Maastricht University. Retrieved from <http://www.nvag.nl/afbeeldingen/2014/Thesis%20Machteld%20Huber.pdf>
- Huber, M., van Vliet, M., Giezenberg M., Winkens, B., Heerkens, Y., Dagnelie, P. C., & Knottnerus, J. A. (2016). Towards a ‘patient-centred’ operationalisation of the new dynamic concept of health: A mixed methods study. *BMJ Open*, 5, e010091. <https://doi.org/10.1136/bmjopen-2015-010091>
- Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior*, 43, 207–222. <https://doi.org/10.2307/3090197>
- Kok, B. E., Coffey, K. A., Cohn, M. A., Catalino, L. I., Vacharkulksemsuk, T., Algoe, S. B., . . . Fredrickson, B. L. (2013a). How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone. *Psychological Science*, 24, 1123–1132. <https://doi.org/10.1177/0956797612470827>
- Kubzansky, L. D., Huffman, J. C., Boehm, J. K., Hernandez, R., Kim, E. S., Koga, H. K., . . . Labarthe, D. R. (2018). Positive psychological well-being and cardiovascular disease. *Journal of the American College of Cardiology*, 72, 1382–1396. <https://doi.org/10.1016/j.jacc.2018.07.042>
- Larson, J. S. (1999). The conceptualization of health. *Medical Care Research and Review*, 56, 123–136. <https://doi.org/10.1177/107755879905600201>
- Lazarus, R. S. (2003). Does the positive psychology movement have legs? *Psychological Inquiry*, 14, 93–109. https://doi.org/10.1207/S15327965PLI1402_02
- Levine, S. (2000). Influence of psychological variables on the activity of the hypothalamic-pituitary-adrenal axis. *European Journal of Pharmacology*, 405, 149–160.
- Lewis, A. (1953). Health as a social concept. *British Journal of Sociology*, 4, 110–115. <https://doi.org/10.2307/587206>
- Lyubomirsky, S. (2007). *The how of happiness: A scientific approach to getting the life you want*. New York, NY: Penguin.
- Lyubomirsky, S. (2013). *The myths of happiness: What should make you happy, but doesn't; What shouldn't make you happy, but does*. New York, NY: Penguin.
- Lyubomirsky, S., Sheldon, K. M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology*, 9, 111–131. <https://doi.org/10.1037/1089-2680.9.2.111>
- Nsamenang, S. A., & Hirsch, J. K. (2015). Positive psychological determinants of treatment adherence among primary care patients. *Primary Health Care Research & Development*, 16, 398–406. <https://doi.org/10.1017/S1463423614000292>

- Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emodiversity and the emotional ecosystem. *Journal of Experimental Psychology: General*, 143, 2057–2066. <https://doi.org/10.1037/a0038025>
- Rathunde, K. (2001). Toward a psychology of optimal human functioning: What positive psychology can learn from the “experiential turns” of James, Dewey, and Maslow. *Journal of Humanistic Psychology*, 41(1), 135–153. <https://doi.org/10.1177/0022167801411008>
- Salvador-Carulla, L., & Garcia-Gutierrez, C. (2011). The WHO construct of health-related functioning (HrF) and its implications for health policy. *BMC Public Health*, 11(Suppl. 4), S9. <https://doi.org/10.1186/1471-2458-11-S4-S9>
- Saracci, R. (1997). The World Health Organisation needs to reconsider its definition of health. *British Medical Journal*, 314, 1409. <https://doi.org/10.1136/bmj.314.7091.1409>
- Scheier, M. F., & Carver, C. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, 4, 219–247. <https://doi.org/10.1037/0278-6133.4.3.219>
- Seligman, M. E. P. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. New York, NY: Free Press.
- Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being*. New York, NY: Free Press.
- Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, 55, 5–14. <https://doi.org/10.1037/0003-066X.55.1.5>
- Seligman, M. E. P., Peterson, C., Barsky, A. J., Boehm, J. K., Kubzansky, L. D., Park, N., & Labarthe, D. (2013). Positive health and health assets: Re-analysis of longitudinal datasets. Positive Health Research. Retrieved from <https://ppc.sas.upenn.edu/sites/default/files/positivehealthassetspub.pdf>
- Smith, E. E. (2013). Meaning is healthier than happiness *The Atlantic*, August 1. Retrieved from <https://www.theatlantic.com/health/archive/2013/08/meaning-is-healthier-than-happiness/278250>
- Smith, R. (2008, July 8) *The end of disease and the beginning of health* [Weblog post]. Retrieved from <https://blogs.bmj.com/bmj/2008/07/08/richard-smith-the-end-of-disease-and-the-beginning-of-health>
- Snyder, C. R., Irving, L. M., & Anderson, J. R. (1991). Hope and health. In C. R. Snyder & D. R. Forsyth (Eds.), *Handbook of social and clinical psychology: The health perspective* (pp. 285–305). Elmsford, NY: Pergamon Press.
- Trudel-Fitzgerald, C., Boehm, J. K., Tworoger, S. S., & Kubzansky, L. D. (2018). Being happier may lead to better health: Positive psychological well-being and lifestyle over 20 years of follow-up. *IPPA Chronicle*, 1. Retrieved from <https://www.ippanetwork.org/2018/07/27/being-happier-may-lead-to-better-health-positive-psychological-well-being-and-lifestyle-over-20-years-of-follow-up>
- World Health Organization. (1946). *Constitution of the World Health Organization*. Retrieved from <http://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf>
- Wood, A. M., Maltby, J., Gillett, R., Linley, P. A., & Joseph, S. (2008). The role of gratitude in the development of social support, stress, and depression: Two longitudinal studies. *Journal of Research in Personality*, 42, 854–871. <https://doi.org/10.1016/j.jrp.2007.11.003>

Wood, A. M., & Tarrier, N. (2010). Positive clinical psychology: A new vision and strategy for integrated research and practice. *Clinical Psychology Review*, 30, 819–829.
<https://doi.org/10.1016/j.cpr.2010.06.003>

Chapter 2

A Critical Reanalysis of Fredrickson et al.'s Study of Genomics and Well-being

Nicholas J. L. Brown¹

Douglas A. MacDonald²

Manoj P. Samanta³

Harris L. Friedman⁴

James C. Coyne

Proceedings of the National Academy of Sciences, 111, 12705–12709 (August 2014)⁵.

DOI: 10.1073/pnas.1407057111

¹ At the time when the manuscript was submitted for publication, my affiliation was with the New School of Psychotherapy and Counselling, London, England.

² University of Detroit Mercy

³ Systemix Institute, Redmond, WA

⁴ Saybrook University

⁵ The published article was accompanied by a Supporting Information document, available at <http://www.pnas.org/content/suppl/2014/08/21/1407057111.DCSupplemental>. Although that document is not part of the present thesis, the Appendix of the present chapter has been extracted from it, as it is of relevance to Chapter 3.

Abstract

Fredrickson et al. (2013a) claimed to have observed significant differences in gene expression related to “hedonic” and “eudaimonic” dimensions of well-being. Having closely examined both their claims and their data, we draw substantially different conclusions. After identifying some important conceptual and methodological flaws in their argument, we report the results of a series of reanalyses of their data set. We first applied a variety of exploratory and confirmatory factor analysis techniques to their self-reported well-being data. A number of plausible factor solutions emerged, but none of these corresponded to Fredrickson et al.’s claimed hedonic and eudaimonic dimensions. We next examined the regression analyses that purportedly yielded distinct differential profiles of gene expression associated with the two well-being dimensions. Using the best-fitting two-factor solution that we had identified, we obtained “effects” almost twice as large as those found by Fredrickson et al. using their questionable Hedonic and Eudaimonic factors. Next, we conducted regression analyses for all possible two-factor solutions of the psychometric data; we found that 69.2% of these gave statistically significant results for both factors, whereas only 0.25% would be expected to do so if the regression process were really able to identify independent differential gene expression effects. Finally, we replaced Fredrickson et al.’s psychometric data with random numbers and continued to find very large numbers of apparently statistically significant “effects.” We conclude that Fredrickson et al.’s widely-publicized claims about the effects of different dimensions of well-being on health-related gene expression are merely artifacts of dubious analyses and erroneous methodology.

Introduction

In their article which has captured the attention and interest of the popular media, such as *The Economist* (M. K., 2013) and CNN (Christensen, 2013), Fredrickson et al. (2013a) claimed to have shown that specific genomic factors for human health are differentially associated with “hedonic” and “eudaimonic” well-being, interpreting their results as suggesting that the way in which individuals seek happiness may have detrimental or beneficial effects on their physical health. Nowhere is this better expressed than in a press article which reported Fredrickson et al.’s findings as having demonstrated that “people who are happy but have little to no sense of meaning in their lives—proverbially, simply here for the party—have the same gene expression patterns as people who are responding to and enduring chronic adversity” (Smith, 2013, para. 14). Given the apparent importance of their findings, which appeared to amount to nothing less than a true breakthrough in behavioral genomics research, we eagerly and with great earnest read the article with the hope that science might finally have been able to illuminate true pathways to “the good life” (or at least help to divert people from a not-so-good life). Unfortunately, what we encountered did not strike us as a breakthrough. In fact, after an extensive reanalysis of Fredrickson et al.’s data, we concluded that their study suffers from numerous problems which render its conclusions unfounded and potentially misleading. In this short commentary, we would like to share our criticisms so that the reader can make an informed decision about the real meaning and value for science of Fredrickson et al.’s study.

This commentary is organized into three sections. The first summarizes what we perceive as some of the theoretical and general methodological shortcomings of Fredrickson et al.’s (2013a) study; it was our observation of these issues, and our consequent skepticism about the plausibility of the authors’ findings, that led us to conduct our detailed reanalyses of their data. The second and third sections focus upon two specific statistical problems, namely the factor analysis of the measure of well-being used by Fredrickson et al., and their use of multiply-iterated regressions to test for the differential relations of hedonic and eudaimonic well-being to gene expression. For the sake of brevity, in the present article we limit our critique to that part of their article that Fredrickson et al. described as their primary analysis, that is, the examination of the relationship between well-being and 53 conserved transcriptional response to adversity (CTRA) genes; we do not discuss their secondary analysis, where the Transcription Element Listening System (TELiS) database was used to derive information about purported differential effects of hedonic and eudaimonic well-being across the entire genome. We anticipate, however, that the publication of the present article

might lead other researchers to examine this secondary analysis in detail; our Supporting Information document contains the results of some preliminary examinations of one aspect of that part of Fredrickson et al.'s study, which suggest that it may also have considerable problems.

Theoretical and methodological issues

Based on their view that “psychological well-being has been shown to forecast future physical health above and beyond its association with current physical health . . . and above and beyond its association with reduced levels of stress, depression, and other negative affect states” (p. 13684), Fredrickson et al. (2013a) set out in their study to identify “molecular signaling pathways that transduce positive psychological states into somatic physiology” (p. 13684). While we do not object to the basic intent of their study—in fact, we see their goals as laudable, since the emergent field of epigenetics has provided compelling reason to see behavior and experience as having a determining influence on gene expression (Cloninger, 2004)—we are critical of Fredrickson et al.'s vagueness in conceptualization and presumed directionality of effect. Other than abductively attributing possible evolutionary significance to a presumed link between positive psychological states and a set of candidate genes implicated in physiological stress response, the authors provided virtually nothing in terms of a theoretically informed directional model to guide their investigation (for example, at the most basic level, there is no clear statement as to whether the researchers see positive states as determining gene expression or vice versa) and, ultimately, the interpretation of their results. For example, the claim made by Fredrickson et al. that hedonic well-being is associated with reports of subjective happiness but is, unbeknownst to people, tied to lack of well-being in other areas of functioning (e.g., physical well-being) is a finding that is actually consistent with research on positive illusions/self-serving bias/self-enhancement bias, a phenomenon which is well documented and has even been characterized as adaptive (Blaine & Crocker, 1993; Shedler, Mayman, & Manis, 1993; Taylor & Brown, 1994). Consequently, their project is reduced to a nebulous and largely exploratory correlational study without any solid founding in available theory and research.

More problematic, however, is the difficulty with the conceptualization of positive psychological states which Fredrickson et al. (2013a) defined in terms of eudaimonic and hedonic well-being. There are two salient weaknesses here. First, despite their assertion of conceptual uniqueness but concurrent acknowledgement of the correlatedness and reciprocal influence of eudaimonic and hedonic well-being on each other, they appeared to give little

consideration as to whether well-being represents a state (i.e., a transient aspect of behavior and experience) or a trait (i.e., a more pervasive and stable aspect of behavior and experience) construct. Given our understanding of these constructs, eudaimonic well-being, generally defined (including by Fredrickson et al.) in terms of tendencies to strive for meaning, appears to be trait-like, since such striving for meaning is typically an ongoing life project, in the sense of Sartre (1956). Conversely, hedonic well-being, typically defined in terms of a person's (recent) affective experiences, appears to be state-like; regardless of the level of meaning in one's life, everyone experiences "good" and "bad" days. Notwithstanding our speculation, the lack of clarity and precision regarding well-being as a state versus trait is unfortunate since it readily translates into potential methodological challenges for a study on functional genomics. For example, if well-being is a state, then a person's level of well-being will change over time and perhaps at a very fast rate. If we only measure well-being at one time point, as Fredrickson et al. did, then unless we obtain a genetic sample at the same time, the likelihood that the well-being score will actually and accurately reflect level of genomic expression will be diminished if not eliminated.

Second, even though the eudaimonic and hedonic well-being constructs have a venerable history in philosophy and psychology, they by no means capture the richness and complexity of the well-being construct domain as manifested in the extant literature. This conceptual limitation leads us to question to whether or not there is a scientifically adequate definition and taxonomy of well-being on which to do research in the first place. The scope of this difficulty becomes painfully obvious when one considers the various incarnations of well-being concepts proffered by several researchers, including general well-being, subjective well-being, psychological well-being, ontological well-being, spiritual well-being, religious well-being, existential well-being, chaironic well-being, emotional well-being, and physical well-being, along with the various constructs which are treated as essentially synonymous with well-being, such as self-esteem, life-satisfaction, and, lest we forget, happiness.

Departing from theoretical concerns and focusing briefly on methodological weaknesses, there are three things about Fredrickson et al.'s (2013a) study that seem to us to merit explicit mention here. First, the sample consisted of just 80, mostly white American adults. (The sample size in the publicly-available data set is, in fact, less than 80; the consequences of this are discussed in more detail in our Supporting Information document.) Although the authors acknowledged that "replication of these findings in other populations ... will be required to gauge generality and consistency of these effects" (ref. 6, p. 13687), this

potential lack of generality was not reported in the majority of reports of this work in major popular new outlets. Also, given the number and type of statistical analyses that were conducted and the number of variables involved, such a small sample size considerably limits the statistical power of the study, especially when running a large number (i.e., 53) of independent regressions with a large number (i.e., 17) of predictor variables. Second, the assessment of well-being was completed through the single administration of one self-report measure and biological samples were obtained at only one time for all participants. Ostensibly, such a state of affairs raises the specter of potential issues with response bias, and confounding situational influences on both self-report data and biological samples. To give just one example, we doubt whether an inventory of minor health problems over the preceding two weeks constitutes an adequate control for the possible effects of an *ongoing* infectious condition on the expression of genes associated with immune system response. Third, the choice of 53 human genes out of some 20,000 potential candidates appears to be based almost exclusively on the prior opinion of one of the authors of Fredrickson et al.'s article, who is also the author, or a co-author, of all of the works cited by Fredrickson et al. in support of their selection of genes. In addition, the decision to apply a weighting of exactly 1.00 or -1.00 to the regression coefficients associated with each of these genes, thus assigning an identical magnitude of effect to every gene, seems to be completely arbitrary. A recent article (Sharon, Tilgner, Grubert, & Snyder, 2013) found the human transcriptome to be extremely complex with over 100,000 distinct transcripts coding for around 20,000 protein-coding genes. In that context, the gene expression model assumed by the authors appears to be rather simplistic.

We turn now to the first of two statistical aspects of Fredrickson et al.'s (2013a) study for which we performed an extensive reanalysis of their data, namely the factor analysis of the measure of well-being they used.

Problems with the factor structure of the Short Flourishing Scale

Fredrickson et al. (2013a) measured well-being using an instrument that they named the Short Flourishing Scale, although this seems to have previously been referred to in the literature as the Mental Health Continuum-Short Form (MHC-SF) (Keyes et al., 2008). According to Keyes (2011), the MHC-SF assesses three forms of well-being: emotional well-being (items SF1–SF3), social well-being (items SF4–SF8), and psychological well-being (items SF9–SF14). This three-dimensional structure has been confirmed in empirical studies (Keyes et al., 2008; Lamers, Westerhof, Bohlmeijer, ten Klooster, & Keyes, 2011). In an

Appendix describing the technical details of the scoring of the MHC-SF, Keyes (2011) applied the word “hedonic” to the first factor and “eudaimonic” (or “positive functioning”) to the second and third, and indicated that a person’s “flourishing” status depends on how many items from each of these two groups are experienced with a minimum specific frequency. However, this split of the MHC-SF items into hedonic and eudaimonic categories appears to have been made principally to simplify the instructions for diagnosing persons as “flourishing”; we were not able to identify any published evidence supporting an underlying psychometric factor in which the previously empirically-demonstrated emotional and social well-being factors combined into one. Indeed, we note that Keyes et al. (2008) referred to these groupings of hedonic and eudaimonic items as “clusters,” an ostensibly neutral term that seems to deliberately avoid the word “factor.” Nevertheless, Fredrickson et al. (2013a) implied, and Cole and Fredrickson (2013b) stated explicitly, that a factor analysis of the psychometric data from the MHC-SF in their study revealed just two distinct factors (“Hedonic” and “Eudaimonic”), corresponding to Keyes’ (2011) “diagnostic question” categories (i.e., items SF1–SF3 for Hedonic and SF4–SF14 for Eudaimonic).

Even though the factor structure of the MHC-SF that Cole and Fredrickson (2013) claimed to have obtained deviates from what has previously been reported for the scale in the published empirical literature and is, thus, already a cause for concern, what specifically raised a red flag for us is the high degree of correlation between the Hedonic and Eudaimonic factors (i.e., $r = .79$, $p < .0001$), as pointed out by Coyne (2013). In conventional psychometric research, such a high intercorrelation would usually be interpreted as suggesting that the two constructs are essentially measuring the same thing. Interestingly, while acknowledging the high correlation, Cole and Fredrickson (2013) argued, based upon the observed reliabilities for the two factors, that about 30% of the variance of each factor is unique and can be used to explore the novel associations of the two forms of well-being to genomic expression. Unfortunately, while their use of their reliabilities to establish how much unique systematic variance remains is not in itself erroneous, their assumption that the 30% of variance is all reflective of meaningful construct variance may be seen as without solid footing. Research has demonstrated that method bias, which is systematic, is pervasive and can have deleterious effects on research findings (Podsakoff, MacKenzie, & Podsakoff, 2012).

Notwithstanding the above issues, we took it upon ourselves to use the data from Fredrickson et al.’s (2013a) study to complete our own exploratory and confirmatory factor analyses to examine the internal structure of the test. Though we only summarize our findings

here, the results of these analyses are reported in detail in the Supporting Information for this article.

In the exploratory factor analyses (EFAs), which we ran using different extraction (e.g., principal axis, maximum likelihood) and rotation (orthogonal, oblique) methods, we found two factors with eigenvalues greater than 1 with all items producing a loading of .50 on at least one factor. Examination of factor loading coefficients consistently showed that the first factor was comprised of elevated loadings from items SF1, SF2, SF3, SF4, SF5, SF9, SF10, SF11, SF12, SF13, and SF14, while the second factor housed high loadings from items SF6, SF7, and SF8. Examination of item content to devise labels for the factors led us to name the first factor “Personal well-being” (PWB) and the second factor—comprised of those MHC-SF items that ask for the respondent’s opinions about society, rather than personal introspection—“Evaluative perception of the social environment” (EPSE). Thus, our EFA findings show that, although it appears that the MHC-SF does seem to emulate a two-factor structure, the distribution of high item loadings across the factors does not conform to what Cole and Fredrickson (2013) reported, nor to what other published studies have found (Keyes et al., 2008; Lamers et al., 2011).

Considering next our confirmatory factor analyses (CFAs), we tested one- and two-factor models to see if the construct of well-being as operationalized by the MHC-SF best fits different theoretically defensible expressions of well-being (e.g., general well-being versus hedonic and eudaimonic well-being, with items assigned to each factor as per Cole and Fredrickson, 2013). In the two-factor model, the latent constructs of hedonic and eudaimonic well-being were permitted to intercorrelate. In all CFAs, factor loadings for all items were found to be statistically significant at $p < .05$ or lower. For the one-factor model, goodness-of-fit statistics indicated grossly inadequate fit ($\chi^2 = 227.64$, $df = 77$, goodness-of-fit index (GFI) = .73, comparative fit index (CFI) = .83, root-mean-square error of approximation (RMSEA) = .154). Although the equivalent statistics for the correlated two-factor model were slightly better, they still came out as poor ($\chi^2 = 189.40$, $df = 76$, GFI = .78, CFI = .87, RMSEA = .135). Thus, even though our findings tended to support the view that well-being is best represented as at least a two dimensional construct, they did not confirm Fredrickson et al.’s (2013a) claim that the MHC-SF produces two factors conforming to hedonic and eudaimonic well-being. Extending from this, we are sure that the reader can appreciate the implications this holds for Fredrickson et al.’s study—if the only measure used to operationalize well-being does not demonstrate factorial validity in a manner consistent with the theory underlying the test, then any analyses and associated assertions based upon those

analyses are rendered highly suspect in their scientific value. As Ryff and Singer (2003) put it, “Lacking evidence of scale validity and reliability, subsequent work is pointless” (p. 276).

As our last point of comment, and perhaps the most significant, we now consider the regression analyses performed by Fredrickson et al. (2013a).

Problems with the regression analyses

Fredrickson et al. (ref. 6, p. 13685; ref. 7, pp. 1–2) described their method of analysis that led to their principal result—namely, that the expression of CTRA genes differs as a function of their two purported well-being dimensions—as follows:

General linear model analyses quantified the association between expression of each of the 53 CTRA contrast genes and levels of hedonic and eudaimonic well-being [each well-being dimension treated as a continuous measure and adjusted for correlation with the other dimension of well-being and for age, sex, race/ethnicity, body mass index (BMI), smoking, alcohol consumption, recent minor illness symptoms, and leukocyte subset prevalence ...]. Contrast coefficient-weighted association statistics were averaged to summarize the magnitude of association over the entire CTRA gene set.

Specifically, the sequence of operations constituting the procedure described above is the following:

1. For each of the 53 CTRA genes of interest (ref. 7, p. 1), an ordinary least-squares linear regression is performed with the gene expression value as the dependent variable. The predictor variables in this regression are, first, the seven demographic variables (age, sex, “race” (white/nonwhite), BMI, alcohol consumption yes/no, smoking yes/no, and number of recent minor illness symptoms); second, the expression levels of the eight “control” genes (*CD3D*, *CD3E*, *CD4*, *CD8A*, *CD19*, *FCGR3A*, *NCAM1*, and *CD14*); third, the two standardized (z-scored) values of the purported Hedonic (SF1–SF3) and Eudaimonic (SF4–SF14) factors from the psychometric data.
2. The above regression generates coefficients for each of the 17 predictor variables, of which the two of interest are those for the Hedonic and Eudaimonic factors. These two coefficients are independently summed, after having first been multiplied by –1 in the case of the 34 genes that are expected to be down-regulated in the CTRA.
3. The average coefficient for the Hedonic and Eudaimonic factors—representing the difference in gene expression attributable to each of these factors—is determined by

dividing the respective sums, obtained in step 2, by the number of CTRA genes (i.e., 53).

4. The averaged coefficients from step 3 are tested for a statistically significant difference from zero using a one-sample *t*-test. The null hypothesis is that there is no difference between the average gene expression differences attributable to hedonic and eudaimonic well-being (ref. 7, p. 2).

Our first reaction to this description was one of surprise that Fredrickson et al. apparently expected to generate meaningful results when analyzing 17 independent variables using just 80 cases. As Tabachnik and Fidell put it: “The cases-to-IVs ratio has to be substantial or the solution will be perfect—and meaningless” (Tabachnik & Fidell, 2001, p. 123). These authors give a formula ($50 + 8m$, where m is number of independent variables) suggesting that the minimum number of cases required in this study for would be at least 186.

We also consider this process—which we refer to henceforth as “RR53,” for “Regression Repeated 53 times”—to be unnecessarily complicated. It seems to us that it would be far simpler to regress the scores for hedonic and eudaimonic well-being on the average expression of the 53 genes of interest, after changing the sign of the values of those genes that were expected to be down-regulated. This would appear to correspond closely to Fredrickson et al.’s (2013b) statement that “[T]he goal of this study is to test associations between eudaimonic and hedonic well-being and average levels of expression of specific sets of genes” (p. 1); it would also have the advantage of greatly reducing the number of dependent variables being predicted from the sample size of 80 participants. We conducted a number of such regressions, using different methods of evaluating the “average level of expression” of the 53 CTRA genes of interest (e.g., taking the mean of their raw values, or the mean of their *z*-scores), but in all cases the model ANOVA was not statistically significant. We therefore set out to analyze and understand in more detail how the RR53 regression procedure could, in contrast to our “naive” regressions, be producing such highly significant results.

Our first step⁶ was to apply the RR53 procedure to reproduce Fredrickson et al.’s (2013a) Figure 2, showing the associations between hedonic well-being and up-regulated CTRA genes, eudaimonic well-being and down-regulated CTRA genes, and the split of these two overall plots into three gene subsets. During this process, which we performed with SPSS

⁶ The data and code needed to reproduce the analyses in this chapter are available at <https://osf.io/xjg53/>

version 18 (and calibrated against several other software packages; see the Supporting Information), we noticed that a substantial number of the regression models for the CTRA genes had a non-significant model ANOVA ($p > .05$ in 22 out of 53 cases). Furthermore, the t tests for the regression coefficients corresponding to the predictor variables of interest, namely hedonic and eudaimonic well-being, were almost all non-significant ($p > .05$ in 104 out of 106 cases; mean $p = .567$, $SD = 0.251$), and in the two remaining cases (gene *FOSL1*, for both “hedonic,” $p = .047$, and “eudaimonic,” $p = .030$), the overall model ANOVA was not statistically significant ($p = .146$). We believe that to draw any conclusions from these coefficients is inappropriate. (The RR53 procedure appears to be exquisitely sensitive to even the smallest variations in the data. Readers are invited to consult the Supporting Information for further details.)

Having reproduced Fredrickson et al.’s (2013a) principal result, and also become aware of the alternative factor structure described in the previous section of the present article, we next proceeded to apply the same statistical techniques to the “Personal well-being/Evaluative perception of the social environment” factor pair. We therefore created two new variables, which we named PWB (corresponding to items SF1–SF5 and SF9–SF14) and EPSE (corresponding to items SF6–SF8). When we applied Fredrickson et al.’s regression procedure using these variables as the two principal predictor variables of interest (replacing the Hedonic and Eudaimonic factor variables), we discovered that the “effects” of this factor pair were about twice as high as those for the Hedonic and Eudaimonic pair (PWB: up-regulation by 13.6%, $p < .001$; EPSE: down-regulation by 18.0%, $p < .001$; see Figures 3 and 4 in the Supporting Information). We found this result rather curious, as it suggests that the participants’ genes are not only expressing “molecular well-being” (ref. 6, p. 13688), but apparently also, and even more vigorously, some other response that we presume Fredrickson et al. might call “molecular social evaluation.”

Curious as to what else these genes might be able to tell us, we wrote a program in R to analyze the effect of applying the RR53 procedure to every possible way of splitting of the psychometric data items SF1–SF14 into two pseudo-factors, which we then used in the RR53 procedure in place of the Hedonic and Eudaimonic factors from Fredrickson et al.’s (2013a) original analysis. Excluding duplicates due to symmetry, there are 8,191 possible such combinations. Of these, we found that 5,670 (69.2%) gave statistically significant results using the method described on pp. 1–2 of Fredrickson et al.’s (2013b) Supporting Information (i.e., the t tests of the fold differences corresponding to the two elements of the pair of pseudo-factors were both significant at the .05 level), with 3,680 of these

combinations (44.9% of the total) having both components significant at the .001 level. Furthermore, 5,566 combinations (68.0%) generated statistically significant pairs of fold difference values that were greater in magnitude than Fredrickson et al.'s (2013a, figure 2A) Hedonic and Eudaimonic factors.

While one possible explanation of these results is that differential gene expression is associated with almost any factor combination of the psychometric data, with the study participants' genes giving simultaneous "molecular expression" to several thousand factors which psychologists have not yet identified, we suspected that there might be a more parsimonious explanation. Therefore, as a further test of the validity of the RR53 procedure, we replaced Fredrickson et al.'s (2013a) psychometric data with random numbers (i.e., every item/respondent cell was replaced by a random integer in the range 0–5) and re-ran the R program. We did this in two different ways. First, we replaced the psychometric data with normally-distributed random numbers, such that the item-level means and standard deviations were close to the equivalent values for the original data. With these pseudo-data, 3,620 combinations of pseudo-factors (44.2%) gave a pair of fold difference values having *t* tests significantly different from zero at the .05 level; of these, 1,478 (18.0% of the total) were both statistically significant at the .001 level. (We note that, assuming independence of up- and down-regulation of genes, the probability of the latter result occurring by chance with random psychometric data if the RR53 regression procedure does indeed identify differential gene expression as a function of psychometric factors, ought to be—literally—one in a million, i.e. 0.001^2 , rather than somewhere between one in five and one in six.) Second, we used uniformly-distributed random numbers (i.e., all "responses" were equally likely to appear for any given item and respondent). With these "white noise" data, we found that 2,874 combinations of pseudo-factors (35.1%) gave a pair of fold difference values having *t* tests statistically significantly different from zero at the .05 level, of which 893 (10.9% of the total) were both significant at the .001 level. Finally, we re-ran the program once more, using the same uniformly distributed random numbers, but this time excluding the demographic data and control genes; thus, the only non-random elements supplied to the RR53 procedure were the expression values of the 53 CTRA genes. Despite the total lack of *any* information with which to correlate these gene expression values, the procedure generated 2,540 combinations of pseudo-factors (31.0%) with a pair of fold difference values having *t* tests statistically significantly different from zero at the .05 level, of which 235 (2.9% of the total) were both significant at the .001 level.

Thus, in all cases, we obtained far more statistically significant results using Fredrickson et al.'s (2013a) methods than would be predicted by chance alone for truly independent variables (i.e., $.05^2 \times 8191 \approx 20$), even when the psychometric data were replaced by meaningless random numbers. To try to identify the source of these puzzling results, we ran simple bivariate correlations on the gene expression variables, which revealed moderate to strong correlations between many of them, suggesting that our significant results were mainly the product of shared variance across criterion variables. We therefore went back to the original psychometric data, and “scrambled” the CTRA gene expression data, reassigning each cell value for a given gene to a participant selected at random, thus minimizing any within-participants correlation between these values. When we re-ran the regressions with these data, the number of statistically significant results dropped to just 44 (.54%).

To summarize: even when fed *entirely random psychometric data*, the RR53 regression procedure generates large numbers of results that appear, according to these authors' interpretation, to establish a statistically significant relationship between self-reported well-being and gene expression. We believe that this regression procedure is, simply put, totally lacking in validity. It appears to be nothing more than a mechanism for producing apparently statistically significant effects from non-significant regression coefficients, driven by a high degree of correlation between many of the criterion variables. In particular, there is no reason to believe that the result shown in Fredrickson et al.'s Figure 2A demonstrates any relationship between hedonic and eudaimonic well-being on the one hand, and differential gene expression on the other.

Conclusion

In this article, we have highlighted the myriad problems with Fredrickson et al.'s (2013a) study, which range from theory and conceptualization, to measurement, to statistics and interpretation of findings. As we said at the beginning, we do not fault these authors for wanting to study the relationship of psychological constructs to functional genomics. However, given the remarkable claims made by Fredrickson et al., as well as the attention their findings have garnered in the popular media, we believe it is vitally important that researchers, practitioners, and laypersons alike should be made aware of the deficiencies of their study and, in particular, of their statistical analyses, which appear to have conjured non-existent “effects” out of thin air. For positive psychology to move into the health arena in a responsible way (Hefferon, 2013), the research on which advice is based must be rigorous

and avoid giving false hope or even causing iatrogenic harm. It seems to us that Fredrickson et al.'s study falls far short of the level of responsible scholarship required in promoting human health and well-being.

References

- Blaine, B., & Crocker, J. (1993). Self-esteem and self-serving biases in reactions to positive and negative events: An integrative review. In R. F. Baumeister (Ed.), *Self-esteem: The puzzle of low self-regard* (pp. 55–85). New York, NY: Springer.
- Christensen, J. (2013, July 30). *Your happiness type matters*. Retrieved from <http://edition.cnn.com/2013/07/30/health/happy-genes>
- Cloninger, C. R. (2004). *Feeling good: The science of well-being*. New York, NY: Oxford University Press.
- Cole, S. W., & Fredrickson, B. L. (2013). Reply to Coyne: Genomic analyses are unthwarted. *Proceedings of the National Academy of Sciences of the United States of America*, 110, E4184. <https://doi.org/10.1073/pnas.1316887110>
- Coyne, J. C. (2013). Highly correlated hedonic and eudaimonic well-being thwart genomic analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 110, E4183. <https://doi.org/10.1073/pnas.1316887110>
- Fredrickson, B. L., Grewen, K. M., Coffey, K. A., Algoe, S. B., Firestone, A. M., Arevalo, J. M. G., . . . Cole, S. W. (2013a). A functional genomic perspective on human well-being. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 13684–13689. <https://doi.org/10.1073/pnas.1305419110>
- Fredrickson, B. L., Grewen, K. M., Coffey, K. A., Algoe, S. B., Firestone, A. M., Arevalo, J. M. G., . . . Cole, S. W. (2013b). Supporting information. *Proceedings of the National Academy of Sciences of the United States of America*, 110. Retrieved from <http://www.pnas.org/lookup/suppl/10.1073/pnas.1305419110/-/DCSupplemental>
- Hefferon, K. (2013). *Positive psychology and the body: The somatopsychic side of flourishing*. Maidenhead, England: Open University Press.
- Keyes, C. L. M. (2011). *Brief description of the mental health continuum short form (MHC-SF)*. Retrieved from <https://www.aacu.org/sites/default/files/MHC-SFEnglish.pdf>
- Keyes, C. L. M., Wissing, M., Potgieter, J. P., Temane, M., Kruger, A., & van Rooy, S. (2008). Evaluation of the mental health continuum-short form (MHC-SF) in Setswana-speaking South Africans. *Clinical Psychology and Psychotherapy*, 15, 181–192. <https://doi.org/10.1002/cpp.572>
- Lamers, S. M. A., Westerhof, G. J., Bohlmeijer, E. J., ten Klooster, P. M., & Keyes, C. L. M. (2011). Evaluating the psychometric properties of the Mental Health Continuum-Short Form (MHC-SF). *Journal of Clinical Psychology*, 67, 99–110. <https://doi.org/10.1002/jclp.20741>
- M. K. (2013). The right kind of happy. *The Economist*, August 1. Retrieved from <http://www.economist.com/blogs/babbage/2013/08/psychosomatic-medicine>
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539–569. <https://doi.org/10.1146/annurev-psych-120710-100452>
- Ryff, C. D., & Singer, B. (2003) Ironies of the human condition: Health and well-being on the way to mortality. In L. G. Aspinwall & U. M. Staudinger (Eds.), *A psychology of human strengths: Fundamental questions and future directions for a positive psychology* (pp. 271–287). Washington, DC: American Psychological Association.

- Sartre, J.-P. (1956). *Being and nothingness: An essay on phenomenological ontology* (H. E. Barnes, Trans.). New York, NY: Philosophical Library.
- Sharon, D., Tilgner, H., Grubert, F., & Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology*, *31*, 1009–1014. <https://doi.org/10.1038/nbt.2705>
- Shedler, J., Mayman, M., & Manis, M. (1993) The illusion of mental health. *American Psychologist*, *48*, 1117–1131. <https://doi.org/10.1037/0003-066X.48.11.1117>
- Smith, E. E. (2013). Meaning is healthier than happiness *The Atlantic*, August 1. Retrieved from <https://www.theatlantic.com/health/archive/2013/08/meaning-is-healthier-than-happiness/278250>
- Tabachnik, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. New York, NY: Pearson.
- Taylor, S. E., & Brown, J. D. (1994). Positive illusions and well-being revisited: Separating fact from fiction. *Psychological Bulletin*, *116*, 21–27. <https://doi.org/10.1037/0033-2909.116.1.21>

Appendix—A note on Fredrickson et al.’s data set

Our task in reanalyzing Fredrickson et al.’s (2013a) study was complicated by the presence of a number of errors in the reporting and content of their data set. These errors are discussed in full in the published supporting information document (SI) that accompanied our published article. Here, I report the most serious of these errors, as it has consequences for Chapter 3 of the present thesis.

In their Supporting Information, Fredrickson et al. (2013b) stated that participants’ race/ethnicity was coded as “white” or “nonwhite”; indeed, the name of this variable in their data set is “white.” In the data set, the value of this variable was coded as 0 (for “non-white”) and 1 (for “white”), and the resulting binary categorical variable was used as a covariate in Fredrickson et al.’s (2013a, 2013b) regression models. However, the race/ethnicity of participant SOBC1-1299 was erroneously coded with the value 4. This is, of course, highly problematic at a statistical level: The presence in the data set of this value of 4 for one participant effectively transforms the variable “white” into a continuous variable with a real value. It seems likely that this error arose because the ethnicity of all the participants was originally coded using another scale; for example, Fredrickson et al.’s (2013a) Table 1 shows that, at some point during the study, participants were identified as being of four different ethnicities (White, Black, Hispanic, and Asian). Perhaps ethnicity was first entered into the data set with a different coding and later manually dichotomized as “nonwhite = 0”/“white = 1.” Whatever the origin of this discrepancy, however, its presence in the data set means that the use of “white” as a covariate variable is a source of error of undetermined (but potentially substantial) magnitude in Fredrickson et al.’s analyses.

More Questions Than Answers: Continued Critical Reanalysis of Fredrickson et al.'s Studies of Genomics and Well-being

Nicholas J. L. Brown

Douglas A. MacDonald¹

Manoj P. Samanta²

Harris L. Friedman³

James C. Coyne

PLoS ONE, 11(6), e0156415 (June 2016)⁴.

DOI: 10.1371/journal.pone.0156415

¹ University of Detroit Mercy

² Systemix Institute, Redmond, WA

³ Saybrook University

⁴ The published article was accompanied by a Supporting Information document, available at <https://doi.org/10.1371/journal.pone.0156415.s001>, which is not part of the present thesis.

Abstract

We critically re-examine Fredrickson et al.'s (2015) renewed claims concerning the differential relationship between hedonic and eudaimonic forms of well-being and gene expression, namely that people who experience a preponderance of eudaimonic well-being have gene expression profiles that are associated with more favorable health outcomes. By means of an extensive reanalysis of their data, we identify several discrepancies between what these authors claimed and what their data support; we further show that their different analysis models produce mutually contradictory results. We then show how Fredrickson et al.'s most recent article on this topic not only fails to adequately address our previously published concerns (Brown, MacDonald, Samanta, Friedman, & Coyne, 2014a) about their earlier related work (Fredrickson et al., 2013), but also introduces significant further problems, including inconsistency in their hypotheses. Additionally, we demonstrate that regardless of which statistical model is used to analyze their data, Fredrickson et al.'s method can be highly sensitive to the inclusion (or exclusion) of data from a single subject. We reiterate our previous conclusions, namely that there is no evidence that Fredrickson et al. have established a reliable empirical distinction between their two delineated forms of well-being, nor that eudaimonic well-being provides any overall health benefits over hedonic well-being.

Introduction

Fredrickson et al. (2013) claimed to have established that individuals who experience a preponderance of hedonic over eudaimonic well-being will have different “Conserved Transcriptional Response to Adversity” (CTRA) gene expression profiles compared to those who have the inverse preponderance. Fredrickson et al. reported that these CTRA profiles differed such that higher hedonic well-being was associated with increased expression of pro-inflammatory genes, typically seen in immune responses to bacterial infections, whereas higher eudaimonic well-being was associated with increased expression of genes involved in type I interferon antiviral responses and IgG1 antibody synthesis (henceforth: “antiviral” genes), typically seen in responses to viral threats. In a previous article (Brown, MacDonald, Samanta, Friedman, & Coyne, 2014a), we demonstrated that Fredrickson et al.’s (2013) study was flawed in several ways, including misidentification of hedonic and eudaimonic well-being factors, questionable data analytic methods based on the unjustified aggregation of non-significant regression coefficients, and an elementary but crucial error in the coding of their data set.

Recently, Fredrickson et al. (2015) published a follow-up article in which they claimed to have reproduced and extended the results from their initial study (Fredrickson et al., 2013), and to have refuted most of the criticisms made by us in our previous article (Brown et al., 2014a). However, we show here that this follow-up article is also flawed in several ways. First, it suffers from the same problems of scale validity as the first article. Second, Fredrickson et al.’s new data analysis model, when applied to their original data, gives results that are inconsistent with those from their previous model, contradicting their earlier conclusions; this new model also gives results that differ between their new sample and their original sample, contradicting the claim that their more recent study is a successful replication and extension of their earlier study. Third, it seems that there has been a substantial (but unacknowledged) shift in the hypotheses across the two articles (Fredrickson et al., 2013, 2015). Fourth, we show that Fredrickson et al. used two versions of the same data set at different points in their new article. Fifth, a recently-discovered issue demonstrates that both of Fredrickson et al.’s (2013, 2015) models are very sensitive to the presence of outliers in their data set. We conclude that these problems make Fredrickson et al.’s results uninterpretable.

Factor structure of the MHC-SF

We previously noted (Brown et al., 2014a) a number of serious flaws with the way in which Fredrickson et al. (2013) measured the crucial distinction between hedonic and eudaimonic well-being in their first article, including the very high intercorrelation between the two measures; the demonstration that the factor structure of the Mental Health Continuum-Short Form (MHC-SF) scale, both in previous independent studies and Fredrickson et al.'s own sample, did not accord with their interpretation of it; and the fact that the references cited by Fredrickson et al. (2013, 2015) as supporting their claims of a two-factor structure did not in reality do so.

We ran essentially the same analyses as before⁵ (Brown et al., 2014a) to test the factor structure of the MHC-SF using the confirmation sample ($N = 122$) from Fredrickson et al. (2015) to determine if problems similar to those that we had found and reported previously (Brown et al., 2014a, 2014b) were still present. Our results (presented in more detail in the Supporting Information) point to continued problems with the measurement of well-being using the MHC-SF; in particular, there is a lack of strong support for two factors that could be labeled “hedonic” and “eudaimonic” well-being. Thus, any assertions about the differential effect of hedonic and eudaimonic well-being on genomic variables based on this measure do not appear to be justified. Indeed, other research (Disabato, Goodman, Kashdan, Short, & Jarden, 2016) suggests that these two forms of well-being are, empirically, essentially indistinguishable from each other.

The mixed effect linear model

Fredrickson et al. (2015) were strongly critical of our reanalysis (Brown et al., 2014a) of the regression method used in their initial study (Fredrickson et al., 2013), which we named “RR53.” In that reanalysis, we demonstrated—by evaluating all possible “factor” pairs (that is, all possible ways of splitting the 14 MHC-SF items into two distinct and non-empty sets)—that the RR53 method would generate statistically significant results, apparently indicating a meaningful differential association of well-being with gene expression, for almost any “factors,” and that this effect was repeated—with only a slightly lower rate of spurious matches—when the MHC-SF item data were replaced by random numbers. Subsequently, other authors (Neuroskeptic, 2014; Barr, 2014) have suggested that the principal flaw in the RR53 regression method lies in its application of a t test to non-

⁵ The data and code needed to reproduce the analyses in this chapter are available at <https://osf.io/xjg53/>

independent data, namely the collected coefficients from multiple separate regressions of the expression of individual genes on measures of well-being. Fredrickson et al. (2015) described our systematic, exhaustive examination of the functioning of their method as “capitaliz[ing] on chance” (p. 11); however, despite their rejection of our criticism of their RR53 method, Fredrickson et al. replaced it with a mixed effect linear model to perform the majority of their analyses in their later article. They applied this model to several data sets. First, they analyzed new data (which they named the “confirmation sample”) collected from a new sample of 122 individuals following the methods and protocols described in their original study (Fredrickson et al., 2013). Second, they applied their new model to an aggregate (“pooled”) data set, which augmented the confirmation sample with the data set from their original study (the “discovery sample”). Third, they analyzed an additional data set (the “generalization sample”), apparently taken from a different study that used a different psychometric instrument (the Ryff Scales of Psychological Well-Being; Ryff, 1989).

The implementation of Fredrickson et al.’s (2015) new model in SAS, as applied to the confirmation and pooled samples, was extensively documented in Fredrickson et al.’s supplemental materials, so we had little difficulty in reproducing it in R, using the `gls()` command from the `nlme` package. Our reproduction produced results that were identical to those reported by Fredrickson et al. (2015, Tables 2 and 3), within the limits of rounding error. We therefore felt confident that our model was a faithful replication of the one built in SAS by Fredrickson et al. (As we were unable to reproduce the model used to analyze the generalization sample, we do not examine that sample further here.)

Fredrickson et al. (2015) did not report the results of the application of their new model to the discovery sample (i.e., the data set from their first study) on its own. Yet, if their aim was to show that their new study had confirmed the results of the first—for which they claimed support with their Figure 1C, showing that their original model produced “similar” results with their new sample—it would seem logical to demonstrate that their new model also produced comparable results when applied to the original sample. We therefore applied our replication of Fredrickson et al.’s (2015) model to the discovery study data (Fredrickson et al., 2013). Here, we found a very different pattern of effects of hedonic and eudaimonic well-being on gene expression, compared to the results obtained by Fredrickson et al. (2015) in their analysis of the confirmation sample. In the latter analysis, Fredrickson et al. noted that the association between hedonic well-being and gene expression was small and non-significant ($b = .085, p = .4829$), whereas the association for eudaimonic well-being and gene

expression was substantial and significant ($b = -.509, p < .0001$). However, our analyses of the discovery sample show precisely the opposite effects. That is, in the discovery sample, hedonic well-being was substantially and significantly associated with CTRA gene expression (bs from .490 to .568, ps from .0031 to .0009), whereas the association between eudaimonic well-being and gene expression was not significant (bs from .134 to .195, ps from .2588 to .4522). Thus, the application of the mixed effect linear model to the two individual samples (discovery and confirmation) gives results that are diametrically opposed to each other; yet, only one of these sets of results (for the confirmation study) was reported by Fredrickson et al. (2015), who instead chose to report the results of combining the discovery and confirmation samples into one aggregate (“pooled”) sample. Our analyses of the discovery sample using the mixed effect linear model are presented in detail, and contrasted with Fredrickson et al.’s analysis of the confirmation sample, in Table D in our Supporting Information.

Apparent change of hypotheses

In their first article on this topic, Fredrickson et al. (2013) consistently reported that hedonic well-being was associated with up-regulation of pro-inflammatory genes and down-regulation of antiviral genes, with the opposite associations applying to eudaimonic well-being. Almost every mention of either form of well-being (hedonic or eudaimonic) in that article was immediately counterbalanced by a mention of the other form, accompanied by a demonstration of an inverse relationship with CTRA gene expression. This symmetry was a recurring theme throughout the whole article.

In their Figure 1C at the start of their more recent article, Fredrickson et al. (2015) appeared to show that the application of their original regression model (Fredrickson et al., 2013) to their new data set (the “confirmation” sample) reproduced the results of their initial study, with up-regulation of pro-inflammatory CTRA genes being associated with hedonic well-being and up-regulation of antiviral genes being associated with eudaimonic well-being. However, the rest of this newer article appeared to espouse a revised version of their theory, in which there is a purported relationship between (only) eudaimonic well-being and CTRA gene expression, with no role for hedonic well-being. No reasons were provided for this change in theoretical direction; indeed, Fredrickson et al. appeared to suggest that no change had even taken place:

The present findings thus *converge with previous results* [emphasis added] in identifying eudaimonic well-being as the primary source of associations between

overall well-being and CTRA gene expression and provide no support for any independent favorable contribution from hedonic well-being. The consistency and robustness of these findings across 3 independent study samples also refutes claims that the initial discovery study findings were somehow spurious or unreplicable.

(Fredrickson et al., 2015, p. 9)

However, at no point in their earlier article did Fredrickson et al. (2013) assert that eudaimonic well-being was the “primary source of associations” between well-being and gene expression. Indeed, as noted above, that article treated hedonic and eudaimonic well-being as symmetrical and of equivalent importance, with the abstract stating, for example, that “hedonic and eudaimonic well-being showed . . . highly divergent transcriptome profiles” and that these two forms of well-being “engage distinct gene regulatory programs” (Fredrickson et al., 2013, p. 13684).

A careful reading of the paragraph quoted above suggests that what the reader is invited to understand here is that only eudaimonic well-being produces “favorable” changes in CTRA gene expression, presumably through the activation of (only) antiviral genes. This implies that the symmetric (and, presumably, “unfavorable”) association of the expression of pro-inflammatory genes with hedonic well-being, identified by Fredrickson et al. (2013) in their initial analyses of their discovery sample, should now be ignored. Such a change of interpretation would also require the reader to ignore our demonstration (above) that the application of Fredrickson et al.’s (2015) new analysis method (i.e., their mixed effect linear model) to the same discovery sample shows that only hedonic (and not eudaimonic) well-being had significant associations with gene expression in that sample. Thus, an examination of the complete available evidence suggests that Fredrickson et al.’s preferred conclusion is by no means the only one supported by it.

Multiple versions of the discovery sample data set

Fredrickson et al. (2015) used their Figure 1C to present the “[p]oint estimates of average association coefficients relating range-spanning variations in hedonic and eudaimonic well-being scores $[-2\text{ SD}, +2\text{ SD}]$ to unstandardized ($\log_2$ metric) gene expression values” (p. 5) in their discovery and confirmation studies. In other words, this figure represents the results of applying their original regression method from Fredrickson et al. (2013) to both their initial ($N = 76$) “discovery” data set and their new ($N = 122$) “confirmation” data set, and, at first glance, appears to show that these results are similar. However, a closer analysis reveals two substantial problems here.

First, as we have previously demonstrated (Brown et al., 2014a, 2014b), Fredrickson et al.'s (2013) original regression method finds large numbers of spurious associations with opposing signs for hedonic and eudaimonic well-being where no real effect exists, even when random numbers are used in place of their psychometric data. Thus, it is no surprise that the application of this method to their new data appears to produce similar results.

Second, Fredrickson et al.'s (2013) original analysis was greatly distorted by one participant (with identification code SOBC1-1299) having an incorrectly coded value of “4” for the variable “White” (a binary indicator of the participant's race). This error effectively transformed this categorical variable into a continuous variable, with dramatic consequences for the regressions in which it was used. When we re-ran our reproduction of Fredrickson et al.'s (2013) regression method after correcting the value of “White” for this participant from “4” to “0,” the fold-difference value for hedonic well-being changed from +8.0% to +3.6%. This elementary data set coding error—which we reported in some detail in our first critique (Brown et al., 2014a, 2014b)—ought to have been corrected before Fredrickson et al.'s (2015) Figure 1C was generated, but the y-axis of that figure shows that it was not (see also our Figure 3.1). However, it appears that Fredrickson et al. (2015) *did* correct this error when constructing their “pooled” data set used for the principal analyses (with their new mixed effect linear model) shown in their Table 3; our reproduction of their new model gives results that are closer to the numbers in their Table 3 when the data set coding error is indeed corrected. This means that Fredrickson et al. (2015) in effect used *two different versions* of the GSE45330 data set in their article; the first (with SOBC1-1299 “White” having the uncorrected value of “4”) was used to demonstrate that their original regression model produced similar results with both the discovery and confirmation samples (as shown in their Figure 1C), while the second (with SOBC1-1299 “White” corrected to “0”) was used as input to their Table 3, as part of the pooled study.

Figure 3.1. Extensions of Fredrickson et al.'s (2015) Figure 1C

Here, we have augmented Fredrickson et al.'s (2015) Figure 1C—from which the first two pairs of bars in our figure are directly taken—with two new pairs of bars. The third pair from the left, labeled “Discovery Study (corrected),” indicates the fold difference values that are produced by the application of Fredrickson et al.'s (2013) repeated-regression method to the discovery sample, after correction of the variable “White” for participant SOBC1-1299 from “4” to “0.” The rightmost pair, labeled “Discovery Study (SOBC1-1293 omitted),” indicates the fold difference values produced by the omission of participant SOBC1-1293 (cf. discussion in the text of the present article).

Sensitivity of analysis methods to individual data points

A preliminary analysis of any data set should always include an evaluation of the extent to which outliers may be driving the results (Osborne & Overbay, 2004), even if the subsequent question of whether to exclude a given data point or participant is often a subjective one. The number of independent variables in Fredrickson et al.'s (2013, 2015) models makes it difficult to determine visually (e.g., using a scatterplot) whether any individual participant's data might be having an undue influence on the overall outcome. We therefore ran these models multiple times, omitting one participant each time, to see whether this would make a disproportionate change to the results. We discovered that the omission of one participant (reference number SOBC1-1293, a male) caused several of the purported

associations between gene expression and both hedonic and eudaimonic well-being to *change sign*, and the remainder to undergo a substantial change in magnitude, whether Fredrickson et al.'s original RR53 regression method or their new mixed effect linear model was used. Specifically, in the case of the RR53 method, the fold difference value for hedonic well-being changes from +3.6% to -8.8% when this participant is excluded, while the value for eudaimonic well-being changes from -7.3% to +5.3%. In the case of the mixed effect linear model, the *b*-coefficients change from 0.536 to -0.120 (hedonic) and from 0.135 to 0.065 (eudaimonic). (See Table E in our Supporting Information for full details of this analysis.)

This improbable-sounding observation is a major problem for several reasons. First, and most obviously, the presence of any given individual in a study where participants were recruited from the community via flyers and e-mail is, essentially, a random event. Had participant SOBC1-1293 not enrolled in the initial study, Fredrickson et al. (2013) would presumably have reached diametrically opposite conclusions about the relative effects of hedonic and eudaimonic well-being on CTRA gene expression. Second, closer inspection shows that this participant's data were in many respects unusual, thus making him a good candidate for exclusion as an outlier; for example, he gave answers at the extreme ends of the scale ("0" or "5") for every item in the MHC-SF, and a substantial number of his gene expression values were near the extremes for the sample. Third, the large difference in the results of their analyses depending on whether this one participant is included or excluded—illustrated in the right-hand panel of Figure 3.1—demonstrates that both Fredrickson et al.'s RR53 regression procedure (2013) and their mixed effect linear model (2015) are highly sensitive to minor variations in the input data. It has recently been suggested that this sensitivity may be due to overfitting, caused by the combination of a large number of predictors and a small sample size (Walker, 2016).

Conclusion

We believe that Fredrickson et al. (2013, 2015) have not provided any substantial evidence to support their hypotheses regarding the relationship between well-being and gene expression, which also appear to have changed across their two articles. In their initial study, they reported that both hedonic and eudaimonic well-being were differentially and symmetrically associated with opposite forms of CTRA gene regulation. We showed previously (Brown et al., 2014a) that their theoretical approach, factor analysis, and regression methods suffered from a variety of problems, and we have demonstrated that the principal result in that initial article were heavily dependent on the data from a single

participant. In their latest article, Fredrickson et al. (2015) presented new evidence for what appear to be a different hypothesis (i.e., that *only* eudaimonic well-being is associated with gene expression), while not addressing many of the issues that we raised in our previous critique (Brown et al., 2014a), although several of those issues continue to apply to this latest article. They changed their statistical model to one which, when applied to their original data set, produced results that were in contradiction with both their conclusions from applying their original model and the results obtained when applying this new model to their new sample. As shown above, they also utilized two versions of the same data set within the same article. All of this methodological confusion makes it impossible to place any meaningful interpretation on their results. To put it as simply as possible: Fredrickson et al.'s (2013, 2015) two studies provide no clear and uncontestable evidence that higher levels of eudaimonic, as compared to hedonic, well-being is associated with healthier patterns of gene expression.

References

- Barr, D. J. (2014). *Anatomy of a statistical artifact: Eudaimonic well-being and genomics* [Weblog post]. Retrieved March 19, 2019 from <https://datahowler.wordpress.com/2014/09/09/anatomy-of-a-statistical-artifact>
- Brown, N. J. L., MacDonald, D. A., Samanta, M. P., Friedman, H. L., & Coyne, J. C. (2014a). A critical reanalysis of the relationship between genomics and well-being. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 12705–12709. <https://doi.org/10.1073/pnas.1407057111>
- Brown, N. J. L., MacDonald, D. A., Samanta, M. P., Friedman, H. L., & Coyne, J. C. (2014b). Supporting information. *Proceedings of the National Academy of Sciences of the United States of America*, *111*. Retrieved from <http://www.pnas.org/content/suppl/2014/08/21/1407057111.DCSupplemental>
- Disabato, D. J., Goodman, F. R., Kashdan, T. B., Short, J. L., & Jarden, A. (2016). Different types of well-being? A cross-cultural examination of hedonic and eudaimonic well-being. *Psychological Assessment*, *28*, 471–482. <https://doi.org/10.1037/pas0000209>
- Fredrickson, B. L., Grewen, K. M., Coffey, K. A., Algoe, S. B., Firestone, A. M., Arevalo, J. M. G., . . . Cole, S. W. (2013). A functional genomic perspective on human well-being. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 13684–13689. <https://doi.org/10.1073/pnas.1305419110>
- Fredrickson, B. L., Grewen, K. M., Algoe, S. B., Firestone, A. M., Arevalo, J. M. G., Ma, J., & Cole, S. W. (2015). Psychological well-being and the human conserved transcriptional response to adversity. *PLoS ONE*, *10*(3), e0121839. <https://doi.org/10.1371/journal.pone.0121839>
- Neuroskeptic. (2014). *(False?) Positive psychology meets genomics* [Weblog post]. Retrieved March 19, 2019 from <http://blogs.discovermagazine.com/neuroskeptic/2014/08/27/false-positive-psychology-genomics>
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, *9*, 1–12. Available: <http://PAREonline.net/getvn.asp?v=9&n=6>
- Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, *57*, 1069–1081. <https://doi.org/10.1037/0022-3514.57.6.1069>
- Walker, J. A. (2016). Monte Carlo simulation of OLS and linear mixed model inference of phenotypic effects on gene expression. *PeerJ*, *4*, e2575. <https://doi.org/10.7717/peerj.2575>

The Elusory Upward Spiral: A Reanalysis of Kok et al. (2013)

Nicholas J. L. Brown

James A. J. Heathers¹

James C. Coyne

Harris L. Friedman²

Psychological Science, 26, 1140–1143 (July 2015)³.

DOI: 10.1177/0956797615572908

¹ University of Sydney

² Saybrook University

³ This chapter is a slightly expanded version of the published article, which had to be kept short to fit the article format imposed by the journal (i.e., as a “commentary”) and did not have an abstract. The cuts that were made for publication affected my contribution more than that of James Heathers. In addition, James was on the job market at the time, and it seemed that a first-author paper in *Psychological Science* was of more importance to him than to me, so I offered him the first-author position for the published version.

Kok et al. (2013a) claimed to have demonstrated that practicing loving-kindness meditation (LKM) generates an “upward spiral” of mutual enhancement among positive emotions, social connectedness, and physical health, with the latter being indexed by vagal nerve tone (VNT). They concluded that “advice about how people might improve their physical health . . . can now be expanded to include self-generating positive emotions” (p. 1131). Close examination of Kok et al.’s study, however, suggests that this claim and conclusion are unwarranted for three reasons which we detail below.

No effect of experimental condition on VNT

Kok et al. (2013a) hypothesized that participants in the LKM group would show greater increases in positive emotions than those in the control group, and that this effect would be moderated by initial (Week 1) high-frequency heart-rate variability (HF-HRV). They further hypothesized that increased positive emotions would lead to increased perceptions of social connections, which in turn would lead to increased endpoint (Week 9) HF-HRV, a series of effects that they described as an “upward spiral.” These hypotheses were represented by what Kok et al. described as a “variant of a mediational, parallel-process, latent-curve model” (p. 1127), which can be considered a mediated moderation model with positive emotions and social connectedness as mediators. Experts in the analysis of such models disagree on whether a significant main effect of the predictor variable on the criterion variable must be demonstrated before testing for the significance of a hypothesized mediator of that effect (e.g., MacKinnon, 2008, pp. 6–11, 67–75). But any demonstration of the existence of an upward spiral, set in motion by LKM training, clearly requires that there be a significant increase in HF-HRV over time for the LKM group (i.e., a significant positive main effect of LKM on change in VNT). However, we found no mention of any data (such as a time $\times$ treatment interaction) supporting such a direct effect in Kok et al.’s article. Instead, these authors’ analyses focused almost entirely on the way in which such an improvement—if it existed—might be mediated using a latent growth-curve model¹. Crucially, however, it is entirely possible for statistically significant effects to be found in this type of mediation model even in the absence of an overall effect (MacKinnon, 2008, p. 9).

¹ Kok et al. (2013a, 2013b) indicated that the use of respiratory sinus arrhythmia—as opposed to HF-HRV—as a proxy for VNT produced a worse-fitting mediation model (i.e., one which did not predict the observed slopes in changes of positive emotions). However, these authors did not discuss why the model based on HF-HRV might be more likely to reflect any underlying psychophysiological reality.

Although Kok et al. (2013a) did not report a main effect of LKM on VNT in their article, in their supplemental material (2013b, p. 3) they did investigate whether experimental condition predicted change in HF-HRV between timepoints. To do so, they (a) dropped from their analysis any participant who did not have HF-HRV data at both Week 1 and Week 9 (reducing N from 65 to 52²), (b) applied a square-root transformation to the HF-HRV values at both timepoints, (c) standardized the resulting transformed values, (d) assessed change in HF-HRV by calculating the residuals from a regression of Week 9 HF-HRV on Week 1 HF-HRV, and (e) entered those residuals into a t test with experimental condition as the independent variable. This analysis yielded a near-significant main effect ($t(36) = -1.93$, $p = .06$, two-tailed), leading Kok et al. to conclude that LKM resulted in an increase in HF-HRV. In an alternative model, which they later discarded, Kok et al. (2013b, p. 9) obtained a similar result that just attained conventional statistical significance ($z = 1.97$, $p = .05$).

However, Kok et al. (2013b) failed to note that both of these purported main effects were partly due to an unexpected and unexplained negative change in HF-HRV in the control group of similar size to the positive change in the LKM group. The mean square-root transformed Week 1 and Week 9 HF-HRV values for the LKM group equaled 0.0350 and 0.0399, whereas for the control group the corresponding values equaled 0.0393 and 0.0345, respectively. When the t test is corrected for the negative change in HF-HRV for the control group, the main effect of experimental condition on change in HF-HRV is clearly not significant ($t = -0.89$, $df = 35$ with unequal variances, $p = .37$). Hence, LKM cannot be said to predict an increase in HF-HRV (or a concomitant improvement in physical health).

Furthermore, Kok et al.'s (2013b) choice of a square-root transformation of HF-HRV to correct for skewness is problematic. HF-HRV values typically exhibit substantial skew, which is nearly always handled with a logarithmic (rather than a square-root) transformation (Ellis, Sollers, Edelstein, & Thayer, 2008; Kuo et al., 1999). The quantile–quantile plots of the application of these two transformations to Kok et al.'s data, shown in Figure 4.1, clearly demonstrate the superiority of the logarithmic transformation in this case. However, when Kok et al.'s (2013b, p. 3) analysis is repeated using the more appropriate logarithmically-transformed HF-HRV data, the result is again clearly non-significant ($t = 1.06$, $df = 43$ with unequal variances, $p = .30$).

² Arguably, two participants (#557004 and #557033) with biologically impossible Week 9 HRV values of 0 and 0.54 ms²/Hz, respectively, should also have been removed.

Figure 4.1. Raw (top), square-root-transformed (middle), and logarithmically-transformed (bottom) HF-HRV data at the start (left) and end (right) of Kok et al.'s (2013a) study.

Use of a non-validated surrogate for physical health

Kok et al. (2013a) relied on HF-HRV as a proxy for VNT, which they in turn interpreted as a surrogate measure for physical health. However, many researchers have cautioned against measuring VNT as an index of physical health because of its complexity and ambiguous clinical significance (e.g., Berntson, Cacioppo, & Grossman, 2007). At least three elements cast doubt on the validity of using VNT (whether measured with HF-HRV or any other proxy) in this role. First, it conflates tonic VNT with the *modulation* of VNT by respiratory sinus arrhythmia (RSA) (e.g., Hedman, Hartikainen, Tahvanainen, & Hakumäki, 1995). Second, increased VNT is far from consistently correlated with “improved physical health;” on an experimental level, an array of unconstrained methodological changes can affect measurement (e.g., Grossman & Taylor, 2007). Third, changes in HF-HRV have been shown to be related to the intensity—rather than the valence—of affect (Frazier, Strauss, & Steinhauer, 2004). Fourth, high HF-HRV has recently shown a strong curvilinear relationship with depression (Kogan, Gruber, Shallcross, Ford, & Mauss, 2013). Fifth, meditative traditions, including LKM, very often include components focused on altering the only aspect of human behavior that exhibits volitional control over the autonomic nervous system, namely breathing. Consequently, meditative or relaxation exercises alter breathing speed, regularity, and tidal volume, all of which are capable of distorting measures of baseline VNT (Brown, Beightol, Koh, & Eckberg, 1993). It is well established that RSA drastically increases when breathing slows below around 0.15Hz, regardless of the level of autonomic regulation (Angelone & Coulter, 1964; Hirsch & Bishop, 1981). This is the most likely explanation for the presence in Kok et al.’s sample of several participants with implausibly high (above 10,000 ms²/Hz) HF-HRV values; such participants would be in the 90th percentile for unfatigued elite athletes during supine rest (Schmitt et al., 2013) if their respiration was controlled. Moreover, one participant recorded a baseline HF-HRV value of 239 ms²/Hz before an endpoint value of 10,600 ms²/Hz; this enormous increase is almost certainly evidence of an altered respiratory pattern, rather than “increased vagal tone,” across the study. Note that the problem of slower respiratory frequencies distorting baseline HF-HRV has been observed, in studies not involving meditation, when persons alter their breathing after being told to relax during baseline measurement (Quintana & Heathers, 2014). Although there is some debate over whether respiratory monitoring is necessary during HF-HRV measurement (Denver, Reed, & Porges, 2007), choosing not to monitor respiration assumes that it will remain within normative frequencies without distortion. In Kok et al.’s (2013) study, however, it seems very likely that this was not the case.

No measure of participant practice of LKM was included

Kok et al. (2013a, p. 1126) indicated that participants reported the number of minutes they spent in “meditation, prayer, or solo spiritual activity.” However, this information, which would seem to be essential for an understanding of how these activities might influence the biological and psychological variables under study, does not appear anywhere in the article or the Supplemental Material, and was not included in the data set the authors provided to us. It is unclear how Kok et al. reached their conclusions about the impact of meditation practice when no measure of the extent of that practice was included in their analysis. In effect, the only independent variable in Kok et al.’s (2013a) study is the assignment of experimental condition to each participant.

Almost half of the participants in the intervention group got worse

In their Figures 4 and 5, which between them took up an amount of space equivalent to around 20% of that which was allocated to the text of their article, Kok et al. (2013a) appeared to show impressive differences between the control and intervention groups, with the members of the former (latter) being in the lower (upper) quartiles of the distribution of the slopes of change of positive emotions versus social connections (Figure 4) or social connections versus VNT (Figure 5). However, these figures—which were not otherwise commented on in the text—are not the only demonstration of the individual-level data that could have been included in their article. In our Figure 4.2, we show a histogram of the difference in the HRV values for the 26 participants in the intervention group who had usable VNT data at the start and end of the study. It can be seen that, first, almost as many participants (12) ended the study with lower (i.e., “worse,” in terms of Kok et al.’s model of VNT as an index of physical health) HRV values than those (14) who ended it with higher values, and, second, that two outliers among the 14 “improvers” are likely to have been driving any overall average improvement that might have been observed.

Figure 4.2. Histogram of Participants in the Intervention Group of Kok et al.'s (2013a) Study Showing Their Change in HRV Values Across the Study.

Conclusion

It is imperative that extraordinary scientific claims be supported with solid evidence, especially when they carry health-related messages that are likely to be widely reported by the popular media (e.g., “Think Yourself Well,” 2012; “Don’t Bother With the Gym Today,” 2013). Close examination of Kok et al.’s (2013a) article, supplementary material, and data set reveals that this evidence is missing because (a) Kok et al.’s claim to have found a significant positive main effect of LKM on change in VNT was based on a misinterpretation of the data; (b) the validity of using VNT as an objective proxy for physical health, and of measuring VNT as HF-HRV, is questionable; and (c) the extent to which participants in the LKM group actually practiced what they had been taught in their LKM training was not considered. Without a significant positive main effect of LKM on change in HF-HRV, Kok et al.’s principal claim that positive emotions enhance physical health, and their bold assertion in the title of their article—that perceived positive social connections account for an upward spiral between positive emotions and VNT—are unsupported by our reanalysis: There appears to be no upward spiral to be explained.

References

- Angelone, A., & Coulter, N. A., Jr. (1964). Respiratory sinus arrhythmia: A frequency dependent phenomenon. *Journal of Applied Physiology*, *19*, 479–482.
- Berntson, G. G., Cacioppo, J. T., & Grossman, P. (2007). Whither vagal tone. *Biological Psychology*, *74*, 295–300. <https://doi.org/10.1016/j.biopsycho.2006.08.006>
- Brown, T. E., Beightol, L. A., Koh, J., & Eckberg, D. L. (1993). Important influence of respiration on human R–R interval power spectra is largely ignored. *Journal of Applied Physiology*, *75*, 2310–2317.
- Denver, J. W., Reed, S. F., & Porges, S. W. (2007). Methodological issues in the quantification of respiratory sinus arrhythmia. *Biological Psychology*, *74*, 286–294. <https://doi.org/10.1016/j.biopsycho.2005.09.00>
- Don't bother with the gym today: Positive outlook on life and making friends could be as good for you as diet and exercise (2013, May 11). *Daily Mail*. Retrieved from <http://www.dailymail.co.uk/health/article-2323054/Dont-bother-gym-today-Positive-outlook-life-making-friends-good-diet-exercise.html>
- Ellis, R. J., Sollers, J. J., III, Edelstein, E. A., & Thayer, J. F. (2008, April). *Data transforms for spectral analyses of heart rate variability*. Paper presented at Rocky Mountain Bioengineering Symposium & International ISA Biomedical Sciences Instrumentation Symposium, Copper Mountain, CO.
- Frazier, T. W., Strauss, M. E., & Steinhauer, S. R. (2004). Respiratory sinus arrhythmia as an index of emotional response in young adults. *Psychophysiology*, *41*, 75–83. <https://doi.org/10.1046/j.1469-8986.2003.00131.x>
- Grossman, P., & Taylor, E. W. (2007). Toward understanding respiratory sinus arrhythmia: Relations to cardiac vagal tone, evolution and biobehavioral functions. *Biological Psychology*, *74*, 263–285. <https://doi.org/10.1016/j.biopsycho.2005.11.014>
- Hedman, A. E., Hartikainen, J. E. K., Tahvanainen, K. U. O., & Hakumäki, M. O. K. (1995). The high frequency component of heart rate variability reflects cardiac parasympathetic modulation rather than parasympathetic ‘tone’. *Acta Physiologica Scandinavica*, *155*, 267–273. <https://doi.org/10.1111/j.1748-1716.1995.tb09973.x>
- Hirsch, J. A., & Bishop, B. (1981). Respiratory sinus arrhythmia in humans: How breathing pattern modulates heart rate. *American Journal of Physiology—Heart and Circulatory Physiology*, *241*, H620–H629.
- Kogan, A., Gruber, J., Shallcross, A. J., Ford, B. Q., & Mauss, I. B. (2013). Too much of a good thing? Cardiac vagal tone’s nonlinear relationship with well-being. *Emotion*, *13*, 599–604. <https://doi.org/10.1037/a0032725>
- Kok, B. E., Coffey, K. A., Cohn, M. A., Catalino, L. I., Vacharkulksemsuk, T., Algoe, S. B., . . . Fredrickson, B. L. (2013a). How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone. *Psychological Science*, *24*, 1123–1132. <https://doi.org/10.1177/0956797612470827>

- Kok, B. E., Coffey, K. A., Cohn, M. A., Catalino, L. I., Vacharkulksemsuk, T., Algoe, S. B., . . . Fredrickson, B. L. (2013b). *Additional analyses for supplemental online material to accompany Kok et al.'s "How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone."* Retrieved from http://pss.sagepub.com/content/suppl/2013/05/06/0956797612470827.DC1/DS_10.1177_0956797612470827.pdf
- Kuo, T. B. J., Lin, T., Yang, C. C. H., Li, C.-L., Chen, C.-F., & Chou, P. (1999). Effect of aging on gender differences in neural control of heart rate. *American Journal of Physiology—Heart and Circulatory Physiology*, 277, H2233–H2239.
- MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Erlbaum.
- Quintana, D. S., & Heathers, J. A. J. (2014). Considerations in the assessment of heart rate variability in biobehavioral research. *Frontiers in Psychology*, 5, 805. <https://doi.org/10.3389/fpsyg.2014.00805>
- Schmitt, L., Regnard, J., Desmarests, M., Mauny, F., Mourot, L., Fouillot, J.-P., . . . Millet, G. (2013). Fatigue shifts and scatters heart rate variability in elite endurance athletes. *PLoS ONE*, 8(8), e71588. <https://doi.org/10.1371/journal.pone.0071588>
- Think yourself well: You can. But it helps to think well of yourself in the first place. (2012, December 8). *The Economist*. Retrieved from <http://www.economist.com/news/science-and-technology/21567876-you-can-it-helps-think-well-yourself-first-place-think-yourself>

Emodiversity: Robust Predictor of Outcomes or Statistical Artifact?

Nicholas J. L. Brown

James C. Coyne

Journal of Experimental Psychology: General, 146, 1372–1377 (September 2017)¹.

DOI: 10.1037/xge0000330

¹ This chapter combines our published article from *Journal of Experimental Psychology: General* with the associated supplemental information document (<https://doi.org/10.1037/xge0000330.supp>). For the purposes of submission to the journal, we moved about half of our manuscript to the supplement in order to keep the main article's word count down to what we considered to be a reasonable level for publication; however, the supplement was included in the manuscript that was sent out for peer review.

Abstract

This article examines the concept of emodiversity, put forward by Quoidbach et al. (2014) as a novel source of information about “the health of the human emotional ecosystem” (p. 2057). Quoidbach et al. drew an analogy between emodiversity as a desirable property of a person’s emotional make-up and biological diversity as a desirable property of an ecosystem. They claimed that emodiversity was an independent predictor of better mental and physical health outcomes in two large-scale studies. Here, we show that Quoidbach et al.’s construct of emodiversity suffers from several theoretical and practical deficiencies, which make these authors’ use of Shannon’s (1948) entropy formula to measure emodiversity highly questionable. Our reanalysis of Quoidbach et al.’s two studies shows that the apparently substantial effects that these authors reported are likely due to a failure to conduct appropriate hierarchical regression in one case, and to suppression effects in the other. It appears that Quoidbach et al.’s claims about emodiversity may reduce to little more than a set of computational and statistical artifacts.

Quoidbach et al. (2014) presented a novel construct, *emodiversity*, to represent the degree to which an individual experiences a wide range of emotions, analogous to the idea of biodiversity in the natural environment. Emodiversity can be conceptualized either on separate axes of positive and negative emodiversity, or as a single overall “global emodiversity.” In two large studies, Quoidbach et al. claimed to have found associations between emodiversity and health outcomes, independent of the associations between the corresponding positive or negative emotions and those outcomes: “Emodiversity is a practically important and previously unidentified metric for assessing the health of the human emotional ecosystem” (Quoidbach et al., 2014, p. 2057).

Quoidbach et al. (2014) operationalized emodiversity in terms of a concept from the field of information theory known as *Shannon entropy* (Shannon, 1948). They claimed that this concept, originally devised by Shannon to represent the information content of a message in a communication system, also “quantifies the number of species and the evenness of species in a biological ecosystem” (p. 2058). In support of this claim, they cited Magurran (2004), who actually cautioned that “most commentators who discuss the relative merits of the various methods of measuring diversity go out of their way to underline the disadvantages of the Shannon index” (Magurran, 2004, p. 106). Leaving aside that debate, however, the validity of Quoidbach et al.’s concept clearly depends on the degree to which it faithfully implements Shannon entropy; its ability to generate reliable, meaningful, and measureable variance across participants; and the behavior of emodiversity when applied to real data. The present article examines whether emodiversity lives up to these requirements.

This article is structured as follows. First, we examine the theoretical underpinnings of emodiversity, especially the applicability of Shannon entropy to the specific context of a multi-item measure of emotions using Likert-type responses. Second, we reanalyze Quoidbach et al.’s (2014) empirical findings to identify where the purported evidence for their remarkable claims about the health benefits of emodiversity might have come from. Finally, we briefly discuss the broader lessons to be learned from this case.

Theoretical issues

Limitations of the analogy with biodiversity

Biodiversity is defined in terms of the *richness* and *evenness* of the variety of species within an ecosystem. Richness is the number of distinct species to be found in a given sample, regardless of how many examples (provided that the number is greater than zero) of that species are detected, while evenness is the degree to which the populations of each

species (or, in some definitions, the corresponding biomass) are similar. In Quoidbach et al.'s (2014) definition of emodiversity, the equivalent of biological richness is the number of different emotions experienced, while the equivalent of evenness is the extent to which the frequency with which a person reports experiencing each emotion is similar across all of the measured emotions. However, both of these dimensions of emodiversity are subject to limitations that are not present in Shannon's (1948) original model or in the Shannon-Wiener index of biodiversity. These limitations severely impair the correspondence between emodiversity and Shannon entropy.

Richness. In a biodiversity setting, the richness of a community is usually unbounded. Certainly, it would be unusual for richness to be subjected to an a priori upper limit imposed by the designer of a field study. With emodiversity, however, the degree to which the richness of participants' emotional experience can be captured is limited by the number of items making up the measure being used. For example, in their Study 1, Quoidbach et al. (2014) used a modified version of the Differential Emotion Scale (mDES), asking participants how often they experienced each of nine positive and nine negative emotions in a recent period. In contrast, the PANAS-X (Watson & Clark, 1994) measures 30 different positive emotions (and 30 negative ones); for example, this scale allows participants to provide distinct reports of the extent to which they feel *calm*, *relaxed*, and *at ease*. Leaving aside for a moment questions of the meaning of taxonomies of emotions (e.g., Kristjánsson, 2003; Solomon, 2002) and the possible discrepancies—whether caused by demand characteristics, social desirability concerns, faulty recollection, or some other issue—between people's actual emotional experiences and how they describe them, especially if they are asked about experiences in the past rather than the present (Robinson & Clore, 2002), this would seem to imply that emodiversity should be measured using a scale that allows the greatest possible number of emotions to be reported. By analogy, sending a field biologist out to report the number of (only) rabbits, mice, rats, voles, and beavers in a given area, while ignoring foxes or wolves because there was no corresponding space on the form, might lead to suboptimal decisions about conservation policy. However, Quoidbach et al. did not address this question, although it appears to be of crucial theoretical importance; in their two empirical studies, they used measures with only nine and 10 positive and negative emotions, respectively.

Evenness. When the Shannon-Wiener index is used in field biology, the count associated with each species is precisely the number of examples of that species observed in the community. However, a self-report by participants in a psychological study of the extent

to which they experienced various emotions will not typically be expressed in terms of a count of discrete occasions. A question such as “How many times did you experience contentment in the last week?” is almost meaningless; for one thing, the existence of 20 distinct occasions of contentment implies 20 transitions from a state of non-contentment, so that somebody who remained contented the whole time would have a very low score. Like most researchers asking about past emotional states, Quoidbach et al. (2014) used measures that asked participants to report the relative frequency with which they experienced each emotion on a Likert-type scale, with the following possible responses: 0: *Never*; 1: *Rarely*; 2: *Some of the time*; 3: *Often*; 4: *Most of the time*². But this decision imposes severe limitations on the potentially measurable (un)evenness of emotions (to go with the limitations noted above on the measurement of richness), because the range of possible values is so small. It is equivalent to a field biologist who uses the Shannon-Wiener index to measure biodiversity reporting the numbers of each species that were observed as *None*, *A few*, *Some*, *Quite a lot*, and *Very many*, with these groupings subsequently being coded 0–4 and the resulting numbers injected into Shannon’s formula. Such an operation would result in the loss of a great deal of information compared to the recording of the exact numbers of each species that were observed.

Could this problem be alleviated by asking participants to report the frequency with which they experienced each emotion on a wider numerical scale, perhaps in the range 0–100? This would seem to allow the relative frequency of emotional experiences to be described with greater (albeit still bounded) precision, thus allowing for more differentiation of the “evenness” component of emodiversity. However, it is unclear whether many individuals are sufficiently aware of their emotional experiences that they could meaningfully assign values of, say, 34 to one and 37 to another. In a classic paper, Miller (1956) described the general problem of people’s limited ability to distinguish between more than about seven levels of a unidimensional construct. Experience from several fields with these kinds of numerical-rating scales (e.g., Berbaum, Dorfman, Franken, & Caldwell, 2002; DeSoto, 2016; Mickes, Wixted, & Wais, 2007) suggests that the most responses will likely be multiples of 10, with a tendency to cluster further around 0, 50, and 100. The problem for emodiversity is

² In fact, Quoidbach et al. (2014) did not report the exact responses corresponding to the values 1, 2, and 3. We have assumed that the version of the mDES that these authors used in Study 1 was similar to that described by Fredrickson (2013), from which the responses shown here are taken, and that the responses for the unspecified 10-item measure used in Study 2 were labeled similarly.

that if each respondent only uses, say, five different numbers from the range of 0–100 to describe their frequencies of emotional experience, then the result is mathematically equivalent to the case where they are constrained to choose from a range of 0–4, and this holds even if every participant chooses a different set of five numbers. Perhaps a better way to measure the range and variety of a person’s emotional experiences, suggested by a reviewer of the present article, might be to use some form of diary, with sampling of emotions at either regular or random intervals (e.g., the Experience Sampling Method; Csikszentmihalyi & Larson, 1987); provided that a sufficient number of “species” of emotion could be defined, this could potentially result in measures of evenness that are more mathematically meaningful.

Numerical effects of bounded richness and evenness

Quoidbach et al.’s (2014) failure to correctly implement the concept of Shannon entropy, described in the preceding section, has immediate and severe consequences for the numerical behavior of the measurement of emodiversity. We enumerated all possible combinations of zero (*Never*) and non-zero responses for a variety of possible numbers of scale items and response formats (0–4 and 0–100). The results, shown in detail in our Supplemental Information, demonstrate that unless a participant responds that they *Never* experience about two-thirds of the emotions being measured—a contingency that would probably raise questions about the validity of the instrument being used—the ratio between the lowest and highest possible emodiversity values is very small (about 1.1:1 for a scale having a 0–4 response format and 1.5:1 for one with a 0–100 response format). Furthermore, even this range of values is only possible under rather implausible circumstances, because the highest and lowest scores for emodiversity are obtained when participants exhibit highly unusual response patterns (all identical responses for the maximum score; alternating extreme responses, or large numbers of *Never* responses, for the minimum score). Normally, however, psychologists do not say that individuals who display such response patterns “have high [or low] emotional diversity.” Rather, they say that such people “are not paying attention,” and typically exclude these participants’ data altogether. Hence, the practical range of emodiversity values from any given measure of emotions is even less than the mathematical limits would suggest. To summarize, we believe that emodiversity—as defined by Quoidbach et al. using the Shannon entropy formula and implemented using short emotion measures with limited ranges of responses—is unlikely to provide any meaningful amount of variance, independent of the underlying emotion measure, to be explained empirically.

Abundance is not measured

Quoidbach et al. (2014) defined people's emodiversity as "the variety and relative abundance of the emotions they experience" (p. 2057). However, it is not clear that the formula these authors used to calculate emodiversity fully corresponds to this definition. The formula generates the highest possible value of emodiversity when all of the responses are equal and non-zero: "If *all* the emotions of the list were evenly experienced, then emodiversity would be maximal" (Quoidbach et al., 2014, p. 2059; emphasis in original). However, this maximum value of emodiversity is attained whatever this non-zero value might be (1, 2, 3, or 4). This means that Alice, who experiences all nine positive emotions (alertness, amusement, etc.) only *Rarely*, has exactly the same emodiversity as Bob, who experiences all of these emotions *Most of the time*, and higher emodiversity than Carol, who experiences four of the nine emotions *Some of the time* and five of them *Often*. Furthermore, any increase in Alice's experience of one of these emotions will result in an immediate *reduction* in her emodiversity. Given Quoidbach et al.'s claims about the health benefits of higher emodiversity, this would appear to be a good reason for Alice not to attempt to increase her currently minimal frequency of, say, amusement to *Some of the time* by watching comedy shows until she is sure that she can simultaneously increase the frequency with which she experiences the other positive emotions to the same extent.

Thus, once a minimum level of frequency of experiencing emotions has been established, there can be no "benefit" (of increased emodiversity) in increasing the frequency with which one experiences any particular emotion, unless this specific emotion is in some way "lagging" behind the others (for example, if one currently experiences eight emotions *Often* and one only *Some of the time*). It is therefore not clear where the "relative abundance" component of Quoidbach et al.'s (2014) definition of emodiversity is to be found, given that a person experiencing all positive emotions only *Rarely* already has the highest possible emodiversity score. This result also suggests that any difference in emodiversity between two participants who report *Never* experiencing the same number of emotions is very likely to be nothing more than noise.

Empirical Issues

In the first part of this article, we have shown that Quoidbach et al. (2014)'s emodiversity is merely a pastiche of Shannon entropy, with a number of conceptual lacunae that make the application of Shannon's (1948) formula invalid. (For reasons of space, we have omitted some other important problems, such as the apparent requirement that

participants be forced to provide responses to all items on the emotion measures being used, or the questionable validity of applying the arithmetic of Shannon's formula to numbers that are merely categorical labels on participants' reports of the frequency with which emotions were experienced; these and other issues are covered in detail in our Supplemental Information.) In fact, we believe that any observed variance in emodiversity is likely to be little more than statistical noise. In view of this, it is necessary to explain how Quoidbach et al. (2014) concluded that emodiversity independently predicted depression (Study 1) or an assortment of physical health outcomes (Study 2) to the substantial degree that they reported in their article. In this part, therefore, we examine how these results might have come about, if this were not due to the claimed predictive power of emodiversity.

Study 1

Dr. Quoidbach (personal communication, November 15, 2015) kindly provided us with the data set for Study 1, which examined the relation between emodiversity and depressive symptoms, while controlling for emotions, in a sample of 35,844 participants who were recruited via a television show. Using SPSS, we were able to reproduce perfectly all of the results reported by Quoidbach et al. (2014) in their article. However, we conducted the majority of our reanalyses in R; our code is available at <https://osf.io/vu4uq/>.

Quoidbach et al. (2014) claimed that the results of their regressions showed that emodiversity substantially and significantly predicted their main outcome variable (depression), over and above the effect of positive or negative emotions themselves. For example, when positive emotion, positive emodiversity, and their interaction were entered into a regression, the reported standardized regression coefficients of these three terms were, respectively, $-.40$, $-.36$, and $-.14^3$. However, simply reporting these coefficients (and their corresponding partial r s) does not demonstrate a substantial effect of emodiversity. First, as our Table 5.1 shows, when positive emodiversity and the interaction term are entered into a hierarchical regression after first entering positive emotion, there is almost no increase in the variance explained at each step; R^2 increases from $.381$ (positive emotion only) to $.394$

³ When we reproduced this regression in SPSS, we obtained the same results reported by Quoidbach et al. (2014). However, when we performed the same operations in R, the standardized regression coefficient for the interaction was $-.094$, although the other two coefficients were identical. Subsequent investigations, after publication of our article in *JEP:G*, revealed that the issue was probably being caused by Quoidbach et al. (2014) having standardized the interaction term independently of its components, which is a meaningless operation.

(addition of positive emodiversity) to .397 (addition of the interaction term). That is, the addition of emodiversity and the interaction term make very little difference to the explanatory power of the model, with just 1.3% of extra variance explained by emodiversity and another 0.3% explained by the interaction. (For negative emodiversity, the amount of additional variance explained is essentially zero, as explained in the section entitled “The effect of adding quadratic terms to the regression models” in our Supplemental Information.) Second, when the interaction term is added, the variance inflation factors (VIFs) of all three terms become quite large, further suggesting that multicollinearity is playing a major role (as could be expected from the correlation of .75 between positive emotion and positive emodiversity). Third, when the interaction term is entered, the standardized regression coefficient (β) for emodiversity more than doubles in magnitude, from $-.175$ to $-.364$, suggesting that substantial confounding effects are emerging at this point.

Table 5.1. Hierarchical linear models of positive, negative, and global emotion and emodiversity predicting depression from Quidbach et al.'s (2014) Study 1.

Model number	Variables	Model R^2	β	SE	VIF
P.1		.381			
	Positive Emotion		-.617	0.017	1.000
P.2		.394			
	Positive Emotion		-.485	0.026	2.298
	Positive Emodiversity		-.175	0.026	2.298
P.3		.397			
	Positive Emotion		-.396	0.010	5.939
	Positive Emodiversity		-.364 ^c	0.018	18.484
	Interaction		-.143 ^r	0.013	9.288
N.1		.460			
	Negative Emotion		.678	0.004	1.000
N.2		.460			
	Negative Emotion		.682 ^s	0.007	2.685
	Negative Emodiversity		-.005 ^s	0.007	2.685
N.3		.466			
	Negative Emotion		.863 ^s	0.011	8.149
	Negative Emodiversity		-.196 ^s	0.012	8.748
	Interaction		-.143	0.007	3.409
G.1		.592			
	Positive Emotion		-.402	0.004	1.219
	Negative Emotion		.507	0.004	1.219
G.2		.595			
	Positive Emotion		-.356	0.005	2.061
	Negative Emotion		.553 ^c	0.005	2.047
	Global Emodiversity		-.070 ^s	0.005	1.961

Notes:

^s : Denotes a coefficient that has either a greater magnitude or a different sign compared to the corresponding zero-order correlation, indicating that suppression has occurred.

^c : Denotes a coefficient that has increased in magnitude from the previous step, indicating that some form of confounding has occurred.

^r : The results in this table were calculated with SPSS. When we performed the same regression in R, we obtained a value of $-.094$ ($SE = 0.008$) for this coefficient; see footnote 3 for a possible explanation of the discrepancy. Only one of the other values in this table differed by more than .002 between SPSS and R, namely the coefficient for the interaction in section N.3, where R reported a value of $-.137$, a difference of 0.006.

Our Table 5.1 also shows the similar results that are obtained when hierarchical techniques are used to explore the other regression analyses in Quoidbach et al.'s (2014) Study 1. For negative emotion and emodiversity, R^2 remains static (within the limitations of rounding) at .460 when emodiversity is added to the basic regression of depressive symptoms on negative emotion, and increases only to .466 when the interaction term is added; furthermore, similar patterns of high VIFs and suppression values are observed as for positive emotion and emodiversity. The addition of global emodiversity to a model predicting depression from positive and negative emotion results in a negligible increase in R^2 (from .592 to .595), and further suppression effects can be observed. Taken together, these results suggest that emodiversity has little independent contribution to make to the prediction of depressive symptoms, over and above the well-established role of positive and negative emotions in that regard.

Study 2

In their Study 2, using a nationally representative sample of 10,000 Belgian adults, Quoidbach et al. (2014) claimed that both positive and negative emodiversity, when entered into a multiple regression together with the corresponding emotion scores and the interaction between these two variables, predicted health outcomes such as doctor's visits, days spent in hospital, and medical expenditure better than factors whose associations with health are well-established, such as diet, exercise, and smoking. If these claims were to be verified, the implications for public health policy would be substantial. In principle, a simple questionnaire asking people how often they experienced particular emotions in the past week would potentially provide better information about the population's health status for a whole year than complex measures of actual behavior.

Unfortunately, the data for Quoidbach et al.'s (2014) Study 2 are not currently publicly available. As a result, we have been unable to determine whether the same issues concerning the lack of additional predicted variance when emodiversity is added to the regression model, seen in Study 1, are also present in Study 2. However, even without access to the data, it can be readily shown that many of the results in Study 2 are the result of statistical suppression effects. For example, when Quoidbach et al. entered positive emodiversity, mean positive emotion, and their interaction into a regression predicting doctor's visits, they reported obtaining a standardized regression coefficient (β) for positive emodiversity of $-.29$. But their Table 2 shows that the zero-order correlation between positive emodiversity and doctor's visits was just $-.05$. The presence, in multiple regression

results, of a β coefficient of greater magnitude (or different sign) to the zero-order correlation between the same variables indicates that suppression has occurred. Suppression is often seen in regression models when two predictors are correlated, with at least one of them having a correlation with the outcome variable that is either zero (a situation referred to as “classical” suppression) or only relatively modest in magnitude compared to the correlation of the other predictor (“negative suppression”). However, suppression will also (always) occur when two predictors are positively correlated with each other—even to a very small extent—while having correlations of opposite sign with the outcome (“reciprocal suppression”).

Inspection of Quoidbach et al.’s (2014) Table 3 suggests that suppression effects are behind the great majority of the results from Study 2. Of the 18 zero-order correlations between positive, negative, and global emodiversity and Quoidbach et al.’s list of six health outcomes, only two (positive emodiversity correlated with doctor’s costs, and negative emodiversity correlated with hospital costs, both at $r = -.07$) are large enough to be statistically significant at the .05 level, even with the sample size of 1,273; this suggests that there is uncertainty about the magnitude, and even the sign, of some of these correlations. However, when these data were used in a regression model, Quoidbach et al. reported standardized regression coefficients (β) of greater magnitude than the zero-order correlations in all 18 cases, with 16 of these being statistically significant at the .05 level. Each of these coefficients appears to be the product of suppression effects. For positive emodiversity, these effects mostly take the form of negative suppression, resulting from the high correlation between positive emotion and emodiversity, combined with the fact that the zero-order correlations between positive emodiversity and the outcome variable are smaller than those between positive emotion and the outcome variable⁴. In contrast, for negative emodiversity the suppression effects mostly take the form of reciprocal suppression, caused by the positive correlations between negative emotion and emodiversity and between negative emotion and the outcome, combined with the negative correlation between negative emodiversity and the outcome (Conger, 1974).

In the absence of a solid theoretical explanation, results based on suppression are typically uninterpretable. Paulhus, Robins, Trzesniewski, and Tracy (2004) presented what

⁴ Specifically, for two predictors X_1 and X_2 and an outcome Y , suppression will occur if the pattern of correlations between these three variables is such that $r_{YX_1}/r_{YX_2} < r_{X_1X_2}$, where r_{AB} is the zero-order correlation between any two variables A and B , and the identifiers X_1 and X_2 are assigned so that $r_{YX_1} < r_{YX_2}$.

they claimed were two reproducible examples of suppression situations in psychology, but there do not seem to be many major effects in the psychological literature that are consistently and reliably explained in terms of theoretically-justified suppression. In particular, the existence of classical, negative, and reciprocal suppression effects in the same study seems difficult to explain theoretically, since these effects result from different patterns of relationships between predictors and outcome (Conger, 1974). It seems to us that by far the most parsimonious explanation here is that most of the statistically significant β coefficients in Quoidbach et al.'s (2014) Study 2 are the result of a combination of noisy data and the high correlation between the predictors (i.e., positive or negative emotion and the corresponding form of emodiversity).

Conclusion

Quoidbach et al. (2014) drew some far-reaching conclusions about the impact of emodiversity on mental and physical health from their two correlational studies. From their Study 1, they concluded that emodiversity has incremental predictive validity, over and above that associated with emotions of the same valence, for depression. From their Study 2, they concluded that emodiversity was at least as good a predictor of physical health as regular exercise, a healthy diet, and refraining from smoking. However, as we have shown, these results are probably nothing more than a statistical mirage. Once the regression analyses in Study 1 were conducted in a hierarchical manner it became clear that the incremental effect of emodiversity in terms of added variance explained was negligible, while in Study 2 the apparently substantial regression coefficients were shown to be the product of suppression effects. In both cases, the fact that most of these regression coefficients and their associated partial correlation coefficients were statistically significant at the traditional .05 level is neither surprising (given the large sample sizes used by Quoidbach et al.) nor very meaningful. With many hundreds or thousands of participants, we feel that the effect size (as measured by the increase in R^2) is a more appropriate indication of the influence of emodiversity than a p value.

This problem appears to have arisen, at least in part, as a result of a misunderstanding of the principles of a mathematical concept that has been imported from another field. As we have shown, Quoidbach et al.'s (2014) application of Shannon entropy is inappropriate, given the constraints imposed by the nature of the measures of emotional experience being used, such as the fact that a fixed number of emotions are measured, corresponding to a fixed list of "species," and a limited number of responses are allowed, corresponding to a limited range of

possible population values. (This is not to suggest that, when applied to data obtained from more appropriate measures, Shannon entropy might not be useful in other areas of psychology or the social sciences more generally; see, for example, Vaquero & Cebrian, 2013). As one of us has pointed out previously (Brown, Sokal, & Friedman, 2013), researchers in psychology and other social sciences who wish to borrow concepts from the natural sciences or mathematics should ensure that they understand all of the conditions for the use of those concepts to be valid. The fact that a set of mathematical formulae (such as those for calculating Shannon entropy, but also those for performing ordinary least squares regression) can be applied to psychological data is no guarantee that the results that emerge from the application of these formulae will have any meaning in the real world.

In conclusion, we do not claim that the idea of emotional diversity is inherently devoid of any possible utility. However, in order for such utility to be demonstrated, it will be necessary to identify ways of operationalizing and measuring this construct that are not compromised by mathematical artifacts and statistical confounds. In its present form, Quoidbach et al.'s (2014) construct of emodiversity does not meet this standard.

References

- Berbaum, K. S., Dorfman, D. D., Franken, E. A. Jr., & Caldwell, R. T. (2002). An empirical comparison of discrete ratings and subjective probability ratings. *Academic Radiology*, 9, 756–763. [https://doi.org/10.1016/S1076-6332\(03\)80344-6](https://doi.org/10.1016/S1076-6332(03)80344-6)
- Brown, N. J. L., Sokal, A. D., & Friedman, H. L. (2013). The complex dynamics of wishful thinking: The critical positivity ratio. *American Psychologist*, 68, 801–813. <https://doi.org/10.1037/a0032850>
- Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, 3, 106–116. <https://doi.org/10.3844/jssp.2007.106.116>
- Conger, A. J. (1974). A revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, 34, 35–46. <https://doi.org/10.1177/001316447403400105>
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease*, 175, 526–536. <https://doi.org/10.1097/00005053-198709000-00004>
- DeSoto, K. A. (2016). Computerized methods for collecting confidence ratings: Task influences in patterns of responding. Manuscript submitted for publication. Preprint available from <https://peerj.com/preprints/1432v1>
- Fredrickson, B. L. (2013). Positive emotions broaden and build. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (Vol. 47), pp. 1–53. Burlington, VT: Academic Press.
- Kristjánsson, K. (2003). On the very idea of “negative emotions.” *Journal for the Theory of Social Behaviour*, 33, 351–364. <https://doi.org/10.1046/j.1468-5914.2003.00222.x>
- Magurran, A. E. (2004). *Measuring biological diversity*. Malden, MA: Blackwell.
- Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, 14, 858–865. <https://doi.org/10.3758/BF03194112>
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97. <https://doi.org/10.1037/h0043158>
- Paulhus, D. L., Robins, R. W., Trzesniewski, K. H., & Tracy, J. L. (2004). Two replicable suppressor situations in personality research. *Multivariate Behavioral Research*, 39, 303–328. https://doi.org/10.1207/s15327906mbr3902_7
- Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emodiversity and the emotional ecosystem. *Journal of Experimental Psychology: General*, 143, 2057–2066. <https://doi.org/10.1037/a0038025>
- Ray, J. J. (1990). Acquiescence and problems with forced-choice scales. *Journal of Social Psychology*, 130, 397–399. <https://doi.org/10.1080/00224545.1990.9924595>
- Robinson, M. D., & Clore, G. L. (2002). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality & Social Psychology*, 83, 198–215. <https://doi.org/10.1037/0022-3514.83.1.198>
- Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>

- Solomon, R. C. (2002). Back to basics: On the very idea of “basic emotions.” *Journal for the Theory of Social Behaviour*, 32, 115–144. <https://doi.org/10.1111/1468-5914.00180>
- Stirling, G., & Wilsey, B. (2001). Empirical relationships between species richness, evenness, and proportional diversity. *The American Naturalist*, 158, 286–299. <https://doi.org/10.1086/321317>
- Vaquero, L. M., & Cebrian, M. (2013). The rich club phenomenon in the classroom. *Nature Scientific Reports*, 3, 1174. <https://doi.org/10.1038/srep01174>
- Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the Positive and Negative Affect Schedule-expanded form*. Iowa City: The University of Iowa.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>

Supplemental Information

This supplemental information document is divided into three parts. Part 1 provides detailed descriptions and results of the numerical analyses of the effects of bounded richness and evenness on Quoidbach et al.'s (2014) construct of emodiversity. Part 2 presents some extended footnotes to correspondingly-named sections in the main article. Part 3 contains some specific standalone points that we omitted from the main article due to lack of space, and because they were less central to our argument.

Part 1: Numerical analyses of the effects of bounded richness and evenness

Table 5.2 shows the range of emodiversity for each possible number of zero-scored responses (corresponding to *Never* experiencing a particular emotion) on three emotion-frequency measures. The first two measures have nine and 20 items, respectively, with possible responses in the range 0–4; the third measure has nine items with possible responses ranging from 0–100. The numerical effects of the severe constraints on the upper bounds on both richness and evenness imposed by Quoidbach et al.'s (2014) formulation of emodiversity can readily be seen in the rather small ranges between the minimum and maximum possible emodiversity values for any given number of non-zero (*Never*) responses. In each case, the lowest emodiversity is obtained when half of the non-zero responses are the lowest possible (i.e., 1) and the other half are the highest possible (i.e., 4 or 100, depending on the scale length); meanwhile, the highest emodiversity occurs when all of the non-zero responses are identical, regardless of which of the possible values in the range 1–4 (or 1–100) these non-zero responses all take.

It can be readily seen that, once the number of zero (*Never*) responses is established, the possible variation in emodiversity is severely constrained, as indicated by the “Ratio” entries in Table 5.2, which show the ratio between the maximum and minimum possible emodiversity values for each number of zero responses. The maximum/minimum ratio increases only slightly as the number of zero responses increases; for the 20-item scale, the ratio progresses from a minimum of 1.069 to a maximum of 1.385 (in the case of a participant who reports *Never* experiencing 18 out of 20 emotions, plus one *Rarely* and one *Most of the time*, which might well be considered a severely atypical pattern of responses), until the entire process breaks down with 19 or 20 zeroes. When the range of possible responses is expanded from 0–4 to 0–100, the ratio between maximum and minimum emodiversity for any given number of non-zero responses becomes somewhat larger; however, for the minimum evenness (and, hence, lowest emodiversity) to be obtained,

participants would need to alternate between responses of 1 and 100 for every pair of emotions on the scale, which again would seem to be a highly unusual pattern, possibly indicative of acquiescence or boredom in participants. Thus, the range of possible “legitimate” emodiversity values needs to be further reduced to take into account that reports of either maximum or minimum emodiversity may be more likely to reflect unthinking response patterns than a sincere reflection of participants’ experiences.

A further problem whose genesis can be glimpsed in Table 5.2, and which we were able to observe more concretely in Quoidbach et al.’s (2014) empirical studies, is the likely presence of a high degree of skewness in emodiversity, because of the strong effect of a single experience of an emotion (moving the response from *Never* to *Rarely*), compared to that of other responses. For example, Quoidbach et al.’s Table 3 shows that in Study 2 the maximum positive emodiversity score was 0.47 standard deviations (SDs) above the mean, whereas the minimum score was 13.1 SDs below the mean; indeed, Quoidbach et al. reported (p. 2062) that they applied a transformation to the emodiversity data in this study to address this skewness. In Study 1 the skewness was less dramatic, in that the maximum positive emodiversity was 0.72 SDs above the mean and the minimum was “only” 4.32 SDs below the mean; nevertheless, the overall skewness for positive, negative, and global emodiversity in Study 1 were -2.52 , -0.92 , and -1.54 , respectively. We investigated the effects of attempting to correct this skewness by squaring (or raising to an even higher power) these emodiversity variables. However, when we did this, the correlation of these new transformed variables for positive and negative emodiversity with their respective emotion measures became even higher than before (e.g., for negative emodiversity, the correlation with negative emotion increased from .79 to .86 when emodiversity was squared). Hence, it seems that this high skewness is likely to be another suboptimal aspect of emodiversity data that researchers will have to cope with in their analyses.

Table 5.2. Minimum and maximum possible emodiversity values, and the ratio of maximum to minimum, for each possible number of zero-scored responses on three different possible scales with different lengths and scoring ranges.

Zeroes	9-item scale (scored 0–4)			20-item scale (scored 0–4)			9-item scale (scored 0–100)		
	Minimum	Maximum	Ratio	Minimum	Maximum	Ratio	Minimum	Maximum	Ratio
0	1.988	2.197	1.105	2.803	2.996	1.069	1.456	2.197	1.510
1	1.887	2.079	1.102	2.744	2.944	1.073	1.442	2.079	1.442
2	1.733	1.946	1.123	2.698	2.890	1.071	1.172	1.946	1.660
3	1.599	1.792	1.121	2.632	2.833	1.077	1.154	1.792	1.552
4	1.390	1.609	1.158	2.580	2.773	1.075	0.776	1.609	2.074
5	1.194	1.386	1.161	2.505	2.708	1.081	0.749	1.386	1.852
6	0.868	1.099	1.266	2.446	2.639	1.079	0.110	1.099	9.978
7	0.500	0.693	1.385	2.361	2.565	1.087	0.056	0.693	12.479
8	0.000	0.000	n/a	2.292	2.485	1.084	0.000	0.000	n/a
9	0.000	0.000	n/a	2.192	2.398	1.094	0.000	0.000	n/a
10				2.110	2.303	1.091			
11				1.988	2.197	1.105			
12				1.887	2.079	1.102			
13				1.733	1.946	1.123			
14				1.599	1.792	1.121			
15				1.390	1.609	1.158			
16				1.194	1.386	1.161			
17				0.868	1.099	1.266			
18				0.500	0.693	1.385			
19				0.000	0.000	n/a			
20				0.000	0.000	n/a			

Notes:

1. The possible emodiversity values for the 9-item scale with 0, 1, 2, etc., zero-scored items are identical to the corresponding values for the 20-item scale with 11, 12, 13, etc. zero-scored items. More generally, the range for an N -item scale ($N < 20$) is obtained by taking the last N items of the 20-item scale.
2. The alternating increase and decrease in the ratio as the number of zeroes increases (particularly noticeable for the 9-item scale scored 0–100) represents the fact that minimum emodiversity increases by a smaller amount when the number of non-zero items on the scale becomes odd (when a value of 1 is added to the list of non-zero item scores) than when it becomes even (when the maximum value or 4 or 100 is added to the list).

Part 2: Extended Footnotes

Notes for “Limitations of the analogy with biodiversity”

Richness vs. Evenness. There is some debate among biologists as to whether richness and evenness should be considered separately, because they are typically closely related empirically (Stirling & Wilsey, 2001). While specific individual measures of both richness and evenness exist, the Shannon entropy formula—sometimes referred to as the Shannon-Wiener index—functions as a composite measure: Once the Shannon-Wiener value (denoted by H') for the biodiversity of a community has been established, an increase in either richness or evenness will produce a larger H' value. This ambiguity could be considered a weakness of attempts to repurpose Shannon’s original concept beyond its originally intended field of communication. In Shannon’s (1948) model, the practical outcome of either a greater number of different characters in the message (“richness”) or a greater variety in the distribution of those characters (“evenness”) is the same, namely an increase in the number of bits required to uniquely encode the message. In other words, when applied to electronic communications, richness and evenness are two sides of the same coin, whereas in biology (and emodiversity), this relation is considerably more complex.

Taxonomy of emotion scales. The nine positive emotions measured by the mDES, used by Quoidbach et al. (2014) in their Study 1, are *alertness*, *amusement*, *awe*, *contentment*, *gratitude*, *hope*, *joy*, *love*, and *pride*. In their Study 2, Quoidbach et al. used a different, unnamed measure, with ten positive emotions: *amusement*, *awe*, *contentment*, *enthusiasm*, *gratitude*, *happiness*, *interest*, *joy*, *pride*, and *serenity* (of which only six are common to this measure and the mDES). Had Quoidbach et al. instead taken their set of emotions from the PANAS (Watson, Clark, & Tellegen, 1988), which certainly seems plausible—for example, the filename of the “emodiversity calculator” spreadsheet that we downloaded from the emodiversity.org website on October 28, 2015 was “Emodiversity Calculator 20 emotion items (like PANAS).xlsx”—their list would have been even more different from the mDES. The PANAS uses adjectives rather than nouns to ask participants how they are (or were) feeling; its list of 10 positive emotions consists of the words *active*, *alert*, *attentive*, *determined*, *enthusiastic*, *excited*, *inspired*, *interested*, *proud*, and *strong*. Thus, the mDES and PANAS have only two positive emotions (*alertness* and *pride*) in common. It is difficult to imagine how one could reliably compare emodiversity values across populations using such different instruments. Yet, the emodiversity calculator spreadsheet invites researchers to use any emotion measure of their choice, as long as the scale is zero-based.

(Note that a base of zero for the responses is a mathematical requirement of the way the emodiversity formula is constructed; any existing scale that uses a response format that starts above zero, such as 1–7, will need to be rebased to zero in order to be used to calculate emodiversity.)

A further complication is that emodiversity values that have been derived from two measures with different numbers of items are not commensurate. The range of emodiversity values for a 10-item scale (maximum emodiversity 2.30) is different from that of an 18-item scale (maximum emodiversity 2.89). Thus, unless a decision is made to standardize on a definitive number of items, there is no prospect of a universal scale of values of emodiversity; any given individual will have a 10-item emodiversity score, an 18-item score, and so on. This represents a further departure from the application of Shannon entropy to communication or biodiversity; in both of those cases, any message or community has exactly one entropy value, and these values can be readily compared across messages or communities.

Notes for “Empirical Issues”

Need for stepwise reporting. Quoidbach et al. (2014) reported the results of their multiple regressions only at the final step, with all variables included. But in the presence of correlated predictors, without an indication of the evolution of the model as variables are added, it is almost impossible to evaluate whether the focal variables of the study are really those that are driving the observed effects. Had Quoidbach et al. followed this recommendation, reporting the extra variance explained at each step, the limited power of emodiversity to explain unique variance (beyond the corresponding measure of emotions) in both of their studies would have been more readily apparent. (We note that Quoidbach et al. did report, in the summary paragraph for their Study 1, that emodiversity explained only about 1% of the variance in depression.)

No apparent checks for automatic responding. We noted in our main article that the highest and lowest values of emodiversity are to be found in cases of extreme response patterns. The participants in Quoidbach et al.’s (2014) Study 1 were recruited through a TV advertisement that appeared in a popular show dedicated to happiness; this advertisement invited viewers to visit the show’s website in order to participate. In addition to the risk of sampling bias, the relatively casual nature of this recruitment process might be expected to include a number of people who might not necessarily spend a lot of time carefully considering their responses. Indeed, of Quoidbach et al.’s 41,723 participants, 637 reported

identical frequencies of experience of all nine positive emotions (288 *Never*, 349 some non-zero frequency), while 1,298 (1,220 *Never*, 78 non-zero) reported identical frequencies of experience of all nine negative emotions. However, it does not appear that any participants were excluded from Quoidbach et al.'s analyses purely on the basis of identical responses.

Part 3: Some further topics

Distortions introduced by the time period under study

The capacity of any of the instruments used by Quoidbach et al. (2014) to measure emodiversity depends on the time period for which participants are invited to report their emotional experiences. This is because, of the possible responses to the scale items, one (*Never*) is an absolute number (i.e., the emotion was experienced on a total of zero occasions during the entire period in question, whatever the length of this period), whereas the others are expressed in terms of an informal proportion of the time available. For example, consider what would happen if the word “week” in the (presumed) question “How often did you experience each of these emotions in the past week?” were to be replaced with the word “month.” Assuming that a person’s experiences of emotions are fairly stable over time, their relative proportion of responses from *Rarely* through *Most of the time* ought to be similar for any given time period; that is, someone who, over the course of a typical week, is alert *Most of the time* and only *Rarely* embarrassed will probably report experiencing those emotions with similar relative frequencies over the course of a month. Indeed, in order for Quoidbach et al.’s claims about the ability of emodiversity scores measured over any given week (which we presume was the time period used in these authors’ empirical studies) to predict long-term mental and physical health outcomes to be valid, such an assumption about the temporal stability of the frequency of emotional experiences would seem to be not just plausible, but necessary. Otherwise, Quoidbach et al.’s inferences about health outcomes over the course of months or years from a single week’s emodiversity scores would appear to be potentially subject to a great deal of random variation depending on the week in question.

In contrast, when the period under consideration is extended from a week to a month (or longer), the number of *Never* responses is likely to be reduced, simply because more time is available for each emotion to have been experienced at least once. Perhaps Dave has not been embarrassed in the past week, but he might have had an awkward moment three weeks ago, thus causing him to respond that he experienced embarrassment *Rarely* (versus *Never*) when the time period is expanded to a month. The importance of this point is illustrated when we examine the influence of zero-coded (*Never*) responses on the calculation of

emodiversity. In the data set for Quoidbach et al.'s (2014) Study 1, we found a correlation of $-.96$ between the count of *Never* responses for each participant to either the positive or negative mDES subscale and the respective measure of emodiversity. (This correlation would be somewhat lower for a 0–100 response format, but only to the extent that participants use a wider range of response values, which—as we mentioned in our main article—appears to be unlikely in practice.) This remarkably high correlation provides a further illustration of the point that we made in our main article regarding the limited range of possible emodiversity values (indeed, it is a mathematical corollary of that limited range). Once the number of *Never* responses is established, emodiversity is almost completely static, so that increasing one's experience of an emotion beyond *Rarely* has almost no effect (and what effect it does have is as likely to be negative as positive); yet, the number of *Never* responses depends on the essentially arbitrary choice of time period used for the measure of emotions. A measure that asked people how many times they had experienced each emotion in the past year—which might be highly valid in an investigation of people's long-term emotional experiences—would likely reveal that many of them had almost the same emodiversity, as most people would not be able to give a response of *Never* to the majority of the items.

Forced responses to all scale items

Shannon's (1948) index of entropy quantifies the diversity among (only) the observed characters in a message. For example, the one-word message *cataract* contains three observations of the letter *a*, two each of *c* and *t*, and one of *r*, and has a Shannon entropy of 1.91^{23} . The fact that the other letters of the alphabet do not appear in this message is irrelevant, as is the length of the alphabet. Similarly, when the Shannon-Wiener index is used in field biology, only the observed species are of interest; the fact that no aardvark or zebra was sighted during a field study of a Welsh bog does not affect the biodiversity index of that community. By definition, there is no such thing as an observation of zero occurrences of a character, or zero members of a species, in these situations. In contrast, the structure of the psychologist's questionnaire—which is equivalent to taking along a pre-printed checklist of

²³ This value is calculated using binary (base 2) logarithms, as used by Shannon (1948). Binary logarithms are appropriate in the case of communication systems because one is typically interested in the number of bits that a coded message will occupy. Natural logarithms, as used by Quoidbach et al. (2014), are arguably more appropriate for other applications of Shannon's formula. When natural logarithms are used, the Shannon entropy of the message *cataract* is multiplied by $\ln(2)$, or 0.6931, to give 1.32.

possible species, rather than a blank sheet of paper, to observe a biological community—means that there is a difference between the absence of any report concerning a particular phenomenon, and a concrete statement that the phenomenon did not occur. The psychological equivalent of the absence of the letter *q* in Shannon’s message, or of the failure to observe any tigers in a field study in biology, is not a report by the participant that an emotion was never experienced; rather, it is the *absence of a response* to the item concerning that emotion. These two situations are clearly distinct. For example, when the mDES is used to calculate a participant’s mean level of positive emotions, a response of *Never*—coded as zero—to a particular emotion item has an effect on the participant’s mean positive emotions score, by increasing the denominator (i.e., the total number of responses), in a way that a failure to respond—which would simply be ignored in the calculation of the mean—would not.

Quoidbach et al. (2014) therefore needed to handle the separate possibilities of a missing response and an explicit zero-coded response²⁴. However, their choice to treat a response of *Never* as making no contribution to emodiversity—by assigning a value of zero for the corresponding ($p_i \times \ln p_i$) term in Shannon’s formula²⁵—effectively transformed *Never* into a missing response. In turn, this meant that a decision had to be made about how to handle cases where the participant did in fact fail to respond to one or more items, because to treat a genuine missing response as being identical to a response of *Never* would represent a serious distortion of the meaning of the items composing the scale. Quoidbach et al. did not discuss this problem in their article, but from an examination of their examples, the data set for their Study 1, and their online emodiversity test at <http://www.emodiversity.org/>, it appears that their solution was to *force* participants to answer every question on the mDES or equivalent measure. For example, 6,611 participant records in the data set for Quoidbach et al.’s Study 1 have no data for any of the items on the nine-item MADR-S depression scale, but a further 848 records have data for between one and eight items, suggesting that no forced choice was imposed for this scale. In contrast, 5,879 participants have no data for the mDES, but none of the records in the data set have partially complete responses for this scale. This suggests to us that responding to all items was mandatory, probably being enforced by the

²⁴ The nature of the $p_i \times \ln p_i$ formula for calculating the contribution of each item to emodiversity means that the scale being used must be (re-)anchored at zero, as noted by Quoidbach and colleagues in their emodiversity calculator spreadsheet.

²⁵ Although the logarithm of zero is undefined, the ($p_i \times \ln p_i$) term for an item whose value (and, hence, p_i term) is zero, is defined (with a value of zero).

computer system used for the survey. (We were unable to check the equivalent numbers for Study 2, as the data set has not been made public.)

In summary, a decision to measure emodiversity would seem to require researchers to take a rather Procrustean (and, potentially, bias-inducing) approach to participants: Only those who are willing and able to provide responses to every item on a scale measuring the frequency of experienced emotions are candidates for inclusion in calculations of emodiversity. However, forcing participants to respond in this way potentially introduces a number of psychometric problems (Ray, 1990). It also implies that emodiversity will be difficult to measure in pencil-and-paper settings, where participants always have the opportunity to not answer a particular item (whether deliberately or not); to exclude those who consciously (or accidentally) omit one or more responses might be a source of bias. Furthermore, the more items in the emotion measurement scale being used, the greater the chance that participants will omit one or more responses and, hence, that their results will not be used at all. This potentially creates a perverse incentive for researchers to use shorter measures, which will not only be less reliable for measuring the emotions that they are supposed to measure, but will also allow for even less variance in emodiversity.

The values of emodiversity are produced by the use of invalid arithmetic

Quoidbach et al. (2014) stated that they obtained the p_i components of the Shannon entropy formula by “divid[ing] the number of times an individual experienced [an emotion] by the total number of times she experienced all types of emotions” (p. 2058). However, it is not correct to say that a measure such as the mDES gives an indication of the “number of times” each emotion was experienced. Instead, an arbitrary integer value is applied to a subjective verbal indicator of the frequency with which each emotion was experienced, ranging from *Never* (scored as 0) to *Most of the time* (scored as 4). But, even if it can be shown empirically that multi-item ordinal (Likert) scales typically behave empirically like interval data when summed and averaged (Carifio & Perla, 2007), the act of dividing an individual number from such a scale by any other number (in this case, the equally-arbitrary total of all the emotion-frequency scores) is a meaningless operation, and subsequently multiplying the result of that operation by its own logarithm merely compounds this error. All that can be said about the results of such a numerical manipulation is that the p_i terms corresponding to more frequent experiences of each emotion will be larger than those corresponding to less frequent experiences, but the magnitude of this relation is essentially arbitrary, and will vary depending on the frequency of experience of other emotions endorsed

by the same participant. Even if the range of possible emotion-frequency responses were to be extended to 0–100, as discussed earlier, it would seem to be difficult to defend the idea that this creates an interval scale, with a value of 60 somehow representing experiencing an emotion “twice as often” as a value of 30.

The equal desirability of all emotions on the scale being used

One of the limitations of Quoidbach et al.’s (2014) operationalization of emodiversity is that it places equal weight on all of the emotions on the scale that is used (which, as we noted in our main article, is already a somewhat arbitrary selection from the possible taxonomy of emotions). Yet, just as not all forms of biological evenness are necessarily desirable (a hectare of African savanna that currently contains an equal biomass of wildebeest and lions is likely to undergo some kind of upheaval in the near future), so it is not clear that it is beneficial for people to experience all emotions, even those typically considered “positive,” with equal frequency. For example, while it might be good for people to experience alertness or contentment *Most of the time*, it might be better for the stress levels of their family and friends if these same individuals experienced pride or awe only *Some of the time*. Yet, Quoidbach et al.’s theoretical approach—which, for any given level of richness, rewards evenness with a higher emodiversity score—implies that all emotions (at least, those of a given valence) are created equal, and that experiencing them all with the same frequency—whatever that frequency might be—is a good thing.

The effect of adding quadratic terms to the regression models

Cortina (1993) noted that when predictors are highly correlated, the effects of their interaction term can become confounded with the effects of higher order (typically quadratic) terms of each predictor. We therefore tested the effects of adding the square of emotion and emodiversity terms to the principal regression models for positive and negative emodiversity in Quoidbach et al.’s (2014) Study 1, using Cortina’s (1993, p. 918) three-step process.

For positive emotion and emodiversity, R^2 was .391 at step 1 (two predictors, cf. Table 5.1, section P.2 in our main article), .397 at step 2 (after adding the two quadratic terms), and .400 at step 3 (after adding the interaction). The added variance explained by the interaction (R^2 increase of .003) for positive emotion and emodiversity was therefore the same as in Quoidbach et al.’s (2014) basic model.

For negative emotion and emodiversity, R^2 was .460 at Cortina’s (1993) step 1, .465 at step 2, and .466 at step 3 (after adding the interaction), so that the added variance explained by the interaction after accounting for the confounding effects of the quadratic term (an

increase in R^2 of .001) was only approximately one-sixth of what was observed by Quoidbach et al. (2014) in their basic model for negative emotion and emodiversity when the interaction term was added. (As sections N.1 and N.2 of Table 5.1 in our main article show, the interaction term was the *only* source of extra explained variance in the regressions for negative emotion and emodiversity; adding negative emodiversity on its own to a regression containing only negative emotion as a predictor produced *zero* increase in explained variance, to three decimal places.)

References

- Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, 3, 106–116. <https://doi.org/10.3844/jssp.2007.106.116>
- Cortina, J. M. (1993). Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. *Journal of Management*, 19, 915–922. <https://doi.org/10.1177/014920639301900411>
- Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emotiversity and the emotional ecosystem. *Journal of Experimental Psychology: General*, 143, 2057–2066. <https://doi.org/10.1037/a0038025>
- Ray, J. J. (1990). Acquiescence and problems with forced-choice scales. *Journal of Social Psychology*, 130, 397–399. <https://doi.org/10.1080/00224545.1990.9924595>
- Stirling, G., & Wilsey, B. (2001). Empirical relationships between species richness, evenness, and proportional diversity. *The American Naturalist*, 158, 286–299. <https://doi.org/10.1086/321317>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>

Chapter 6

**Does Twitter Language Reliably Predict
Heart Disease? A Commentary on
Eichstaedt et al. (2015a)**

Nicholas J. L. Brown

James C. Coyne

PeerJ, 6, e5656 (September 2018).

DOI: 10.7717/peerj.5656

Abstract

We comment on Eichstaedt et al.'s (2015a) claim to have shown that language patterns among Twitter users, aggregated at the level of U.S. counties, predicted county-level mortality rates from atherosclerotic heart disease (AHD), with “negative” language being associated with higher rates of death from AHD and “positive” language associated with lower rates. First, we examine some of Eichstaedt et al.'s apparent assumptions about the nature of AHD, as well as some issues related to the secondary analysis of online data and to considering counties as communities. Next, using the data files supplied by Eichstaedt et al., we reproduce their regression- and correlation-based models, substituting mortality from an alternative cause of death—namely, suicide—as the outcome variable, and observe that the purported associations between “negative” and “positive” language and mortality are reversed when suicide is used as the outcome variable. We identify numerous other conceptual and methodological limitations that call into question the robustness and generalizability of Eichstaedt et al.'s claims, even when these are based on the results of their ridge regression/machine learning model. We conclude that there is no good evidence that analyzing Twitter data in bulk in this way can add anything useful to our ability to understand geographical variation in AHD mortality rates.

Eichstaedt et al. (2015a) claimed to have demonstrated that language patterns among Twitter users, aggregated at the level of U.S. counties, were predictive of mortality rates from atherosclerotic heart disease (AHD) in those counties, with “negative” language (expressing themes such as disengagement or negative relationships) being associated with higher rates of death from AHD and “positive” language (e.g., upbeat descriptions of social interactions or positive emotions) being associated with lower AHD mortality. Eichstaedt et al. examined a variety of measures to demonstrate the associations between Twitter language patterns and AHD, including (a) the frequency of usage of individual words associated with either positive or negative feelings or behaviors, (b) the tendency of Twitter users to discuss “positive” (e.g., skilled occupations) or “negative” (e.g., feelings of boredom) topics, and (c) an omnibus model incorporating all of their Twitter data, whose performance they compared with one using only “traditional” predictors such as health indicators and demographic and socioeconomic variables.

The claims made by Eichstaedt et al. (2015a) attracted considerable attention in the popular media (e.g., Izadi, 2015; Jacobs, 2015; Singal, 2015), with many of these articles being based in large part on the Association for Psychological Science’s (2015) own press release. However, a close examination of Eichstaedt et al.’s article and data appears to reveal a number of potential sources of distortion and bias in its assumptions about the nature of AHD, the use of Twitter data as a proxy for the socioemotional environment and people’s health, and the use of counties as the unit of analysis. Some of these problems are immediately obvious from reading Eichstaedt et al.’s article, while others only manifested themselves in the testing of the relevant data that we undertook.

Here, we first present a selection of the main problems that we identified when reading Eichstaedt et al.’s (2015a) article. We then expose further problems that emerged after we scrutinized the original data and also cross-validated the application of Eichstaedt et al.’s models to a different mortality outcome variable, namely suicide. We end with a discussion of some broader implications for the use of large-scale sources of data to draw conclusions about health and human behavior based on sophisticated computer models.

Issues related to the idea of psychological causes of AHD

Perhaps a large part of the appeal of Eichstaedt et al.’s (2015a) claims about the potential for community-level psychological factors—notably, those that purportedly lead Twitter users to either make aggressive or otherwise anti-social outbursts, or, conversely, express upbeat and positive sentiments—to somehow affect local rates of death from AHD is

that these claims echo the common belief, repeated in the first sentence of Eichstaedt et al.'s abstract, that hostile or highly stressed individuals are more susceptible to cardiovascular problems. The notion that a loudmouthed, dominating, aggressive person is somehow more likely to suddenly drop dead from a heart attack is widespread in lay or pop psychology thinking (e.g., Riggio, 2014), perhaps at least in part because it provides some comfort to people on the receiving end of such individuals' behavior. Indeed, although Eichstaedt et al. did not use the term "Type A personality" in their article, this stereotype—characterized by a tendency towards aggression and a variety of other negative interpersonal characteristics, as well as greater susceptibility to cardiovascular problems—is a staple part of the popular culture surrounding the relation between mental and physical health (e.g., Wilson, 2009). Yet, despite initial promising findings suggesting a link between Type A behavior pattern (TABP) and cardiac events and mortality in small samples (Friedman & Rosenman, 1959), an accumulation of evidence from more recent large-scale studies has consistently failed to show reliable evidence for such an association (Kuper, Marmot, & Hemingway, 2002). Appearances of positive findings could be generated using a range of distress or negative affectivity variables (Smith, 2006). However, it was then recognized that negative affectivity could not readily be separated from a full range of antecedent and concurrent biological, psychological, and social factors. At best, negative affectivity is likely to be no more than a non-informative risk marker (Ormel et al., 2004), not a risk factor for AHD. Its apparent predictive power is greatly diminished with better specification and measurement of confounds (Smith, 2011).

A recent large scale study of over 150,000 Australians (Welsh et al, 2017) provides a typical example of the problem. Significant associations between levels of psychological distress and incidence of ischemic heart disease (a superordinate category in the ICD-10 classification that includes AHD as a subcategory) decreased as adjustments were made for demographic and behavioral characteristics until measurement error and residual confounding seemed to account for any remaining association. As the authors put it, "A substantial part of the distress–IHD association is explained by confounding and functional limitations Emphasis should be on psychological distress as a marker of healthcare need and IHD risk, rather than a causative factor" (Welsh et al., 2017, p. 1).

In contrast to TABP, socioeconomic conditions have long been identified as playing a role in the development of AHD. For example, Clark, DesMeules, Luo, Duncan, and Wielgosz (2009) noted the importance of access to good-quality healthcare and early-life

factors such as parental socioeconomic status. However, no county-level measure of either of those variables appeared in Eichstaedt et al.'s (2015a) model or data set.

Issues related to the etiology of AHD

As the single most common ICD-10 cause of mortality in the United States, AHD might have seemed like a natural choice of outcome variable for a study such as that of Eichstaedt et al. (2015a). However, it is important to take into account some aspects of the nature and course of AHD. It is a chronic disease that typically develops asymptotically over decades. The first recognition of AHD often follows an event such as an acute myocardial infarction or some other sudden incident, reflecting the fact that the cumulative build-up of plaque over time has finally caused a blockage of the arteries (Funk, Yurdagul, & Orr, 2012) rather than any specific change in the immediately preceding time period. Indeed, disagreement among physicians as to whether the cause of death is AHD or some other cardiac event is common (Mant et al., 2006). A definitive post-mortem diagnosis of AHD may require an autopsy, yet the number of such procedures performed in the United States has halved in the past four decades (Hoyert, 2011).

In contrast to AHD, there is another cause of death for which the existence of an association with the victim's recent psychological state is widely accepted, namely suicide (Franklin et al., 2017; O'Connor & Nock, 2014). Although suicide can be the result of long-term mental health problems and other stressors, a person's psychological state in the months and days leading up to the point at which they take their own life clearly has a substantial degree of relevance to their decision. Hence, we might expect any county-level psychological factors that act directly on the health and welfare of members of the local community to be more closely reflected in the mortality statistics for suicide than those for a chronic disease such as AHD.

Issues related to the secondary analysis of data collected online

In the introduction to their article, Eichstaedt et al. (2015a) invoked Google's use of search queries related to influenza, "providing earlier indication of disease spread than the Centers for Disease Control and Prevention" (p. 160) as a (presumably positive) example of how "digital epidemiology can support faster response and deeper understanding of public-health threats than can traditional methods" (p. 160). However, both Google's project to infer a relation between searches for certain terms and the immediate prevalence of an acute infectious disease, and Eichstaedt et al.'s attempt to correlate certain kinds of communication with chronic cardiovascular conditions, suffer from the fundamental problem that they are

attempting to extract some kind of signal from what may very well be a large amount of noise. In fact, before it was quietly shut down in August 2015, Google Flu Trends (GFT) failed—in some cases spectacularly—to correctly predict the seasonal spread of influenza in the United States on several occasions. As Lazer, Kennedy, King, and Vespignani (2014) put it, in their review of the problems that beset GFT, “The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis” (p. 1203)—a remark that clearly also applies to the notionally random selection of tweets that constitute the Twitter “Garden Hose” data, especially since, as we show below, this selection may not have been entirely random.

As well as the limitations noted in Eichstaedt et al.’s article (2015a) and Supplemental Material (2015b)—such as the difference in Twitter user density across counties, and the fact that 7% of tweets were mapped to the wrong county—we note that there is also a potential source of bias in the geographical coding of their Twitter data, namely the assumption that the users who provided enough information to allow their tweets to be associated with a county represented an unbiased sample of Twitter users in that county. This requires people of all ages, genders, socio-economic status levels, and ethnic backgrounds to be equally likely to either list their city and state in their profile, or to enable geotagging of their Tweets. However, it seems entirely plausible that certain categories of individuals might be more likely to self-censor their profile information than others (for example, it could be that people who do not wish to reveal their location are more or less restrained in their use of hostile language). Given that only 16% of tweets could be mapped to counties, any bias in this area could have substantial consequences.

Issues associated with considering counties as communities

As Eichstaedt et al. (2015a, p. 160) themselves noted, “Counties are the smallest socioecological level for which most CDC health variables and U.S. Census information are available.” Thus, these authors seem to have acknowledged that their use of counties as their unit of analysis was driven principally by (eminently reasonable) practical constraints. However, Eichstaedt et al.’s subsequent interpretation of their results (e.g., “language expressed on Twitter revealed several community-level psychological characteristics that were significantly associated with heart-disease mortality risk,” p. 164) requires that counties also constitute meaningful communities. Indeed, this interpretation also implies that any psychological mechanism that might account for the relationship between the vocabulary of Twitter users and the health outcomes of the wider population within any given county works

in a similar way across all counties. Yet there seem to be several reasons to question such assumptions. First, the size and population of U.S. counties varies widely; both their land areas (ranging from 1.999 to 144,504.789 square miles), and their 2010 populations (from 82 to 9,818,605) span five orders of magnitude. Second, across the United States, the political and economic importance of counties as a level of government (between the municipality and state level) is highly variable, being generally greater in the South (Benton, 2002, p. 7; Haider-Merkel, 2009, p. 723); indeed, in Connecticut, Rhode Island, and much of Massachusetts, counties have no governmental function whatsoever. Third, it is not clear that many Americans identify at all closely with their county as a source of community (Kilgore, 2012). Fourth, within counties, socioeconomic and other factors often vary enormously:

[I]n New York County, New York, . . . neighborhoods range from the Upper East Side and SoHo to Harlem and Washington Heights. . . . [I]n San Mateo County, California, . . . neighborhoods range from the Woodside estates of Silicon Valley billionaires to the Redwood City bungalows of Mexican immigrants. (Abrams & Fiorina, 2012, p. 206)

Given such diversity in the scale and sociopolitical significance of counties, we find it difficult to conceive of a county-level factor, or set of factors, that might be associated with both Twitter language and AHD prevalence with any degree of consistency across the United States. Eichstaedt et al. (2015a, p. 166) cited two meta-analyses (Leyland, 2005; Riva, Gauvin, & Barnett, 2007), which they claimed provided support for the idea that “the aggregated characteristics of communities . . . account for a significant portion of variation in health outcomes,” but both of those meta-analyses were based predominantly on small and relatively homogeneously-sized geographical areas (and Leyland’s study examined only health-related *behaviors*, not outcomes). The approximate population of each area in Leyland’s study was 5,000, while Riva et al. reported a median area population of 8,600; compare these with Eichstaedt et al.’s mean county population of 206,000 and median of 78,000. As Beyer, Schultz, and Rushton (2008, p. 40) put it, “The county often represents an area too large to use in determining true, local patterns of disease.”

Working with aggregated data sets, such as Eichstaedt et al.’s (2015a) county-level data, also raises the question of whether specific data items can be aggregated in a meaningful way to create a collective “characteristic” (cf. Subramanian, Jones, Kaddour, & Krieger, 2008). For example, it is difficult to know what interpretation to place on the median individual household income of a county, especially one with a highly diverse population. It

is also worth noting that, as reported by Eichstaedt et al. in their Supplemental Tables document (2015c), the “county-level” data for all of the variables that measure “county-level” health in their study (obesity, hypertension, diabetes, and smoking) are in fact statistical estimates derived from state-level data using “Bayesian multilevel modeling, multilevel logistic regression models, and a Markov Chain Monte Carlo simulation method” (p. DS7). However, Eichstaedt et al. provided no estimates of the possible errors or biases that the use of such techniques might introduce.

Based on the above observations, we decided to reproduce a number of Eichstaedt et al.’s (2015a) analyses, using their original data files in conjunction with other sources of information, to see whether the assumptions made by these authors stand up to a critical examination.

Method

We were able to download all of Eichstaedt et al.’s (2015a) data files from the relevant Open Science Foundation (OSF) repository, and the majority of the analyses that follow are based on these data²⁶. We did not have access to the original Twitter “Garden Hose” data set, so our analyses rely on the summaries of word usage provided by Eichstaedt et al. in their data files.

Next, we downloaded county-aggregated, age-adjusted mortality data for 2009 and 2010 for the ICD-10 category I25.1 (atherosclerotic heart disease) from the Centers for Disease Control and Prevention (CDC) online public health database, known as Wonder (<http://wonder.cdc.gov/>), in order to check that we could reproduce Eichstaedt et al.’s (2015a) data set exactly. We also downloaded comparable mortality data for the ICD-10 categories X60–X84, collectively labeled “Intentional self-harm”—in order to test the idea that suicide might be at least as well predicted by Twitter language as AHD—as well as the data for several other causes of death (including all-cause mortality) for comparison purposes.

²⁶ An earlier version of the present article, available in preprint form at <https://psyarxiv.com/dursw>, stated that we had not been able to obtain Eichstaedt et al.’s (2015a) code because it had not been made available in the same OSF repository as the data. We are happy to acknowledge here that Eichstaedt and colleagues had in fact made their code available on the website of their Differential Language Analysis ToolKit (DLATK) software project, a fact that they have now documented in their recent preprint (Eichstaedt et al., 2018). We followed the installation instructions for DLATK and were able to reproduce the analyses described by Eichstaedt et al. (2015a, p. 161) under the heading of “Predictive models.”

Finally, we obtained data about the precise geographical locations of counties from the US Census Bureau (http://www.census.gov/geo/www/gazetteer/files/Gaz_counties_national.zip). All of our statistical analyses were performed in R; a copy of the analysis code can be found at <https://osf.io/g42dw>.

Results

Using data downloaded from the CDC Wonder database, we were able to reproduce Eichstaedt et al.'s (2015a) principal data file, named `countyoutcomes.csv`, exactly. We were also able to reproduce Eichstaedt et al.'s reported numerical results from their article to within rounding error using our own R code (for the dictionary and topic language variables) or by running their modeling software in a virtual machine environment (for the analyses under the heading of “Predictive models”).

In the following sections, we report a number of findings that we made when exploring Eichstaedt et al.'s (2015a) data set and the other data that we obtained. The order of these sections follows the same structure that we used in the introduction of the present article, showing how these findings relate to the concerns that we expressed earlier under three broad headings: (a) the use of mortality from AHD as the outcome variable; (b) the use of county-level aggregated data; and (c) the use of patterns of language in posts to Twitter as the principal predictor of the outcome.

Variability in ICD-10 coding of cause of death

The validity of the mortality data from the CDC, and in particular the ICD-10 coding of the cause of death, is crucial to establishing the validity of Eichstaedt et al.'s (2015a) findings. Examination of the mortality figures from the counties included in Eichstaedt et al.'s study shows that death rates from AHD in 2009–2010 ranged from 13.4 per 100,000 people (1.55% of all recorded deaths) in East Baton Rouge Parish, LA to 185.0 per 100,000 people (26.11% of all recorded deaths) in Richmond County, NY. It is not clear that any county-level environmental, economic, or social factors that might contribute to the development of AHD would be sufficient to explain why this condition—which, as Eichstaedt et al. noted, is the single most prevalent ICD-10 cause of mortality in the United States—appears to cause 13.8 times more deaths in one county than another. The two counties just mentioned are both among the 10% most populous counties in the US, with a total of 7,001 and 6,907 deaths, respectively, being recorded there in 2009–2010; this suggests that the large difference in recorded mortality from AHD is unlikely to be a statistical fluke due to a limited sample. In contrast, for deaths from cancer, the range of per-

county mortality rates across the 1,347 counties included by Eichstaedt et al. runs from 102.4 to 326.7 per 100,000 people—a factor of 3.19 from lowest to highest, which is little different from the range for all-cause mortality (478.8 to 1390.3 per 100,000 people, a factor of 2.90 from lowest to highest). Eichstaedt et al.’s acknowledgement that “the coding on death certificates may be inconsistent” (p. 166) would thus appear to be somewhat of an understatement, at least as far as AHD is concerned. Indeed, it seems possible that at least part of the variance in AHD mortality (whether this is to be explained by “community-level psychological characteristics” or some other mechanism) might be due, not to differences in the *actual* prevalence of AHD as the principal cause of death, but rather to variations in the propensity of local physicians to *certify* the cause of death as AHD (McAllum, St. George, & White, 2005; Messite & Stellman, 1996; Stausberg, Lehmann, Kaczmarek, & Stein, 2006). Researchers who intend to study AHD mortality using county-level data may wish to take this possibility into account (cf. Roth et al., 2017).

Use of mortality from AHD as the outcome variable

The chronic nature of AHD implies that, to the extent that its prevalence may be affected by deleterious or protective lifestyle and social factors—including Eichstaedt et al.’s (2015a, p. 164) purported “indicators of community-level psychosocial health”—it will have been necessary for these factors to have exerted their effects over a long period. However, there are two potential sources of discrepancy between the current psychosocial state of a person’s county of residence and its past effects on their health. First, the socioeconomic climate of an area can change substantially in less than a generation. The decline in the fortunes of the city of Detroit provides a recent dramatic example of this (LeDuff, 2014), but economic development can also bring rapid positive change to parts of a state within quite a short time. Second, individuals tend not to stay in one place; the county where someone spends his or her childhood may be a long way from where he or she ends up working, and possibly different again from where the same person is living when symptoms of AHD emerge later in life, which might be after retirement. Data from the U.S. Census Bureau (2011) show that in 2010 approximately 3.5% of Americans moved either to another county in the same state, or to another state altogether, a figure that appears from an examination of comparable data from other years to be relatively constant over time. Thus, it seems likely that a substantial proportion of the people who die from AHD each year in any given county may have lived in one or more other counties during the decades when AHD was developing,

and thus been exposed to different forms, both favorable and unfavorable, of Eichstaedt et al.'s purported community-level psychological characteristics during that period.

We suggested earlier that if a county-level psychological factor was sufficiently strong to influence mortality from AHD, it might also be expected to influence local rates of suicide. We therefore examined the relationship of the set of causes of death listed by the CDC as “self-harm” with Twitter language usage, using the procedures reported in the first subsections entitled “Language variables from Twitter” and “Statistical analysis” of Eichstaedt et al.'s (2015a, p. 161) Method section. Because of a limitation of the CDC Wonder database, whereby county-aggregated mortality data are only returned for any given county when at least 10 recorded deaths per year, on average, match the requested criteria for the period in question, data for self-harm were only available for 741 counties; however, these represented 89.9% of the population of Eichstaedt et al.'s set of 1,347 counties.

In the “Dictionaries” analysis, we found that mortality from self-harm was *negatively* correlated with all five “negative” language factors, with three of these correlations (for anger, negative-relationship, and negative-emotion words) being statistically significant at the .05 level (see our Table 6.1). That is, counties whose residents made greater use of negative language on Twitter had lower rates of suicide, or, to borrow Eichstaedt et al.'s (2015a, p. 162) words, use of negative language was “significantly protective” against self-harm; this statistical significance was unchanged when income and education were added as covariates. In a further contrast to AHD mortality, two of the three positive language factors (positive relations and positive emotions) were *positively* correlated with mortality from self-harm, although these correlations were not statistically significant at the conventional .05 level.

Table 6.1. Correlations Between Self-Harm and Twitter Language Measured by Dictionaries.

Language variable	<i>r</i>	<i>p</i>	95% CI
Risk factors			
Anger	−0.169	<.001	[−0.238, −0.099]
Negative relationships	−0.095	.010	[−0.166, −0.023]
Negative emotions	−0.102	.005	[−0.173, −0.030]
Disengagement	0.008	.831	[−0.064, 0.080]
Anxiety	−0.045	.219	[−0.117, 0.027]
Protective factors			
Positive relationships ^a	−0.001	.976	[−0.073, 0.071]
Positive emotions	0.059	.110	[−0.013, 0.131]
Positive engagement	−0.031	.393	[−0.103, 0.041]

^a Following Eichstaedt et al. (2015a), the word *love* was removed from the dictionary for this variable. See discussion in the text.

Next, we analyzed the relationship between Twitter language and self-harm outcomes at the “Topics” level. Among the topics most highly correlated with increased risk of self-harm were those associated with spending time surrounded by nature (e.g., *grand, creek, hike*; $r = .214$, $CI^{27} = [.144, .281]$), romantic love (e.g., *beautiful, love, girlfriend*; $r = .176$, $CI = [.105, .245]$), and positive evaluation of one’s social situation (e.g., *family, friends, wonderful*; $r = .175$, $CI = [.104, .244]$). There were also topics of discussion that appeared to be strongly “protective” against the risk of self-harm, such as baseball (e.g., *game, Yankees, win*; $r = -.317$, $CI = [-.381, -.251]$), binge drinking (e.g., *drunk, sober, hungover*; $r = -.249$, $CI = [-.316, -.181]$), and watching reality TV (e.g., *Jersey, Shore, episode*; $r = -.200$, $CI = [-.269, -.130]$). All of the correlations between these topics and self-harm outcomes, both positive and negative, were significant at the same Bonferroni-corrected significance level (i.e., $.05/2,000 = .000025$) used by Eichstaedt et al. (2015a), and remained significant at that level after adjusting for income and education. That is, several topics that were ostensibly

²⁷ We report 95% CIs here for consistency with Eichstaedt et al. (2015a); however, given their use of a Bonferroni-corrected significance threshold, it could be argued that Eichstaedt et al. should have reported 99.9975% CIs.

associated with “positive,” “eudaimonic” approaches to life predicted higher rates of county-level self-harm mortality, whereas apparently hedonistic topics were associated with lower rates of self-harm mortality, and the magnitude of these associations was at least as great—and in a few cases, even greater—than those found by Eichstaedt et al. These topics are shown in “word cloud” form (generated at <https://www.jasondavies.com/wordcloud/>) in our Figure 6.1 (cf. Eichstaedt et al.’s Figure 1).

This would seem to pose a problem for Eichstaedt et al.’s (2015a, p. 166) claim to have shown the existence of “community-level psychological factors that are important for the cardiovascular health of communities.” Apparently the “positive” versions of these factors, while acting via some unspecified mechanism to make the community as a whole less susceptible to developing hardening of the arteries, also simultaneously manage to make the inhabitants of those counties more likely to commit suicide, and vice versa. We suggest that more research into the possible risks of increased levels of self-harm might be needed before “community-level psychological factors” were to be made the focus of intervention, as Eichstaedt et al. suggested in the final sentence of their article.²⁸

²⁸ In their recent preprint, Eichstaedt et al. (2018) noted that the relation between self-harm and positive Twitter language disappeared when they added measures of county-level rurality and elevation as covariates. We do not find this surprising. Our point is not that Twitter language actually predicts county-level mortality from suicide (or AHD); rather, it is that with a sufficient number of predictors, combined with unreliability in the measurement of these, one can easily find spurious relations between variables (cf. Westfall & Yarkoni, 2016).

Twitter Topics Positively Correlated With County-Level Self-Harm Mortality

A word cloud featuring various positive words in different colors and sizes. The words include: 'thankful' (large, dark red), 'loving' (large, yellow), 'family' (large, yellow), 'friends' (large, yellow), 'amazing' (large, dark red), 'wonderful' (large, dark red), 'blessed' (large, yellow), 'grateful' (medium, blue), 'caring' (medium, green), 'daughter' (medium, red), 'husband' (medium, red), 'boyfriend' (medium, dark red), 'supportive' (medium, purple), 'lucky' (medium, dark red), 'fortunate' (medium, blue), 'truly' (medium, purple), 'helped' (medium, blue), 'surrounded' (medium, blue), and 'incredible' (small, green).

A word cloud of adjectives describing a daughter. The words are arranged in a circular pattern. The most prominent words are 'beautiful' (large, dark blue) and 'love' (large, orange). Other words include 'perfect' (purple), 'sweet' (purple), 'girlfriend' (pink), 'wonderful' (dark red), 'daughter' (red), 'girl' (orange), 'inside' (dark blue), 'loving' (dark red), 'amazingly' (pink), 'talented' (dark blue), 'woman' (dark red), 'struggle' (green), 'wife' (blue), 'angel' (blue), 'adore' (green), 'obsess' (green), 'concrete' (green), and 'person' (green).

trail hike headed springs park
grandcreek
national forest lake river
falls valley mountains
oak cave hiking total
hills
adventure

Twitter Topics Negatively Correlated With County-Level Self-Harm Mortality

jersey

episode

ronnie

watchin

reunion

pump

shore

finale

tshirt

situation

checked

ron

nj

cast

fist

chump

ronnie

italia

practices

rivals

A word cloud featuring various terms related to drinking and hangovers. The words are arranged in a dynamic, overlapping layout. The most prominent words are 'drunk' in large yellow letters at the top right, 'hungover' in large dark red letters on the left, and 'sober' in large tan letters in the center. Other visible words include 'drunken', 'drinking', 'wasted', 'tipsy', 'hangover', 'slightly', 'bar', 'parties', 'drink', 'stoned', 'fucked', 'alcohol', 'whilst', 'partying', 'piss', and 'drank'. The colors used for the words include yellow, dark red, tan, blue, green, and purple.

Baseball, $r = -.317$

Bias caused by selection of counties

As noted above, the CDC Wonder database returns county-aggregated mortality data for any given cause of death only for those counties where at least 10 deaths from that cause were recorded per year, on average, during the period covered by the user's request. This cutoff means that Eichstaedt et al.'s (2015a) data set is skewed toward counties having higher rates of AHD. For example, AHD mortality was sufficiently high in Plumas County, CA (2010 population 20,007; 25 deaths from AHD in 2009–2010) for this county to be included, whereas the corresponding prevalence in McKinley County, NM (2010 population 71,492; 17 deaths from AHD in 2009–2010), as well as 130 other counties with higher populations than Plumas County but fewer than 20 deaths from AHD, was not. Thus, the selection of counties tends to include those with higher levels of the outcome variable, which has the potential to introduce selection bias (Berk, 1983). For example, counties with a 2010 population below the median (78,265) in Eichstaedt et al.'s data set had significantly higher AHD mortality rates than counties with larger populations (53.21 versus 49.87 per 100,000; $t(1344.6) = 3.12$, $p = .002$, $d = 0.17$). The effect of such bias on the rest of Eichstaedt et al.'s analyses is hard to estimate, but given that one of these authors' headline results—namely, that their Twitter-only model predicted AHD mortality “significantly better” than a “traditional” model, a claim deemed sufficiently important to be included in their abstract—had an associated p value of .049, even a very small difference might be enough to tilt this result away from traditional statistical significance.

Problems associated with county-level aggregation of data

We noted earlier that the diversity among counties made it difficult to imagine that the relation between Twitter language and AHD would be consistent across the entire United States. Indeed, when the counties in Eichstaedt et al.'s (2015a) data set are split into two equal subsets along the median latitude of their centroids (38.6484 degrees North, a line that runs approximately through the middle of the states of Missouri and Kansas), the purported effect of county-level Twitter language on AHD mortality as measured by Eichstaedt et al.'s dictionaries seems to become stronger in the northern half of the US than for the country as a whole, but mostly disappears in the southern half (see our Table 6.2). There does not appear to be an obvious theoretical explanation for this effect; if anything one might expect the opposite, in view of the observation made previously that counties may play a greater role in people's lives in the South.

Table 6.2. Partial Correlations Between Atherosclerotic Heart Disease (AHD) Mortality and Twitter Language Measured by Dictionaries, Across the Northern and Southern Halves of the United States.

Language variable	North			South		
	Partial r	p	95% CI	Partial r	p	95% CI
Risk factors						
Anger	0.240	<.001	[0.168, 0.310]	−0.020	.604	[−0.095, 0.056]
Negative relationships	0.156	<.001	[0.081, 0.229]	0.060	.121	[−0.016, 0.135]
Negative emotions	0.108	.005	[0.032, 0.182]	0.028	.462	[−0.047, 0.104]
Disengagement	0.166	<.001	[0.092, 0.239]	−0.012	.750	[−0.088, 0.063]
Anxiety	0.017	.654	[−0.058, 0.093]	0.104	.007	[0.028, 0.178]
Protective factors						
Positive relationships ^a	−0.032	.411	[−0.107, 0.044]	0.111	.004	[0.035, 0.185]
Positive emotions	−0.166	<.001	[−0.238, −0.091]	0.082	.034	[0.006, 0.156]
Positive engagement	−0.192	<.001	[−0.264, −0.119]	0.041	.288	[−0.035, 0.116]

Notes:

Partial r : partial correlation coefficients obtained from a regression predicting AHD from the Twitter theme represented by the language variable, with county-level education and income as control variables.

^a : Following Eichstaedt et al. (2015a), the word *love* was removed from the dictionary for this variable. See discussion in the text.

A further issue with Eichstaedt et al.'s (2015a) use of data aggregated at the level of counties is that it resulted in an effective sample size that was much smaller than these authors suggested. For example, Eichstaedt et al. compared their model to “[t]raditional approaches for collecting psychosocial data from large representative samples . . . [which] are based on *only* [emphasis added] thousands of people” (p. 166). This suggests that these authors believed that their approach, using data that was generated by an unspecified (but implicitly very large) number of different Twitter users, resulted in a more representative data set than one built by examining the behaviors and health status of “only” a few thousand individuals. However, by aggregating their Twitter data at the county level, and merging it with other county-level health and demographic information, Eichstaedt et al. reduced each of their variables of interest to a single number for each county, regardless of that county’s population. In effect, Eichstaedt et al.’s data set contains a sample of only 1,347 individual units of analysis, each of which has the same degree of influence on the conclusions of the study. A corollary of this is that, despite the apparently large number of participants overall, a very small group of voluble Twitter users could have a substantial influence in smaller counties. For example, in tweets originating from Jefferson County, WA (population 29,872) just nine instances of the word “fuck” (or derivatives thereof) appear in the entire data set, whereas the same word and its variants appear 9,271 times in the tweets of the residents of Montgomery County, NY (population 50,219), a per-inhabitant rate that is almost 600 times larger. Perhaps community standards of polite discourse vary rather more widely across rural areas of the United States than most people might imagine, but it seems at least equally likely that just a few angry people in mid-state New York are responsible for this avalanche of social media profanity, and that their input may consequently have had rather more impact on Eichstaedt et al.’s results than if those same few individuals had been living in Los Angeles County, CA (population 9,818,605). This aggregation into counties calls into question Eichstaedt et al.’s claim (p. 166) that an analysis of Twitter language can “generate estimates based on 10s of millions of people”; indeed, it could be that their results are being driven by just a few hundred particularly active Twitter users, particularly those living in smaller counties¹.

¹ Examination of Eichstaedt et al.’s (2015a) data set shows that for 267 counties (19.8%), less than 100,000 words were included in the database, which corresponds to around 5,000 tweets. However, no indication is available of the number of unique Twitter users per county.

Apparent censorship of the Twitter data

An examination of the words and phrases that appear in Eichstaedt et al.'s (2015a) dictionaries suggests that some form of censorship may have been applied to the Twitter data. For example, the list of frequencies of each word in the dictionary contains entries for the words “nigga,” “niggas,” and “niggaz” (which, between them, appeared 1,391,815 times in the analyzed tweets, and at least once in all but five of the 1,347 included counties), but not “nigger” or “niggers.” It seems highly unlikely that the most common spelling² of this word would not appear even once in any of the 148 million tweets that were included in Eichstaedt et al.'s analyses. A number of other common epithets for ethnic and religious groups are also absent from the dictionary–frequency table—although a search for such words on Twitter suggests that they are in relatively common usage—as are the words “Jew[s]” and “Muslim[s].” In Appendix B of their recent preprint, Eichstaedt et al. (2018) have indicated that they were previously unaware of this issue, which we therefore presume reflects a decision by Twitter to bowdlerize the “Garden Hose” data set. Such a decision clearly has substantial consequences for any attempt to infer a relation between the use of hostile language on Twitter and health outcomes, which requires that the tweets being analyzed are truly representative of the language being used. Indeed, it could be argued that there are few better examples of language that expresses interpersonal hostility than invective directed towards ethnic or religious groups³.

Potential sources of bias in the “Topics” database

A further potential source of bias in Eichstaedt et al.'s (2015a) analyses, which these authors did not mention in their article or their supplemental documents (Eichstaedt et al., 2015b, 2015c), is that their “Topics” database was derived from posts on Facebook (i.e., not Twitter) by a completely different sample of users, as can be seen at the site from which we downloaded this database (<http://wwbp.org/data.html>). Furthermore, some of the topics that

² The spelling “nigga” is often used by African Americans in a neutral or positive sense; the form “nigger” is the one typically used by members of other groups as a racial slur (Goudet, 2013).

³ We assume that the apparent omission of “Jew[s]” and “Muslim[s]” was motivated by concerns that at least some of the tweets mentioning these words might be expressing hostility towards these groups.

were highlighted by Eichstaedt et al. in the word clouds⁴ in their Figure 1 contain words that appear to directly contradict the topic label (e.g., “famous” and “lovers” in “Hate, Interpersonal Tension,” left panel; “enemy” in “Skilled Occupations,” middle panel; “painful” in “Positive Experiences,” left panel). There are also many incorrectly spelled words, as in topic #135 (“can’t, wait, afford, move, believe [*sic*], concentrate”), topic #215 (“wait, can’t, till, tomorrow, meet, tomorrow [*sic*]”), topic #467 (“who’s, guess, coming, thumbs, guy, whos [*sic*], idea, boss, pointing”), and many topics make little sense at all, such as #824 (“tooo, sooo, soooo, alll, sooooo, toooo, goood, meee, meeee, youuu, gooo, soooooo, allll, goood, ohhh, ughhh, ohhhh, gooooo, mee, soooooo”) and #983 (“ur, urself, u’ll, coz, u’ve, cos, urs, bcoz, wht, givin”). The extent to which these automatically extracted topics from Facebook really represent coherent psychological or social themes that might appear with any frequency in discussions on Twitter seems to be questionable, in view of the different demographics and writing styles on these two networks.

Flexibility in interpretation of dictionary data

A problem for Eichstaedt et al.’s (2015a) general argument about the salutary effects of “positive” language was the fact that the use of words expressing positive relationships appeared, in these authors’ initial analyses, to be positively correlated with AHD mortality. To address this, Eichstaedt et al. took the decision to eliminate the word *love* from their dictionary of positive-relationship words. Their justification for this was that “[r]eading through a random sample of tweets containing *love* revealed them to be mostly statements about loving things, not people” (p. 165). However, similar reasoning could be applied to many other dictionary words—including those that featured in results that did not contradict Eichstaedt et al.’s hypotheses—with the most notable among these being, naturally, *hate*. In fact, it turns out that *hate* dominated Eichstaedt et al.’s negative relationships dictionary (41.6% of all word occurrences) to an even greater degree than *love* did for the positive relationships dictionary (35.8%). We therefore created an alternative version of the negative relationships dictionary, omitting the word *hate*, and found that, compared to the original, this version was far less likely to produce a statistically significant regression model when predicting AHD mortality (e.g., regressing AHD on negative relationships, controlling for income and education: with *hate* included, partial $r = .107$, $p < .001$, 95% CI = [.054, .159];

⁴ It appears that the relative size of the words in Eichstaedt et al.’s (2015a, Figure 1) word clouds is determined by the relative frequency of these words in the Facebook data from which the topics were derived, and does not represent the prevalence of these words in the Twitter data.

with *hate* excluded, partial $r = -.005$, $p = .849$, 95% CI = $[-.057, .048]$). However, we do not have access to the full Twitter “Garden Hose” data set, and are thus unable to establish whether, analogously to what Eichstaedt et al. did for *love*, a similar examination of a “random sample of tweets” containing the word *hate* might “reveal” that they are mostly statements about hating things rather than people, thus providing an equivalent justification for dropping this word from the analyses.

In their Note 6, Eichstaedt et al. (2015a, p. 167) mentioned another justification for removing the word *love* from the positive relationships dictionary. They noted that “in lower-SES areas, users [may] share more about personal relationships on Twitter, which distorts the results obtained when using the original positive-relationships dictionary.” Of course, it *might* be true of personal relationships, or indeed any other aspect of people’s lives, that those who live in lower-SES areas—or, for that matter, those who are married, or smoke, or suffer from diabetes—tend to communicate more (or, indeed, less) about that topic on Twitter. But the factor analysis in Eichstaedt et al.’s Note 6 does not provide any direct evidence for their claim of a possible relation between residence in a lower SES area and a tendency to tweet about personal relationships⁵.

Comparison of Twitter-based and “traditional” prediction models

Eichstaedt et al. (2015a, p. 161) reported that they “created a single model in which all of the word, phrase, dictionary, and topic frequencies were independent variables and the AHD mortality rate was the dependent variable.” It is not clear exactly how this model was constructed or what weighting was given to the various components, even though the numbers of each category (words, phrases, dictionary entries, and topics) vary widely. In their Results section, under the heading “Predictions,” these authors compared the performance of what we might call their “Twitter omnibus” model with what they described as “traditional” models (i.e., those based on their demographic, socioeconomic, and health status variables), and claimed that the performance of the Twitter-based model was “significantly” better, based on p values of .026 and .049. However, the degrees of freedom here (1,346) are sufficiently numerous, and thus the statistical power to detect an effect sufficiently high, that these p values arguably constitute quite strong evidence *in favor* of the null hypothesis of no effect (cf. Lakens & Evers, 2014). It is also unclear from Eichstaedt et

⁵ Eichstaedt et al. (2018) have recently explained that the exclusion of “love” from their positive relationships dictionary in their earlier article (Eichstaedt et al., 2015a) was the result of discussions with a reviewer of that article.

al.'s (2015b) Supplemental Material exactly how many predictors were finally included in their model.

How similar are the comparative maps?

Figure 3 of Eichstaedt et al.'s (2015a) article shows two maps of the counties of the northeastern United States, with those counties that had sufficient cases of AHD mortality to be included in their sample being color-coded according to either CDC-reported or Twitter-predicted AHD mortality. Eichstaedt et al. (2015a, p. 164) claimed that “a high degree of agreement is evident” between the two maps. We set out to evaluate this claim by examining the color assigned to each county on each of the two maps, and determining the degree of difference between the mortality rates corresponding to those colors. To this end, we wrote a program to extract the colors of each pixel across the two maps, convert these colors to the corresponding range of AHD mortality rates, and draw a new map that highlights the differences between these rates using a simple color scheme. The color-based scale shown at the bottom of Eichstaedt et al.'s Figure 3 seems to imply that the maps are composed of 10 different color codes, each representing a decile of per-county AHD mortality, but in fact this scale is somewhat misleading. In fact, 14 different colors (plus white) are used for the counties in the maps in Eichstaedt et al.'s Figure 3, with each color apparently (assuming that each color corresponds to an equally-sized interval) representing around 7 percentile rank places, or what we might call a “quattuordecile,” of the AHD mortality distribution. For more than half (323 out of 608, or 53.1%) of the counties that have a color other than white in Eichstaedt et al.'s maps, the difference between the two maps is three or more of these 7-point “color intervals,” as shown in our Figure 6.2. Within these counties, the mean discrepancy is at least 29.5 percentage points, and probably considerably higher⁶. It is therefore questionable whether it is really the case that “[a] high degree of agreement is

⁶ More precisely, 200 counties (32.9%) had a discrepancy of three or more color intervals, while 123 counties (20.2%) had a discrepancy of six or more. Assuming for simplicity that rounding error is uniformly distributed, a difference of three intervals corresponds to a mean difference of 21.4 percentage points, and a difference of six intervals to a mean difference of 42.8 percentage points. Thus, even with the extremely conservative simplifying assumption that “three (six) or more color intervals” actually means “exactly three (six) intervals,” the mean discrepancy across these counties is $((200 \times 21.4) + (123 \times 42.8)) / 323 = 29.5$ percentage points. Note also that the possible extent of the discrepancy is bounded at between 7 and 13 color intervals, depending on the relative positions of the two counties along the 1–14 scale.

evident” (Eichstaedt et al., 2015, p. 164) between the two maps, such that one might use the Twitter-derived value to predict AHD mortality for any practical purpose.

Figure 6.2. Difference Between the Two Maps of AHD Mortality Rates (CDC-reported and Twitter-predicted) From Eichstaedt et al.’s (2015a) Figure 3.

Note: Green indicates a difference of 0–2 color-scale points (see discussion in the text) between the two maps; yellow, a difference of 3–5 points; red, a difference of 6 or more points. Of the 608 colored areas, 123 (20.2%) are red and 200 (32.9%) are yellow.

Discussion

In the preceding paragraphs, we have examined a number of aspects of Eichstaedt et al.'s (2015a) claims about the county-level effects of Twitter language on AHD mortality, using for the most part these authors' own data. We have shown that many of these claims are, at the least, open to alternative interpretations. The coding of their outcome variable (mortality from AHD) is subject to very substantial variability; the process that selects counties for inclusion is biased; the same regression and correlation models "predict" suicide at least as well as AHD mortality but with almost opposite results (in terms of the valence of language predicting positive or negative outcomes) to those found by Eichstaedt et al.; the Twitter-based dictionaries appear not to be a faithful summary of the words that were actually typed by users; arbitrary choices were apparently made in some of the dictionary-based analyses; the topics database was derived from a completely different sample of participants who were using Facebook, not Twitter; there are numerous problems associated with the use of counties as the unit of analysis; and the predictive power of the model, including the associated maps, appears to be questionable. While we were able to reproduce—at a purely computational level—the results of Eichstaedt et al.'s advanced prediction model, based on ridge regression and *k*-fold cross-validation, we do not believe that this model can address the problems of validity and reliability posed by the majority of the points just mentioned. In summary, the evidence for the existence of community-level psychological factors that determine AHD mortality better than traditional socioeconomic and demographic predictors seems to be considerably less strong than Eichstaedt et al. claimed.

A tourist, or indeed a field anthropologist, driving through Jackson County (2010 population 42,376) and Clay County (2010 population 26,890) in southern Indiana might not notice very much difference between the two. According to the US Census Bureau (<https://factfinder.census.gov/>) these counties have comparable population densities, ethnic make-ups, and median household incomes. It seems unlikely that there would be a large variation in the "norms, social connectedness, perceived safety, and environmental stress, that contribute to health and disease" (Eichstaedt et al., 2015a, pp. 159–160) between these two rural Midwestern communities. Yet, according to Eichstaedt et al.'s data, the levels of anger and anxiety expressed on Twitter were more than 12 and 6 times, respectively, higher in Jackson County than in Clay County. These differences are not easy to explain; any community-level psychological characteristics that might be driving them must obey some strange properties. Such characteristics would have to operate, at least partly, in ways that are

not accounted for by variables for which Eichstaedt et al. applied statistical controls (such as income and smoking prevalence), yet presumably they must have some physical manifestation in order to be able to have an effect on people's expressed feelings. It is difficult to imagine how such a characteristic might have gone unnoticed until now, yet be able to cause people living in very similar socioeconomic conditions less than 100 miles apart in the same state to express such varying degrees of negative emotionality. A more parsimonious explanation, perhaps, is that there is a very large amount of noise in the measures of the meaning of Twitter data used by Eichstaedt et al., and these authors' complex analysis techniques (involving, for example, several steps to deal with high multicollinearity) are merely modeling this noise to produce the illusion of a psychological mechanism that acts at the level of people's county of residence. Certainly, the different levels of "negative" Twitter language between these two Indiana counties appear to have had no deleterious differential effect on local AHD mortality; indeed, at 45.4 deaths per 100,000 inhabitants, "angry" Jackson County's AHD mortality rate in 2009–2010 was 23.7% *lower* than "laid-back" Clay County's (59.5 per 100,000 inhabitants). As we showed in our analysis of Eichstaedt et al.'s comparative maps, this failure of Twitter language to predict AHD mortality with any reliability is widespread.

In a recent critique, Jensen (2017) examined the claims made by Mitchell, Frank, Harris, Dodds, and Danforth (2013) regarding the ability of Twitter data to predict happiness. Jensen argued that "the extent of overlap between individuals' online and offline behavior and psychology has not been well established, but there is certainly reason to suspect that a gap exists between reported and actual behavior" (p. 2). Jensen went on to raise a number of other points about the use of Twitter's "garden hose" data set that appear to be equally applicable to Eichstaedt et al. (2015a), concluding that "When researchers approach a data set, they need to understand and publicly account for not only the limits of the data set, but also the limits of which questions they can ask . . . and what interpretations are appropriate" (p. 6). It is worth noting that Mitchell et al. were attempting to predict happiness only among the people who were actually sending the tweets that they analyzed. While certainly not a trivial undertaking, this ought to be considerably less complex than Eichstaedt et al.'s attempt to predict the health of one part of the population from the tweets of an entirely separate part (cf. their comment on p. 166: "The people tweeting are not the people dying"). Hence, it would appear likely that Jensen's conclusions—namely, that the limitations of secondary data analyses and the inherent noisiness of Twitter data meant that Mitchell et al.'s claims about their ability to predict happiness from tweets were not reliably supported by the evidence—

would be even more applicable to Eichstaedt et al.'s study, unless these authors could show that they took steps to avoid the deficiencies of Mitchell et al. On a related theme, Robinson-Garcia, Costas, Isett, Melkers, and Hicks (2017) warned that bots, or humans tweeting like bots, represent a considerable challenge to the interpretability of Twitter data; this theme has become particularly salient in view of recent claims that a substantial proportion of the content on Twitter and other social media platforms may not represent the spontaneous output of independent humans (Varol, Ferrara, Davis, Menczer, & Flammini, 2017).

The principal theoretical claim of Eichstaedt et al.'s (2015a) article appears to be that the best explanation for the associations that were observed between county-level Twitter language and AHD mortality is some geographically-localized psychological factor, shared by the inhabitants of an area, that exerts⁷ a substantial influence on aspects of human life as different as vocabulary choice on social media and arterial plaque accumulation, independently of other socioeconomic and demographic factors. However, Eichstaedt et al. did not provide any direct evidence for the existence of these purported community-level psychological characteristics, nor of how they might operate. Indeed, we have shown that the same techniques that predicted AHD mortality could equally well have been used to predict county-level suicide prevalence, with the difference that *higher* rates of self-harm seem to be associated with "positive" Twitter language. Of course, there is no suggestion that the study of the language used on Twitter by the inhabitants of any particular county has any real predictive value for the local suicide rate; we believe that such associations are likely to be the entirely spurious results of imperfect measurements and chance factors, and to use Twitter data to predict which areas might be about to experience higher suicide rates is likely to prove extremely inaccurate (and perhaps ethically questionable as well). We believe that it is up to Eichstaedt et al. to show convincingly why these same considerations do not apply to their analyses of AHD mortality; as it stands, their article does not do this. Taken in conjunction

⁷ Eichstaedt et al. (2015a) included a disclaimer about causality on p. 166 of their article. However, we feel that this did not adequately compensate for some of their language elsewhere in the article, such as "Local communities *create* [emphasis added] physical and social environments that *influence* [emphasis added] the behaviors, stress experiences, and health of their residents" (p. 166; both of the italicized words here seem to us to imply causation at least as strongly as our word "exerts"), and "Our approach . . . could bring researchers closer to understanding the community-level psychological factors that are important for the cardiovascular health of communities and should become the focus of intervention" (p. 166, seemingly implying that an intervention to change these psychological factors would be expected to lead to a change in cardiovascular health).

with the pitfalls (Westfall & Yarkoni, 2016) of including imperfectly-measured covariates (such as Eichstaedt et al.’s county-level measures of smoking and health status, as described above) in regression models, and the likely presence of numerous substantial but meaningless correlations in any data set of this type (the “crud factor”; Meehl, 1990⁸), it seems entirely possible that Eichstaedt et al.’s conclusions might be no more than the result of fitting a model to noise.

Conclusions

It appears that the existence of community-level psychological characteristics—and their presumed valence, either being “protective” or “risk” factors for AHD mortality—was inferred by Eichstaedt et al. (2015a) from the rejection of a series of statistical null hypotheses which, though not explicitly formulated by these authors, appear to be of the form “There is no association between the use of [a ‘positive’ or ‘negative’ language element] by the Twitter users who live in a given county, and AHD-related mortality among the general population of that county.” Yet, the rejection of a statistical null hypothesis cannot in itself justify the acceptance of any particular alternative hypothesis (Dienes, 2008)—especially one as vaguely specified as the existence of Eichstaedt et al.’s purported county-level psychological characteristics that operate via some unspecified mechanism—in the absence of any coherent theoretical explanation. Indeed, it seems to us that Eichstaedt et al.’s results could probably equally well be used to justify the claim that the relation between Twitter language and AHD mortality is being driven by county-level variations in almost any phenomenon imaginable. To introduce a new psychological construct without a clear definition, and whose very existence has only been inferred from a correlational study—as Eichstaedt et al. did—is a very risky undertaking indeed.

⁸ For example, using data from the CDC for the 2009–2010 period, county-level mortality from assault is strongly correlated with county-level mortality from cancer ($r = .55$), but completely uncorrelated with county-level mortality from AHD ($r = .00$). There seems to be no obvious theoretical explanation for these results.

References

- Abrams, S. J., & Fiorina, M. P. (2012). "The Big Sort" that wasn't: A skeptical reexamination. *PS: Political Science & Politics*, 45, 203–210. <https://doi.org/10.1017/S1049096512000017>
- Association for Psychological Science. (2015). *Language on Twitter tracks rates of coronary heart disease*. Retrieved from <http://www.psychologicalscience.org/news/releases/twitter-usage-can-predict-rates-of-coronary-heart-disease.html>
- Benton, J. E. (2002). *Counties as service delivery agents: Changing expectations and roles*. Westport, CT: Praeger.
- Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, 48, 386–398.
- Beyer, K. M. M., Schultz, A. F., & Rushton, G. (2008). Using ZIP® codes as geocodes in cancer research. In G. Rushton, M. P. Armstrong, J. Gittler, B. R. Greene, C. E. Pavlik, M. M. West, & D. L. Zimmerman (Eds.), *Geocoding health data: The use of geographic codes in cancer prevention and control, research, and practice* (pp. 37–67). Boca Raton, FL: CRC Press.
- Clark, A. M., DesMeules, M., Luo, W., Duncan, A. S., & Wielgosz, A. (2009). Socioeconomic status and cardiovascular disease: Risks and implications for care. *Nature Reviews Cardiology*, 6, 712–722. <https://doi.org/10.1038/nrcardio.2009.163>
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015a). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26, 159–169. <https://doi.org/10.1177/0956797614557867>
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015b). *Supplemental Method*. Retrieved from <http://journals.sagepub.com/doi/suppl/10.1177/0956797614557867>
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015c). *Supplemental Tables*. Retrieved from http://wwbp.org/papers/PsychSci2015_HeartDisease_Tables.pdf
- Eichstaedt, J. C., Schwartz, H. A., Giorgi, S., Kern, M. L., Park, G., Sap, M., . . . Ungar, L. H. (2018). *More evidence that Twitter language predicts heart disease: A response and replication*. Retrieved from <https://psyarxiv.com/p75ku>
- Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., . . . Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, 143, 187–232. <https://doi.org/10.1037/bul0000084>
- Friedman, M., & Rosenman, R. (1959). Association of specific overt behaviour pattern with blood and cardiovascular findings. *Journal of the American Medical Association*, 169, 1286–1296. <https://doi.org/10.1001/jama.1959.03000290012005>
- Funk, S. D., Yurdagul, A., Jr., & Orr, A. W. (2012). Hyperglycemia and endothelial dysfunction in atherosclerosis: Lessons from type 1 diabetes. *International Journal of Vascular Medicine*, 2012, 569654. <https://doi.org/10.1155/2012/569654>
- Goudet, L. (2013). Alternative spelling and censorship: The treatment of profanities in virtual communities. In D. Jamet & M. Jobert (Eds.), *Aspects of linguistic impoliteness* (pp. 209–222). Cambridge, England: Cambridge Scholars

- Haider-Merkel, D. P. (2009). *Political encyclopedia of U.S. states and regions*. New York, NY: CQ Press.
- Hoyert, D. L. (2011). *The changing profile of autopsied deaths in the United States, 1972–2007 (NCHS data brief no. 67)*. Hyattsville, MD: National Center for Health Statistics. Retrieved from <https://www.cdc.gov/nchs/data/databriefs/db67.pdf>
- Izadi, E. (2015, January 21). Tweets can better predict heart disease rates than income, smoking and diabetes, study finds. *The Washington Post*. Retrieved from <https://www.washingtonpost.com/news/to-your-health/wp/2015/01/21/tweets-can-better-predict-heart-disease-rates-than-income-smoking-and-diabetes-study-finds>
- Jacobs, T. (2015, January 21). Happier tweets, healthier communities. *Pacific Standard*. Retrieved from <https://psmag.com/environment/happier-tweets-healthier-communities-98710>
- Jensen, E. A. (2017). Putting the methodological brakes on claims to measure national happiness through Twitter: Methodological limitations in social media analytics. *PLoS ONE*, 12(9), e0180080. <https://doi.org/10.1371/journal.pone.0180080>
- Kilgore, E. (2012, March 20). Counties as communities. *Washington Monthly*. Retrieved from http://www.washingtonmonthly.com/political-animal-a/2012_03/counties_as_communities036175.php
- Kuper, H., Marmot, M., & Hemingway, H. (2002). Systematic review of prospective cohort studies of psychosocial factors in the aetiology and prognosis of coronary heart disease. *Seminars in Vascular Medicine*, 2, 267–314. <https://doi.org/10.1055/s-2002-35401>
- Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, 9, 278–292. <https://doi.org/10.1177/1745691614528520>
- Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, 343, 1203–1205. <https://doi.org/10.1126/science.1248506>
- LeDuff, C. (2014). *Detroit: An American autopsy*. New York, NY: Penguin.
- Leyland, A. H. (2005). Socioeconomic gradients in the prevalence of cardiovascular disease in Scotland: The roles of composition and context. *Journal of Epidemiology & Community Health*, 59, 799–803. <https://doi.org/10.1136/jech.2005.034017>
- Mant, J., Wilson, S., Parry, J., Bridge, P., Wilson, R., Murdoch, W., . . . Warfield, A. (2006). Clinicians didn't reliably distinguish between different causes of cardiac death using case histories. *Journal of Clinical Epidemiology*, 59, 862–867. <http://dx.doi.org/10.1016/j.jclinepi.2005.11.021>
- McAllum, C., St. George, I., & White, G. (2005). Death certification and doctors' dilemmas: A qualitative study of GPs' perspectives. *The British Journal of General Practice*, 55, 677–683.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195–244. <https://doi.org/10.2466/pr0.1990.66.1.195>
- Messite, J., & Stellman, S. D. (1996). Accuracy of death certificate completion: The need for formalized physician training. *Journal of the American Medical Association*, 275, 794–796. <https://doi.org/10.1001/jama.1996.03530340058030>

- Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. (2013). The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), e64417. <https://doi.org/10.1371/journal.pone.0064417>
- O'Connor, R. C., & Nock, M. K. (2014). The psychology of suicidal behaviour. *The Lancet Psychiatry*, 1, 73–85. [https://doi.org/10.1016/S2215-0366\(14\)70222-6](https://doi.org/10.1016/S2215-0366(14)70222-6)
- Ormel, J., Rosmalen, J., & Farmer, A. (2004). Neuroticism: A non-informative marker of vulnerability to psychopathology. *Social Psychiatry and Psychiatric Epidemiology*, 39, 906–912. <https://doi.org/10.1007/s00127-004-0873-y>
- Riggio, R. E. (2014, August 14). Are you a type A, B, or D Personality? *Psychology Today*. Retrieved from <https://www.psychologytoday.com/blog/cutting-edge-leadership/201408/are-you-type-b-or-d-personality>
- Riva, M., Gauvin, L., & Barnett, T. A. (2007). Toward the next generation of research into small area effects on health: A synthesis of multilevel investigations published since July 1998. *Journal of Epidemiology & Community Health*, 61, 853–861. <https://doi.org/10.1136/jech.2006.050740>
- Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—about journal articles. *PLoS ONE*, 12(8), e0183551. <https://doi.org/10.1371/journal.pone.0183551>
- Roth, G. A., Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Naghavi, M., . . . Murray, C. J. L. (2017). Trends and patterns of geographic variation in cardiovascular mortality among US counties, 1980–2014. *JAMA*, 317, 1976–1992. <https://doi.org/10.1001/jama.2017.4150>
- Singal, J. (2015, January 22). Can angry tweets predict heart-disease rates? *New York Magazine*. Retrieved from <http://nymag.com/scienceofus/2015/01/can-angry-tweets-predict-heart-disease-rates.html>
- Smith, T. W. (2006). Personality as risk and resilience in physical health. *Current Directions in Psychological Science*, 15, 227–231. <https://doi.org/10.1111/j.1467-8721.2006.00441.x>
- Smith, T. W. (2011). Toward a more systematic, cumulative, and applicable science of personality and health: lessons from type D personality. *Psychosomatic Medicine*, 73, 528–532. <https://doi.org/10.1097/PSY.0b013e31822e095e>
- Stausberg, J., Lehmann, N., Kaczmarek, D., & Stein, M. (2006). Reliability of diagnoses coding with ICD-10. *International Journal of Medical Informatics*, 75, 50–57. <https://doi.org/10.1016/j.ijmedinf.2006.11.005>
- Subramanian, S. V., Jones, K., Kaddour, A., & Krieger, N. (2008). Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology*, 38, 342–360. <https://doi.org/10.1093/ije/dyn359>
- U.S. Census Bureau. (2011, May 23). *Geographical Mobility: 2009 to 2010*. Retrieved from <https://www.census.gov/data/tables/2010/demo/geographic-mobility/cps-2010.html>
- Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017, May). *Online human-bot interactions: Detection, estimation, and characterization*. Paper presented at the 11th International AAAI Conference on Web and Social Media, Montréal, Canada. Retrieved from <https://arxiv.org/abs/1703.03107>

- Welsh, J., Korda, R. J., Joshy, G., Butterworth, P., Brown, A., & Banks, E. (2017). Psychological distress and ischaemic heart disease: cause or consequence? Evidence from a large prospective cohort study. *Journal of Epidemiological and Community Health*, 71, 1074–1079. <https://doi.org/10.1136/jech-2017-209535>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11(3), e0152719. <https://doi.org/10.1371/journal.pone.0152719>
- Wilson, S. (2009). Type A and B personalities. *The Guardian*, March 7. Retrieved from <https://www.theguardian.com/lifeandstyle/2009/mar/07/personality-type-a-type-b>

Easy as (Happiness) Pie? A Critical Evaluation of a Popular Model of the Determinants of Well-Being

Nicholas J. L. Brown

Julia M. Rohrer¹

Journal of Happiness Studies, online publication ahead of print (May 2019).

DOI: 10.1007/s10902-019-00128-4

¹ Max Planck Institute for Human Development, Berlin, Germany

Abstract

An underlying principle behind much of the research in positive psychology is that individuals have considerable leeway to increase their levels of happiness. In an influential article that is frequently cited in support of such claims, Lyubomirsky, Sheldon, and Schkade (2005) put forward a model (subsequently popularized under the name of the “happiness pie”) in which approximately 50% of individual differences in happiness are due to genetic factors and 10% to life circumstances, leaving 40% available to be changed via volitional activities. We re-examined Lyubomirsky et al.’s claims and found several apparent deficiencies in their chain of arguments on both the empirical and the conceptual level. We conclude that there is little empirical evidence for the variance decomposition suggested by the “happiness pie,” and that even if it were valid, it is not necessarily informative with respect to the question of whether individuals can truly exert substantial influence over their own chronic happiness level. We believe that our critical re-examination of Lyubomirsky et al.’s seminal article offers insights into some common misconceptions and pitfalls of scientific inference, and we hope that it might contribute to the construction of a more rigorous and solid empirical basis for the field of positive psychology.

Introduction

One of the key claims of the positive psychology movement is that most individuals are able to exercise substantial control over their own levels of happiness². On this account, people can become durably happier—in Seligman’s (2002, p. ix) words, “go from plus two to plus seven” in their lives—by adopting a variety of strategies, such as expressing gratitude (Emmons & McCullough, 2003), applying one’s character strengths in novel ways (Seligman, Steen, Park, & Peterson, 2005), writing about positive experiences (Burton & King, 2004), or practicing loving-kindness meditation (Fredrickson, 2013). Over the past two decades or so, these ideas have been widely publicized both within scholarly circles and in the popular media (e.g., Scelpo, 2017).

Arguably the most popular model of the factors influencing well-being, and hence people’s ability to improve their own happiness, is the one presented by Lyubomirsky, Sheldon, and Schkade (2005) that has become widely known as the “happiness pie.” This model, reproduced here as Figure 7.1, purports to show the percentage of variance in well-being that is explained by each of three sources of variation (which Lyubomirsky et al. referred to as “factors”): genetic predisposition (which was labeled the “happiness set point” by Lyubomirsky et al.; see also Lykken, 1999; Lykken & Tellegen, 1996), life circumstances, and volitional (intentional) activities. It should be noted that this model attempts to explain “chronic happiness levels,” that is, an experience that is more enduring than momentary fluctuations in well-being and can be captured by retrospective summary judgments or by averaging momentary judgments over a period of months. Lyubomirsky et al. claimed that up to 40% of the variance in well-being levels according to this definition was explained by volitional activities, suggesting further that intentional activity was “arguably [the] most promising means of altering one’s happiness level” (p. 118). According to Lyubomirsky et al., 50% of the variance in well-being is explained by genetic factors and just 10% by “life

² For simplicity, throughout the present article, we assume that the terms “happiness” and “[subjective] well-being” represent the same construct, which can to some extent be captured by “life satisfaction.” Treating these terms as (sometimes) interchangeable appears to be a common practice in the positive psychology literature, as shown by chapter titles such as “Happiness: Also known as ‘life satisfaction’ and ‘subjective well-being’” (Veenhoven, 2011). A discussion of the extent to which these constructs might in fact differ and how they are related is beyond the scope of the present article; see Diener, Scollon, and Lucas (2009) for an overview of this topic.

circumstances,” a term whose meaning and scope we explore in some depth in the present article.

Figure 7.1. The “Happiness Pie” (after Lyubomirsky et al., 2005, p. 116). Original caption: “Three primary factors influencing the chronic happiness level.”

Since the publication of Lyubomirsky et al.’s (2005) article, which according to Google Scholar had 3,217 citations as of August 18, 2019, the “happiness pie”—including the relative proportions of its components, and especially the claim that intentional activity is a far more important determinant of well-being than life circumstances—has become a key element of the discourse surrounding well-being. In the academic sphere, this model has been cited by scholars working not only on topics that might be considered as the traditional domain of positive psychology, such as employee well-being (Page & Vella-Brodrick, 2009), the relation between pleasure and engagement (Schueller & Seligman, 2010), or interventions to enhance happiness in normal populations (Proyer, Ruch, & Buschor, 2013), but also in

research areas as diverse as psychotherapy (Rashid, 2008), aging (Butler & Ciarrochi, 2007), consumer behavior (Hudders & Pandelaere, 2012), negative emotions (Jordan et al., 2011), self-care for psychologists (Wise, Hersh, & Gibson, 2012), heart disease (Huffman et al., 2011), tourism (Nawijn & Mitas, 2012), and intellectual disability (Dykens, 2006). In wider society, the happiness pie has been a prominent element in the popularization of positive psychology, especially in the areas of happiness-related coaching, self-help books (e.g., Achor, 2010; Fredrickson, 2009), and other “personal development” products and services that have sprung up on the periphery of this new subfield. For example, a large cherry pie cut into two parts (60%–40%) features prominently on the front of the hardcover edition of Lyubomirsky’s (2007) book *The How of Happiness: A Scientific Approach to Getting the Life You Want*, and the same image was used by a training company to promote a “Positive Psychology Master Class” with the same author (MentorCoach, 2008). The 40% figure has been widely touted as an empirical fact, as an Internet search for the terms “happiness” and “40%” will quickly reveal. As the New York-based British writer Ruth Whippman put it:

This 40 per cent figure is much quoted in the positive-psychology literature in both academic and popular texts, and represents the field’s great marketing opportunity.

This is the 40 per cent that anyone with a book to sell, a course of coaching to offer or a happiness technique to promote is hoping to co-opt. (Whippman, 2016, p. 195)

Although some researchers have written brief critiques of the “happiness pie” in blog posts (e.g., Bergink, 2015; Kashdan, 2015; Krueger, 2015), it seems that until now no prominent peer-reviewed articles have been published in the formal psychological literature that discuss the origins and validity of Lyubomirsky et al.’s model. Such an exploration is the focus of the present article. First, we consider some conceptual issues that arise when using population-level variance decomposition to make claims about individual-level potential for change. Second, we discuss weaknesses of the “happiness equation” that is implied by the “happiness pie” and Lyubomirsky et al.’s claim that up to 40% of the variance in happiness can be attributed to volitional activities. Third, we re-examine the numerical estimates that Lyubomirsky et al. assigned to the genetic set point and circumstances. We conclude with a summary of our findings and some recommendations for future research directions.

Between-Subjects Variance Decomposition Versus Individual-Level Potential for Change

Lyubomirsky et al.'s (2005) examination of the determinants of chronic happiness levels draws on the well-known logic of variance decomposition. For example, psychologists frequently report how the percentage of variance of a certain outcome that can be explained by a set of predictors (typically using the coefficient of determination, R^2), or the percentage of variance that a new construct can explain over and above established measures (incremental R^2). However, such a variance decomposition can only be interpreted in terms of variation within a certain population (e.g., "10% of the between-subject variance in happiness can be explained by life circumstances"), and cannot be translated into individual-level variation (e.g., "10% of a person's variance in happiness can be explained by that person's life circumstances"), unless it was explicitly established on the within-subject level (which, as we will see later, is not the case for the studies underlying Lyubomirsky et al.'s model)

Lyubomirsky et al. (2005) seem to have been aware of this distinction, as they pointed out that previous research has been dominated by cross-sectional, between-subjects studies (p. 112). However, these authors' language was not always consistent throughout their article; for example, they wrote that their derivation of the 50:10:40 ratio "implies that *one's* [emphasis added] chronic happiness during a particular life period can be increased" (p. 117), which could be taken to suggest that each individual necessarily has a substantial margin within which to improve their well-being. In line with this interpretation, popular self-help books and web sites often use the 50:10:40 ratio as evidence that (all) *individuals* can choose to affect their own well-being substantially. The 40% of a person's happiness that, on this account, can be changed (implicitly for the better) by one's own choices is typically described using phrases such as "providing an excellent opportunity to increase our level of happiness" (Cygan, 2013, p. 22) or "a substantial amount of control to have over one's own well-being" (Vondruska, 2017, p. 11), demonstrating that regardless of what Lyubomirsky et al.'s intended interpretation may have been, many writers have decided to go with the within-subject version. (Later in this article, we discuss these issues in more detail for the benefit of researchers who are interested in investigating well-being at the level of individual subjects.)

Furthermore, even if the numbers used to partition the happiness pie held on the within-subject level, they would still not necessarily imply anything about the *potential* of possible volitional activities or interventions, for two reasons. First, the variance that can be explained by certain factors will necessarily be constrained by the variability of these factors

in the population. For example, imagine that a highly effective method was found that allowed individuals to intentionally raise their chronic happiness levels by multiple points. If this method gained some popularity—say, being adapted by half of the population—then it would, by definition, increase the amount of variance in chronic happiness levels attributable to volitional activities. However, if almost everybody in the population adapted the method, it would no longer explain much variance in chronic happiness levels, because it would have consistently raised the level of the whole population. Hence, the variance attributable to this hypothetical intervention would change, even if its potential to increase an individual’s well-being remained constant. Second, there is no evidence for the proposition that the activities that individuals choose to undertake can actually explain all of the “missing” 40% of the variance in their well-being; this leads us to another major issue with the general idea of the happiness pie, which we explore in the next section.

Does the “Happiness Pie” Add Up?

Lyubomirsky et al. (2005) based their pie chart on the idea that, if 50% of variance in well-being is explained by genetic factors and 10% by circumstances, this “leaves as much as 40% of the variance for intentional activity” (p. 116). While these authors’ arithmetic cannot be faulted, it is not clear that the rest of their reasoning is as impeccable. The variance decomposition suggested by Lyubomirsky et al. implies an underlying “happiness equation,” which bears a resemblance to the formula for multiple linear regression:

$$H = \beta_1 S + \beta_2 C + \beta_3 V \quad (1)$$

where H is happiness (or well-being), S (“set point”) is the genetic component³, C is circumstances, and V is volitional activities. However, equation (1) suffers from several defects, which we investigate in the following sections.

Genes, circumstances, and volitional activities are not independent additive factors

The conceptualization of the happiness pie and its underlying variance decomposition is only correct when all three factors are independent (i.e., they do not interact and do not covary). As Krueger (2015) has pointed out, this assumption is unlikely to hold. As an

³ The meaning of S remains conceptually somewhat unclear throughout Lyubomirsky et al.’s (2005) article, as the authors referred to it as the “happiness set point” but then simply used heritability estimates to gauge its impact. A conceptual discussion of the relation between the concept of an individual “set point” and the percentage of population-level variance in well-being explained by genetic factors is beyond the scope of the present article.

example of an interaction, lower well-being due to anxiety might be the result of the interaction of childhood stress with a genetic predisposition (cf. Swann & Seyle, 2005).

Considering covariation between the factors influencing happiness, the simple additive model ignores the possibility that the effects of genes on happiness might be largely mediated by circumstantial factors and volitional activities (e.g., Lykken, 1999, p. 81). Lyubomirsky et al. (2005) acknowledged that genes influence happiness indirectly through experiences and environments, but they seemed to interpret this only in terms of mitigation of *negative* genetic influences, as in the claim that “unwanted effects of genes could be minimized by active efforts” (p. 114). It seems equally plausible, however, that *positive* genetic influences on happiness might be mediated through behaviors that could be classified as volitional activities. For example, Lyubomirsky et al. mentioned exercising regularly and striving for important personal goals as volitional activities, but both the inclination to perform daily physical activity and the tendency to persevere in one’s efforts seem to be partly heritable (Moore-Harrison & Lightfoot, 2010; Rimfeld, Kovas, Dale, & Plomin, 2016). This is neither particularly surprising (cf. Turkheimer, 2000), nor does it constitute support for genetic determinism, but merely demonstrates that the simple breakdown of the three factors into proportions that sum to 100% is unjustified from an empirical point of view.

Furthermore, the distinction between the slices of the happiness pie might reify common misunderstandings regarding the nature of heritability estimates. While it seems popular to assume that high heritability implies low malleability, this is not the case: A trait can be both highly heritable and malleable at the same time. For example, the heritability of general intelligence is high, but education reliably increases intelligence (Plomin & Deary, 2015; Ritchie & Tucker-Drob, 2018).

Lack of an error term

To the extent that equation (1) corresponds to a multiple regression formula, an element that is obviously missing is an error term. Given the vicissitudes of human subjects, it seems rather optimistic to expect that this would ever be zero⁴, even if someone were to actually conduct a single study to examine the relative contributions of genetic factors, circumstances, and volitional activities to chronic levels of well-being. At the very least,

⁴ Note that assuming that the error term is zero implies that *all* differences in chronic happiness can be explained in a deterministic manner—a very strong claim, to which most psychologists probably would not want to subscribe.

measurement error in H would need to be considered⁵. For example, if one measured H with a reliability of .80, the three factors S , C , and V would add up to 80% at best—the remaining 20% of variance in happiness would be attributable to measurement error and thus unable to be explained. *Assuming* the additive model held true, and *assuming* the estimates for 50% and 10% were true, this would immediately decrease the amount of variance *potentially* attributable to volitional activities to $80\% - 50\% - 10\% = 20\%$.

It is, of course, possible that Lyubomirsky et al. (2005) conceptualized their formula as a thought experiment (cf. Werner, 2018) on the latent variable level (or that they were thinking about estimates of chronic happiness that were averaged over many measurement occasions, thus reducing the measurement error), so that one could assume that the 100% of the “pie” refers to the reliable variance in chronic happiness levels. However, in such a case, the estimates for S and C would need to be based on models that take into account measurement error (which was not the case for the studies from which the estimates were drawn). Holding everything else equal, using such models would revise estimates of the respective percentages for the impact of genes and life circumstances upwards, since measurement error attenuates observed associations and heritability estimates (see Kendler, Karkowski, & Prescott, 1999, for a demonstration of this phenomenon). Consequently, the “leftover” variance attributable to volitional activities would need to be revised downwards once again.

Potentially omitted determinants of happiness

Lyubomirsky et al.’s (2005) argument regarding the potential of volitional activities to influence levels of happiness assumes that genetic factors, circumstances, and volitional activities are the only three possible causes of variance in well-being. This depends on the use of “circumstances” as a catch-all term to include all of the vicissitudes of life that are outside of one’s personal control and that “*happen to people*” (p. 118; emphasis in original). Thus, assuming this definition, factors like *in utero* development effects would need to be considered in the estimate for C . Because a multitude of such factors has likely been omitted due to the nature of the surveys on which the estimate is based, this estimate would almost certainly need to be revised upwards. Again, under the assumption that the subtraction logic of the “happiness pie” is valid, this would thus lead to a smaller estimate of V . Note that

⁵ If the outcome measures are different in each case (e.g., “happiness” and “well-being”), then to the extent that these constructs do not overlap (cf. footnote 2), a certain amount of variance due to the difference in these constructs will also need to be accounted for in the overall “error term.”

Lyubomirsky et al. did not explicitly claim that their choice of three factors (*S*, *C*, and *V*) was exhaustive; however, once *S* and *C* have been defined, the third section of the pie chart (*V*) ought logically to be labeled “anything that is not circumstances and set point, including volitional activities.” Of course, this would result in a much less convincing thought experiment, as it does not allow us to infer how promising volitional activities, on their own, potentially are.

Adding up percentages of variance explained from different studies is questionable

Lyubomirsky et al.’s (2005) drew their estimates of the variance explained by genetic factors and life circumstances from different sources, who studied samples drawn from different populations. It seems questionable that these estimates can simply be combined and then subtracted from 100% to derive an estimate of the influence of something that has not been measured. The percentage of variance that can be explained by any single construct, or set of constructs, is always specific to a certain population. Thus, it seems ill-advised to combine estimates across studies without ensuring that they are referring to the same populations.

Re-examining the Numerical Estimates of the Effect of Genes and Circumstances

How much variance in chronic happiness levels can be explained by genetic factors?

It appears that Lyubomirsky et al. (2005) derived their figure of 50% for the heritability of well-being by combining, in a way that was not clearly specified, the results from three studies (Braungart, Plomin, DeFries, & Fulker, 1992; Lykken & Tellegen, 1996; Tellegen et al., 1988). Lykken and Tellegen estimated that “the heritability of the stable component of subjective well-being approaches 80%” (p. 186). Tellegen et al. examined general personality traits in adult twins and determined that heritability ranged from 39% to 58% with an estimate of 48% for the personality trait labeled “Well-Being.” In contrast, Braungart et al. measured the influence of genetic factors on externally-reported developmental and emotional behaviors in children aged 1 or 2 and found heritability ranging from 35% to 57%. Even taking the lower bound of the estimated range of heritability from this study (which did not measure adult well-being), we arrive at an average figure of around 61%, whether or not the studies are weighted to take into account their sample sizes. About the best that can be said, then, for the figure of 50% for genetic factors is that it is surrounded by considerable uncertainty, and appears—based on the studies cited by Lyubomirsky et al.—likely to be a lower bound. Of course, any upward adjustment of this number would necessarily imply a downward adjustment of the percentage of variance attributable to

volitional activities, if we accept the subtraction logic suggested by Lyubomirsky et al. Note that current overviews of the literature locate the heritability of *overall* (i.e., including momentary) happiness at 32–41%, but for the *stable* component of happiness (i.e., the chronic levels that are the outcome of interest in Lyubomirsky et al.’s model) heritability is reported in the 70–80% range (Nes & Røysamb, 2017).

Origins of the claim that only 10% of the variance in well-being is due to life circumstances

Lyubomirsky (2007, p. 41) claimed that the figure of 10% for the amount of variance in happiness explained by life circumstances “represents an average from many past investigations, which reveal that all life circumstances and situations put together account for only about 10 percent in how happy different people are.” Lyubomirsky et al. (2005) cited two sources in support of the 10% claim: a book chapter by Argyle (1999) and an article by Diener, Suh, Lucas, and Smith (1999). Argyle in fact gave two figures on his pp. 353–354: “less than 10 percent of the variance,” which he attributed to Andrews and Withey (1976), and 15%, which he reported as having been proposed by Diener (1984). The first of these figures appears to correspond to the following quote from Andrews and Withey (1976, p. 109): “The demographic variables, either singly or jointly, account for very little of the variance in perceptions of global well-being (less than 10 percent),” while the second appears to refer to this: “individual demographic variables rarely account for more than a few percent of the variance in SWB, and taken together probably do not account for much more than 15% of the variance” (Diener, 1984, p. 558). Notice that both of these quotes refer to *demographic variables*, which, crucially, are not the same as life circumstances, as we discuss below.

In the second article cited by Lyubomirsky et al. (2005), Diener et al. (1999, pp. 278–279) reported three candidate percentages for the variance in happiness explained by circumstances: 20% according to Campbell, Converse, and Rodgers (1976), 8% according to Andrews and Withey (1976), and the same 15% mentioned by Argyle that was cited separately by Lyubomirsky et al. (which was, of course, Diener’s original informal estimate from 1984). Later in their article, Lyubomirsky et al. stated that “all circumstances combined account for only 8% to 15% of the variance in happiness levels” (p. 117), again with citations of Argyle (1999) and Diener et al. (1999); for some reason they omitted to mention the figure of 20%, derived from Campbell et al.’s study, which Diener et al. had cited in the sentence immediately preceding their discussion of the 8% and 15% figures. Lyubomirsky et al. did not report how they finally settled on their “headline” figure of 10%, but we tentatively

assume that this was derived by some sort of informal interpolation between 8% and 15%. In summary, it seems that there is a considerable lack of clarity about how the (subsequently highly influential) figure of 10% was arrived at. Hence, we turned to the original sources to look for this information.

First, we examined Andrews and Withey's (1976) book, which describes the results of a large, multiple-occasion study involving an overall total of 5,142 US Americans. Of interest here are the surveys that were conducted in May 1972 ($N = 1,297$), and April 1973 ($N = 1,433$), which were the ones used by Andrews and Withey (1976, pp. 138–142) to illustrate their discussion of the limited value of “classification variables” in predicting overall life satisfaction. We found two tables (Exhibit 4.6 on p. 139 and Exhibit 4.7 on p. 141) showing that the percentage of variance explained by “six classification variables” was 8% in the May 1972 survey and 11% in the April 1973 survey. These “classification variables” are age, family income, education, race, sex, and “family life-cycle stage,” with the last of these being a categorical variable based on “the respondent's own age and marital status, and the age of the youngest child living in the family” (Andrews & Withey, 1976, p. 138). These variables, which Andrews and Withey (1976, p. 138) described as “demographic or social characteristics,” were used mainly as the basis of classifying their subjects into subgroups; it is clear, using Lyubomirsky et al.'s (2005) own definition (which we discuss below), that they do not constitute an exhaustive inventory of “life circumstances,” sufficient to allow conclusions to be drawn about the effect of such circumstances on an individual's well-being.

Turning to Campbell et al.'s (1976) book, which was based on a survey of 2,147 US Americans conducted in 1971, we found a number of tables (e.g., those on pp. 230, 241, 253, 279, and 302) that explored the relation between satisfaction with an individual domain of life (e.g., “housing” or “life in the United States”) on the one hand, and participants' subjective assessments of that domain and their “personal characteristics” on the other. The percentage of variance explained by the combination of subjective assessments and personal characteristics in these tables is typically in the range of 20–30%, with personal characteristics typically accounting for around 8–10%, but—as with Andrews and Withey's (1976) “classification variables”—the lists of these characteristics were short (five or six items), with race and (where present) age and sex typically having the largest partial correlations with the relevant satisfaction measure. Indeed, Diener et al. made it clear that their estimates of the variance explained in these two books were based on “demographic factors,” not an extensive inventory of life circumstances:

Campbell, Converse, and Rodgers (1976) found that demographic factors (e.g., age, sex, income, race, education, and marital status) accounted for less than 20% of the variance in SWB. Andrews and Withey (1976) could only account for 8% by using these variables. (Diener et al., 1999, pp. 278–279)

In summary, although it is not clear exactly how Lyubomirsky et al. (2005) arrived at their estimate that only 10% of variance in well-being is explained by life circumstances, the available evidence suggests that this well-publicized figure may have been derived from an erroneous conflation of “classification variables” or “demographic factors” with the much wider category of “life circumstances,” in a few tables of results in two books describing studies that were conducted by researchers at the same institution⁶ (the Institute for Social Research, Ann Arbor, Michigan) in the early 1970s. Whether or not this constitutes “many past investigations,” as claimed by Lyubomirsky (2007) in the sentence quoted at the start of this section, is perhaps a matter of opinion. However, given that these sources provided the empirical evidence that was ultimately used by Lyubomirsky et al. to support their claims about the effect of life circumstances on well-being, it would seem interesting to examine the evidence provided by these studies, while not losing sight of the limitations just mentioned. We were able to obtain the data sets on which Andrews and Withey (1976) and Campbell et al. (1976) based their books, and reanalyze them to see what percentage of variance in these authors’ respective life satisfaction outcomes was explained by a set of 15 to 18 variables (depending on the exact questions asked in each survey) that were agreed by external evaluators to correspond to “life circumstances.” With the same statistical techniques that had been used by the original authors to arrive at the figure of 8–10% for demographic variables, we obtained values of between 18.13% and 26.47% for life circumstances. (A more modern approach using cross-validation results in a wide range of estimates, between 1.92% and 17.90%, highlighting how such estimates depend both on the statistical methods employed and on the variables available in the datasets.) Further details are provided in the Supplementary Information of the present article.

⁶ Indeed, Andrews and Withey’s (1976) book refers to the studies conducted by their colleagues Campbell et al. (1976) at numerous points, and even describes (p. 244) an exploratory analysis using a combination of their own data and those of Campbell et al.

Measures of well-being are specific to given populations and periods of time

Even if one were to assume that the variance decomposition suggested by Lyubomirsky et al. (2005) was valid (which we questioned above), it is not clear that these numbers still hold today (or, indeed, whether they did so when Lyubomirsky et al.'s article was published). Campbell et al. (1976) collected their data in 1971, while Andrews and Withey (1976) collected theirs in 1972–1973. This was a time when baby boomers were reaching adulthood and the Vietnam War was at its height. The relevance of these studies to the modern social landscape probably cannot be taken for granted. For example, the participants were between 87% and 90% White, with the rest being mostly reported as Black and no more than about 1% described as being of another race, while the demographic questions make it clear that the default expectation of the researchers was that cohabiting adults would be heterosexual and probably married. It is not clear whether the same results would be found were these studies to be repeated today, after forty years of social change in the United States and with considerably greater diversity in the population and people's lifestyles. Any estimate of the percentage of variance that can be explained by a certain combination of predictors is only meaningful with respect to the population from which the sample has been drawn. Of course, in the period spanning almost half a century since these studies were conducted, other researchers have collected data that show the relation between demographic variables and well-being (see Diener, Lucas, & Oishi, 2018, for a review), but as we have already noted, such relations tell us little about the influence of variables that fall under the much wider category of life circumstances.

A further caveat to note here is that when estimating the amount of variance in an outcome that can be attributed to certain predictors, analyses are necessarily constrained by the amount of variability in those predictors (as already explained for the case of volitional activities above). For example, broader economic circumstances are often constant within a survey at a given time point in a given sample. However, changes in these circumstances can have remarkable effects on well-being, as indicated by the World Happiness Report (United Nations, 2015), which is based on the Gallup World Poll. This report periodically measures the self-reported well-being of the inhabitants of most of the countries of the world on a "ladder" from 0 to 10 (Cantril, 1965). The 2015 report showed that between the measurement periods of 2005–2007 and 2012–2014, self-reported well-being in Greece fell from 6.327 to 4.857, a drop that represents almost one and a half steps down the ladder from the best possible life towards the worst. This reduction of 1.470 points also corresponds to 31.0% of the range of country-average well-being scores in the 2015 report from the highest

(Switzerland, at 7.587) to the lowest (Togo, at 2.839). It seems unlikely that more than a trivial part of this decline in well-being over the better part of a decade can be accounted for by a change in the genetic characteristics of the population of Greece, or a decision by Greek citizens to eschew *en masse* the forms of intentional activity that might contribute to their individual happiness. Rather, the drastic change in their (economic) circumstances forced on the majority of Greeks by the ongoing financial crisis affecting their country would seem to be by far the most plausible candidate to explain this drop in national well-being (cf. Rhodes Hatzimalonas, 2017). Of course, one could argue that these changes are in some sense transitory, and ought to fade over the course of the years as the Greek economy recovers. However, if such a multi-year criterion is to be used to distinguish between mere fluctuations in circumstance and true changes to people's chronic level of well-being, it should logically also be applied to intervention studies claiming to produce sustainable changes in happiness.

The wide range of scores between the top and bottom countries on the *World Happiness Report* list is another indication that there might be many circumstantial factors that affect well-being beyond those that were (or, indeed, could have been) included by Andrews and Withey (1976) and Campbell et al. (1976) in their surveys of US Americans. On a continuum from the smaller democracies of northern Europe at one end through to the poorest countries of sub-Saharan Africa at the other, there are considerable variations in political freedom or repression, institutional transparency or corruption, and the availability or lack of basic services such as electricity and clean drinking water, to name just a few examples of factors that might be expected to influence well-being, but would not be expected to vary much among any sample drawn from just one country. Indeed, Lyubomirsky et al. (2005, p. 117) acknowledged that "Happiness-relevant circumstances may include the national, geographical, and cultural region in which a person resides," but these authors then failed to limit their own claims to the specific population from which the data were drawn (i.e., US Americans from the 1970s). We suggest that future research into the effects of life circumstances on well-being should take into account potential cross-cultural concerns and acknowledges the limitations of the samples being studied. For example, it has recently been suggested that researchers should explicitly limit conclusions to specific target populations to avoid unjustified generalizations (Simons, Shoda, & Lindsay, 2017).

Likewise, researchers should carefully reconsider the usage of terms such as "volitional activities." What might be a matter of choice for some people can be a question of circumstances for others. For example, middle-class Western parents with a comfortable household income can exercise a wide range of choice over the diet that they and their

children consume; they can choose to drive their car to the farmer's market instead of taking the bus to the discount grocery store, and afford to place organic blueberries rather than canned, syrup-laden fruit cocktail in their shopping cart. They can also enjoy a wide range of recreational activities, such as vacations to awe-inspiring or culturally important destinations, enabling them to enjoy the benefits of feeling that they are part of a meaningful world (Gilovich & Kumar, 2015). Indeed, Lyubomirsky et al. (2005) themselves described the types of volitional activities that they were considering in decidedly middle-class terms: "rather than running on a track, a fitness-seeking wilderness lover might instead choose to run on a trail through the woods . . . rather than learning classical pieces, a jazz-loving piano student might instead choose to work on jazz standards" (p. 122). A subsistence farmer working from dawn to dusk in a developing nation, a worker who spends 14 hours per day assembling electronic devices in a south-east Asian factory, or a single Western parent living in less fortunate circumstances than the family just mentioned, may not be in a position to make such choices. As with many phenomena studied by primarily Western psychologists (cf. Henrich, Heine, & Norenzayan, 2010), the list of volitional activities that are available to improve one's well-being may be less than universal, not just for cultural reasons but also precisely because of one's life circumstances.

Moving Forward: Between-Subjects Versus Within-Subject Designs and Causal Inference

We noted earlier that there appeared to be a mismatch between the evidence (such as it was) provided by Lyubomirsky et al. (2005) for their model of variance decomposition—based on between-subjects studies—and the suggestion that individuals can improve their happiness with volitional activities, which is inherently a within-subject effect. Such limitations of positive psychology research have been noted numerous times in the past (e.g., Lazarus, 2003; Nickerson, 2007, 2014), and calls for more longitudinal, within-subject research are not uncommon even in the discussion sections of studies based on cross-sectional, between-subjects data. In fact, it appears—at least from an informal examination—that within-subject designs may be becoming more popular within positive psychology. In early December 2018, we examined the 50 most recently published articles in the *Journal of Happiness Studies* that reported empirical data, and compared these with a random selection of 50 articles published in the same journal approximately ten years earlier. Whereas in 2007–2008 only seven out of 50 articles used within-subjects designs, by 2017–2018 this

number had risen to 16⁷, with a variety of methods being used. For example, on a shorter timescale, data collected with the Experience Sampling Method (Csikszentmihalyi & Hunter, 2003) were used to investigate how detachment from workplace stresses affect the quality of interactions in romantic relationships (Debrot, Siegler, Klumb, & Schoebi, 2018), while on a longer timescale, data collected as part of a panel study of 1,500 households in Nepal were used to investigate the cost of coping on different forms of well-being (Chindarkar, Chen, & Gurung, 2018).

Part of the appeal of within-subject designs is that they can, if properly analyzed, control stable inter-individual differences, rendering a causal interpretation of associations more plausible (although such designs do not automatically warrant causal conclusions, as time-varying confounders can offer alternative explanations). Causal inference might not always be the *explicitly* stated goal of empirical investigations of subjective well-being, but it often seems to be what authors are most interested in; indeed, Lyubomirsky et al. (2005, p. 116) referred to the genetic set point, life circumstances, and intentional activity as the “three primary types of factors that we believe *causally* [emphasis added] affect the chronic happiness level.” For example, if there is a between-subjects association between a certain activity and happiness (i.e., on average, people who do X are happier), but no within-subject association (i.e., a person who starts doing X does not become happier), we would be quick to dismiss the between-subjects association as spurious and refrain from advising people to do more X to become happier.

However, this advantage becomes a complication when the aim is to give a holistic picture of *all* of the determinants that explain happiness (whether this is measured over an extended period of time or as short-term fluctuations). Many inter-individual differences, such as personality or socio-economic status, hardly vary within individuals, making it difficult to establish any within-subject association with the outcome. Even research on fairly common life events such as marriage or childbirth makes use of the massive sample sizes of ongoing panel studies to ensure that sufficient individuals have actually experienced the event in question while they were part of the study.

Of course, the above comments merely underscore the self-evident truth that no single type of research design can answer all questions. Within-subject designs may pave the way

⁷ Although it would not be appropriate to perform an inferential test here, we note in passing that the *p* value associated with a chi-square test of independence for these data is .057. Full details of the method used for this informal study are given in the Supplementary Information of the present article.

for interesting new research questions with a different scope, such as whether individuals reliably differ in the volatility of their well-being, or in their reactivity to stressors, and which factors shape these inter-individual differences in intra-individual variation. However, the complex nature of human experience is such that data collected over a range of timescales, from a few minutes to many decades, will likely be needed to fully identify the determinants of well-being as individuals experience daily existence, major life milestones, and the general aging process (as well as any psychological interventions that we might subject them to). To both enhance and complement these efforts, it could be fruitful if researchers working on subjective well-being considered more explicitly whether they are trying to draw causal conclusions (the answer might be “yes” more often than not), and, if so, made use of systematic and established approaches from other fields of research. For example, counterfactuals (e.g., Morgan & Winship, 2015; Neyman, 1923/1990), directed acyclic graphs (e.g., Kuipers, Moffa, Kuipers, Freeman, & Bebbington, 2018; Pearl, 2000), and various types of natural experiments (e.g., Dunning, 2008; Rutter, 2007) might have been under-utilized in psychological well-being research (likely because they are not routinely taught to psychologists). A more careful consideration of the admittedly complex issue of causality might lead to the sobering conclusion that a complete decomposition of happiness into genes, circumstances, and intentional activities is not only an elusive research goal, but not even a well-defined one.

Concluding Remarks

Since the publication of Lyubomirsky et al.’s (2005) model, research on subjective well-being has progressed considerably, as documented extensively by Diener et al. (2018). However, the “happiness pie” is still widely cited, and many of the problematic arguments underlying its claims—for example, misconceptions about the generalizability of between-subjects variance decompositions and about the interpretation of heritability estimates—can be found in contemporary research endeavors, both within and outside of positive psychology. Our somewhat lengthy search—documented earlier in the present article—to track down the claim that only 10% of the variance in happiness is due to life circumstances is a good example of the need, when integrating previous findings, to pay close attention to the details of the empirical studies to ensure that the summarized literature does in fact support the claim being made.

As we have shown in this article, there is a considerable list of issues with the popular “happiness pie” decomposition of variance in well-being. First, a variance decomposition in

the population does not necessarily inform us about the potential for individual level changes, although such conclusions are frequently drawn, and not just by statistically uninformed writers in the popular media. Second, the decomposition of the variance in chronic levels of happiness into three factors suffers from a major conceptual issue (the three factors do not additively affect well-being in isolation) and certain omissions (error term, other potential determinants) that bias the analysis towards assigning higher percentages to volitional activities. Third, the estimates assigned to life circumstances and genetic factors are questionable. The empirical evidence cited indirectly by Lyubomirsky et al. (2005) in support of the idea that individuals' life circumstances account for only 10% of the variance in well-being seems to be based on a misunderstanding of what the relevant studies actually investigated (many factors describing people's life circumstances, versus the narrower category of demographic variables). Likewise, there is only very limited evidence to place the figure for the heritability of well-being as low as (precisely) 50%. Consequently, there is little reason to believe that 40% is a reliable estimate of the variance in chronic happiness attributable to intentional activity—for example, if Lyubomirsky et al. had chosen a different (but, in our view, at least equally plausible) set of estimates, they might just as easily have concluded that as little as $(100\% - 80\% - 15\% = 5\%)$ of variance in chronic happiness can be attributed to volitional activities. Of course, because of the conceptual issues that we have discussed, we do not believe that such an estimate would necessarily be more reasonable or more interpretable.

Although our analyses in the present article have been focused on the “happiness pie,” we feel that is useful to point out that the issues we have addressed—namely, misconceptions about the interpretation of between-subjects variance decomposition; misconstruction of genes, environment, and volition as orthogonal factors; and misreading or misinterpretation of the findings from previous studies—are certainly not limited to the study by Lyubomirsky et al. (2005). In fact, such issues seem to recur throughout many research fields in psychology. However, the influence on the popular imagination of the “happiness pie” model particularly underscores the need for psychologists to be circumspect in their claims, including when they venture beyond academic writing into popular books and other ways of disseminating their work to a broader public. While this form of outreach ought to be an admirable development, it comes with the responsibility to avoid making claims beyond what the evidence allows for.

Several writers (e.g., Di Martino, Eiroa-Orosa, & Arcidiacono, 2017; Ehrenreich, 2009; Whippman, 2016) have criticized positive psychology for placing too much emphasis

on the purported ability of individuals to improve their psychological well-being regardless of the social conditions under which they live. The idea that only 10% of one's happiness is due to circumstances, with four times as much "capacity" being available to people to forge their own well-being, is attractive and fits into a pervasive contemporary "feel-good" narrative in which self-improvement is claimed to be not only desirable, but also eminently achievable (see Moreau, Macnamara, & Hambrick, 2019, for a discussion of several other current topics of psychological research where claims about the malleability of traits appear to have been overstated). At this point we take no particular view of the political questions that might be implicitly raised by such claims; our principal interest is to point out some ways in which they might not be justified by the available evidence. Indeed, it seems to us to be vital at this point for positive psychologists to critically re-examine the evidence base for their claims about the ability of people to improve their own happiness. We hope that our discussion of the conceptual limitations of Lyubomirsky et al.'s (2005) model may generate helpful insights for future research programs to ensure a solid empirical foundation of positive psychology.

References

- Achor, S. (2010). *The happiness advantage: The seven principles of positive psychology that fuel success and performance at work*. New York, NY: Crown.
- Argyle, M. (1999). Causes and correlates of happiness. In D. Kahneman, E. Diener, & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 353–373). New York, NY: Russell Sage.
- Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being: Americans' perceptions of life quality*. New York, NY: Plenum.
- Bergink, J. (2015, September 17). *Happiness: It's not just your genes, stupid!* [Weblog post]. Retrieved from <http://www.forastateofhappiness.com/tag/50-10-40-formula>
- Braungart, J. M., Plomin, R., DeFries, J. C., & Fulker, D. W. (1992). Genetic influence on tester-rated infant temperament as assessed by Bayley's Infant Behavior Record: Nonadoptive and adoptive siblings and twins. *Developmental Psychology*, 28, 40–47. <https://doi.org/10.1037/0012-1649.28.1.40>
- Burton, C. M., & King, L. A. (2004). The health benefits of writing about intensely positive experiences. *Journal of Research in Personality*, 38, 150–163. [https://doi.org/10.1016/S0092-6566\(03\)00058-8](https://doi.org/10.1016/S0092-6566(03)00058-8)
- Butler, J., & Ciarrochi, J. (2007). Psychological acceptance and quality of life in the elderly. *Quality of Life Research*, 16, 607–615. <https://doi.org/10.1007/s11136-006-9149-1>
- Campbell, A., Converse, P. E., & Rodgers, W. L. (1976). *The quality of American life: Perceptions, evaluations, and satisfactions*. New York, NY: Russell Sage.
- Cantril, H. (1965). *The pattern of human concerns*. New Brunswick, NJ: Rutgers University Press.
- Cygan, D. S. (2013). *The joy of financial security: The art and science of becoming happier, managing your money wisely, and creating a secure financial future*. Albuquerque, NM: Sage Future Press.
- Chindarkar, N., Chen, Y. J. & Gurung, Y. (2018). Subjective well-being effects of coping cost: Evidence from household water supply in Kathmandu Valley, Nepal. *Journal of Happiness Studies*. Advance online publication. <https://doi.org/10.1007/s10902-018-0060-6>
- Csikszentmihalyi, M., & Hunter, J. (2003). Happiness in everyday life: The uses of experience sampling. *Journal of Happiness Studies*, 4, 185–199. <https://doi.org/10.1023/A:1024409732742>
- Debrot, A., Siegler, S., Klumb, P. L., & Schoebi, D. (2018). Daily work stress and relationship satisfaction: Detachment affects romantic couples' interactions quality. *Journal of Happiness Studies*, 19, 2283–2301. <https://doi.org/10.1007/s10902-017-9922-6>
- Di Martino, S., Eiroa-Orosa, F. J., & Arcidiacono, C. (2018). Community psychology's contributions on happiness and well-being: Including the role of context, social justice, and values in our understanding of the good life. In N. J. L. Brown, T. Lomas, & F. J. Eiroa-Orosa (Eds.), *The Routledge international handbook of critical positive psychology* (pp. 99–116). London, England: Routledge.
- Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, 95, 542–575. <https://doi.org/10.1037/0033-2909.95.3.542>
- Diener, E., Lucas, R. E., & Oishi, S. (2018). Advances and open questions in the science of subjective well-being. *Collabra: Psychology*, 4(1), 15. <https://doi.org/10.1525/collabra.115>

- Diener, E., Scollon, C. N., & Lucas, R. E. (2009). The evolving concept of subjective well-being: The multifaceted nature of happiness. In E. Diener (Ed.), *Assessing well-being: The collected works of Ed Diener* (pp. 67–100). Dordrecht, Netherlands: Springer.
- Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin*, 125, 276–302. <https://doi.org/10.1037/0033-2909.125.2.276>
- Dunning, T. (2008). Improving causal inference: Strengths and limitations of natural experiments. *Political Research Quarterly*, 61, 282–293. <https://doi.org/10.1177/1065912907306470>
- Dykens, E. M. (2006). Toward a positive psychology of mental retardation. *American Journal of Orthopsychiatry*, 76, 185–193. <https://doi.org/10.1037/0002-9432.76.2.185>
- Ehrenreich, B. (2009). *Bright-sided: How the relentless promotion of positive thinking has undermined America*. New York, NY: Henry Holt.
- Emmons, R. A., & McCullough, M. E. (2003). Counting blessings versus burdens: An experimental investigation of gratitude and subjective well-being in daily life. *Journal of Personality and Social Psychology*, 84, 377–389. <https://doi.org/10.1037/0022-3514.84.2.377>
- Fredrickson, B. L. (2013). *Love 2.0: Finding happiness and health in moments of connection*. New York, NY: Hudson Street Press.
- Gilovich, T., & Kumar, A. (2015). We'll always have Paris: The hedonic payoff from experiential and material investments. *Advances in Experimental Social Psychology*, 51, 147–187. <https://doi.org/10.1016/bs.aesp.2014.10.002>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hudders, L., & Pandelaere, M. (2012). The silver lining of materialism: The impact of luxury consumption on subjective well-being. *Journal of Happiness Studies*, 13, 411–437. <https://doi.org/10.1007/s10902-011-9271-9>
- Huffman, J. C., Mastromauro, C. A., Boehm, J. K., Seabrook, R., Fricchione, G. L., Denninger, J. W., & Lyubomirsky, S. (2011). Development of a positive psychology intervention for patients with acute cardiovascular disease. *Heart International*, 6, e13. <https://doi.org/10.4081/hi.2011.e13a>
- Jordan, A. H., Benoît Monin, B., Dweck, C. S., Lovett, B. J., John, O. P., & Gross, J. J. (2011). Misery has more company than people think: Underestimating the prevalence of others' negative emotions. *Personality and Social Psychology Bulletin*, 37, 120–135. <https://doi.org/10.1177/0146167210390822>
- Kashdan, T. (2015, August 19). What really makes you a happy person? *Psychology Today*. Retrieved from <https://www.psychologytoday.com/blog/curious/201508/what-really-makes-you-happy-person>
- Kendler, K. S., Karkowski, L. M., & Prescott, C. A. (1999). Fears and phobias: Reliability and heritability. *Psychological Medicine*, 29, 539–553. <https://doi.org/10.1017/S0033291799008429>
- Krueger, J. I. (2015, March 4). Happy pie: Intend to become happier and do something about it. *Psychology Today*. Retrieved from <https://www.psychologytoday.com/blog/one-among-many/201503/happy-pie>

- Kuipers, J., Moffa, G., Kuipers, E., Freeman, D., & Bebbington, P. (2018). Links between psychotic and neurotic symptoms in the general population: an analysis of longitudinal British National Survey data using directed acyclic graphs. *Psychological Medicine*, 49, 388–395. <https://doi.org/10.1017/S0033291718000879>
- Lazarus, R. S. (2003). Does the positive psychology movement have legs? *Psychological Inquiry*, 14, 93–109. https://doi.org/10.1207/S15327965PLI1402_02
- Lykken, D. (1999). *Happiness: What studies on twins show us about nature, nurture, and the happiness set point*. New York, NY: Golden Books.
- Lykken, D., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. *Psychological Science*, 7, 186–189. <https://doi.org/10.1111/j.1467-9280.1996.tb00355.x>
- Lyubomirsky, S. (2007). *The how of happiness: A scientific approach to getting the life you want*. New York, NY: Penguin.
- Lyubomirsky, S., Sheldon, K. M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology*, 9, 111–131. <https://doi.org/10.1037/1089-2680.9.2.111>
- MentorCoach. (2008). *The how, what, and why of happiness: The positive psychology master class with Sonja Lyubomirsky, Ph.D.* Retrieved from <http://mentorcoach.com/courses/how-of-happiness2.htm>
- Moore-Harrison, T., & Lightfoot, J. T. (2010). Driven to be inactive?—The genetics of physical activity. *Progress in Molecular Biology and Translational Science*, 94, 271–290. <https://doi.org/10.1016/B978-0-12-375003-7.00010-8>
- Moreau, D., Macnamara, B. N., & Hambrick, D. Z. (2019). Overstating the role of environmental factors in success: A cautionary note. *Current Directions in Psychological Science*, 28, 28–33. <https://doi.org/10.1177/0963721418797300>
- Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). New York, NY: Cambridge University Press.
- Nawijn, J., & Mitás, O. (2012). Resident attitudes to tourism and their effect on subjective well-being: The case of Palma de Mallorca. *Journal of Travel Research*, 51, 531–541. <https://doi.org/10.1177/0047287511426482>
- Nes, R. B., & Røysamb, E. (2017). Happiness in behaviour genetics: An update on heritability and changeability. *Journal of Happiness Studies*, 18, 1533–1552. <https://doi.org/10.1007/s10902-016-9781-6>
- Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. (D. M. Dabrowska and T. P. Speed, Trans.). *Statistical Science*, 5, 465–472. (Original work published 1923.) <https://doi.org/10.1214/ss/1177012031>
- Nickerson, C. (2007). Theory/analysis mismatch: Comment on Fredrickson and Joiner’s (2002) test of the broaden-and-build theory of positive emotions. *Journal of Happiness Studies*, 8, 537–561. <https://doi.org/10.1007/s10902-006-9030-5>
- Nickerson, C. A. (2014). No empirical evidence for critical positivity ratios. *American Psychologist*, 69, 626–628. <https://doi.org/10.1037/a0036961>
- Page, K. M., & Vella-Brodrick, D. A. (2009). The ‘what’, ‘why’ and ‘how’ of employee well-being: A new model. *Social Indicators Research*, 90, 441–458. <https://doi.org/10.1007/s11205-008-9270-3>

- Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.
- Plomin, R., & Deary, I. J. (2015). Genetics and intelligence differences: Five special findings. *Molecular Psychiatry*, 20, 98–108. <https://doi.org/10.1038/mp.2014.105>
- Proyer, R. T., Ruch, W., & Buschor, C. (2013). Testing strengths-based interventions: A preliminary study on the effectiveness of a program targeting curiosity, gratitude, hope, humor, and zest for enhancing life satisfaction. *Journal of Happiness Studies*, 14, 275–292. <https://doi.org/10.1007/s10902-012-9331-9>
- Rashid, T. (2008). Positive psychotherapy. In S. J. Lopez (Ed.), *Positive psychology: Exploring the best in people* (Vol. 4), pp. 187–217. Westport, CT: Praeger.
- Rhodes Hatzimalonas, M. (2018). Philotimo: Vices and virtues of a moral archetype. In N. J. L. Brown, T. Lomas, & F. J. Eiroa-Orosa (Eds.), *The Routledge international handbook of critical positive psychology* (pp. 474–487). London, England: Routledge.
- Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of Personality and Social Psychology*, 111, 780–789. <https://doi.org/10.1037/pspp0000089>
- Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, 29, 1358–1369. <https://doi.org/10.1177/0956797618774253>
- Rutter, M. (2007). Proceeding from observed correlation to causal inference: The use of natural experiments. *Perspectives on Psychological Science*, 2, 377–395. <https://doi.org/10.1111/j.1745-6916.2007.00050.x>
- Scelpo, J. (2017, April 5). Get happy: Four well-being workouts. *The New York Times*. Retrieved from <https://www.nytimes.com/2017/04/05/education/edlife/get-happy-four-well-being-workouts.html>
- Schueller, S. M., & Seligman, M. E. P. (2010). Pursuit of pleasure, engagement, and meaning: Relationships to subjective and objective measures of well-being. *Journal of Positive Psychology*, 5, 253–263. <https://doi.org/10.1080/17439761003794130>
- Seligman, M. E. P. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. New York, NY: Simon & Schuster.
- Seligman, M. E. P., Steen, T. A., Park, N., & Peterson, C. (2005). Positive psychology progress: Empirical validation of interventions. *American Psychologist*, 60, 410–421. <https://doi.org/10.1037/0003-066X.60.5.410>
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, 12, 1123–1128. <https://doi.org/10.1177/1745691617708630>
- Swann, W. B., & Seyle, C. (2005). Personality psychology's comeback and its emerging symbiosis with social psychology. *Personality and Social Psychology Bulletin*, 31, 155–165. <https://doi.org/10.1177/0146167204271591>
- Tellegen, A., Lykken, D. T., Bouchard, T. J., Wilcox, K. J., Segal, N. L., & Rich, S. (1988). Personality similarity in twins reared apart and together. *Journal of Personality and Social Psychology*, 54, 1031–1039. <https://doi.org/10.1037/0022-3514.54.6.1031>

- Turkheimer, E. (2000). Three laws of behavior genetics and what they mean. *Current Directions in Psychological Science*, 9, 160–164. <https://doi.org/10.1111/1467-8721.00084>
- United Nations. (2015). *World happiness report 2015*. Retrieved from https://s3.amazonaws.com/happiness-report/2015/WHR15_Sep15.pdf
- Veenhoven, R. (2011). Happiness: Also known as “life satisfaction” and “subjective well-being.” In K. C. Land, A. C. Michalos, & M. J. Sirgy (Eds.) *Handbook of social indicators and quality of life research*, pp. 63–77. Dordrecht, Netherlands: Springer.
- Vondruska, B. (2017). *The optimal life experience: What it is, why it matters, and how to maximize well-being*. Aurora, OH: Aurora Park.
- Werner, K. [kaitlynmwerner] (2018, July 18). Side note from Sonja Lyubomirsky: the happiness pie was more of a thought experiment from 18 years ago. Based on a very limited body of work and they’re open to re-interpretations! [Tweet]. Retrieved from <https://twitter.com/kaitlynmwerner/status/1019608693468467201>
- Whippman, R. (2016). *America the anxious: How our pursuit of happiness is creating a nation of nervous wrecks*. New York, NY: St. Martin’s Press.
- Wise, E. H., Hersh, M. A., & Gibson, C. M. (2012). Ethics, self-care and well-being for psychologists. *Professional Psychology: Research and Practice*, 43, 487–494. <https://doi.org/10.1037/a0029446>

Supplementary Information

Re-examination of the 10% claim

It turns out that, beyond the simple demographic variables such as age and ethnicity that seem to form the basis of the numbers calculated by Diener and colleagues that were cited by Lyubomirsky, Sheldon, and Schkade (2005), both Andrews and Withey (1976) and Campbell, Converse, and Rodgers (1976) did in fact measure a substantial number of other variables that appear to fall into the category of “life circumstances.” Prima facie examples of variables that are candidates for this status include questions such as whether the participant is currently in employment, how many years of education he or she completed, and whether he or she has any health problems⁸. Indeed, these variables appear to meet Lyubomirsky et al.’s own definition of life circumstances:

This category consists of happiness-relevant circumstantial factors, that is, the incidental but relatively stable facts of an individual’s life. Happiness-relevant circumstances may include the national, geographical, and cultural region in which a person resides, as well as demographic factors such as age, gender, and ethnicity . . . Circumstantial factors also include the individual’s personal history, that is, life events that can affect his or her happiness, such as having experienced a childhood trauma, being involved in an automobile accident, or winning a prestigious award. Finally, circumstantial factors include life status variables such as marital status, occupational status, job security, income, health, and religious affiliation.
(Lyubomirsky et al., 2005, p. 117)

Thus, both Andrews and Withey’s (1976) and Campbell et al.’s (1976) data include a number of variables that ought to allow us to estimate the variance in life satisfaction that these authors might have reported as being explained by life circumstances (versus “classification variables” or “demographic factors”), had this been one of the purposes of their respective studies. Using the original data sets and accompanying documentation for the studies reported by Andrews and Withey (1976) and Campbell et al. (1976), we retained 15 predictors for Andrews and Withey’s (1976) May 1972 survey, 15 predictors for the same

⁸ Of course, as mentioned earlier, a participant’s scores on these items might result from an interaction between life circumstances and genetic factors, but for the purpose of our re-analysis we assume that the underlying logic of Lyubomirsky et al.’s (2005) variance decomposition is correct.

authors' April 1973 survey, and 18 predictors for Campbell et al.'s 1971 survey. Using software that was a direct descendant of the mainframe programs used by these authors more than 40 years ago, we determined that the percentage of variance explained by either Multiple Classification Analysis, as originally used by Andrew and Withey, or ordinary least squares regression, as used by Campbell et al., was at least $R^2 = 18.15\%$ (Andrews & Withey / May), 18.13% (Andrews & Withey / April), and 26.47% (Campbell et al.). That is, had those researchers set out to study the amount of variance explained by life circumstances, they would likely have reported numbers that were on average at least twice as large as Lyubomirsky et al.'s (2005) figure of 10%. Full details of our method, as well as links to all of the information to reproduce our analyses, are provided in the following sections, together with a more conservative re-analysis that uses cross-validation to avoid the dangers of overfitting in models in which all the predictors were treated as categorical variables. This re-analysis also shows that, with an appropriate choice of predictors, the amount of variance robustly predicted by life circumstances can be considerably higher than 10%.

Figure 7.2. A typical question from Campbell et al.'s (1976, p. 540) survey.

IF R IS A HOUSEWIFE:

F44. Different people feel differently about taking care of a home. I don't mean taking care of children, but things like cooking, sewing, and keeping house. Some women look on these things as just a job that has to be done; other women really enjoy them. How do you feel about this?

Method

In order to estimate the percentage of variance explained by life circumstances (henceforth LC) in the original studies on which Lyubomirsky et al.'s (2005) article was ultimately based, we downloaded the data sets and accompanying documentation for the study reported by Campbell et al. (1976), and the May 1972 and April 1973 studies reported by Andrews and Withey (1976), from the web site of the Inter-university Consortium for Political and Social Research (ICPSR), as follows:

- For Campbell et al.: <https://doi.org/10.3886/ICPSR03508.v1>
- For Andrews and Withey: <https://doi.org/10.3886/ICPSR03636.v1>

We opened the downloaded data files with SPSS, then exported them into comma-separated value (CSV) files for our R code to read. This code then either generated data files for the MicroSIRIS Multiple Classification Analysis (MCA) software, or performed our equivalent regression analyses (both “classical” OLS and cross-validated) directly.

The data sets that we downloaded contained a large number of variables: 176 from Andrews and Withey's (1976) survey conducted in May 1972, 247 from the same authors' survey conducted in April 1973, and 567 from Campbell et al.'s (1976) survey conducted in 1971, making a total of 990 survey items. However, it was immediately clear that many of these could not possibly be described as measuring LC. Therefore, the first author of the present article conducted an initial coding exercise in which he assigned many of these “obviously not LC” items to one of several categories, as follows:

- “Administrative” (206 variables across the three surveys): items used by the investigators to structure the organization of the studies (e.g., “Week of Interview”). Typically, this information would be determined by the interviewer who was conducting the survey in the participant's home, rather than being provided by the participants themselves.
- “Choice” (26 variables): items that appear to reflect a clear volitional choice by the participant, such as “Do you belong to a sports team?”. In terms of the “happiness pie” model, such items fall under the 40% of variance in well-being that Lyubomirsky et al. (2005) appeared to claim can be enhanced with the appropriate choice of activities.
- “Feeling” (323 variables): these are items that reflect participants' expressed opinion or sentiment about some aspect of their lives or their circumstances. Examples are “Are most people trustworthy?”, “In what ways is life in the US getting

better/worse?,” and “How do you feel about your job?” It might be argued that the last of these is to some extent a circumstance; someone who has had a privileged upbringing and become, say, a lawyer might look more favorably on their job than someone who was brought up in poverty and left school early to perform manual labor. However, the subject matter of this item seems to be adequately covered by items such as those asking about pay and promotion chances at the respondent’s job, which were put to the panel as candidates to be included as LC.

- “Outcome” (49 variables): items that measure the participant’s satisfaction (either with their life overall, or with a particular domain) or some other aspect of their well-being (e.g., “How happy are you these days?”). These are not included because they are the dependent variables that the authors of the original surveys were attempting to explain. Notably, each survey contains one measure of overall life satisfaction that we used as our dependent variable; this was given the specific coding “Focal” to make it easy for our analysis code to pick up.
- “Race–Sex” (6 variables): Lyubomirsky et al. (2005) included “gender” and “ethnicity” in their definition of LC. Hence, we included the variables for these two characteristics in each survey in our analyses.
- “Recode” (128 variables): items for which the responses were not given directly by the participant, but were instead derived by recoding, rescaling, or combining other responses.
- “Uninterpretable” (68 variables): items that did not seem to have any obvious psychological interpretation (e.g., “Country where respondent’s father was born”).

This left 184 variables across the three studies that appeared to be LC candidates. Our aim was, for each study, to further reduce the number of predictors, including only those would be considered as representing LC by a majority of neutral observers. Therefore, as our next step, we recruited six independent coders to participate in a Delphi-like panel (Adler & Ziglio, 1996; Stitt-Gohdes & Crews, 2004) in order to classify the remaining 184 variables into LC or “Other.” Among the coders were one undergraduate, two doctoral candidates, two postdoctoral researchers, and one assistant professor, all working in psychology or related disciplines. Although they were not explicitly told of our hypotheses, they were aware that the purpose of the exercise was to examine whether Lyubomirsky et al.’s (2005) claims about the percentage of variance explained by LC were supported by the available data. They were also given copies of the original code books from the Andrews and Withey (1976) and

Campbell et al. (1976) studies, which they could consult if they needed clarification of the meaning of any particular item.

Coding by the Delphi panel proceeded in three rounds. In the first round, each coder indicated, for each variable, whether they considered it to correspond to LC or “Other.” Coders could also leave comments, for example to explain their thought processes in borderline cases. Once all of the results had been received, the first author collated them. The 34 variables on which there was unanimous agreement were set to one side, and the remainder (150 out of 184) were sent out to the coders for a second round, along with the list of votes from the first round and any comments that had been made. The coders reconsidered and in some cases changed their coding of each variable, and returned their results, which were again collated, resulting in the reduction of the number of variables without unanimous agreement from 150 to 73. (Two coders dropped out of the process after the first round, so the scores in the second and third rounds were based on the results of four people, rather than six.) The process was then repeated for a third round of coding, at the end of which 53 variables remained without unanimous agreement on their coding. In these cases, those with only one dissenting vote ($N = 49$) were treated as if the result had been unanimous, while in the remaining four cases, the tie was broken by the first author, who decided that three variables corresponded to LC and one did not. At all times during this process, all of the coders (who included the second author of the present article) were unaware of the others’ identities, except that each coder was identified by the same unique letter throughout the process. The second author of the present article became a co-author late in the coding process; however, she received no more information about the ongoing process of data collection than any other panel member.

At the end of three rounds of coding by the panel, only eight of the 184 variables that had been identified by the first author before the process had been eliminated. It was clear that we needed to reduce this number further, partly to avoid overfitting with classical models (MCA and linear regression) due to an excessive number of predictors, partly to avoid attrition of the effective sample size due to missing data on some items (MCA, in particular, only allows listwise deletion in the case of missing values on a single predictor), and partly because we wanted to establish reasonable *minimum* values for the amount of variance explained by the models that the original authors could have chosen to run, had they set out specifically to investigate the effects of life circumstances (LC) on well-being. We therefore applied a further selection to reduce the number of variables, most notably from Campbell et al.’s (1976) data set, in order to reduce model complexity and to ensure some comparability

between the three different samples. This reduced the number of LC variables to 15 (out of 32 that had been identified at the end of the Delphi panel exercise as measuring LC) for Andrews and Withey's (1976) May survey, 15 (out of 35) for the same authors' April survey, and 18 predictors (out of 119) for Campbell et al.'s (1976) survey. Within the retained variables, we also pooled response categories that were very uncommon or appeared to be excessively numerous (for example, we bracketed length of marriage into five-year groups, to avoid having 40 or more categories).

We believe that our process for selecting and recoding LC variables was appropriately conservative. Those decisions taken unilaterally by the authors were always in the direction of reducing the number of predictors; the decisions to include (rather than exclude) variables from the original surveys were taken by majority or unanimous votes by the Delphi panel. None of the three variables for which the first author's casting vote resulted in them having LC status at the end of the Delphi process survived our subsequent simplification. In particular, we accepted that our approach might, on occasion, omit a predictor of well-being that could have been included, under the heading of LC, by another team of researchers. Any such omission would, of course, be likely to result in us underestimating the amount of variance explained.

Having identified a subset of variables that we believed unambiguously corresponded to LC, we then computed the variance explained by those variables in the principal outcome variable for life satisfaction in each study. First, we applied Multiple Classification Analysis (MCA; Lolle, 2008) with the MicroSiris software package (Van Eck and Van Eck, 2015). MCA was the technique used exclusively by Andrews and Withey (1976) for their analyses; indeed, MicroSiris is a direct descendant of the OSIRIS mainframe software used by these authors more than 40 years ago. Second, we applied multiple regression analyses that mirror those used by Campbell et al. (1976). These two analytic approaches (which are almost equivalent, because MCA is essentially a technique for automatically performing the necessary dummy coding to allow categorical predictors in multiple regression) allowed us to answer the question "If these authors had tried to estimate the percentage of variance explained by LC, what numbers would they have found?" Third, since the use of MCA or multiple regression with dummy coded predictors (as performed in the original studies) will likely result in inflated estimates because these highly flexible models might overfit the data on which they are estimated (see Yarkoni & Westfall, 2017, for an introduction to the issue of overfitting), we applied the technique of cross-validation to our regression models to arrive at more realistic/reliable estimates of the percentage of variance that can be explained by LC.

All of the information (R code and other documentation) needed to reproduce our analyses is available at <https://osf.io/423sr>.

Results

The lists of items measuring LC that were retained at the end of the classification exercise are shown in Table 7.1. The MicroSirius software indicated that the percentage of variance explained by the retained variables using MCA was $R^2 = 18.15\%$ (Andrews & Withey / May), 18.13% (Andrews & Withey / April), and 26.47% (Campbell et al.). Running multiple regression analyses in R, in which all predictors were treated as categorical variables, resulted in virtually the same estimates, highlighting the conceptual similarity between the MCA approach and the more common multiple regression framework. The percentage of variance explained according to these models were $R^2 = 18.19\%$ (Andrews & Withey / May), 18.32% (Andrews & Withey / April), and 26.47% (Campbell et al.). However, because treating all predictors as categorical variables results in extremely flexible models, these numbers might reflect overestimations: Hence, we further reduced the flexibility of the regression models by pooling rare categories and treating predictors such as years of education as continuous variables. We then applied repeated K -fold cross-validation (10 repetitions, 10 folds) to estimate how well the regression models performed on data that had not been used to estimate the model (i.e., *out-of-sample performance*). Using this more conservative approach, the *predicted* variance in life satisfaction was 1.92% (Andrews & Withey / May), 6.39% (Andrews & Withey / April), and 17.90% (Campbell et al.). It should be noted that, while the N s (1,297, 1,433, and 2,147) of the samples that we re-analyzed exceed the sample sizes of many psychological studies, they are comparably small for the research question at hand, which involves a multitude of potentially intercorrelated predictors, as well as an outcome variable that might be substantially affected by measurement error because of the brief nature of the measurement. This issue could account for the large drop in variance explained when moving from in-sample performance to out-of-sample performance and highlights the need for large-scale investigations to arrive at reliable and credible estimates.

Discussion

Perhaps somewhat ironically, our most conservative estimates of the variance in well-being explained by LC are even below 10% in two of the three samples. However, the variables included by Campbell et al. (1976) clearly outperformed the 10% figure, even when applying a method that is more conservative than the current domain standard in psychology

(Yarkoni & Westfall, 2017). This can probably be attributed to the fact that Campbell et al.'s study included a broader range of LC, including, for example, social factors such as the reported number of close friends. Once those variables that were essentially identical across the three studies are taken into account, our selection of predictors for the Campbell et al. study includes just six—out of around 100—variables that were unique to that study (i.e., not included by Andrews & Withey, 1976, in either of their studies), suggesting that a few well-chosen variables tapping LC can robustly predict a considerable amount of variance.

Change in the number of within-subject articles in *Journal of Happiness Studies*

On December 19, 2018, we visited the web site of the *Journal of Happiness Studies* and used a null search (<https://link.springer.com/search/?facet-journal-id=10902>) to retrieve all of the articles from the journal, sorted “Newest First.” The first article retrieved was Ingenfeld, Wolbring, and Bless (2018). We selected articles in chronological order down the list until we had obtained 50 abstracts from which we could determine whether the article described an empirical study, and if so, whether within-subject methods were used within that article. After checking 52 abstracts, the last of which was Tavor, Gonen, Weber, and Spiegel (2018) on page 3 of the results, we had obtained 50 empirical articles, of which 16 described within-subject methods. We next sorted the articles “Oldest First” and chose a page that contained articles from 2007–2008 (<https://link.springer.com/search/page/12?facet-journal-id=10902&sortOrder=oldestFirst>), the first of which was Waterman, Schwartz, and Conti (2008). Again, we moved down the list checking the abstracts (81 in total) until we had obtained 50 empirical articles, the last of which was Collins, Sarkisian, and Winner (2009) on page 16. Of those 50 articles, seven described within-subject methods. A list of all of the articles that we examined can be found at <https://osf.io/423sr>.

References

- Adler, M., & Ziglio, E. (Eds.) (1996). *Gazing into the Oracle: The Delphi method and its application to social policy and public health*. London, England: Jessica Kingsley.
- Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being: Americans' perceptions of life quality*. New York, NY: Plenum.
- Campbell, A., Converse, P. E., & Rodgers, W. L. (1976). *The quality of American life: Perceptions, evaluations, and satisfactions*. New York, NY: Russell Sage.
- Collins, A. L., Sarkisian, N. & Winner, E. (2009). Flow and happiness in later life: An investigation into the role of daily and weekly flow experiences. *Journal of Happiness Studies*, 10, 703–719. <https://doi.org/10.1007/s10902-008-9116-3>
- Ingenfeld, J., Wolbring, T. & Bless, H. (2018). Commuting and life satisfaction revisited: Evidence on a non-linear relationship. *Journal of Happiness Studies*. Advance online publication. <https://doi.org/10.1007/s10902-018-0064-2>
- Lolle, H. (2008). Multiple Classification Analysis (MCA): Unfortunately, a nearly forgotten method for doing linear regression with categorical variables. In *Symposium i anvendt statistik 2008* [Symposium in applied statistics 2008] (pp. 103–122). Copenhagen, Denmark: Copenhagen Business School Press.
- Lyubomirsky, S., Sheldon, K. M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology*, 9, 111–131. <https://doi.org/10.1037/1089-2680.9.2.111>
- Stitt-Gohdes, W. L., & Crews, T. B. (2004). The Delphi technique: A research strategy for career and technical education. *Journal of Career and Technical Education*, 20, 55–67. <https://doi.org/10.21061/jcte.v20i2.636>
- Tavor, T., Gonen, L. D., Weber, M., & Spiegel, U. (2018). The effects of income levels and income inequalities on happiness. *Journal of Happiness Studies*, 19, 2115–2137. <https://doi.org/10.1007/s10902-017-9911-9>
- Van Eck, N., & Van Eck, S. (2015). *MicrOsiris statistical and data management software system, Version 24*. Retrieved March 19, 2019 from <https://www.microsiris.com/MicrOsiris%20User%20Guide.pdf>
- Waterman, A. S., Schwartz, S. J. & Conti, R. J. (2008). The implications of two conceptions of happiness (hedonic enjoyment and eudaimonia) for the understanding of intrinsic motivation. *Journal of Happiness Studies*, 9, 41–79. <https://doi.org/10.1007/s10902-006-9020-7>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12, 1100–1122. <https://doi.org/10.1177/1745691617693393>

Table 7.1. Survey Questions Included in Our Multiple Classification Analysis and Regression Models.

Andrews & Withey (1976), May survey		Andrews & Withey (1976), April survey		Campbell et al. (1976)	
Q#	Text	Q#	Text	Q#	Text
11	Age of respondent	191	Respondent's age bracket	381	Respondent's age at time of interview
147	Sex of Respondent	192	Respondent's sex	46	Respondent's sex
148	Race of Respondent	185	Respondent's ethnic group	423	Respondent's race
142	Education of Respondent	136	Highest year of school/college completed by respondent	149	What is the highest level of education attained by respondent?
138	Total Family Income Bracket	170	Total family income	270	Respondent's total family income last year
16	Number of children under 18 in respondent's family unit	131	Number of people less than 18 years old in household	24	Number of children in dwelling unit
41	Are you working now, unemployed or laid off, retired and not working, or what?	141	Respondent's present employment status	163	Respondent's current employment status
42	What is your main occupation?	159	Respondent's main occupation	516	Respondent's census occupation classification
44	Do you work for someone else, or yourself, or what?	144	Is respondent employed by someone else, or self-employed?	166	Is respondent self-employed or does s/he work for someone else?
139	Are you married, widowed, divorced, separated, or have you never been married?	178	Respondent's marital status	288	Respondent's marital status
140	How long have you been married?			289	How long has respondent been married?
17	Age of youngest child under 18 in respondent's family unit	231	Age of respondent's youngest child		
18	Age of oldest child under 18 in respondent's family unit				
46	Are you the head of the family unit?	160	Is respondent head of family?		
		171	Does family own house/apartment, or rent it?	108	Does respondent own or rent the dwelling unit?
34	Would you say that you are better off or worse off financially than you were a year ago?				
		180	Do you have a telephone here at home?		
		44	Do you (or anyone else here in your family) own or lease a car?		
				104	How adequate is dwelling unit for respondent and family?
				265	Respondent's religious instruction when growing up
				266	Does respondent have problems with health?
				277	How many good friends does respondent have?
				426	Respondent's apparent intelligence
				387	Did respondent live with father & mother until age 16?

Note:

Q#: Question number from the original survey. Questions on the same line are considered to be equivalent.

Chapter 8

An Introduction to Criticality for Students of Positive Psychology

Nicholas J. L. Brown

Originally published as Chapter 12 in: Brown, N. J. L., Lomas, T. & Eiroa-Orosa, F. J. (Eds.). (2018). *The Routledge handbook of critical positive psychology*. London, England: Routledge.

Foreword

At about the same time that I was applying to the PhD program at Groningen, I was invited by my colleague Tim Lomas from the University of East London to help put together a book with the working title of *Critical Positive Psychology*. Initially, we thought that this might make for a 15-chapter trade paperback, but we received far more offers than we expected from colleagues who wanted to contribute chapters. We added a third editor, Francisco Jose Eiroa-Orosa of the University of Barcelona, and, in a surprisingly short time, found ourselves with a contract to produce the *Routledge International Handbook of Critical Positive Psychology*, which ended up having 36 chapters, 290,000 words, and nearly 600 pages.

As we noted in the introduction to the book, several of our colleagues were skeptical, with one remarking, “Critical positive psychology? Isn’t that an oxymoron?”. Nevertheless, we feel that the resulting book gives an idea of what positive psychology (PP) could be if it were to move away from what we see as a scientific and social dead end of attempting to improve the self-reported happiness levels of people who can afford to spend money on self-help books, coaching, and other interventions, and become, instead, a true psychology of human strengths. Such a psychology would acknowledge that “happiness” in any readily-definable sense is difficult to pursue, that simple, upbeat (“magic bullet”) psychological solutions for major issues such as depression or physical health problems are probably not available, and that US-American conceptions of well-being are not the only ones worth taking into account.

My first exposure to psychology in any form was my Master’s course in Applied Positive Psychology (MAPP) at the University of East London. Even when I graduated from that program, though, I had little idea of what psychology, *qua* academic discipline, was really about; that knowledge has been accumulating over the past five years. This topic therefore seemed like a natural thing for me to write about in a book dedicated to changing the image of PP, especially since my introduction to the subject—via a Master’s course aimed at older adults, many of whom will have had no previous formal training in psychology of any kind—was typical of how a large proportion of PP practitioners are trained. Hence, this chapter can be seen as a list of things that I wish I had known when I was taking my own Master’s course, in the hope that it might enable readers who are following similar courses to exercise appropriate degree of scientific skepticism towards some of the claims made by PP advocates.

Introduction

Since the Master of Applied Positive Psychology (MAPP) course at the University of Pennsylvania admitted its first cohort of students in 2005, more than a thousand people have graduated from the various master's level and other graduate diploma programs in positive psychology at universities worldwide, such as those at the University of East London (UEL) and Bucks New University in the UK, the University of Aarhus in Denmark, or the University of Melbourne in Australia. Even in France, where psychoanalysis of both the Freudian & Lacanian varieties still holds sway (see Kalkin, this volume, for an intriguing exploration of the relation between psychoanalysis and positive psychology), a small group of students started studying for the country's first postgraduate positive psychology diploma at the Université Pierre Mendès-France in Grenoble in the fall of 2015.

MAPP programs typically do not impose a requirement for previous training in psychology, statistical methods, or the philosophy of science as a prerequisite for admission (e.g., University of Pennsylvania, 2015). Such a policy certainly has some potential advantages in terms of diversity of viewpoints and life experience among students;¹ however, a drawback of this relaxed approach to formal prior knowledge is that students may not have the time during their one or two years of intensive study, which they often combine with a full-time job, to develop a sufficiently scientific (or, to put it another way, skeptical) approach to evaluating the evidence that is placed in front of them. A glance at the websites of a selection of MAPP graduates who work as coaches, for example, suggests that some students are able to graduate with very little understanding of the limitations of what statistically-based research methods—especially those involving single, underpowered studies of undergraduates—can tell us about the behavior of individuals in the real world, or indeed the ethicality of applying interventions, even when these are called “positive,” to people who might have substantial underlying psychological problems. Others arrive on MAPP courses with an assortment of pseudoscientific beliefs in discredited phenomena, including in some cases the “crystal healing and aromatherapy” derided by Seligman and Csikszentmihalyi (2000, p. 7), and leave with those same beliefs intact and unchallenged. A Google search for “positive psychology” combined with terms such as “reiki,” “chakra

¹ Without such a policy, this chapter, and perhaps even this book in its present form, might never have seen the light of day. Until I was admitted to the MAPP program at the University of East London in 2011, I could not name more than one living psychologist, and if asked what a *t* test was, I would probably have guessed that it involved a comparative tasting of warm beverages.

energy,” or “law of attraction” turns up a worrying number of MAPP graduates around the world who appear to be using their qualification as “scientists” principally to legitimize their businesses catering to the vulnerable (and the gullible) with “essential oils” and “neuro-linguistic programming.”

Of course, you, gentle reader, are different. If you are reading this Handbook of Critical Positive Psychology at all, it is because you are personally committed to good scientific practice, not averse to questioning what you are told, and prepared to walk boldly into the kitchen and demand to see how the sausage is made. In that case, this chapter is for you. It contains a number of insights that I have gleaned since I started my own MAPP journey, most of which would have been useful for me to have known right from the start. Hopefully it will get you started on your path of open-mindedness and critical thinking, which, as Seligman (2002, p. 142) put it, “is a significant part of the healthy trait of not confusing your own wants and needs with the facts of the world.”

Is psychology the hardest science?

Psychology is, of course, regarded as one of the “soft” sciences, but as the play on words in the title of this section suggests, it is a field in which making genuine progress is not easy, because its objects of study (i.e., human beings) are so complex and unpredictable. Writers such as Breckler (2005) and Banyard (2015) have noted that psychologists, in the 150 years or so since they first set up their own laboratories and journals, have not yet come up with many “technological” breakthroughs that can be reliably applied on an industrial scale to revolutionize how society functions, in the way that the physical and biomedical sciences have. Despite this, however, psychological researchers do not seem to lack ambition; it is not uncommon for the concluding paragraph of an empirical article or a grant application to suggest that the solution to any given problem of suboptimal human functioning that has frustrated philosophers for centuries is no more than a couple of studies away.

Figure 8.1. Abstract representations of the signal-to-noise ratio in the physical, biological, and psychological sciences.

Figure 8.1 summarizes the challenges facing psychologists. Each panel depicts a highly simplified representation of the limitations on knowledge facing researchers in the physical, biological (including medical), and social sciences. The inner circle represents the “signal”—what we might provisionally call the “truth,” if writers such as Feyerabend (1975) can be persuaded to look the other way for a moment—that scientists are trying to detect, while the fuzzy outer circles depict different types of “noise” that frustrate this aim. (Note that the relative sizes of the circles are not meant to suggest the relative scale of the signal-to-noise ratio for any given problem; in any of the fields shown, the “truth” can be as obvious as the sun in the sky, or as hidden as a single pulsar in a distant galaxy.) As one moves from left to right, new problems are introduced. The physicist or chemist does not generally have to worry too much about the properties of the matter under study being substantially different from one sample to the next. The cell biologist or medical researcher does have to cope with that problem, but typically does not have to deal with what happens when their samples decide to walk out of the laboratory. The psychologist, however, like other social scientists, must consider the full range of human behaviors at every stage of an experiment. Maybe our subjects who are currently filling out a questionnaire on positive emotions are paying attention, but we have no way of knowing whether they are in fact thinking about lunch or tomorrow’s exam. Furthermore, in psychology, the researcher’s own biases are likely to play

a substantial role; even if those biases can never be completely eliminated in other fields, a proton or a cancer cell is unlikely to try and guess, from a moment of eye contact across the laboratory, what the experimenter's real intention is, and react accordingly.

Further complicating the task of psychologists is the lack of information produced by their experiments. At one end of the information scale, the particle detectors and other instruments in the Large Hadron Collider generate more than a billion bytes of data every second (Pralavorio, 2015), and it may take months or even years of data collection to arrive at any particular conclusion about the behavior of a particle that is identical everywhere in the universe. In contrast, the information contained in all of the responses by 80 American undergraduate subjects to a 20-item psychological scale with responses in the range 1–5, such as the PANAS (Watson, Clark, & Tellegen, 1988) can be compressed into less than 500 bytes—less than is required to store all of the words in this paragraph—and yet many articles in positive psychology (and their accompanying press releases and subsequent popular media coverage) draw generalized conclusions about human well-being based on little more than this amount of information. These conclusions typically depend on the rejection—with a probability that lies on the “right” side of a boundary that was set arbitrarily over 80 years ago (Gigerenzer, 1994)—of a null hypothesis that is little better than a straw man (Berger & Sellke, 1987; Lindley, 2014), often with minimal consideration of the size of the observed effect, or whether the study has adequate statistical power to detect it.

Of course, many psychologists are aware of these issues, but frank discussions in the literature of the limitations to what psychology can achieve seem to be rare. Perhaps this is not surprising; in most fields of human endeavor, people do not generally spend a lot of time questioning the legitimacy of their particular field, especially if their involvement in it is accompanied by a regular monthly income. But an awareness of the limitations of psychology can be useful in order to detect and counterbalance some of the overhyped claims that are made from time to time. The thoughtful articles on the status and role of psychology by Ferguson (2015) and Koch (1981) are good starting points for this, as are several articles written by Paul Meehl towards the end of his career (e.g., Meehl, 1978, 1990). In a slightly longer form, one of the best introductions to the whole question of what we can know—and how we come to know it—in psychology is *Understanding Psychology as a Science* by Zoltán Dienes (2008). As for how these issues relate specifically to positive psychology: If my co-editors and I have done our job properly, you have one of the best available collections of material on that subject in your hands right now.

Beware of “sciencey-ness”

A corollary to the discussion in the previous section is that (positive) psychologists should exercise great caution in attempting to draw parallels between their field and the hard sciences. However, the temptation to do just that—sometimes humorously characterized as being due to a condition known as “physics envy”—seems to recur from generation to generation². Sixty years ago, the Russian-American founder of the sociology department at Harvard, Pitirim Sorokin, coined the term “quantophrenia” to describe the inappropriate application of mathematical techniques in the social sciences:

[W]hen the true quantitative method is replaced by pseudomathematical imitations; when the method is misused and abused in various ways; when it is applied to phenomena which, so far, do not lend themselves to quantification; and when it consists . . . in the mere transcription of mathematical formulae on paper without tying them to the relevant psychosocial units—then the approach misfires. Under these conditions, use of mathematical method becomes a mere quantophrenic preoccupation having nothing in common with mathematics and giving no recognition of the psychosocial world. (Sorokin, 1956, p. 103)

The attraction of such an approach to human behavior was summed up a few years later by the Polish-born British sociologist Stanislaw Andreski:

[Q]uantophrenic psychology gratifies the bureaucrats, big business, and the advertisers, by telling them what they want to hear. For it would make the job easier for the bureaucrat or any other kind of manipulator if everybody (except himself) resembled a push-button automaton, with perfectly predictable outputs following known inputs. (Andreski, 1972, p. 139)

Yet, echoes of the errors that Sorokin and Andreski were warning about decades ago can still be found today, especially at the intersection of positive psychology, business studies, and certain forms of neuroscience. A recent example was provided by Kilburg and Donohue (2014), who attempted to model business organizations in terms of thermodynamics and several other concepts abstracted from physics. These authors asserted that crystalline structures (which are defined by the endless repetition of identical unit cells in all directions)

² A good general principle for any student of psychology is to read one article from before 1970 for every two or three articles from the contemporary literature. It is remarkable how often the same ideas (and pitfalls) come round again and again, in only slightly different forms.

were good analogies for hierarchies (which are defined by differentiation at all levels), and likened the volatile behavior of a crowd to a combustible gas such as hydrogen or methane, which “under pressure can literally explode in the presence of a simple spark,” ignoring the inconvenient (for their analogy) fact that this explosion, and indeed the combustible nature of the gas, also requires the presence of plenty of oxygen—a necessity for life! Although such exercises might appear to be relatively harmless—any attempt to implement any sort of policy on the basis of these analogies would surely collapse immediately under the weight of its own absurdity—they nevertheless risk, by virtue of their status as “science” published in presumably reputable journals, comforting the bureaucrats and other manipulators referred to by Andreski in the above quotation.

Perhaps the best-known example of an inappropriate use of concepts drawn from the natural sciences in positive psychology is Fredrickson and Losada’s (2005) claim to have discovered a critical minimum positivity ratio—part of a “General Theory of Positivity” (p. 685), no less—on the basis of a single study purporting to show that the behavior of participants in a business meeting could be modeled by a set of equations abstracted from fluid mechanics. My co-authors and I (Brown, Sokal, & Friedman, 2013, 2014a) showed, through extensive analyses, both that Fredrickson and Losada’s theoretical model was completely invalid and that their empirical study also provided no support for their theory. However, to some extent, these results came as no surprise, simply on the basis of the absurd degree of determinism implied by Fredrickson and Losada’s theoretical claims.

In a more recent (and less extreme) example, Quoidbach et al. (2014) presented a new construct that they named “emodiversity,” which—analogously to the phenomenon of ecodiversity in the natural environment—supposedly represents an individual’s ability to experience a wide range of emotions, either on positive or negative axes, or combined into a single emodiversity score spanning both valences. Quoidbach et al. claimed to have observed favorable associations between emodiversity and health outcomes, which in the case of positive or negative emodiversity were independent of any concomitant associations between those outcomes and the frequency with which the equivalent emotions were experienced. They represented emodiversity using a formula devised by Shannon (1948) to describe the “entropy” (a concept analogous to physical entropy) of an information source in communication systems theory. However, because of their fundamental misunderstanding of the limits of the applicability of Shannon’s concept—for example, that it can only be used with actual counts of real-world phenomena, not with Likert-type scale scores that describe broad (and subjective) estimates of the relative frequency with which emotions are

experienced—Quoidbach et al. ended up with a construct that was entirely devoid of any predictive value (Brown & Coyne, 2017). As Blue and Brierley (2015, p. 2) put it, “Doing measurement is not necessarily doing science.”

So, be careful when you see claims in the media that “scientists” have discovered “how people work,” especially when these seem to suggest that the behavior of individuals can be predicted from their results. Remember that reliable effects in psychology are typically small, and that most studies observe the differences of means across groups, within which some people quite likely showed the opposite effect. Sciencey-looking formulas for happiness, whether this is “Feeling Good = $(M \times 16 + C \times 1 + L \times 2) + (T \times 5 + N \times 2 + B \times 33)$ ” (Robertson, 2013) or the more modest “ $H = C + S + V$ ” (Seligman, 2002, p. 45), should also be treated with skepticism until the author has explained what every component means, including the = and + signs.

Solutions in search of problems

Scientists typically start out by observing phenomena and then trying to explain them. Ever since humans first observed a frozen lake, it has been known that ice floats on water (unlike most solids, which are denser in solid than liquid form and therefore tend to sink), but it took until the 1920s and the discovery of the hydrogen bond (Smith, 1994) to explain this. Traditionally, the observation has come first, followed by the development of a theory to explain that observation, and then controlled experimentation that allows the theory to be refined until it is either (provisionally) accepted as a better explanation of the universe than what went before, or rejected.

However, some theories in positive psychology sometimes seem to be in need of a phenomenon to explain. For example, the idea of a positivity ratio (Fredrickson, 2013a), in the absence of any rigorous, consistent, and universally agreed-upon definition of what should be counted as one unit of experience of positive or negative emotions—assuming that those adjectives can even be defined, as Lazarus (2003) and Held (2018) have called into question—seems to be little more than post hoc numerology (Brown, Sokal, & Friedman, 2014b). Here are three further concepts from positive psychology that, I suggest, may be good candidates for discussion on similar lines.

“Fixed” and “growth” mindsets. Dweck’s (2006) theory of “growth” versus “fixed” mindsets is extremely influential in the education and child development literature. Yet, as pointed out by Alexander (2015), there seems to be very little real-world evidence for the purported dramatic effects of mindset shown in laboratory settings. When studies are

published that compare people with “growth” versus “fixed” mindsets, the former group consistently have better outcomes under laboratory conditions (e.g., Mueller & Dweck, 1998). Yet, such studies have been conducted at all levels of the population, notably at various stages in the education system. If growth mindset is really the key to success at every stage of development in education or sports, why are there any people with a fixed mindset available to be tested in, say, calculus classes at “a highly selective university in the Northeast United States” (Good, Rattan, & Dweck, 2012, p. 702), or among the professional sports stars that Dweck (2006) contrasted with Michael Jordan? Should these people not all have fallen by the wayside by now?

Upward spirals. A recurring idea in positive psychology is the notion of “upward spirals” of mutually reinforcing desirable outcomes, whose common theme is the power of positive emotions. For example, according to Fredrickson’s (2013b) upward spiral theory of lifestyle change, people can experience an ongoing virtuous cycle, in which the experience of positive emotions leads to an individual adopting a range of “wellness behaviors” (such as increasing their degree of physical activity or number of social interactions), which leads to improvements in physical and psychological health resources (such as better cardiac vagal nerve tone or greater mindfulness), which in turn lead to more positive emotions, thus sustaining the spiral. This appears to be a development of earlier upward spiral models from the same research program, such as that proposed by Fredrickson and Joiner (2002), who claimed to have shown that an upward spiral of positive affect and broad-minded coping led to improved emotional well-being, or Garland et al.’s (2010) model of psychological flourishing in which the ingredients in the upward spiral were positive emotions, optimism, and resilience. Garland et al. contrasted such upward spirals with the downward spiral of negative emotion, social withdrawal, and concomitant health problems (such as neuroendocrine disruption) that, they argued, characterizes the psychopathology of depression³.

However, there seem to be a number of problems with this model. First, several parameters of the purported upward spiral are missing. Exactly how much experience of positive emotions will bring about a one-point change (whatever that might mean) in

³ It is not clear whether, or to what extent, the psychopathology of depression is in fact characterized by an ongoing downward spiral. For example, in a meta-analysis of wait-list controlled trials and observational cohort studies, Whiteford et al. (2013) found that 53% of people in their study with untreated major depression experienced spontaneous remission within 12 months.

wellness behaviors? How many hours spent watching TV need to be replaced with jogging or social interaction to obtain a measurable improvement in health outcomes? How much healthier does a person have to become before they start to feel happier? Furthermore, what is the timescale for all this? If I decide to kick-start my upward spiral tomorrow by using loving-kindness meditation to self-generate positive emotions (Kok et al., 2013), how long will it take before my improved wellness behaviors and physical health indicators are keeping the upward spiral going on their own?

Second, when and how does the upward spiral stop? Figure 8.2, reproducing Garland et al.'s (2010) Figure 2, shows the “spring” that represents the spiral becoming larger and larger over time, both in the volume that it encloses and the thickness of the “material” that makes up the spiral. This picture, perhaps unwittingly, illustrates the fact that the effort required to continually improve one's situation will require a constantly increasing amount of energy. For example, if playing bridge twice a week builds an amount X of social connections, which in turn leads me to be healthier by some amount Y and experience some amount Z of extra positive emotions, then in order to make further progress with my social connections, I will presumably have to start playing bridge three times a week, or perhaps join a book club as well, while the next step in increasing my health will require me to get up even earlier in order to increase the length of my morning jog. At some point, I am going to encounter the limit of the number of hours in the day.

Figure 8.2. Upward spiral of flourishing (Garland et al, 2010). Reprinted with permission of Elsevier BV.

Third, leading on from the previous point, if all of this benefit can be brought about by something as simple as positive emotions, which do not appear to be in especially short supply, where are all the people who are already experiencing one or other (or, indeed, all) of these upward spirals? As Nickerson (2007, p. 555) commented on Fredrickson and Joiner's (2002) version of the upward spiral, "why is the world not divided into manics with superb coping skills, and all the rest of us?" Or, considering Fredrickson's (2013b) latest upward spiral model, why do none of us know anyone who has increased their happiness, social activities, circle of friends, daily jog distance, and oxytocin levels on a monthly basis for the past ten years?

The "happiness pie." Lyubomirsky, Sheldon, and Schkade (2005) suggested that no more than 10% of the variance in individuals' well-being is accounted for by circumstances, with 50% being due to genetics (the "set point" of happiness), and the remaining 40% being

available for people to make a difference to their well-being through their choice of actions. But in that case, why do people in Denmark report happiness levels of over 7.5 out of 10 on average (United Nations, 2013) while those in Togo report scores of less than 3? Is there something radically different in the genetic makeup of the inhabitants of Togo compared to those of Denmark, or are the former simply not trying hard enough? Alternatively, perhaps such things as safe drinking water, affordable healthcare, free education, a wide choice of consumer goods, and the protection of human rights might have more than a small impact on well-being?

To summarize: When thinking about theoretical claims, ask yourself what it would mean for them to be taken to their logical conclusion. In particular, look for start and end points: Ask how effect X is set in motion (e.g., by event Y), and the conditions under which it is sustained (e.g., as long as condition Z is satisfied). Ask how much of a difference effect X is expected to make, and in what populations. Physicists do not content themselves with saying “Light is fast”; we expect them to tell us *how* fast, just as we expect medical researchers to tell us, at least approximately, how many milligrams of this new drug are typically required to reduce the size of a tumor by 50%. Without clear statements of their expected limits, psychological theories can easily turn out to be unfalsifiable (Popper, 1962).

Correlation is not causation, no matter how appealing it might be

When a correlation is observed between two variables, A and B, there are generally considered to be three possible explanations of the relation between the two: A might cause B, or B might cause A, or A and B might have a common cause, C. (For simplicity, effects such as mediation by other variables are ignored here.) Our status as meaning-making, pattern-detecting organisms means that humans tend to overestimate the likelihood that the explanation is one of the first two possibilities, because A and B are frequently more salient phenomena than any of the many possibilities for C. This has obvious evolutionary advantages, especially when the cost of a false positive is relatively small and the cost of a false negative is relatively high. For example, if I think that there is a possibility that the rustling noise behind me might be caused by a predator, it is probably advisable to flee the scene immediately, even if most of the time the rustling turns out to be caused only by the wind, or a friend coming to greet me.

This strategy, however, is the exact opposite of what is required in scientific research, where the discovery of new direct causal relations is not a daily occurrence. Yet, despite Lazarus’s (2003) admonitions, positive psychology research continues to be largely

correlational in nature; we observe B, we think that A might be causing it, and we set out to “prove” it, despite Meehl’s (1990) observation that “in the social sciences everything is correlated with everything” (p. 210), giving rise to what he called the “crud factor” that introduces noise into every study. Of course, the limitations of such correlational studies are generally mentioned in the discussion sections of the articles that describe them, but there is often the feeling that the flagging-up of these limitations is done in a rather *pro forma* manner. Thus, for example, Eichstaedt et al. (2015), in the conclusion of their article that described a purported relation between expressions of anger or boredom on Twitter and mortality from heart disease within U. S. counties, noted that “associations between language and mortality do not point to causality” (p. 166). But this was arguably little more than a fig leaf, coming as it did at the end of an article whose title claimed that “Psychological Language on Twitter *Predicts* [emphasis added] County-Level Heart Disease Mortality” and that also contained claims such as “Language patterns reflecting . . . negative emotions—especially anger—emerged as *risk factors* [emphasis added]” (p. 159), and “Local communities create physical and social environments that *influence* [emphasis added] the behaviors, stress experiences, and health of their residents” (p. 166). The typical message in such cases is clear: While the authors cannot *prove* a causal relation from A to B, they can write up their article in such a way that the reader might very well get the *impression* that such a relation has been established. This process, which often leads to borderline or non-existent effects being transformed into popular media articles that open with the words “Scientists have discovered that people [behave in a certain way],” was well documented by Barr (2014), who named it “the innuendo machine.”

You can’t control for everything

Another common source of error in psychological research, including positive psychology, is the belief that the confounding effects of complex social phenomena can be eliminated from an analysis by taking a simple measure of each such phenomenon (or, perhaps, a particularly salient aspect of it) and incorporating this in a regression model or ANCOVA. However, while the inclusion of a statistical control of this type can potentially show that a confounding factor likely accounts for much or all of the apparent relation between the independent and dependent variables, thus demonstrating that the independent variable does not have any meaningful predictive value, the opposite is not the case. That is, the fact that additional variance appears to be associated with the independent variable of

interest, after “controlling” for various confounds, is not sufficient to demonstrate that this relation exists. There are two main reasons for this.

First, it is only possible to control for a finite number of confounding factors in the analysis. This number is constrained, among other things, by the researcher’s imagination in establishing the list, and also by the resources available to obtain the data for the selected variables. For example, a number of authors (e.g., Nevin, 2000; Reyes, 2007; although see Pinker, 2012, for counterarguments) have suggested that the drop in crime rates observed across the Western world since the early 1990s might be explained by the removal, around two decades earlier, of lead-based additives from gasoline. But in, say, 1960, how many researchers investigating crime levels in various neighborhoods would have thought to include exposure to automobile exhaust emissions (a factor likely to be closely related to socio-economic status, for example through housing density) in their models, or, if they had thought of that, would have been in a position to measure, or obtain data on, atmospheric lead conditions?

Second, many statistical controls use only an approximate measure of the underlying variable. Of course, this is true for any variable measured by social scientists, but in many cases, far more effort is expended on attempts to ensure the validity of the principal predictor and outcome variables (for example, by the use of scale validation methods such as Cronbach’s alpha, or by the careful design of experiments to eliminate extraneous influences on participants) than on equivalent measures to guarantee the validity and relevance of the covariates. Thus, for example, Cole et al. (2015) reported that “[a] dummy variable for smoking history was coded 1 if participants reported smoking ≥ 100 cigarettes during their lifetime” (p. 12), meaning that someone who smoked five cigarettes a week in their senior year in high school fifty years previously would have the same coding as a current smoker of two packs a day; it is not clear how such a measure might accurately predict the effect of smoking on any aspect of the health or psychology of participants in Cole et al.’s study. The general problem here is that the inclusion of a (necessarily imperfect) measure of a variable will always leave some true variance due to that variable being unexplained. With a sufficiently large sample size, the p value associated with this remaining variance can be statistically significant at the .05 level even if the imperfection in the measure is very small.

This last point has another important aspect. In the case of the main variables of interest (for example, the principal independent or dependent variables in a study), the researcher has an obvious interest in designing the study to squeeze the maximum amount of “variance juice” from each of those variables. In the case of controls, however, the opposite is

true. Indeed, there is a perverse incentive *not* to design one's control variables too carefully, lest they explain too much of the variance that the researcher had hypothesized was due to the principal independent variable(s). No conscious dishonesty is required for this phenomenon to have an effect; it is simply one of the many biases that researchers need to be aware of and attempt to eliminate (Feynman, 1985, pp. 338–346; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003).

In summary, when reading reports of studies in which various possible confounding factors were “controlled for,” it is advisable to pay close attention to the details. Ask yourself: why were these controls, and only these controls, chosen? How well do they operationalize the phenomenon being controlled for, in the context of this study with this population? What does the analysis look like without the controls; for example, are all of the steps of a hierarchical regression model shown? After all, if it were possible to control statistically for every possible confound, it would arguably not be necessary for researchers to go to the trouble of running complex and expensive randomized controlled trials. For a more rigorous analysis of some of the problems raised in this section, see Miller and Chapman (2001), Schisterman, Cole, and Platt (2009), or Westfall and Yarkoni (2016).

Be alert to possible conflicts of interest

The question of conflicts of interest is one of the major ethical issues in contemporary scientific publishing. Cases where researchers have failed to disclose industry funding or other financial interests still turn up regularly in the biomedical literature, in particular, despite the widespread awareness of the influence of funding on study outcomes (Lundh, Sismondo, Lexchin, Busuioac, & Bero, 2012) and the extensive requirements for disclosure imposed by publishers when articles are submitted to journals. However, conflicts of interest can be much more subtle than simply taking money from a pharmaceutical company to fund a test of its new drug. Aside from the near-universal problems of academia around funding and tenure, as well as questions of prestige and awards (as one ascends the seniority scale), positive psychology, in particular, also has to deal with the fact that several of its leading researchers derive a substantial personal income from writing and speaking about the field, with their own research often figuring prominently. There is, understandably, a large public appetite for knowledge about ways to be happier, and some of the leading “gurus” of positive psychology can command sums in excess of US\$10,000 for a single keynote speech or corporate morale-boosting session, as an online search for the names of some leading figures in the field with the term “public speaking” will readily show. Others leverage their name and

fame by writing popular books (frequently with a substantial “self-help” component), or offer online courses to help coaches get up to speed on “the latest cutting-edge findings” in the field. Of course, we are all entitled to earn a living; as one of the editors of this book, I expect to receive some very modest royalties, and there is nothing illegitimate about people applying their knowledge of positive psychology in areas such as business coaching, including presentations to large groups of paying listeners. Yet, once a scientist has become famous for a particular discovery, with an income at least partly dependent on people paying to hear about it, the understandable desire to maintain that income stream has the potential to affect the degree to which such a scientist is prepared to “lean over backwards” (cf. Feynman, 1985, p. 341) to show how his or her conclusions might be wrong.

Group-level studies do not always translate to individual-level outcomes

Positive psychology does not just aim to help individuals flourish; Seligman and Csikszentmihalyi (2000) noted that “people and experiences are embedded in a social context. . . . positive psychology needs to take the positive community and positive institutions into account” (p. 8). Yet, arguably, the improvement of communities and institutions cannot make much progress without change among individuals; as the former British Prime Minister, Margaret Thatcher, famously once said, “There is no such thing as society. There are individual men and women, and there are families.” While Thatcher’s words may not have been especially politically astute or compassionate, they at least make it clear that any institution is composed, on a day-to-day basis, of the people who live and work within it, and who agree, often tacitly, to respect the rules that define that institution’s existence on a formal or legal basis. One can change the rules, but unless the people are prepared to change as well, little improvement will occur; indeed, in many organizations, the rules are not the biggest obstacle to flourishing. While community and institutional flourishing undoubtedly requires leadership, those leaders are also individuals who will probably need to change, perhaps more than others in the organization. Thus, whatever the outcome that is being sought, individual-level change is a key requirement.

A great deal of positive psychology research is grounded in the traditions of social psychology⁴, in which the effects of interest are generally those that take place on average

⁴ I would argue that the publication of Peterson and Seligman’s (2004) *Character Strengths and Virtues: A Handbook and Classification*—which included concepts such as “judgment,” “self-regulation,” “perspective,” and “teamwork”—represented the implicit annexation by the positive psychology movement of a large part of the domain traditionally studied by social psychologists.

across the group being studied; hence, it is deemed appropriate to report that, say, the mean amount of money offered by participants in a dictator game differed significantly according to their experimental condition. However, such research is usually aimed, not at evaluating interventions at the individual level, but at understanding how people (either generally, or in specific pre-defined groups) tend to behave under certain circumstances. As a corollary, proposed actions that might result from such research also tend to be at a group or environmental level, such as a change to the physical arrangement of a store to encourage (most) consumers to buy more items, or the rewording of a sign to get people to wash their hands more often after using the toilet. Using the results of group-level analyses to support the design and delivery of individual-level interventions can be problematic, because some consideration will need to be given to the distribution of outcomes among the participants. For example, Kok et al. (2013) concluded, on the basis of a complex mathematical model rather than a simple comparison of group means (which would have not given a statistically significant result; Heathers, Brown, Coyne, & Friedman, 2015), that the practice of loving-kindness meditation led to an overall improvement in the vagal nerve tone (which was used in Kok et al.'s study as a surrogate for physical health) of their participants, compared to those in a wait-list control condition. However, what was not reported was that at the individual level, the same measure of vagal nerve tone of 12 out of 26 participants in the experimental condition was *worse* at the end of the study than at the start (Heathers, 2015).

Consequently, if a planned intervention has the potential to have a non-trivial negative impact on some participants, the benefits of this intervention to some participants must be carefully weighed against the potential costs to others. Furthermore, even if the group-level research appears to show that negative effects are rare, and the intervention is considered to be harmless at worst, it may not be appropriate to spend finite resources on administering it to a large group of people if only a few are likely to obtain benefit from it; perhaps those resources could have been better targeted at those most in need. The ethical issues involved here are sometimes complex, yet they are potentially of more than just academic interest for MAPP students, with demand for interventions by recent positive psychology graduates (e.g., from the U.S. Army's Comprehensive Soldier and Family Fitness program) being high. For

Interestingly, among the outspoken critics of positive psychology, mainstream social psychologists seem to be somewhat underrepresented. Only a cynic, however, would suggest that this might be related to a perception that an association with positive psychology might bring prestige, media exposure, and other personal benefits (cf. Stapel, 2014).

further exploration of these issues, informed by medical and international development contexts, see Brock and Wikler (2006).

An interesting approach for researchers wishing to investigate the effect of individual-level interventions is the single-case experimental design, also called N-of-1 design (Hersen, 1990). Studies using N-of-1 designs allow researchers to examine change within individuals in outcomes, and also typically simulate the clinical setting of an intervention better than between-group designs. For example, Woodworth, O'Brien-Malone, Diamond, and Schüz (2016) used an N-of-1 design to measure individual change in subjective well-being (SWB) in 15 participants, who received three well-known positive psychology interventions (PPIs), namely "Gratitude visit," "Three good things," and "Using signature strengths in a new way," over a period of 9–10 weeks. These authors found little evidence of any effect of the interventions on participants' SWB, and concluded that "Although such interventions may produce between-group changes in well-being, this does not necessarily mean they produce clinically relevant intra-individual changes" (p. 28). It remains to be seen what testing at the individual level will reveal about other PPIs.

Be negative when necessary

Positive psychology naturally tends to attract people who have (or, at least, express) an upbeat and positive approach to life, and who enjoy spreading this approach to others. However, this can lead to difficulties when one has doubts about any aspect of the field, because to criticize the work of others seems to require taking a "negative" viewpoint. The would-be critic has to take into account not only whether such criticism risks creating external friction with colleagues, but also a certain degree of internal friction in the form of cognitive dissonance (Festinger & Carlsmith, 1959), especially if one has made a substantial public (or financial) commitment to what might turn out to be a more nuanced subject than was first thought. Yet, sometimes a policy of *reculer pour mieux sauter* is necessary for real progress to be made.

For example, consider the superficially attractive idea that some countries are collectively "happier" than others and, hence, the implication that some factor(s) in those countries' ways of organizing society might be "causing" this extra happiness and could potentially be adopted by other countries. Surveys promoted by such prominent organizations as the United Nations (2013) and the Organisation for Economic Cooperation and Development (Wolchover, 2011) consistently show that Denmark is among the "happiest" countries in the world. Yet, a little prodding below the surface reveals some contradictory

messages. To go with its high levels of happiness, Denmark also has the highest reported rates of violence against women and girls in the European Union (European Union Fundamental Rights Agency, 2014). Furthermore, along with Iceland, Australia, Canada, and Sweden—all countries that also traditionally figure among the world’s happiest—Denmark is among the highest consumers of antidepressants per head in the Western world (Organisation for Economic Cooperation and Development, 2013, p. 103). Of course, it is possible to identify at least partial explanations for these phenomena: Perhaps Danish women are (several times) more likely than those in other countries to discuss their experiences of violence with researchers, or maybe the citizens of some countries are generally happier because their country has good care for those with mental health problems (at least, to the extent that treating depression with drugs is considered desirable; cf. Seligman, 2011, ch. 3). But if one seeks alternative explanations for those statistics related to negative phenomena, why should one not also seek reasons why self-reported happiness might not reflect the underlying reality? To ask this kind of question might involve challenging some of the central narratives of positive psychology. Such an exercise will, however, stand the student in good stead when it comes to considering the impact of articles such as that by Liu et al. (2015) who, in a prospective study of over 700,000 women in the United Kingdom, found no evidence for any effect of happiness on physical health.

Eat your own dog food

“Dogfooding” is a term from the software industry, which refers to the practice of personally using the programs that one designs or builds. (The word “dogfooding” allegedly originates from an American television advertisement for dog food in which the CEO of the manufacturer showed how good his firm’s product was by eating it.) Given the claimed benefits of PPIs, one might expect positive psychologists to be enthusiastic and frequent practitioners of these interventions. However, a number of conversations (only some of which took place in the bar at conferences), combined with some observations of their behavior under naturalistic conditions, lead me to believe that a surprising number of positive psychologists do not, in fact, spend very much time putting their recommendations into practice for themselves. This may be partly a question of the typical training process for psychologists. While psychoanalysts have traditionally always had their own analyst throughout their training (and, in many cases, their entire career), and other schools of psychotherapy also prescribe regular therapy—or, at a minimum, professional supervision from their peers—for practicing therapists, psychologists in training do not always spend

much time on the other side of the consulting room. This is, perhaps, reasonable in the world of “psychology as usual,” where practice is aimed at helping people with serious psychopathologies that are seriously limiting their functioning; by definition, such interventions are unlikely to apply to psychologists themselves unless they formally seek help as patients. With positive psychology, however, things are different: PPIs are generally modest in scale and not too onerous to undertake; they typically claim to be able to improve at least some aspect of the well-being of almost everybody; and the psychological risks involved are, in principle, small.

Thus, an excellent exercise for the positive psychology student is to take a selection of PPIs and conduct them on yourself. (On some MAPP programs, a project to do exactly this may be a coursework assignment.) Put yourself in the position of a client and keep your gratitude diary, or determine what your signature strengths are and use them in novel ways. Identify a measure of the outcome that the PPI is intended to change, and apply it at the beginning and end of each PPI, and again three months later. Of course, it’s not easy to step outside of your own skin and be a neutral experimenter at the same time as you are a participant in your own study; but this is perhaps not much more than practitioners ask of real clients when suggesting that they undergo a PPI, with their levels of well-being or gratitude or optimism being measured at the start and end of the process to show them how they have “improved.” Such appraisals, especially at the end of the PPI when the client probably has a good idea of the progress he or she is supposed to have made, may be colored by demand characteristics or and social desirability effects (Williams, Edwards, Patterson, & Chamow, 2011, pp. 236–239). Clients have to try to separate how they really feel from, first, how they themselves would like to feel (having invested their time and possibly public commitment, as well as a certain amount of their own or their employer’s money), and second, how they think the practitioner—who is probably a very nice person, whom the client is eager to please—would like them to feel. First-hand, honest reflection on this process can help the future practitioner understand the reality of how clients might experience it. It is also worth setting a reminder to “check in” with yourself a month or two after the end of a PPI to see whether its effects have lasted, in terms of either your feelings or your behavior. If this result is negative, be careful of the temptation to dismiss it because “I’ve been busy studying the subject at the same time, everything will have gotten mixed up.” Clients have complex lives too; after all, that’s often one reason why they look for help with their well-being. If PPIs are to have value, they have to work for real people in real situations. And if, by some chance, any given

PPI doesn't seem to make much of a difference to you in your situation, then perhaps there is something to be learned from that experience.

Read the primary literature

As one of my psychology teachers told me: “An average of seven people will read your article when it’s published, and one of those is your mother⁵.” Especially for the newcomer to the field, it can sometimes seem much easier to read summaries of research in handbook chapters (!), popular books, news articles, and blogs. Yet, empirical research articles published in peer-reviewed journals are the basis for the legitimacy of the personal, group, and educational interventions that positive psychology practitioners perform, and also that of the popular books, online courses, corporate speaking engagements, and public policy statements (cf. Diener & Diener, 2011, pp. 64–67) of the leaders of the field. Without this scientific basis, some aspects of positive psychology’s public face would be hard to distinguish from a succession of well-meaning platitudes, many of which would not be out of place in a Sunday school or personal development retreat (cf. Ehrenreich, 2010).

How, then, should the would-be critical reader approach an empirical article, perhaps with a famous positive psychology researcher as first (or, perhaps more likely, last) author? Just reading the abstract of an article and taking away the main points from it will not suffice; nor will reading only the Discussion section, where it is widely accepted that authors may speculate quite freely on the meaning of their study. The secret, it turns out, is to read the Method and Results sections in detail, and not just skim the latter looking for the *p* values. Make sure you understand what took place and why; a good way to do this is to try and explain it to a layperson, or—if you can’t find a layperson prepared to sit still for long enough—a classmate, perhaps in return for a similar favor. (In some cases, literally acting out the experimental scenario can be useful; start with the description of the method given by the authors and see if you can trace what the experimenter and participants actually did, step by step.) If the study was performed in a laboratory, do the variables really correspond to something of interest in the real world, and is there anything about the participants that might not be representative of the wider population (Henrich, Heine, & Norenzayan, 2010)? Do the stated hypotheses sound like something one would really specify in advance, in a funding application or a request for ethical approval, or might they instead have been “adjusted” after

⁵ Presumably she was not referring to those mothers who take a professional interest in their offspring’s publications.

the experiment so that they would appear to predict the results that were obtained (Kerr, 1998)? Are any of the results very strongly suggestive of a real effect, or do the p values all cluster around .02–.04 (Lakens, 2015; Simonsohn, Nelson, & Simmons, 2013)? In short, does this study make sense?

The primary literature in psychology contains many errors in the reporting of results, and more often than not, these errors favor the authors' hypotheses (Bakker & Wicherts, 2011). Consider, for example, the classic study by Rodin and Langer (1977) of the effects of choice and personal responsibility on residents in a retirement home. This article is part of almost every positive psychology course's introduction, not least for its striking claim that residents who were given a plant and told that they had to look after it themselves—versus those who were given a similar plant but told that the nursing home staff would look after it—were significantly more likely to be alive 18 months after the end of the study. Here, new positive psychology students are told, is scientific proof—or, at least, scientific evidence, the teacher might add with a wink—that a positive outlook on life is good for you.

However, there are at least two major problems with this study. First, Rodin and Langer's (1977) discovery was made post hoc, not hypothesized in advance; this raises the question of whether the p value associated with the z score that resulted from the comparison of the two groups has any meaning at all, given that it was not associated with the rejection of a null hypothesis specified a priori (cf. de Groot, 1956). As the authors themselves described it:

In the subsequent 18-month period following the intervention, only 7 of the 47 subjects (15%) in the responsibility-induced group died, whereas 13 of 44 subjects (30%) in the comparison group had died. Using the arcsine transformation for frequencies, this difference is reliable ($z = 3.14, p < .01$). Because these results *were so startling* [emphasis added], we assessed other factors that might have accounted for the differences. (Rodin & Langer, 1977, pp. 900–901)

Second, Rodin and Langer (1977) did not calculate their z score correctly. An alert reader of the article recalculated the correct value at $z = 1.73$, which has a p value of .096, and notified the editor of the journal. (Alternatively, one could build a contingency table containing the numbers of residents and deaths reported by Rodin and Langer and perform Fisher's exact test, which gives a p value of .129.) A correction was published, noting that “a more cautious interpretation of the mortality findings than originally given is necessary” (Rodin & Langer, 1978, p. 462), but this correction is almost never cited; the typical reader

who downloads the PDF file of the original 1977 article from the journal's website will not even be aware of the existence of this correction.

Another issue to consider is whether the study actually tests the stated hypotheses at all. For example, Fredrickson's (e.g., 2013b) broaden-and-build theory states that the experience of positive emotions will broaden a person's thought-action repertoire⁶ and build psychological resources over time. As such, it is clearly a statement about ongoing change within individuals, and therefore any test of this theory should examine participants over multiple time points. Yet, as pointed out by Nickerson (2007, 2014, 2018), much of the research that is claimed to support the broaden-and-build theory does not, in fact, do so. Indeed, in some cases, the claimed results—e.g., in Fredrickson and Losada (2005) where people who scored higher on positive emotions were more likely to be categorized as “flourishing”—are hard to distinguish from the circular reasoning of “happier people have higher levels of happiness” (cf. Snæbjörnsdóttir, 2010).

How can you get some practice with critical reading of the literature? Well, if your program or department has a journal club, join it; and if it doesn't, consider starting one. Ask your teachers if they will offer their assistance occasionally—although, once you have set the wheels in motion, it can sometimes be even more informative to try to make sense of an article without an experienced researcher present. A group of people, even those who are relatively new to the field, can often find things that the journal's reviewers and editors (and, in some cases, the researchers themselves) may have missed, simply by deploying many eyes and not accepting that “if it's been published in a peer-reviewed journal, it must be right.” And if, among a group of intelligent and motivated students, nobody understands what is happening in an article, that's often a sign of a problem. It's reasonable for a journal article not to be the most exciting thing you read all week; after all, academic writing sometimes has to trade liveliness for precision. But if an article makes no sense to several graduate students in the field, that's the author's problem, not the readers'.

You can make a difference

I want to end this chapter on a positive note. One of the remarkable things about positive psychology as a field is its openness (perhaps because of the feeling that nobody is likely to come to harm), which in turn leads to there being a very short distance from the

⁶ See Domachowska et al. (2016) and Gable and Harmon-Jones (2008) for empirical demonstrations that certain forms of positive affect may actually result in the opposite effect.

master's students to the top of the field. Almost all positive psychologists are more than happy to answer questions about their work (if only to have confirmation that someone other than their mother has read it), so with the possible exception of one or two “superstars” who may be too busy, your interest will almost certainly be rewarded with a constructive reply. Indeed, from first-hand experience, I can attest that a simple e-mail asking for clarification of a point in an article can be the beginning of a fruitful long-term research collaboration. Seligman (2011, ch. 4) referred to “The Magic of MAPP”; in my experience, that magic can be real, although it may perhaps take you in directions other than those that Seligman envisioned.

References

- Alexander, S. (2015). *No clarity around growth mindset yet* [Web log post]. Retrieved from <http://slatestarcodex.com/2015/04/08/no-clarity-around-growth-mindset-yet>
- Andreski, S. (1972). *Social sciences as sorcery*. London, England: André Deutsch.
- Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, 43, 666–678. <https://doi.org/10.3758/s13428-011-0089-5>
- Banyard, P. (2015, August 12). Where is psychology's non-stick frying pan? *The Psychologist*. Retrieved from <https://thepsychologist.bps.org.uk/where-psychologys-non-stick-frying-pan>
- Barr, D. J. (2014, September 21). *The innuendo machine: Turning scientific lead into journalistic gold* [Web log post]. Retrieved from <https://datahowler.wordpress.com/2014/09/21/the-innuendo-machine-turning-scientific-lead-into-journalistic-gold>
- Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82, 112–122. <https://doi.org/10.1080/01621459.1987.10478397>
- Blue, B., & Brierley, G. (2015). “But what do you measure?”: Prospects for a constructive critical physical geography. *Area*, 48, 190–197. <https://doi.org/10.1111/area.12249>
- Breckler, S. J. (2005). Psychology and the frontier of science. *Monitor on Psychology*, 36, 61.
- Brock, D. W., & Wikler, D. (2006). Ethical issues in resource allocation, research, and new product development. In D. T. Jamison, J. G. Breman, A. R. Measham, G. Alleyne, M. Claeson, D. B. Evans, . . . P. Musgrove (Eds.), *Disease control priorities in developing countries* (2nd ed., pp. 259–270). Washington, DC: World Bank.
- Brown, N. J. L., & Coyne, J. C. (2017). *Emodiversity: Robust predictor of outcomes or statistical artifact?* *Journal of Experimental Psychology: General*, 146, 1372–1377. <https://doi.org/10.1037/xge0000330>
- Brown, N. J. L., Sokal, A. D., & Friedman, H. L. (2013). The complex dynamics of wishful thinking: The critical positivity ratio. *American Psychologist*, 68, 801–813. <https://doi.org/10.1037/a0032850>
- Brown, N. J. L., Sokal, A. D., & Friedman, H. L. (2014a). The persistence of wishful thinking. *American Psychologist*, 69, 629–632. <https://doi.org/10.1037/a0037050>
- Brown, N. J. L., Sokal, A. D., & Friedman, H. L. (2014b). Positive psychology and romantic scientism. *American Psychologist*, 69, 636–637. <https://doi.org/10.1037/a0037390>
- Cole, S. W., Levine, M. E., Arevalo, J. M., Ma, J., Weir, D. R., & Crimmins, E. M. (2015). Loneliness, eudaimonia, and the human conserved transcriptional response to adversity. *Psychoneuroendocrinology*, 62, 11–17. <https://doi.org/10.1016/j.psyneuen.2015.07.001>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302. <https://doi.org/10.1037/h0040957>
- de Groot, A. D. (1956). De betekenis van “significantie” bij verschillende typen onderzoek [The meaning of “significance” for different types of research]. *Nederlands Tijdschrift voor de Psychologie*, 11, 398–409. English translation by E. J. Wagenmakers et al. (2014), *Acta Psychologica*, 148, 188–194. <https://doi.org/10.1016/j.actpsy.2014.02.001>
- Diener, E., & Diener, C. (2011). Monitoring psychosocial prosperity for social change. In R. Biswas-Diener (Ed.), *Positive psychology as social change* (pp. 53–71). Dordrecht, Netherlands: Springer.

- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Basingstoke, England: Palgrave Macmillan.
- Domachowska, I., Heitmann, C., Deutsch, R., Goschke, T., Scherbaum, S., & Bolte, A. (2016). Approach-motivated positive affect reduces breadth of attention: Registered replication report of Gable and Harmon-Jones (2008). *Journal of Experimental Social Psychology*, 67, 50–56. <https://doi.org/10.1016/j.jesp.2015.09.003>
- Dweck, C. S. (2006). *Mindset: The new psychology of success*. New York, NY: Random House.
- Ehrenreich, B. (2010). *Smile or die: How positive thinking fooled America and the world*. London, England: Granta.
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26, 159–169. <https://doi.org/10.1177/0956797614557867>
- European Union Fundamental Rights Agency. (2014). *Violence against women: An EU-wide survey—Main results*. Vienna, Austria: Author. Retrieved from http://fra.europa.eu/sites/default/files/fra-2014-vaw-survey-main-results-apr14_en.pdf
- Ferguson, C. J. (2015). “Everybody knows psychology is not a real science”: Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public. *American Psychologist*, 70, 527–542. <https://doi.org/10.1037/a0039405>
- Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, 58, 203–210. <https://doi.org/10.1037/h0041593>
- Feyerabend, P. (1975). *Against method: Outline of an anarchist theory of knowledge*. London, England: New Left Books.
- Feynman, R. P. (1985). “Surely you’re joking, Mr. Feynman!”: *Adventures of a curious character*. New York, NY: Norton.
- Fredrickson, B. L. (2013a). Updated thinking on positivity ratios. *American Psychologist*, 68, 814–822. <https://doi.org/10.1037/a0033584>
- Fredrickson, B. L. (2013b). Positive emotions broaden and build. *Advances in Experimental Social Psychology*, 47, 1–53. <https://doi.org/10.1016/B978-0-12-407236-7.00001-2>
- Fredrickson, B. L., & Losada, M. F. (2005). Positive affect and the complex dynamics of human flourishing. *American Psychologist*, 60, 678–686. <https://doi.org/10.1037/0003-066X.60.7.678>
- Gable, P. A., & Harmon-Jones, E. (2008). Approach-motivated positive affect reduces breadth of attention. *Psychological Science*, 19, 476–482. <https://doi.org/10.1111/j.1467-9280.2008.02112.x>
- Garland, E. L., Fredrickson, B., Kring, A. M., Johnson, D. P., Meyer, P. S., & Penn, D. L. (2010). Upward spirals of positive emotions counter downward spirals of negativity: Insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. *Clinical Psychology Review*, 30, 849–864. <https://doi.org/10.1016/j.cpr.2010.03.002>
- Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics*, 33, 587–606. <https://doi.org/10.1016/j.socec.2004.09.033>

- Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, 102, 700–717. <https://doi.org/10.1037/a0026659>
- Heathers, J. (2015). *A primer on Kok et al. 2013. Part 1: The paper* [Web log post]. Retrieved from <https://medium.com/@jamesheathers/a-primer-on-kok-et-al-2013-part-1-the-paper-54f6cbfed682>
- Heathers, J. A. J., Brown, N. J. L., Coyne, J. C., & Friedman, H. L. (2015). The elusory upward spiral: A reanalysis of Kok et al. (2013). *Psychological Science*, 26, 1140–1143. <https://doi.org/10.1177/0956797615572908>
- Held, B. S. (2018). Positive psychology's *a priori* problem. *Journal of Humanistic Psychology*, 58, 313–342. <https://doi.org/10.1177/0022167817739409>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hersen, M. (1990). Single-case experimental designs. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed., pp. 175–210). New York, NY: Plenum.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Kilburg, R. R., & Donohue, M. D. (2014). Leadership and organizational behavior: A thermodynamic inquiry. *Consulting Psychology Journal: Practice and Research*, 66, 261–287. <https://doi.org/10.1037/cpb0000013>
- Koch, S. (1981). The nature and limits of psychological knowledge: Lessons of a century qua “science.” *American Psychologist*, 36, 257–269. <https://doi.org/10.1037/0003-066X.36.3.257>
- Kok, B. E., Coffey, K. A., Cohn, M. A., Catalino, L. I., Vacharkulksemsuk, T., Algae, S. B., . . . Fredrickson, B. L. (2013a). How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone. *Psychological Science*, 24, 1123–1132. <https://doi.org/10.1177/0956797612470827>
- Lakens, D. (2015). On the challenges of drawing conclusions from *p*-values just below 0.05. *PeerJ*, 3, e1142. <https://doi.org/10.7717/peerj.1142>
- Lazarus, R. (2003). Does the positive psychology movement have legs? *Psychological Inquiry*, 14, 93–109. https://doi.org/10.1207/S15327965PLI1402_02
- Lindley, D. V. (2014). *Understanding uncertainty*. Hoboken, NJ: Wiley.
- Liu, B., Floud, S., Pirie, K., Green, G., Peto, R., & Beral, V. (2015). Does happiness itself directly affect mortality? The prospective UK Million Women Study. *The Lancet*, 387, 874–881. [https://doi.org/10.1016/S0140-6736\(15\)01087-9](https://doi.org/10.1016/S0140-6736(15)01087-9)
- Lundh, A., Sismondo, S., Lexchin, J., Busuioc, O. A., & Bero, L. (2012). Industry sponsorship and research outcome. *The Cochrane database of systematic reviews*, 12, MR000033. <https://doi.org/10.1002/14651858>
- Lyubomirsky, S., Sheldon, K. M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology*, 9, 111–131. <https://doi.org/10.1037/1089-2680.9.2.111>

- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195–244. <https://doi.org/10.2466/pr0.1990.66.1.195>
- Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, 110, 40–48. <https://doi.org/10.1037/0021-843X.110.1.40>
- Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology*, 75, 33–52. <https://doi.org/10.1037/0022-3514.75.1.33>
- Nevin, R. (2000). How lead exposure relates to temporal changes in IQ, violent crime, and unwed pregnancy. *Environmental Research*, 83, 1–22. <https://doi.org/10.1006/enrs.1999.4045>
- Nickerson, C. (2007). Theory/analysis mismatch: Comment on Fredrickson and Joiner's (2002) test of the broaden-and-build theory of positive emotions. *Journal of Happiness Studies*, 8, 537–561. <https://doi.org/10.1007/s10902-006-9030-5>
- Nickerson, C. A. (2014). No empirical evidence for critical positivity ratios. *American Psychologist*, 69, 626–628. <https://doi.org/10.1037/a0036961>
- Nickerson, C. A. (2018). There is no empirical evidence for critical positivity ratios: Comment on Fredrickson (2013). *Journal of Humanistic Psychology*, 58, 284–312. <https://doi.org/10.1177/0022167817740468>
- Organisation for Economic Cooperation and Development. (2013). *Health at a glance 2013: OECD indicators*. Paris, France: Author. Retrieved from http://www.oecd-ilibrary.org/pharmaceutical-consumption_5k3wgvb76ljk.pdf
- Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. Washington, DC: American Psychological Association.
- Pinker, S. (2012). *Pinker on lead and crime*. Retrieved from http://stevenpinker.com/files/pinker/files/pinker_comments_on_lead_removal_and_declining_crime.pdf
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88, 879–903. <https://doi.org/10.1037/0021-9010.88.5.879>
- Popper, K. (1962). *Conjectures and refutations: The growth of scientific knowledge*. New York, NY: Basic Books.
- Pralavorio, C. (2015). *LHC Season 2: CERN computing ready for data torrent*. Retrieved from <http://home.cern/about/updates/2015/06/lhc-season-2-cern-computing-ready-data-torrent>
- Purvis, A. (2006). The h-index: Playing the numbers game. *Trends in Ecology and Evolution*, 21, 422. <https://doi.org/10.1016/j.tree.2006.05.014>
- Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emodiversity and the emotional ecosystem. *Journal of Experimental Psychology: General*, 143, 2057–2066. <https://doi.org/10.1037/a0038025>
- Reyes, J. W. (2007). Environmental policy as social policy? The effect of childhood lead exposure on crime. *The B. E. Journal of Economic Analysis & Policy*, 7(1), 51. <https://doi.org/10.3386/w13097>

- Robertson, E. (2013, July 26). The formula for a happy life? Stay curious, live in the moment and look after your health. *Daily Mail*. Retrieved from <http://www.dailymail.co.uk/health/article-2378821/The-formula-happy-life-Stay-curious-live-moment-look-health.html>
- Rodin, J., & Langer, E. J. (1977). Long-term effects of a control-relevant intervention with the institutionalized aged. *Journal of Personality and Social Psychology*, 35, 897–902. <https://doi.org/10.1037/0022-3514.35.12.897>
- Rodin, J., & Langer, E. J. (1978). Erratum to Rodin and Langer. *Journal of Personality and Social Psychology*, 36, 462. <https://doi.org/10.1037/h0084387>
- Schisterman, E. F., Cole, S. R., & Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*, 20, 488–495. <https://doi.org/10.1097/EDE.0b013e3181a819a1>
- Seligman, M. E. P. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. New York, NY: Free Press.
- Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being*. New York, NY: Free Press.
- Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, 55, 5–14. <https://doi.org/10.1037/0003-066X.55.1.5>
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General*, 143, 534–547. <https://doi.org/10.1037/a0033242>
- Smith, D. A. (1994). *Modeling the hydrogen bond*. Washington, DC: American Chemical Society.
- Snæbjörnsdóttir, B. (2010). *The a priori nature of Fredrickson's theory of emotions* (Bachelor's thesis, University of Reykjavik, Iceland). Retrieved from <http://skemman.is/stream/get/1946/5286/15843/1/BS.pdf>
- Sorokin, P. (1956). *Fads and foibles in modern sociology*. Chicago, IL: Henry Regnery.
- Stapel, D. (2014). *Faking science: A true story of scientific fraud*. (N. J. L. Brown, Trans.). Originally published in Dutch as *Ontsporing* (2012). Retrieved from <http://nick.brown.free.fr/stapel>
- United Nations. (2013). *World happiness report 2013*. New York, NY: United Nations Sustainable Development Solutions Network. Retrieved from <http://unsdsn.org/resources/publications/world-happiness-report-2013>
- University of Pennsylvania. (2015). *Master of Applied Positive Psychology: Admissions and eligibility*. Retrieved from <https://www.sas.upenn.edu/lps/graduate/mapp/admissions>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54, 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11(3), e0152719. <https://doi.org/10.1371/journal.pone.0152719>
- Whiteford, H. A., Harris, M. G., McKeon, G., Baxter, A., Pennell, C., Barendregt, J. J., & Wang, J. (2013). Estimating remission from untreated major depression: A systematic review and meta-analysis. *Psychological Medicine*, 43, 1569–1585. <https://doi.org/10.1017/S0033291712001717>
- Williams, L., Edwards, T. M., Patterson, J., & Chamow, L. (2011). *Essential assessment skills for couple and family therapists*. New York, NY: Guilford Press.

- Wolchover, M. (2011, October 13). Denmark is happiest country, U.S. misses top 10. *LiveScience*. Retrieved from <http://www.livescience.com/16534-denmark-happiest-country-misses-top-10.html>
- Woodworth, R. J., O'Brien-Malone, A., Diamond, M. R., & Schüz, B. (2016). Happy days: Positive psychology interventions effects on affect in an N-of-1 trial. *International Journal of Clinical and Health Psychology*, 16, 21–29. <https://doi.org/10.1016/j.ijchp.2015.07.006>

Chapter 9

General Discussion

Brief summary of Chapters 2–7

In Chapter 2, we demonstrated how the repeated regression model that had been derived by Fredrickson et al. (2013), which involved collecting the beta coefficients from numerous regressions and comparing them with zero, produced essentially identical results when the independent variables were replaced with random numbers. In the article that was the subject of Chapter 3, Fredrickson et al. (2015) attempted to remedy this weakness by using a mixed-effect linear model, but this merely shifted the problem elsewhere. From the 2013 to the 2015 article, Fredrickson et al. also changed their hypotheses: Whereas in 2013 they framed their arguments in terms of equal and opposite effects of hedonic and eudaimonic well-being, by 2015 this had become a claim that only eudaimonic well-being would be expected to have an effect on health. (Somewhat bizarrely, they also claimed that their later article still provided confirmatory support for the hypotheses of the earlier one.) Chapters 2 and 3 demonstrated how this combination of problems means that Fredrickson et al.'s conclusions about the differential effects of hedonic and eudaimonic well-being—to the extent that these can be separated at all (Disabato, Goodman, Kashdan, Short, & Jarden, 2016)—are not supported by the evidence that they presented, and do not in fact represent the empirical basis for an objective approach to moral philosophy, as has been suggested by Fredrickson (cf. Downie, 2014).

In Chapter 4, we showed that Kok et al.'s (2013) investigation of the relation between meditation practice, positive emotions, and vagal nerve tone was conceptually and methodologically flawed, due to their use of a non-standard transformation of their main outcome variable (when a model using the standard transformation failed to achieve statistical significance). Additionally, there was no significant improvement in the intervention condition over time; the apparent difference between the groups was principally due to a decline in the scores of the control group, rather than an improvement in the intervention group (within which 12 out of 26 participants had worse physical outcomes at the end of the study than at the beginning). We also showed that in any case the methods of analysis used by Kok et al. were unable to ascribe any improvement that might have occurred to the intervention. Although Kok et al.'s statistical model was rather complex (a mediated moderation model, with two mediators), in effect the only independent variable was assignment to the intervention or control group.

In Chapter 5, we determined that Quoidbach et al.'s (2014) operationalization of their novel construct of “emodiversity,” using concepts borrowed from information theory and field ecology, was conceptually flawed because psychological constructs are typically not

measured in an open-ended manner, which is a requirement of the mathematical tools used to determine entropy in those other fields. We also showed that Quoidbach et al.'s technique of using the same instrument to measure both emotions and emodiversity was statistically flawed, as the two will necessarily always be strongly correlated with each other, thereby almost guaranteeing inflated beta coefficients when emodiversity is added to the emotion measure in a multiple regression model. Hence, it appears that the purported relations between emodiversity and health outcomes are nothing more than (systematic) statistical artifacts.

In Chapter 6, we showed that Eichstaedt et al.'s (2015) article has numerous methodological deficiencies that are not all readily visible on a first reading, and which make it difficult to have confidence in the results. In particular, we found that when the outcome variable (mortality from atherosclerotic heart disease) is replaced with mortality from self-harm, an unfavorable relation with positive emotions emerges; that is, higher county-level positive emotions, as expressed in Twitter language, are associated with higher rates of suicide. Furthermore, the idea of using Twitter language as a reliable detector of incipient cardiovascular problems in a community is likely to be impractical because of the large number of confounding factors. (For reasons of space, our published article did not contain our analysis showing that the single best predictor of AHD mortality in a county in any given year is AHD mortality in the same county in the previous year, with a correlation of about .86).

Finally, in Chapter 7, we demonstrated that the model of variance decomposition used by Lyubomirsky, Sheldon, and Schkade (2005)—namely, adding two approximate numbers from different study populations and subtracting their total from 100%—is inadequate for drawing conclusions about the effect of volitional activities on people's well-being, in part because it failed to consider measurement uncertainty or interaction effects. We also showed that the idea that activities undertaken by individuals might explain as much as 40% of the variance in well-being was likely to be greatly exaggerated. In particular, the claim by Lyubomirsky et al. that only around 10% of variance in well-being is due to "life circumstances" is probably based on a misreading of previous sources, with the true figure likely to be considerably higher.

Drawing the chapters together

Despite the wide variety of subject matter in the articles that were analyzed in Chapters 2 through 7, it is possible to discern a number of common themes that run through

most (or all) of them. In the following sub-sections, I review three of these themes. Some of these are also touched upon, in more general terms, in Chapter 8.

Use of inappropriate analytical methods. A common theme of most of the chapters in this thesis is the use by the original authors of inappropriate—and often quite complex—analytical and statistical methods to draw conclusions about the relation between variables that in some cases bordered on causal claims, mostly without any kind of experimental manipulation. Fredrickson et al. (2013, cf. Chapter 2) amalgamated the coefficients from 53 regressions into a single t test, but the non-independence of these results (which corresponded to the expression levels of groups of genes that are known to have similar functions) meant that the chance of Type I error was extremely high (Walsh, 1947); this issue was later explored in more detail by Barr (2014), Neuroskeptic (2014), and Nickerson (2017). Furthermore, as we showed, Fredrickson et al.’s model was so sensitive to single cases that the absence of one particular participant from their sample would have resulted in both of their main effects changing sign. In their later article, Fredrickson et al. (2015, cf. Chapter 3) made the arguably more appropriate choice to use a mixed-effects model, but this could not alter the fact that they were fitting models that in effect had 71 independent variables with only 76, 122, or (combining the two data sets) 198 participants (Walker, 2016). The only true experiment involving random assignment to condition was the study by Kok et al. (2013), discussed in Chapter 4; however, here once again a highly complex model was used, with the differences between the groups being inferred from the extent to which slopes of change were correlated (rather than simple pre–post comparisons in the outcome variable between the groups) after an unusual transformation had been applied to the data, while ignoring the fact that the purported physical health indicator for almost half of the participants in the intervention group got worse over the course of the study. Quoidbach et al. (2014, cf. Chapter 5) adopted a formula from computer science (whose use is somewhat controversial even in the empirical research area where it is principally applied, namely field ecology; cf. Magurran, 2004, p. 106), without apparently noticing that the most basic assumptions of that formula were not met by their data. Eichstaedt et al. (2015, cf. Chapter 6) performed their key analyses using a ridge regression/machine learning model that consists of several thousand lines of computer code and produced focal p values of .026 and .049, which—given their sample size of 1,347 counties—arguably provides greater support for the null hypothesis than for any alternative (Lakens & Evers, 2014). In all of these cases, it seems questionable whether the reviewers of the articles can have fully understood the methods that were being used.

In contrast to all of the complexities just mentioned, Lyubomirsky et al. (2005, cf. Chapter 7) arguably made the opposite error, by greatly oversimplifying their model of variance decomposition. Their implicit claim was that if two sources (genetic factors and life circumstances) could be shown to explain 60% of the overall variance in a model, that would mechanically leave 40% to be explained by exactly one other factor. However, they presented no evidence to support this claim, which was essentially the result of adding the two other percentages and subtracting the total from 100. That this was accepted by the reviewers could be explained to some extent by the fact that Lyubomirsky et al.'s article was mostly theoretical in nature, but the discourse around this article within positive psychology circles—at least, until the appearance of our preprint—has been framed as if it presented empirically-supported realities.

Lack of plausible causal mechanisms. Several chapters in the present thesis examine articles in which the authors appeared to make suggestions of a causal relation between mental states and physical health outcomes, even if at other points they acknowledged that their (mostly cross-sectional) research designs precluded such claims. Fredrickson et al. (2013, cf. Chapter 2) appeared to assume the existence of a mechanism whereby well-being could affect gene expression when they noted that the biological basis for the observed relationship between psychological and physical well-being “remains poorly understood, in part because of a paucity of information regarding the molecular signaling pathways that transduce positive psychological states into somatic physiology” (p. 13684). While Kok et al. (2013, cf. Chapter 4) at least had a longitudinal study design that was potentially capable of showing causality, their results—even if they had otherwise been statistically convincing—could only have established a change in heart-rate variability (HRV) as a result of their intervention; however, HRV is regarded by most researchers as an epiphenomenon (e.g., Routledge, Campbell, McFetridge-Durdle, & Bacon, 2010), so that any such change would not necessarily have been associated with other, more directly meaningful health improvements. Quoidbach et al. (2014, cf. Chapter 5) noted that their study was only cross-sectional and could not be used to draw causal inferences, although they then went on to beg the question of causality by asserting that “future research is needed to document the precise pathway by which emotion diversity improves outcomes” (p. 2064). Eichstaedt et al. (2015, cf. Chapter 6) also noted that the associations they claimed to have identified between Twitter language and heart disease could not be used to infer causality, but they then appeared to imply just such a relation when they discussed “community-level psychological factors that are important for the cardiovascular health of communities and should become the focus of

intervention” (p. 166) or noted that “Local communities create . . . environments that influence the behaviors, stress experiences, and health of their residents” (p. 166).

Several of the above examples are reminiscent of what McBee (in press, p. 7) described as “the talisman-like disclaimers that are found in so many non-experimental journal articles, in which authors piously disclaim any notion that their findings provide evidence of causation,” before noting that “The abstract, introduction, discussion, and implications sections of manuscripts are almost always leveraged on the causal interpretation that the authors forbear.”

Coverage in the popular media. Most of the articles that we critiqued in Chapters 2 through 7 have been reported on in popular media outlets, either as “science news” when the article first appeared, or subsequently as support for “lifestyle” pieces or mass-market (e.g., “self-help”) books. An example of the latter is Huffington (2014), who cited both Fredrickson et al.’s (2013, cf. Chapter 2) study of genomics and well-being (p. 239) and Kok and Fredrickson’s (2013, cf. Chapter 4) study of loving-kindness meditation (p. 45) as scientific support for her claims about human thriving.

Much of this (overwhelmingly favorable) media coverage went beyond what the authors had written (or, perhaps, been able to persuade the reviewers to accept) in their articles. Thus, for example, Smith (2013), describing the implications of Fredrickson et al.’s (2013; cf. Chapter 2) study, noted that the bodies of people who—according to Fredrickson et al.—had higher levels of hedonic (versus eudaimonic) well-being “are preparing them for bacterial threats by activating the pro-inflammatory response. Chronic inflammation is, of course, associated with major illnesses like heart disease and various cancers” (para. 14). Only a short step is required for a lay reader to infer from such a statement that Fredrickson et al. might in fact have demonstrated that not having the right balance of well-being could directly lead people to suffer from cancer or heart disease. (See Haber et al., 2018, for an examination of how media coverage can often “stretch” mere statistical associations into causal relations.) Regarding Eichstaedt et al.’s (2015, cf. Chapter 5) study, media coverage often ignored the limited inferences that could be drawn from the county-level nature of the data, with claims such as “Twitter Can Tell If *You’re* [emphasis added] More Likely To Get Heart Disease” (Ramsey, 2015) or “cities, towns and neighborhoods where people tended to express happiness and elation via their tweets were also more likely to enjoy better heart health” (Hays, 2015, para. 5).

Aftermath of the publication of our articles

The publication of the articles (or, in the case of Chapter 7, the preprint) corresponding to Chapters 2 through 7 was followed in each case by further developments that appear relevant to the interpretation of those chapters. I give a brief overview of these developments here, to the extent that they are supported by written evidence such as e-mail traces. In some cases there is plenty more that could be said, but the absence of a full evidential trail reduces this to little more than hearsay.

Following the publication of Chapter 2 in *PNAS*, in December 2014 we wrote to the Editor-in-Chief of the journal to request the retraction of Fredrickson et al.'s (2013) article, on the basis that we had shown its results to be essentially meaningless. We had still received no acknowledgement of this request by April 2015, at which point we discovered that Fredrickson et al. (2015) had used a different version of their 2013 data set (reference GSE45330), in which a coding error from the previous article had been corrected (see p. 48 of this thesis), in their subsequent article in *PLoS ONE*. We did not receive a substantive reply until February 2016, at which time the editorial board of *PNAS* indicated that they did not propose to correct or retract the Fredrickson et al. (2013) article, and considered the matter closed. To this date, then, the version of record of Fredrickson et al.'s (2013) article contains no indication that the results presented therein do not correspond (even after taking into account the deficiencies in the methods of analysis, discussed in Chapter 2) to the publicly available data set.

Turning to the second of the “genomics and well-being” articles, Fredrickson et al. (2015), which is the subject of our critique in Chapter 3, it is apparent that the left-hand panel of Figure 1C of this article is identical to Figure 2A of Fredrickson et al.'s earlier (2013) article. This implies that Figure 1C was produced using the uncorrected version of the GSE45330 data set mentioned in the previous paragraph, whereas the rest of Fredrickson et al.'s (2015) article (e.g., their Table 3) used the corrected version of this data set. Interestingly, in each case the version used provides the best support for the hypotheses being discussed at the respective point of the article. A correction was issued (Fredrickson et al., 2016), but this only applied to the 2015 (*PLoS ONE*) article, even though the same problem also affected the 2013 (*PNAS*) article. This correction also failed to point out the consequences of the change to Figure 1C for the authors' hypotheses. Despite these weaknesses, Fredrickson et al.'s (2013, 2015) articles continue to accumulate citations (179 and 68, respectively, since 2017, according to Google Scholar on August 18, 2019) at more than five times the rate of our rebuttals (31 and 10, respectively).

The article that is the subject of Chapter 4 (Kok et al., 2013) was defended by its authors in a reply to our critique that was published simultaneously with our article (Kok & Fredrickson, 2015). However, it was later revealed in a corrigendum (Psychological Science, 2016) that the data set used by Kok et al. (2013) had previously been used to publish another article (Kok & Fredrickson, 2010) whose results partially contradicted those of the later article, leaving doubt in the mind of anyone who read both articles as to which of them corresponded to the reality of what took place during the study. A subsequent article by Nickerson (2018) delved deeper into these issues and further demonstrated the fragility of the apparent effects that had been claimed by Kok et al. (2013).

In response to our article that became Chapter 5, the authors of the emodiversity article published a follow-up article in the same journal (Quoidbach et al., 2018) in which they rebutted our analyses and reported that they had reproduced the results of their Study 2 in a separate sample. However, given that a key part of our critique was that the purported effects of emodiversity are a statistical artifact that arises as a necessary consequence of the way the concept is defined, any such reproduction is unlikely to be meaningful. A subsequent article (Ong, Benson, Zautra, & Ram, 2017) used a different formula to compute emodiversity, but did not address any of our other concerns about the statistical problems inherent in adding a “recycled” version of a measure to a regression model that already contains the original measure.

In response to the appearance of the preprint version of Chapter 6, Eichstaedt and colleagues produced a reply, also in preprint form (Eichstaedt et al., 2018), which was the basis of some of Eichstaedt’s comments on our manuscript during the two rounds of peer review that it underwent; this preprint also included new analyses that appeared to show similar results to Eichstaedt et al.’s (2015) article. As with Quoidbach et al.’s response to Chapter 5, however, this reply by Eichstaedt and colleagues does not really address our most fundamental point, which is that there is no plausible reason why there should be a county-level psychological factor than can explain wide variations in atherosclerotic heart disease mortality across otherwise demographically similar counties, especially given the wide variety in the apparent (as certified by attending physicians) prevalence of this cause of death. Some of these issues were discussed in the British Psychological Society’s Research Digest by Jon Brock (2018).

Finally, although the ideas in Chapter 7 only became available in preprint form in May 2018 (with the article itself not appearing for another year), they seem to already be having some impact. For example, well-being researcher Maike Luhmann has recently

modified the “happiness pie” chart that she uses in her presentations to reflect our results (Rose, 2018), while in a conference discussion following the publication of the preprint, Lyubomirsky described the “happiness pie” model as merely a “thought experiment” that was “[b]ased on a very limited body of work” (Werner, 2018), which appears to represent a substantial change of emphasis.

Some recommendations

Based on the foregoing summary, I have formulated some recommendations for researchers in positive psychology who are investigating the relation between psychological states and physical health (cf. Chapter 8, where I make some recommendations for graduate students who are attempting to understand the positive psychology literature).

First, it is important that researchers be aware of the limits to what can be determined from observational data, especially when the purported effects are only detectable with multiple regression analyses or even more complex modeling techniques (cf. Nisbett, 2016). This can be especially problematic in the presence of measurement error in the covariates (Kahneman, 1965; Westfall & Yarkoni, 2016). Health outcomes are typically determined by a large number of factors, many of which will be unmeasured in any given study. These unmeasured variables may correspond to main effects or interactions, some of which may explain all or part of the variance in the variable of interest.

Second, it should be borne in mind that “Causal inference requires external information” (McBee, in press, p. 7). The three-variable causal diagrams for mediation and confounding differ only in the direction of one arrow (MacKinnon, Krull, & Lockwood, 2000), and any line in a causal path diagram can potentially be replaced with an additional variable from which two new causal paths emerge (Arah, 2008); strong theory is normally required to determine the correct selection of variables and the directions of causal influence. Sir Ronald Fisher (e.g., 1958) was skeptical, on statistical grounds, about the causal link between smoking and lung cancer, at least until Cornfield et al. (1959) demonstrated mathematically, using what became known as “Cornfield’s inequality”, that any other cause would need to be at least as strongly correlated with lung cancer as cigarette use, although it was clear that no candidate for such a variable was apparent. In contrast, it is rare to find reliably measureable psychological predictors that are more strongly associated with health outcomes than any other possible explanatory variables.

Third, researchers should take a principled position in dealing with the media. They need to explain what can and cannot be reliably concluded from their research, especially

from a single study (cf. Cohen, 1994), and emphasize where their findings fit into, and build incrementally upon, the existing literature. Of course, researchers cannot completely control what any particular journalist might choose to write about their article, although Sumner et al. (2014) noted that the majority of the inflation of claims (e.g., correlational studies being reported using language that implied causality) that they found in media articles about published scientific articles “was already present in the text of the press releases produced by academics and their establishments” (p. 4). But if a researcher’s rigorously reported results are misrepresented in the media—for example, if a statement that “It would be irresponsible to take [a model of Facebook data] and use it to say: ‘You’re depressed, you’re not depressed’” becomes a headline that reads “Your Facebook Posts Can Reveal If You’re Depressed” (Gonzalez, 2015)—then he or she could consider not including that article in the “media mentions” section of their personal or laboratory web pages.

Finally, researchers ought to be fully transparent in disclosing their interests. It is remarkable that, while authors are typically obliged to report the sources of all of the funding that enables them to conduct their studies, there seems to be no such obligation to disclose any downstream financial or business interests that might be tied in with that research (cf. Chivers, 2019). This is a particular concern for positive psychology, since its very nature—as a field that purports to be able to help almost everybody—means that its leading personalities can earn substantial amounts of money for public speaking (e.g., AAE Speakers, 2018), providing online seminars to large numbers of paying attendees (MentorCoach, 2013), or writing best-selling popular books based on their research. When submitting a manuscript for review, authors should consider including any financial or other benefits that they might reasonably expect to derive as a consequence of the publication of their research in the “Conflict of Interests” section of the submission form.

I have one other recommendation, this time for journal editors. It seems that there are no coherent standards for whether authors should take part in the peer review process of critiques of their articles. Journal policies in this regard appear to vary between “The critiqued authors are never invited to be reviewers,” “It’s up to the action editor,” and “We always invite the critiqued authors to be reviewers” (the last of which is the policy at *PeerJ*, for example). However, I believe that it is almost never appropriate for the author(s) of a critiqued article to be the reviewer(s) of the critique. The argument from editors in favor of doing this seems to be that the original authors are the best placed to spot flaws in the critics’ arguments, but I feel that any advantage that this might confer is outweighed by the obvious degree of self-interest that the original authors have in defending their work. Furthermore, if a

study or theoretical argument is so subtle and nuanced that only the original authors can have a full understanding of it, that would appear to call into question its value for science, given that the original authors are not going to be around for ever. Hence, I suggest that editors should adopt an explicit policy that, while critiqued authors may have the right of reply, they should not be full-rank reviewers of critiques of their work (and, likewise, the authors of the critique should not be involved in reviewing the reply).

A possible solution for better rigor and integrity: The open science movement

As part of my PhD program, I was required to complete a course with the title “Ethics of Research and Scientific Integrity for Researchers.” During this course I developed an understanding of the extent to which integrity is about more than avoiding specific unethical practices such as treating participants unethically, fabricating data, or using other questionable research practices in studies (John, Loewenstein, & Prelec, 2012). An important aspect of scientific integrity involves conducting research from the start with sufficient rigor that the results are likely to reflect true facts about the world that can be reproduced by others (cf. Feynman, 1974), rather than narrowly adhering to a set of rules that specify what constitutes a minimum standard of acceptable conduct. I would argue that several of the articles that we examined in Chapters 2 through 7 fail to live up to this ideal. Fredrickson et al. (2013; cf. Chapter 2) set out to explain the difference between two well-being constructs (hedonic and eudaimonic well-being) whose intercorrelation (.79) was the same as the one-week test–retest reliability of a leading measure of positive emotions (PANAS; Watson, Clark, & Tellegen, 1988), suggesting that there was almost no reliable variance to be explained whatever model these authors could have decided to use. Quoidbach et al. (cf. Chapter 5) devised (2014) and continue to advocate for (2018) a concept that we showed is, almost out of mathematical necessity, merely a statistical artifact. The discovery that Lyubomirsky’s (cf. Chapter 7) “happiness pie” is regarded by its inventor as no more than a “thought experiment,” as described in the previous section, will probably come as a surprise to the many graduates of positive psychology programs who have been taught it as something close to empirical fact.

Although the authors of the seminal article that arguably signaled the start of the burgeoning “open science” movement in psychology (Simmons, Nelson, & Simonsohn, 2011) have stated publicly (e.g., Simonsohn, 2018) that the title of their article (“False-Positive Psychology”) was not intended to suggest that their criticisms applied specifically or especially to positive psychology, it appears that positive psychology has been slower than

some other subfields of psychology to embrace that movement's aims of shifting psychological research towards the production of more rigorous and reproducible science. For example, the 2018 conferences of the European Health Psychology Society (EHPS, 2018, p. 28), the European Association of Personality Psychology (EAPP, 2018, p. 9), and the Association for Psychological Science (Fried, 2018) all featured talks, symposia, or workshops on reproducibility and other open science practices, while a group of developmental psychologists recently organized an entire conference dedicated to these topics (DEVSEC, 2018). In contrast, the book of abstracts (ECP, 2018) of the most recent European Conference on Positive Psychology contains descriptions of around 400 sessions in the form of symposia, workshops, oral presentations, and posters, but not one of these descriptions mentions anything to do with reproducibility, pre-registration of hypotheses or analysis plans, or any other aspect of open science. Yet arguably there are few subfields of psychology that currently have as much influence on public opinion and policymakers as positive psychology, so that considerable harm (if only, at best, in the form of wasted time and resources) could be caused by relying on poorly conducted studies involving many "research degrees of freedom" (Simmons et al., 2011)¹. If positive psychology is to make the world a better place and help people to flourish mentally or physically, it will need to show that its research methods are reliable and can be expected to lead to reproducible findings.

Limitations of the present thesis

This thesis is unusual in that it presents no empirical hypotheses or inferential statistics, and any (tentative) conclusions that can be drawn are at a meta-scientific level. Most of my recommendations are neither especially novel (cf. Cohen, 1990, 1994; Meehl, 1967, 1990), nor are the problems that they address unique to positive psychology research. However, the facts that these issues seem to recur time and again in the history of psychology makes me think that there may be room for them to be revisited in this particular context.

¹ A recent example is provided by Sutton (2017), who reported on a proposed intervention with elderly people who are about to receive a seasonal influenza vaccine, namely showing them humorous video clips. This was based on the results from a single study (Ayling et al., 2018) that showed that a self-reported positive mood on the day of the vaccination was associated with a higher level of subsequent seroprotection. However, that study was only correlational, and its principal results were based on regression analyses that did not control for marital status, income, education, tobacco use, BMI, or existing medical conditions, even though measures of all of those variables were available in the data set.

The principal theoretical limitation of this thesis is that it is difficult to establish a negative—in this case, that there is no good evidence demonstrating, after controlling for confounding demographic factors such as socioeconomic status or health-related behaviors such as smoking or exercise, that positive emotions lead, causally, to better physical health. About all that can be said with certainty is that the articles that my co-authors and I have critiqued in the preceding chapters do not seem to offer much in the way of reliable evidence to that effect. Of course, it cannot be excluded that such evidence, complete with plausible psychophysiological mechanisms, might emerge in the future.

It might also be objected that the articles that were the subject of earlier chapters are just an arbitrary sample among many others that could have been chosen. However, an important point about these articles, in particular, is that they represent the research output of some of the leading figures in the positive psychology movement. In five of the six cases, the senior author of the article in question is one of the five figures depicted on the “Mount Rushmore of Positive Psychology,” a visual homage to the “pioneers and prime movers in an explosive new field of psychology” (Mann, 2013, para. 20), while in the remaining case (Quoidbach et al., 2014, cf. Chapter 5), the lead author was the host of a nationally networked TV show in France on the subject of happiness (Deroy, 2012). These articles have been cited (in chapter order) 377, 90, 488, 71, 297, and 3,217 times, according to Google Scholar (August 18, 2018), and the average 2016 impact factor of the journals in which they appeared is 5.099. The impact of these articles on perceptions of psychology and health among both scholars and members of the public is thus potentially substantial. I am unaware of any other recent articles that make comparable claims for the health benefits of positive psychological states, and which have had equivalent coverage within scientific circles or in the popular media; again, though, this comes down to demonstrating the absence of something.

Closing thoughts

The authors of the highly-cited “False positive psychology” article—which may have had as much impact on the movement to reform psychology as the Bem/psi debacle or the discovery of Stapel’s fraud—have recently declared that psychology is entering a renaissance period: “Although [good research practices] are not yet common enough, it is clear that the Middle Ages are behind us, and the Enlightenment is just around the corner” (Nelson, Simmons, & Simonsohn, 2018, p. 17.19). I am somewhat skeptical of this claim. Certainly, under the collective banner of open science, the tools and methods that are required for psychology to improve are being developed at a rapid pace, and there are numerous

initiatives to make people across the academic spectrum aware of what needs to change. But the former international bureaucrat (*malgré moi*) in me wonders exactly what is different now—in terms of the power structures and incentives that govern people’s behavior in the day-in, day-out world of hiring, promotion, tenure, grant applications, and everything else that makes up a working scientist’s life apart from the occasional opportunity to get to the bench or write up some theoretical ideas—to the situation 40, 50, or 60 years ago, when writers such as Andreski (1972), de Groot (1956), and Meehl (1967) were pointing out much the same statistical, methodological, and organizational issues as the mostly younger scholars who are calling for reform today. There is a degree of irony in the idea that one of the biggest obstacles to progress in psychology is the tendency of people (in this case, psychologists themselves) to respond to incentives in rather predictable ways. However, despite my inherent pessimism, I intend to continue to make the best use that I can of my slightly unusual circumstances to try to help psychology to improve in the future. I am fortunate in that, being retired, I do not need to keep anybody happy for personal economic reasons, while being a relative newcomer to the field means that I also do not have a large amount of past research “baggage” of my own. My challenge is to mitigate the consequences of the corollary of those advantages, namely my still-limited expertise. Writing this thesis has been a step along the road to overcoming that.

The writing of much of this thesis has taken place in the shadow of recent world events, whether in the form of specific milestones such as the UK’s referendum vote to leave the European Union (“Brexit”) and the election of Donald Trump as President of the United States, or wider issues such as the rise of anti-science and even anti-Enlightenment tendencies in Western societies (often, although not exclusively, associated with right-wing populism). The latter tendencies are exemplified by the continued denial in some quarters of the reality of man-made climate change, but also by the worrying decrease in the take-up rate of vaccinations, in some cases with apparent endorsement from elected political leaders (Pianigiani, 2018). In the face of this denialism, it is important for scientists to stand up and speak out in favor of science as a method for understanding the world; however, as I and others have argued (Lewandowsky, Mann, Brown, & Friedman, 2016), it is also vital for scientists to act in ways that retain the trust of the public. If science does not self-correct, someone else may come along and try to correct it for us, and we may not like the way in which they do it.

References

- AAE Speakers. (2018). *Dr. Martin Seligman speaker profile*. Retrieved from <https://www.allamericanspeakers.com/celebritytalentbios/Dr.+Martin+Seligman/5005>
- Andreski, S. (1972). *Social sciences as sorcery*. London, England: Andre Deutsch.
- Ayling, K., Fairclough, L., Tighe, P., Todd, I., Halliday, V., Garibaldi, J., . . . Vedhara, K. (2018). Positive mood on the day of influenza vaccination predicts vaccine effectiveness: A prospective observational cohort study. *Brain, Behavior, and Immunity*, 67, 314–323. <https://doi.org/10.1016/j.bbi.2017.09.008>
- Arah, O. A. (2008). The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect: Covariate selection in the analysis of observational studies. *Emerging Themes in Epidemiology*, 5, 5. <https://doi.org/10.1186/1742-7622-5-5>.
- Barr, D. J. (2014). *Anatomy of a statistical artifact: Eudaimonic well-being and genomics* [Weblog post]. Retrieved from <https://datahowler.wordpress.com/2014/09/09/anatomy-of-a-statistical-artifact>
- Brock, J. (2018, April 23) *Are tweets a goldmine for psychologists or just a lot of noise? Researchers clash over the meaning of social media data* [Weblog post]. Retrieved from <https://digest.bps.org.uk/2018/04/23/are-tweets-a-goldmine-for-psychologists-or-just-a-lot-of-noise-researchers-clash-over-the-meaning-of-social-media-data>
- Chivers, T. (2019). Does psychology have a conflict-of-interest problem? *Nature*, 571, 20–23. <https://doi.org/10.1038/d41586-019-02041-5>
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304–1312. <https://doi.org/10.1037/0003-066X.45.12.1304>
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22, 173–203. <https://doi.org/10.1093/jnci/22.1.173>
- Disabato, D. J., Goodman, F. R., Kashdan, T. B., Short, J. L., & Jarden, A. (2016). Different types of well-being? A cross-cultural examination of hedonic and eudaimonic well-being. *Psychological Assessment*, 28, 471–482. <https://doi.org/10.1037/pas0000209>
- de Groot, A. D. (1956). De betekenis van “significantie” bij verschillende typen onderzoek. [The meaning of “significance” for different types of research; E. J. Wagenmakers et al., Trans.] *Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden*, 11, 398–409. Translation retrieved from http://www.ejwagenmakers.com/inpress/DeGroot1956_TA.pdf
- Deroy, G. (2012). J’ai décidé d’être heureux, la nouvelle émission de M6 [“I’ve decided to be happy,” the new show from M6]. *Cosmopolitan*. Retrieved from <https://www.cosmopolitan.fr/j-ai-decide-d-etre-heureux-la-nouvelle-emission-de-m6,2147,1857879.asp>
- DEVSEC. (2018). *Conference on the use of secondary and open source data in developmental science: Program guide*. Retrieved from http://srcd.org/sites/default/files/documents/devsec_program_guide.pdf
- Downie, A. [DownieAndrea] (2014, July 2). Great to listen to Barbara Fredrickson again this time on Molecular genetics and well being #ECPP2014 [Tweet]. Retrieved from <https://twitter.com/DownieAndrea/status/484288005860233216>

- EAPP. (2018). *19th European conference on personality: Program*. Retrieved from <http://conference.unizd.hr/ecp19/wp-content/uploads/sites/11/2018/07/Program.pdf>
- ECPP. (2018). 9st [sic] European conference on positive psychology: Book of abstracts. Retrieved from https://static.akcongress.com/downloads/ecpp/ecpp2018_book_of_abstract.pdf
- EHPS (2018). *32nd conference of the EHPS: Conference programme*. Retrieved from https://ehps.net/2018/wp-content/uploads/2018/08/EHPS2018_Handbook_14082018.pdf
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26, 159–169. <https://doi.org/10.1177/0956797614557867>
- Eichstaedt, J. C., Schwartz, H. A., Giorgi, S., Kern, M. L., Park, G., Sap, M., . . . Ungar, L. H. (2018). *More evidence that Twitter language predicts heart disease: A response and replication*. Retrieved from <https://psyarxiv.com/p75ku>
- Feynman, R. P. (1974). *Cargo cult science*. California Institute of Technology. Retrieved from <http://calteches.library.caltech.edu/51/2/CargoCult.pdf>
- Fisher, R. (1958). Cigarettes, cancer, and statistics. *Centennial Review of Arts & Science*, 2, 151–166.
- Fredrickson, B. L., Grewen, K. M., Coffey, K. A., Algae, S. B., Firestone, A. M., Arevalo, J. M. G., . . . Cole, S. W. (2013). A functional genomic perspective on human well-being. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 13684–13689. <https://doi.org/10.1073/pnas.1316887110>
- Fredrickson, B. L., Grewen, K. M., Algae, S. B., Firestone, A. M., Arevalo, J. M. G., Ma, J., & Cole, S. W. (2015). Psychological well-being and the human conserved transcriptional response to adversity. *PLoS ONE*, 10(3), e0121839. <https://doi.org/10.1371/journal.pone.0121839>
- Fredrickson, B. L., Grewen, K. M., Algae, S. B., Firestone, A. M., Arevalo, J. M. G., Ma, J., & Cole, S. W. (2016). Correction: Psychological well-being and the human conserved transcriptional response to adversity. *PLoS ONE*, 11(6), e0157116. <https://doi.org/10.1371/journal.pone.0157116>
- Fried, E. (2018, May 30). *Reflections on APS 18: open science, transparency, and inclusion* [Weblog post]. Retrieved from <http://eiko-fried.com/reflections-on-aps-18-open-science-transparency-and-inclusion>
- Gonzales, R. (2018). Your Facebook posts can reveal if you're depressed. *Wired*, October 16. Retrieved from <https://www.wired.com/story/your-facebook-posts-can-reveal-if-youre-depressed>
- Haber, N., Smith, E. R., Moscoe, E., Andrews, K., Audy, R., Bell, W., . . . Suarez, E. A. (2018). Causal language and strength of inference in academic and media articles shared in social media (CLAIMS): A systematic review. *PLoS ONE*, 13(5), e0196346. <https://doi.org/10.1371/journal.pone.0196346>
- Hays, B. (2015, January 21). *Twitter knows when you're going to have a heart attack*. Retrieved from <https://www.upi.com/4631421876406>
- Huffington, A. (2014). *Thrive: The third metric to redefining success and creating a life of well being wisdom and wonder*. New York, NY: Harmony.

- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524–532. <https://doi.org/10.1177/0956797611430953>
- Kahneman, D. (1965). Control of spurious association and the reliability of the controlled variable. *Psychological Bulletin*, 64, 326–329. <https://doi.org/10.1037/h0022529>
- Kok, B. E., Coffey, K. A., Cohn, M. A., Catalino, L. I., Vacharkulksemsuk, T., Algoe, S. B., . . . Fredrickson, B. L. (2013). How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone. *Psychological Science*, 24, 1123–1132. <https://doi.org/10.1177/0956797612470827>
- Kok, B. E., & Fredrickson, B. L. (2010). Upward spirals of the heart: Autonomic flexibility, as indexed by vagal tone, reciprocally and prospectively predicts positive emotions and social connectedness. *Biological Psychology*, 85, 432–436. <https://doi.org/10.1016/j.biopsycho.2010.09.005>
- Kok, B. E., & Fredrickson, B. L. (2015). Evidence for the upward spiral stands steady: A response to Heathers, Brown, Coyne, and Friedman (2015). *Psychological Science*, 26, 1144–1146. <https://doi.org/10.1177/0956797615584304>
- Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, 9, 278–292. <https://doi.org/10.1177/1745691614528520>
- Lewandowsky, S., Mann, M. E., Brown, N. J. L., & Friedman, H. (2016). Science and the public: Debate, denial, and skepticism. *Journal of Social and Political Psychology*, 4, 537–553. <https://doi.org/10.5964/jspp.v4i2.604>
- Lyubomirsky, S., Sheldon, K. M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology*, 9, 111–131. <https://doi.org/10.1037/1089-2680.9.2.111>
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science*, 1, 173–181. <https://doi.org/10.1023/A:1026595011371>
- Magurran, A. E. (2004). *Measuring biological diversity*. Malden, MA: Blackwell.
- Mann, J. D. (2013, October 11). *Happiness revolution*. Retrieved from <https://www.livehappy.com/science/positive-psychology/happiness-revolution>
- McBee, M. (in press). *Causal analysis*. London, England: SAGE.
- Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103–115. <https://doi.org/10.1086/288135>
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195–244. <https://doi.org/10.2466/pr0.1990.66.1.195>
- MentorCoach. (2013). *Love 2.0—The master class*. Retrieved from <https://www.mentorcoach.com/class/love-2-0-master-class>
- Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology’s renaissance. *Annual Review of Psychology*, 69, 17.1–17.24. <https://doi.org/10.1146/annurev-psych-122216-011836>
- Neuroskeptic. (2014). *(False?) Positive psychology meets genomics*. [Weblog post]. Retrieved from <http://blogs.discovermagazine.com/neuroskeptic/2014/08/27/false-positive-psychology-genomics>

- Nickerson, C. A. (2017). No evidence for differential relations of hedonic well-being and eudaimonic well-being to gene expression: A comment on statistical problems in Fredrickson et al. (2013). *Collabra: Psychology*, 3(1), 11. <https://doi.org/10.1525/collabra.81>
- Nickerson, C. A. (2018). Overstating the effects of loving-kindness meditation: Comment on Kok et al. (2013). *Psychological Science*, 29, 463–466. <https://doi.org/10.1177/0956797616682946>
- Nisbett, R. (2016). The crusade against multiple regression analysis. *Edge*, January 21. Retrieved from https://www.edge.org/conversation/richard_nisbett-the-crusade-against-multiple-regression-analysis
- Ong, A. D., Benson, L., Zautra, A. J., & Ram, N. (2017). Emodiversity and biomarkers of inflammation. *Emotion*, 18, 3–14. <https://doi.org/10.1037/emo0000343>
- Pianigiani, G. (2018, July 5). Proof of children's vaccinations? Italy will now take parents' word for it. *New York Times*. Retrieved from <https://www.nytimes.com/2018/07/05/world/europe/italy-vaccines.html>
- Psychological Science. (2016). Corrigendum: How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone. *Psychological Science*, 27, 931. <https://doi.org/10.1177/0956797616647346>
- Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emodiversity and the emotional ecosystem. *Journal of Experimental Psychology: General*, 143, 2057–2066. <https://doi.org/10.1037/a0038025>
- Quoidbach, J., Mikolajczak, M., Gruber, J., Kotsou, I., Kogan, A., & Norton, M. I. (2018). Robust, replicable, and theoretically-grounded: A response to Brown and Coyne's (2017) commentary on the relationship between emodiversity and health. *Journal of Experimental Psychology: General*, 147, 451–458. <https://doi.org/10.1037/xge0000400>
- Ramsey, L. (2015). Twitter can tell if you're more likely to get heart disease. *Popular Science*, January 23. Retrieved from <https://www.popsoci.com/twitter-can-tell-if-youre-more-likely-get-heart-disease>
- Rose, N. (2018, May 10). Re-Slicing the happiness pie: How much of our well-being can be influenced through intentional activities? *Mappalicious*. Retrieved from <https://mappalicious.com/2018/05/10/re-slicing-the-happiness-pie-how-much-of-our-well-being-can-be-influenced-through-intentional-activities>
- Routledge, F. S., Campbell, T. S., McFetridge-Durdle, J. A., & Bacon, S. L. (2010). Improvements in heart rate variability with exercise therapy. *Canadian Journal of Cardiology*, 26, 303–312. [https://doi.org/10.1016/S0828-282X\(10\)70395-0](https://doi.org/10.1016/S0828-282X(10)70395-0)
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Simonsohn, U. [uri_sohn]. (2018, July 17). The title has nothing to do with positive psychology [Tweet]. Retrieved from https://twitter.com/uri_sohn/status/1019235237971288064
- Smith, E. E. (2013). Meaning is healthier than happiness *The Atlantic*, August 1. Retrieved from <https://www.theatlantic.com/health/archive/2013/08/meaning-is-healthier-than-happiness/278250>

- Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Venetis, C. A., Davies, A., . . . Chambers, C. D. (2014). The association between exaggeration in health related science news and academic press releases: Retrospective observational study. *BMJ*, *349*, g7015. <https://doi.org/10.1136/bmj.g7015>
- Sutton, J. (2017). “Imagine the brave new world... these are small psychological interventions capable of big effects.” *The Psychologist*, *30*, 58–16. Retrieved from <https://thepsychologist.bps.org.uk/volume-30/november-2017/imagine-brave-new-world-these-are-small-psychological-interventions-capable>
- Walker, J. A. (2016). Monte Carlo simulation of OLS and linear mixed model inference of phenotypic effects on gene expression. *PeerJ*, *4*, e2575. <https://doi.org/10.7717/peerj.2575>
- Walsh, J. E. (1947). Concerning the effect of intraclass correlation on certain significance tests. *Annals of Mathematical Statistics*, *18*, 88–96. <https://doi.org/10.1214/aoms/1177730495>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*, 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Werner, K. [kaitlynmwerner] (2018, July 18). Side note from Sonja Lyubomirsky: the happiness pie was more of a thought experiment from 18 years ago. Based on a very limited body of work and they’re open to re-interpretations! [Tweet]. Retrieved from <https://twitter.com/kaitlynmwerner/status/1019608693468467201>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, *11*(3), e0152719. <https://doi.org/10.1371/journal.pone.0152719>

Chapter 10

Other Projects

Over the past four years, while working on the articles that became the principal chapters of this thesis, I have also been involved with a number of other projects that, although not always directly related to my principal theme of the relation between positive psychology and physical health, are consistent with the broader theme of the critical examination of published (mostly psychological) science. Some of these investigations have been conducted via the peer-reviewed literature, while others have been based on the use of newer forms of scholarly communication such as preprints and the judicious use of social media to spread awareness.

Reconstructing data: GRIM and SPRITE

With my colleague James Heathers (cf. Chapter 4), I have been working on ways to learn as much as possible about the data underlying a published article without having access to the data set itself. This project started when we were reading an article (Guéguen, 2015) and realized that we could reconstruct the entire data set from the summary statistics, given the extreme regularity that was present in the reported table of results (see Table 10.1). This technique, which we formalized under the name of Granularity-Related Inconsistency of Means (GRIM), is based on an extremely simple idea: If an integer measure, such as a Likert-type scale, is administered to N participants, then the decimal part of the mean of their scores, rounded to two decimal places, can only take on N of the numerically possible 100 values. Thus, for example, with $N = 41$, a mean of 2.63 or 2.66 in response to a single-item scale is consistent with correct rounding, but a mean of 2.64 or 2.65 is not. Checking that a mean is consistent in this way is simple; a more complex procedure (known as GRIMMER) is needed to check that a reported standard deviation is consistent. Using GRIM, we were able to demonstrate that the descriptive statistics in several other articles by the same author were either highly implausible (e.g., Guéguen, 2012a) or mathematically impossible (e.g., Guéguen, 2012b).

Table 10.1, reproduced from Table 1 of Guéguen (2015). Original caption: Mean of participants' helping score (SD in parentheses) according to experimental condition and the sex of pedestrians.

Gender of participant	Hairstyle condition			Total
	Natural	Ponytail	Bun	
Male pedestrians ($N = 90$)	2.80 (0.41)	1.80 (0.76)	1.80 (0.76)	2.13
Female pedestrians ($N = 90$)	1.80 (0.41)	1.60 (0.81)	1.60 (0.50)	1.67
Total	2.30	1.70	1.70	

Given the additional information (reported in the text of the article) that there were exactly 30 participants in each condition (Hairstyle $\times$ Gender) and that helping behaviors were measured on a scale from 1 to 3, it is possible to show that for each condition, there is only one combination of 30 values that gives the reported mean and SD when applying correct rounding to two decimal places. Furthermore, the number of participants giving each response (1, 2, or 3) is always a multiple of 6 in every condition.

Extending our investigations to the broader literature led to the discovery that a large number of articles published in three leading psychology journals of between 2011 and 2015 contained errors at the most reporting level of the descriptive statistics (Brown & Heathers, 2017). Since then, we have extended the idea of GRIM to a new technique called Sample Parameter Reconstruction via Iterative TEchniques (SPRITE; Heathers, Anaya, van der Zee, & Brown, 2018), a software-based approach that allows users to reconstruct plausible distributions of data from reported sample characteristics (mean, standard deviation, range, and sample size). Briefly, SPRITE works by (1) generating a sample of integers having the required size and range of values, first two of those properties; (2) adjusting those values upwards or downwards until the desired mean is obtained; and (3) adjusting pairs of values (adding 1 to one and subtracting 1 from another) in an iterative fashion until the desired standard deviation is obtained. I wrote the most widely-used implementation of SPRITE, which is available for anyone to download on their own computer or use as a web-based application (<http://shiny.ieis.tue.nl/sprite/>).

Questioning “newsworthy” claims in other fields beyond (positive) psychology

Although the chapters in this thesis focus on positive psychology, I have also been involved in a number of collaborative projects that examined spectacular claims in other areas of psychology and beyond. For example, in 2015 I worked with several colleagues (Holcombe, Brown, Goodbourn, Etz, & Geukes, 2016) on a collaborative reanalysis of an article in *Psychological Science* (Thorstensen, Pazda, & Elliott, 2015) that made some remarkable claims about the effect of negative emotions on people’s perceptions of different colors. In particular, in one of their two studies, Thorstenson et al. appeared to show that people’s ability to distinguish blue from yellow was severely impaired after watching a sadness-inducing video clip with a duration of just over two minutes. This finding was widely reported in outlets such as *Time* magazine (Basu, 2015) and the *Washington Post* (Kaplan, 2015). While examining the data set for this study, I found a number of highly regular patterns that could not reasonably have been the results of observing human subjects. Thorstenson et al.’s article was retracted after being published online, but before being assigned a page number in the paper version of *Psychological Science*.¹

In 2016, together with my colleagues Casper Albers and Stuart Ritchie, I was involved in the critical reanalysis of an article in *Nature* (Dong, Milholland, & Vijg, 2016) that claimed to have found evidence that human lifespan has “plateaued” and will likely never increase beyond 115 years. This article had also attracted considerable interest in the popular media. In our commentary (Brown, Albers, & Ritchie, 2017), we noted that Dong et al. had made some basic errors in their analyses, including using the existence of a single outlier to infer conclusions about the population. They had also chosen to use what was

¹ This case provided an interesting example of a theme that I have observed in numerous situations in psychological research, namely the question of how the effects observed by psychologists—whether through experiments in the laboratory, or extracted from regression analyses or more complex types of linear models—can be so powerful, yet nobody seems to have noticed these effects in everyday life (cf. Lakens, 2017). This theme is worth considering when examining some of the claims made by positive psychology researchers, such as the idea that emotional diversity is a better predictor of health than smoking status, diet, and exercise (Chapter 5) or that volitional activities explain four times more variance in well-being than life circumstances (Chapter 7). Although many of the leading discoveries of science—in particular, in the physical sciences—seem to defy common sense (otherwise we would arguably not need the scientific method), when considering claims about human behavior there is sometimes a good case to be made for standing back for a moment and considering what certain theoretical claims might imply (cf. Andreski, 1972).

essentially the onset of a period of regression to the mean following this outlier to make their principal point, thus violating an elementary methodological principle by using the same data to generate and confirm their hypotheses. Four other commentaries were published at the same time as ours, making this one of the most critiqued articles in *Nature* for several years.

In similar vein, in 2017–2018 I took part in a project (Scheel, Ritchie, Brown, & Jacques, 2018) to reanalyze an article in *Current Biology* (Reid et al., 2017) that claimed to show that fetuses *in utero* tend to turn their heads towards a pattern of lights that are arranged in an inverted triangle shape (i.e., with the apex pointing downwards) analogous to a human face, but away from an arrangement with the apex of the triangle pointing upwards. Once again, this article had been covered widely in the popular media, including in outlets with an anti-abortion stance such as the *Catholic Herald* (Staff Reporter, 2017), even though the research was carried out on third-trimester fetuses. We showed that not only were Reid et al.'s statistical analyses subject to a considerable number of “researcher degrees of freedom” (Simmons, Nelson, & Simonsohn, 2011), but also that the physiological basis of their experiment was severely flawed. Indeed, it is not clear that, even if the fetuses had their eyes open when tested (which was not clearly established in any individual case), many of them would not have seen anything but a fuzzy blob of light.

Investigating misconduct

The revelation in September 2011 that Diederik Stapel, who was at the time Dean of Social Sciences at Tilburg University, was a perpetrator of research fraud on a large scale was one of the defining events of the beginning of the “replication crisis.” This occurred just a couple of weeks before I attended the first weekend of my Master's course in London, although I was completely unaware of it at the time. While the commission of inquiry was investigating Stapel's fraud and the circumstances that made it possible, Stapel wrote a book in Dutch, which he entitled *Ontsporing* (Derailment). The book contains an account of Stapel's life and academic career, but it also reveals an insider's view of the research process and the meaning of (social) psychology that—even with the obvious caveats associated with a book written by someone who had admitted to falsifying or fabricating data on numerous occasions—I found extremely helpful for my own understanding of the field, especially as someone who has had no daily contact with researchers in a laboratory setting. Because I have a reasonably good command of the Dutch language, I offered to work with Stapel to produce an English-language translation of the book, which we made available as a free PDF

download under the title of *Faking Science: A True Story of Scientific Fraud* (<http://nick.brown.free.fr/stapel>). Over 5,000 people have downloaded this translation to date.

Since finishing the translation of *Faking Science*, I have been directly involved in the investigation of (what appear to be) two major cases of scientific misconduct. First, starting in November 2015, I was the principal author of an extensive dossier (Brown & Heathers, 2018) examining a sample of 10 of the many dozens of field studies published by the psychologist Nicolas Guéguen, many of which appear to describe scenarios that would be almost impossible to carry out in the real world (in addition to exhibiting numerous strange patterns in their tables of results, as previously mentioned in the description of GRIM). In December 2018, Guéguen's university (Université de Bretagne-Sud, in France) issued a statement, confirming that Guéguen has conducted his research in an unacceptable manner and requiring him to retract two of his studies. Second, since December 2016, with several colleagues, I have been investigating the work of Brian Wansink, a professor at Cornell University whose publications on the topic of consumer behavior, particularly as applied to healthier eating, have had a very large impact on popular opinion and public policy in the United States. The work coming from Wansink's lab is characterized by a remarkably high degree of methodological sloppiness; for example, we (van der Zee, Anaya, & Brown, 2017) discovered 150 statistical reporting errors in just four articles that were drawn from the same data set. However, the problems with Wansink's work go far beyond simple numerical issues. I found multiple cases of duplicate publication of both text and data, with one table of results appearing in two articles describing two completely different samples. In another case, the demographic characteristics of the participants in a study published in *JAMA Pediatrics* (Wansink, Just, & Payne, 2012) were inconsistently reported on multiple occasions (Lee, 2017). On September 20, 2018, Cornell announced that Wansink had been found to have committed multiple forms of research misconduct, including "misreporting of research data, problematic statistical techniques, failure to properly document and preserve research results, and inappropriate authorship" (Kotlikoff, 2018). As of March 19, 2019, 17 of Wansink's articles had been retracted and another 15 corrected.

As a result of my work on the translation of Stapel's book, and the subsequent investigations into possible misconduct just mentioned, I have developed an interest in the question of the extent to which actively fraudulent behavior—as opposed to the various forms of "softer" misconduct that are generally classified under the rather euphemistic title of "Questionable Research Practices" (QRPs; e.g., John, Loewenstein, & Prelec, 2012)—might be responsible for some of the failures of important findings in psychology to replicate. Until

now, the modal attitude of commentators has tended to be that while some research may be conducted somewhat sloppily, cases of outright fraud are extremely rare—although, as Uri Simonsohn has remarked (Rationally Speaking, 2016), this might be mostly because people would like to *believe* that fraud is rare, not because they have any firm evidence for that position. My own belief is that while the outright fabrication of entire studies, such as was done by Stapel, may be rare, the prevalence of data manipulation may be rather higher than is currently acknowledged. For example, as Chambers (2017, p. 122) has pointed out, simply swapping the condition variable for a few participants in an experiment is essentially undetectable; it will not affect any of the other statistics that might otherwise become suspiciously distorted if the researcher were to start changing the values of the dependent variable, and it could be done by the first person in the research team to get hold of the data set, without the other members being any the wiser. My belief is that the preparedness of most people—including scientists—to bend or break the rules is probably distributed along a continuum. Given the incentives and pressures to which researchers are subjected in the quest to obtain publishable results, I would be surprised if their apparently high willingness (John et al., 2012) to indulge in a range of QRPs—from, say, adding a covariate to turn a p value of .054 into .048 at one end, through to highly selective outlier removal at the other—were somehow to be distinguished from the willingness to commit “entry-level” fraud (e.g., changing one or two raw numbers in a data set to obtain a similar improvement in a p value) by a large moral gap into which hardly any scientist dares to tread. Stapel’s (2016, pp. 102–103) comment is perhaps apt here: “Even with my various ‘gray’ methods for ‘improving’ the data, I wasn’t able to get the results the way I wanted them. I couldn’t resist the temptation to go a step further.” Whether or not Stapel did, indeed, slide gradually down a slippery slope in this way as his career progressed has been questioned (Levelt, 2012), but it seems reasonable to imagine that the temptation that he describes is a common one for researchers alone in their lab with their data. Indeed, John et al. (2012, Figure 1) suggested, using some modest assumptions about the relation between people’s own admission rates and their estimates of others’ admission rates, that the actual lifetime prevalence rate of data falsification among psychologists could be as high as 35%.

Thoughts on the place of this work in psychology’s “replication crisis”

When my first peer-reviewed article (Brown, Sokal, & Friedman, 2013) was published, I was only just becoming aware that there was any kind of general problem with research in psychology. Indeed, I had started drafting that article in November 2011, before I

had even handed in the first assignment on my Master's course, which (as mentioned in my Acknowledgements section) was my first exposure to psychology in any form. By the summer of 2013, as I was getting ready to hand in my Master's thesis, I had heard of the fraudulent activities of Diederik Stapel, but hadn't appreciated the impact that his unmasking had had on psychologists, in the Netherlands and beyond. I had read the influential article by Simmons, Nelson, and Simonsohn (2011) in which they demonstrated that many research practices commonly used by psychologists are likely to lead to a large number of false positive results, but I had not appreciated the extent to which the methods they described had pervaded psychological research up to that point. I was also dimly aware of the controversy surrounding Bem's (2011) claims to have demonstrated the existence of "psi" phenomena using the standard methodologies of psychology, but I had not appreciated that the rebuttal of Bem's article by Wagenmakers, Wetzels, Borsboom, and van der Maas (2011)—which, at the time, I found too technical to fully understand—would signal the start of an extensive period of self-examination in the field, which is still ongoing today.

The Brown et al. (2013) article was a critique (or, more accurately, a refutation) of a key piece of positive psychology theory known as the "positivity ratio" (Fredrickson & Losada, 2005). At the time, it was not clear that this story was related to the then-nascent replication crisis. Many researchers outside positive psychology had never heard of the positivity ratio, and when I explained to them how it was supposed to work, a common reaction was a shaking of the head in disbelief that anyone in psychology could have fallen for something "so obviously wrong." However, since the publication of our 2013 article, I have been arguing that the problems that I have been finding in positive psychology research—namely, the making of claims about the purportedly substantial benefits of certain behaviors, attitudes, or states for one's mental and (especially) physical health—are just another aspect of the issues that have led to the replication crisis. The remarkable claims made by Fredrickson and Losada—briefly, that human emotions evolve over time in a way that is precisely governed by the same set of precise differential equations that determine the convection flow of fluids—were perhaps an extreme example of this, but the chapters in the present volume show that there are also theoretical and methodological issues in a number of lines of positive psychology research that make less spectacular claims. Much of the attention of proponents of the replication crisis has been focused on QRPs, such as *p*-hacking (Head, Holman, Lanfear, Kahn, & Jennions, 2015) or HARKing (Kerr, 1998). However, at least in the case of the articles examined earlier in this volume, they can also be due to poor research design and inappropriate analytical methods—issues that, on their own, do not seem to be

raised as often as QRPs in discussions about reproducibility. Why the latter should be the case is beyond the scope of the present exercise, but it could be, for example, that journal reviewers are reluctant to criticize unfamiliar methods, especially if these are combined with mathematical analyses that they do not fully understand, as I presume was the case with the article by Fredrickson and Losada.

References

- Basu, T. (2015, September 3). Feeling sad turns your world gray—literally. *Time*. Retrieved from <http://time.com/4018860/sadness-color-perception>
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 1–19. <https://doi.org/10.1037/a0021524>
- Brown, N. J. L., Albers, C. A., & Ritchie, S. J. (2017). Contesting the evidence for limited human lifespan. *Nature*, *546*, E6. <https://doi.org/10.1038/nature22784>
- Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, *8*, 363–369. <https://doi.org/10.1371/10.1177/1948550616673876>
- Brown, N. J. L., & Heathers, J. A. J. (2018). *A commentary on some articles by Dr. Nicolas Guéguen*. Retrieved from <https://www.dropbox.com/s/t61v1ypndryts63/NB-JH%20commentary%20on%20articles%20-%2016%20January%202018.pdf?dl=0>
- Brown, N. J. L., Sokal, A. D., & Friedman, H. L. (2013). The complex dynamics of wishful thinking: The critical positivity ratio. *American Psychologist*, *68*, 801–813. <https://doi.org/10.1037/a0032850>
- Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.
- Dong, X., Milholland, B., & Vigg, J. (2016). Evidence for a limit to human lifespan. *Nature*, *538*, 257–259. online. <https://doi.org/10.1038/nature19793>.
- Fredrickson, B. L., Grewen, K. M., Coffey, K. A., Algae, S. B., Firestone, A. M., Arevalo, J. M. G., . . . Cole, S. W. (2013). A functional genomic perspective on human well-being. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 13684–13689. <https://doi.org/10.1073/pnas.1316887110>
- Fredrickson, B. L., Grewen, K. M., Algae, S. B., Firestone, A. M., Arevalo, J. M. G., Ma, J., & Cole, S. W. (2015). Psychological well-being and the human conserved transcriptional response to adversity. *PLoS ONE*, *10*(3), e0121839. <https://doi.org/10.1371/journal.pone.0121839>
- Fredrickson, B. L., & Losada, M. F. (2005). Positive affect and the complex dynamics of human flourishing. *American Psychologist*, *60*, 678–686. <https://doi.org/10.1037/0003-066X.60.7.678>
- Guéguen, N. (2012a). Dead indoor plants strengthen belief in global warming. *Journal of Environmental Psychology*, *32*, 173–177. <https://doi.org/10.1016/j.jenvp.2011.12.002>
- Guéguen, N. (2012b). Gait and menstrual cycle: Ovulating women use sexier gaits and walk slowly ahead of men. *Gait & Posture*, *35*, 621–624. <https://doi.org/10.1016/j.gaitpost.2011.12.011>
- Guéguen, N. (2015). Women’s hairstyle and men’s behavior: A field experiment. *Scandinavian Journal of Psychology*, *56*, 637–640. <https://doi.org/10.1111/sjop.12253>
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106. <https://doi.org/10.1371/journal.pbio.1002106>
- Heathers, J., Anaya, J., van der Zee, T., & Brown, N. J. L. (2018). Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (SPRITE). *PeerJ Preprints*, *6*, 26968v1. <https://doi.org/10.7287/peerj.preprints.26968v1>

- Holcombe, A. O., Brown, N. J. L., Goodbourn, P. T., Etz, A., & Geukes, S. (2016). Does sadness impair color perception? Flawed evidence and faulty methods. *F1000 Research*, 5, 1778. <https://doi.org/10.12688/f1000research.9202.1>
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23, 524–532. <https://doi.org/10.1177/0956797611430953>
- Kaplan, S. (2015, September 3). The scientific reason your world brightens up when you do. *The Washington Post*. Retrieved from <https://www.washingtonpost.com/news/morning-mix/wp/2015/09/03/feeling-sad-could-literally-make-the-world-seem-grayer-study-finds>
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2, 196–217. https://doi.org/10.1207/s15327957pspr0203_4
- Kotlikoff, M. (2018, September 20). *Provost issues statement on Wansink academic misconduct investigation*. Retrieved from <http://news.cornell.edu/stories/2018/09/provost-issues-statement-wansink-academic-misconduct-investigation>
- Lee, S. M. (2017, October 18). Here’s how a controversial study about kids and cookies turned out to be wrong—and wrong again. *BuzzFeed News*. Retrieved from <https://www.buzzfeednews.com/article/stephaniemlee/who-really-ate-the-apples-though>
- Levelt, W. J. M. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*. Retrieved from https://www.tilburguniversity.edu/upload/3ff904d7-547b-40ae-85fe-bea38e05a34a_Final%20report%20Flawed%20Science.pdf
- Rationally Speaking (2016, March 20). RS 155—Uri Simonsohn on “Detecting fraud in social science.” Retrieved from <http://bit.ly/rstp155>
- Reid, V. M., Dunn, K., Young, R. J., Amu, J., Donovan, T., & Reissland, N. (2017). The human fetus preferentially engages with facelike visual stimuli. *Current Biology*, 27, 1825–1828. <https://doi.org/10.1016/j.cub.2017.05.044>
- Scheel, A. M., Ritchie, S. J., Brown, N. J. L., & Jacques, S. L. (2018). Methodological problems in a recent study of fetal visual perception. *Current Biology*, 28, R594–R596. <https://doi.org/10.1016/j.cub.2018.03.047>
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22, 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Staff Reporter. (2017, June 10). Babies recognise faces while still in the womb, research shows. *Catholic Herald*. Retrieved from <http://www.catholicherald.co.uk/news/2017/06/10/babies-recognise-faces-while-still-in-the-womb-research-shows>
- Thorstenon, C. A., Pazda, A. D., & Elliot, A. J. (2015). Sadness impairs color perception. *Psychological Science*. Advance online publication. <https://doi.org/10.1177/0956797615597672>
- van der Zee, T., Anaya, J., & Brown, N. J. L. (2017). Statistical heartburn: An attempt to digest four pizza publications from the Cornell Food and Brand Lab. *BMC Nutrition*, 3(54), 1–22. <https://doi.org/10.1186/s40795-017-0167-x>

- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432.
<https://doi.org/10.1037/a0022790>
- Wansink, B., Just, D. R., & Payne, C. R. (2012). Can branding improve school lunches? *Archives of Pediatric and Adolescent Medicine*, *166*, 967–968.
<https://doi.org/10.1001/archpediatrics.2012.999>

Chapter 11

Samenvatting¹

¹ Deze samenvatting bestaat uit een amalgaam van Hoofdstuk 1 en het eerste onderdeel (“Summary of Chapters 2–7”) van Hoofdstuk 9. Ik heb de vertaling zelf gemaakt, met correcties en suggesties van drie Nederlandse collega’s. Alle resterende (taal)fouten zijn geheel voor mijn rekening.

De wetenschappelijke beweging die zichzelf *positieve psychologie* noemt (Seligman & Csikszentmihalyi, 2000) heeft de laatste 20 jaar een grote invloed uitgeoefend op de perceptie van de psychologie in het algemeen onder het brede publiek. Één van de belangrijkste boodschappen van de positieve psychologie is dat personen niet meer “gestoord” hoeven te zijn om baat te hebben bij een psychologische interventie. Verschillende op het grote publiek gerichte boeken (bijv. Duckworth, 2016; Fredrickson, 2009, 2013; Lyubomirsky, 2007, 2013; Seligman, 2002, 2011) geven een wetenschappelijke glans aan de zelfhulpafdeling van boekenwinkels, en een behoorlijk aantal coaches die een groot gedeelte van hun aanpak op de positieve psychologie hebben gebaseerd bieden hun diensten aan in de vorm van websites, seminaria, en terugtrekweekendjes.

In het wereldbeeld van de positieve psychologie geldt het als vanzelfsprekend dat de traditionele splitsing van mensen tussen degenen die een diagnosticeerbaar mentaal gezondheidsprobleem (zoals depressie of angst) hebben—mensen die tot nu toe door klinische psychologen werden behandeld—en de anderen niet meer voldoende is. Immers, als het niet mogelijk is om “normale” mensen van elkaar te onderscheiden, dan wordt het moeilijk om aan te tonen dat een interventie om het geestelijke welbevinden van iemand te verhogen een effect heeft gehad. Daarom is het misschien niet geheel verbazingwekkend dat voorstanders van de positieve psychologie menen dat er verschillende niveaus van gezond mentaal functioneren bestaan. Seligman (2002), bijvoorbeeld, schreef: “Lying awake at night, you probably ponder, as I have, how to go from plus two to plus seven in your life”, waarmee hij sterk impliceert dat het welbevinden van iemand op een numerieke schaal (die misschien van nul of min 10 tot plus 10 loopt) kan worden gemeten, en dat het én mogelijk én wenselijk is om diens persoonlijke “score” te verbeteren. Keyes (2002) had het idee om niet-depressieve personen in twee categorieën in te delen, degenen die *kwijnend* (“languishing”) zijn en degenen die *florissant* (“flourishing”) door het leven gaan, en het duurde niet lang voordat onderzoekers in de positieve psychologie begonnen te onderzoeken hoe deze twee psychologische toestanden (kwijnend en florissant) van elkaar te onderscheiden zouden zijn, en hoe de mensen van de ene toestand naar de andere overgebracht zouden kunnen worden. Fredrickson en Losada (2005) beweerden zelfs dat zij de precieze wiskundige formule hadden gevonden om de grens tussen de twee toestanden te bepalen. Seligman (2011) noemde zijn boek, waarin hij zijn herziene model van welbevinden (de opvolger van *Authentic Happiness* uit 2002) presenteerde, simpelweg *Flourish*.

Een vaak voorkomend onderwerp in deze literatuur is dat, om te *floreren*, mensen regelmatig positieve emotionele toestanden moeten ervaren. De *verbreed-en-bouw* theorie

(“broaden-and-build”) theorie van Fredrickson (2009) stelt voor dat het ervaren van positieve emoties leidt tot een breder repertoire van denken en gedrag en tot de ontwikkeling van psychologische vaardigheden zoals veerkracht; volgens Fredrickson is hiervoor een verhouding van tenminste drie positieve emoties tegenover elke negatieve emotie optimaal. Seligman (2011) gebruikte het acroniem “PERMA”, waarvan de eerste letter “Positieve emoties” vertegenwoordigt, om de vijf onderdelen van zijn model van een florissant leven aan te duiden.

Tegen het einde van het eerste decennium van de 21ste eeuw had de positieve psychologie dus zijn doel (het welbevinden van iedereen, niet alleen mensen met depressie of angst), het brandmerk van dit gewenste uitkomst (floreren), en de belangrijkste middel om het doel te bereiken (positieve emoties) bepaald. Bovendien, omdat de doelbevolking niet uit kwetsbare mensen bestaat, zouden de hulpverleners een minder gespecialiseerde opleiding nodig hebben. Mede daarom werd in 2005 de eerste *Master’s in Applied Positive Psychology* (MAPP) opleiding opgezet aan de University of Pennsylvania (Seligman, 2011), een traject dat inmiddels meerdere gelijkgenaaamde navolgers aan verschillende universiteiten over de hele wereld kent. Veel afgestudeerden van deze programma’s verlenen nu zelf op positieve psychologie gebaseerde training of coaching. Een aantal alumni van de opleiding aan de University of Pennsylvania werkt voor het “Comprehensive Soldier and Family Fitness”-programma van het Amerikaanse leger dat elk jaar positieve psychologie interventies aan honderdduizenden soldaten en hun familieleden verleent (Brown, 2015).

Een aantal onderzoekers in de positieve psychologie heeft de laatste jaren artikelen gepubliceerd waarin werd geïmpliceerd dat mensen met beter psychologisch welbevinden—d.w.z., de “florissanten” t.o.v. de “kwijnenden”, of degenen die op een andere manier een beter gestructureerd emotioneel leven hebben—ook objectief een betere lichamelijke gesteldheid hebben. Het idee dat “positief” mentaal functioneren van zichzelf gepaard zou gaan met betere lichamelijke gezondheidsuitkomsten is uiteraard aantrekkelijk; het is immers bijna onomstreden dat diagnoses van psychische aandoeningen negatief gecorreleerd zijn met levensverwachting (bijv. Chesney, Goodwin, & Fazel, 2014). Maar men kan niet zomaar extrapoleren van deze verhouding onder klinische patiënten naar populaties die nog niet bestudeerd zijn. Daarom heeft men in een aantal recente empirische artikelen geprobeerd om een dergelijke verhouding te constateren binnen de algemene bevolking. De kern van dit proefschrift bestaat uit onderzoek naar een groep van deze artikelen, waarin beweerd wordt dat er een positieve—zo niet monotonische of zelfs lineaire—verhouding bestaat tussen

welbevinden (het ervaren van wenselijke emotionele toestanden) en een betere lichamelijke gezondheid, die door verschillende directe of indirecte metingen wordt vastgesteld.

Hoofdstukken 2 en 3 zijn nauw met elkaar verbonden, omdat beide een kritische beschouwing bevatten van een en dezelfde onderzoeksvraag. Wij bespreken de beweringen van Fredrickson et al. (2013, 2015) omtrent de ogenschijnlijk verschillende effecten op de lichamelijke gezondheid—vertegenwoordigd door de differentiële uitdrukking van een aantal genen waarvan is bekend dat zij op phenotypisch niveau met het menselijk immuunsysteem geassocieerd zijn—van de “hedonische” en “eudaimonische” vormen van welbevinden. Fredrickson en haar collega’s beweerden dat mensen die een verhoudingsgewijs hoger niveau van eudaimonisch welbevinden hebben (een “betekenisvol leven” ervaren), t.o.v. mensen met een hoger niveau van hedonisch welbevinden (een “plezierig leven”), hogere expressies hebben van genen die geassocieerd zijn met het bestrijden van virussen, terwijl mensen in de omgekeerde situatie (dus, met meer hedonisch welbevinden) hogere expressies hebben van pro-inflammatoire of anti-bacteriële genen. De evolutionaire redenering hiervoor is dat mensen met meer eudaimonisch welbevinden ook meer sociale contacten zullen hebben, en dus meer besmettelijke virale infecties zullen oplopen, terwijl hedonisch ingestelde mensen bij meer conflicten zullen worden betrokken en dus bacteriële infecties zullen krijgen als gevolg van hun wonden (Smith, 2013). In deze twee hoofdstukken laten wij zien dat deze resultaten zeer onwaarschijnlijk zijn. De twee vormen van welbevinden zijn sterk met elkaar gecorreleerd, de metingen van genexpressie zijn niet onafhankelijk van elkaar, en in hun artikel uit 2015 hebben Fredrickson et al. enkele hun theorie onwelgevallige uitkomsten niet gerapporteerd. In Hoofdstuk 2 laten wij zien dat het “herhaalde regressie”-model van Fredrickson et al. (2013), waarin de beta-coëfficiënten van genexpressie uit 53 regressies waren vergeleken met nul, ongeveer dezelfde resultaten geeft indien de voorspellende variabelen vervangen worden door willekeurige getallen. Het belangrijkste probleem, dat wij misschien niet voldoende hebben onderstreept in ons artikel van 2014, was dat de coëfficiënten die corresponderen met de expressie van 53 verschillende genen niet van elkaar onafhankelijk waren (de genen vormden drie groepen, elk met een specifiek fysiologisch effect), zodat de *t* toetsen van Fredrickson et al. ongeldig waren. In het artikel dat het onderwerp van Hoofdstuk 3 vormt hebben Fredrickson et al. (2015) geprobeerd dit zwakke punt op te lossen door gemengde modellen te gebruiken, maar deze beslissing heeft het probleem slechts naar elders verschoven.

In Hoofdstuk 4 kijken wij kritisch naar een artikel (Kok et al., 2013) waarin wordt beweerd dat het uitvoeren van een specifieke meditatiepraktijk, de zogenaamde *loving-*

kindness meditation (metta-meditatie, in het Nederlands), resulteerde in betere lichamelijke gezondheid van deelnemers, met als bewijs een vermeende verbetering in de tonus van hun nervus vagus (*vagal nerve tone* of VNT). Wij tonen aan dat dit artikel onder drie grote problemen lijdt. Ten eerste is de manier waarop Kok et al. de VNT gebruikte uit zowel conceptueel als statistisch oogpunt onjuist; zo hebben zij bijvoorbeeld een ongewone transformatie van hun belangrijkste afhankelijke variabele toegepast, terwijl de bij dit soort onderzoek gebruikelijke transformatie juist geen statistisch significant resultaat geeft. Ten tweede blijkt het verschil tussen de deelnemers te maken hebben met een achteruitgang onder de mensen in de controlegroep; er was geen verbetering onder de mensen in de interventiegroep. Ten derde waren de gebruikte analysemethoden ongeschikt om een eventuele verbetering aan de interventie toe te schrijven. In een later gepubliceerde correctie (Psychological Science, 2016) werd zelfs onthuld dat de gegevens van Kok et al. eerder al gebruikt waren om een ander artikel (Kok & Fredrickson, 2010) te publiceren, waarvan de resultaten deels in strijd waren met die van Kok et al.

Hoofdstuk 5 geeft een beschouwing van een artikel (Quoidbach et al., 2014) betreffende een nieuwe constructie genaamde emodiversiteit (*emodiversity*). Het idee is dat, naast een overwicht van positieve (t.o.v. negatieve) emoties, het ook belangrijk is voor een individu om zoveel mogelijk verschillende emoties, zowel positief als negatief, te ervaren. Volgens de auteurs heeft de emodiversiteit in twee grote studies, met duizenden deelnemers, tot betere mentale (Studie 1) en lichamelijke (Studie 2) gezondheidsuitkomsten geleid, bovenop het effect van de intensiteit van de emoties zelf. Quoidbach et al. beweerde zelfs dat emodiversiteit de lichamelijke gezondheid beter voorspelde dan standaard maten van diët, lichaamsbeweging, of rookgedrag. Echter, onze analyses laten zien dat de manier waarop emodiversiteit in wiskundige vorm werd uitgedrukt zeer gebrekkig is. Bovendien zorgt de verhouding tussen emoties en emodiversiteit ervoor dat de relatie tussen emodiversiteit en gezondheidsuitkomsten slechts een (onvermijdelijk) statistisch artefact is.

Het onderwerp van Hoofdstuk 6 is een artikel de bewering dat het niveau van sterfte aan slagaderziekte (*atherosclerotic heart disease* of AHD) per Amerikaanse *county* te voorspellen is aan de hand van de graad van positiviteit of negativiteit die door de bevolking van dezelfde county op Twitter wordt uitgedrukt. Deze stelling van Eichstaedt et al. (2015), die uitvoerige aandacht in de media heeft gekregen, lijkt niet ondersteund te zijn door het beschikbare bewijsmateriaal. Het artikel van Eichstaedt et al. bevat een aantal methodologische tekortkomingen die niet allemaal eenvoudig te identificeren zijn bij de eerste doorlezing, maar die het moeilijk maken om veel vertrouwen in de resultaten te

hebben. In het bijzonder tonen wij aan dat de tegenovergestelde relatie (d.w.z., positieve emoties geven een slechtere uitkomst) bestaat wanneer de sterftcijfers van slagaderziekte vervangen worden door de equivalente statistieken betreffend zelfmoord.

In Hoofdstuk 7 gaat het om de “gelukstaart” (*happiness pie*), een model van de determinanten van het welbevinden dat werd voorgesteld door Lyubomirsky, Sheldon, en Schkade (2005). Lyubomirsky en haar collega’s beweerde dat maar liefst 40% van de variantie in het welbevinden van individuen de resultaat is van activiteiten die mensen zelf kiezen te doen (*volitional activities*), terwijl 50% door genetische factoren en slechts 10% door levensomstandigheden bepaald worden. Hoewel er in het artikel van Lyubomirsky et al. geen directe discussie van lichamelijke gezondheid bevat (in tegenstelling tot Lyubomirskys (2007) populaire boek over hetzelfde onderwerp), dit hoofdstuk verdient een plaats in dit proefschrift omdat het een aantal punten van wat men het *business model* van de positieve psychologie zou kunnen noemen illustreert, zoals het idee dat mensen verantwoordelijk zijn voor een groot deel van hun eigen (on)geluk. In onze commentaar demonstreren wij dat het idee dat de mensen tot 40% van hun geluk kunnen beïnvloeden zeer optimistisch lijkt, en dat de veronderstellingen achter dit idee tamelijk naïef zijn.

Hoofdstuk 8 bestaat uit mijn contributie aan een recent verschenen handboek (Brown, Lomas, & Eiroa-Orosa, 2017) waarvan ik ook de hoofdredacteur was. Met dit boek hebben wij geprobeerd om in enige mate de discussie van de positieve psychologie terug te draaien naar zijn oorspronkelijke focus op het optimale functioneren van de mens (Rathunde, 2001), i.p.v. wat wij (als redacteurs) zagen als een nadruk op het geluk en welbevinden van welgestelde westerlingen (Coyne, 2013). Mijn hoofdstuk presenteert, onder meer, een aantal manieren waarmee studenten van positieve psychologie (bijv. deelnemers aan MAPP-opleidingen) hun eigen kritiek op dit onderzoeksveld kunnen brengen.

Een gemeenschappelijk thema van de meeste hoofdstukken van dit proefschrift is het gebruik van ongepaste (en vaak zeer ingewikkelde) analytische en statistische methoden. Vaak worden causale verbanden sterk gesuggereerd, zonder dat er (in de meeste gevallen) een experiment met manipulatie van één of meer variabelen werd uitgevoerd². Toch gaat het hier om een steekproef van de belangrijkste recente theoretische en empirische vindingen in

² Het enige echte experiment vond plaats in het artikel van Kok et al. (2013), dat in Hoofdstuk 4 werd besproken. Helaas was in deze studie de toewijzing van deelnemers aan één van de twee condities de enige gemanipuleerde variabele, want de graad van naleving van de deelnemers in de interventiegroep aan hun meditatie-training werd niet gemeten.

de positieve psychologie, vooral wat positieve emoties en lichamelijke gezondheid betreft. In vijf van de zes gevallen staat de senior auteur op de *Mount Rushmore of Positive Psychology*, een visuele eerbetoon aan de hoofdpersonages van deze beweging (Mann, 2013). Deze artikelen zijn respectievelijk 377, 90, 488, 71, 297, en 3.217 maal geciteerd geweest volgens Google Scholar (cijfers van 18 augustus 2019), en de tijdschriften waarin ze gepubliceerd waren hadden in 2016 een gemiddelde *impact factor* van 5,099.

Meerdere schrijvers van buiten de positieve psychologie, zoals Ehrenreich (2010), Schneider (2011), en Whippman (2016, hfst. 8) hebben reeds twijfels geuit over in hoeverre de empirische beweringen van positieve psychologen bestand tegen zorgvuldig analytisch onderzoek zouden blijken. Misschien geven onze resultaten deze auteurs tenminste deels gelijk.

Referenties

- Brown, N. J. L. (2015). A critical examination of the U.S. Army's Comprehensive Soldier Fitness program. *The Winnower*, 3, e143751.17496. <https://doi.org/10.15200/winn.143751.17496>
- Brown, N. J. L., Lomas, T., & Eiroa-Orosa, F. J. (2018). *The Routledge international handbook of critical positive psychology*. London, England: Routledge.
- Chesney, E., Goodwin, G. M., & Fazel, S. (2014). Risks of all-cause and suicide mortality in mental disorders: A meta-review. *World Psychiatry*, 13, 153–160. <https://doi.org/10.1002/wps.20128>
- Coyne, J. C. (2013, August 21). *Positive psychology is mainly for rich white people* [Weblog post]. Retrieved from <https://mindthebrain.blog/2013/08/21/positive-psychology-is-mainly-for-rich-white-people>
- Duckworth, A. (2016). *Grit: The power of passion and perseverance*. New York, NY: Scribner.
- Ehrenreich, B. (2010). *Smile or die: How positive thinking fooled America and the world*. London, England: Granta.
- Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., . . . Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26, 159–169. <https://doi.org/10.1177/0956797614557867>
- Fredrickson, B. L. (2009). *Positivity: Groundbreaking research reveals how to embrace the hidden strength of positive emotions, overcome negativity, and thrive*. New York, NY: Crown.
- Fredrickson, B. L. (2013). *Love 2.0: Finding happiness and health in moments of connection*. New York, NY: Hudson Street Press.
- Fredrickson, B. L., Grewen, K. M., Coffey, K. A., Algae, S. B., Firestone, A. M., Arevalo, J. M. G., . . . Cole, S. W. (2013). A functional genomic perspective on human well-being. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 13684–13689. <https://doi.org/10.1073/pnas.1316887110>
- Fredrickson, B. L., Grewen, K. M., Algae, S. B., Firestone, A. M., Arevalo, J. M. G., Ma, J., & Cole, S. W. (2015). Psychological well-being and the human conserved transcriptional response to adversity. *PLoS ONE*, 10(3), e0121839. <https://doi.org/10.1371/journal.pone.0121839>
- Fredrickson, B. L., & Losada, M. F. (2005). Positive affect and the complex dynamics of human flourishing. *American Psychologist*, 60, 678–686. <https://doi.org/10.1037/0003-066X.60.7.678>
- Heathers, J. A. J., Brown, N. J. L., Friedman, H. L., & Coyne, J. C. (2015). The elusory upward spiral: A reanalysis of Kok et al. (2013). *Psychological Science*, 26, 1140–1143. <https://doi.org/10.1177/0956797615572908>
- Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior*, 43, 207–222. <https://doi.org/10.2307/3090197>
- Kok, B. E., Coffey, K. A., Cohn, M. A., Catalino, L. I., Vacharkulksemsuk, T., Algae, S. B., . . . Fredrickson, B. L. (2013). How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone. *Psychological Science*, 24, 1123–1132. <https://doi.org/10.1177/0956797612470827>

- Kok, B. E., & Fredrickson, B. L. (2010). Upward spirals of the heart: Autonomic flexibility, as indexed by vagal tone, reciprocally and prospectively predicts positive emotions and social connectedness. *Biological Psychology*, 85, 432–436.
<https://doi.org/10.1016/j.biopsycho.2010.09.005>
- Lyubomirsky, S. (2007). *The how of happiness: A scientific approach to getting the life you want*. New York, NY: Penguin.
- Lyubomirsky, S. (2013). *The myths of happiness: What should make you happy, but doesn't; What shouldn't make you happy, but does*. New York, NY: Penguin.
- Lyubomirsky, S., Sheldon, K. M., & Schkade, D. (2005). Pursuing happiness: The architecture of sustainable change. *Review of General Psychology*, 9, 111–131.
<https://doi.org/10.1037/1089-2680.9.2.111>
- Mann, J. D. (2013, October 11). *Happiness revolution*. Retrieved from
<https://www.livehappy.com/science/positive-psychology/happiness-revolution>
- Psychological Science. (2016). Corrigendum: How positive emotions build physical health: Perceived positive social connections account for the upward spiral between positive emotions and vagal tone. *Psychological Science*, 27, 931. <https://doi.org/10.1177/0956797616647346>
- Quoidbach, J., Gruber, J., Mikolajczak, M., Kogan, A., Kotsou, I., & Norton, M. I. (2014). Emdiversity and the emotional ecosystem. *Journal of Experimental Psychology: General*, 143, 2057–2066. <https://doi.org/10.1037/a0038025>
- Rathunde, K. (2001). Toward a psychology of optimal human functioning: What positive psychology can learn from the “experiential turns” of James, Dewey, and Maslow. *Journal of Humanistic Psychology*, 41(1), 135–153. <https://doi.org/10.1177/0022167801411008>
- Schneider, K. (2011). Toward a humanistic positive psychology: Why can't we just get along? *Existential Analysis*, 22, 32–38.
- Seligman, M. E. P. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. New York, NY: Free Press.
- Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being*. New York, NY: Free Press.
- Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, 55, 5–14. <https://doi.org/10.1037/0003-066X.55.1.5>
- Smith, E. E. (2013). Meaning is healthier than happiness *The Atlantic*, August 1. Retrieved from
<https://www.theatlantic.com/health/archive/2013/08/meaning-is-healthier-than-happiness/278250>
- Walker, J. A. (2016). Monte Carlo simulation of OLS and linear mixed model inference of phenotypic effects on gene expression. *PeerJ*, 4, e2575. <https://doi.org/10.7717/peerj.2575>
- Whippman, R. (2016). *America the anxious: How our pursuit of happiness is creating a nation of nervous wrecks*. New York, NY: St. Martin's Press.

Chapter 12

Acknowledgements

Life is a series of highly contingent events. Most people, especially of my age, can probably trace several major changes in their lives back to very small things that might equally well not have happened, rather than the implementation of carefully-considered plans. If I hadn't read a newspaper advertisement for a job at the Council of Europe when sitting in my office in Bahrain in the spring of 1990, I would not have spent the last 29 years living in Strasbourg. If the person standing in front of me in the queue to get a book signed by the British psychologist Richard Wiseman almost ten years to the day before the defence of this thesis hadn't decided to give up waiting, leaving me to get the last book on the pile, I would probably never have had the brief conversation with him that led to me returning to full-time study and entering the world of psychology shortly before my 51st birthday. If Alan Sokal hadn't revisited his pile of e-mails one day in early 2012 to find my request for help with an obscure piece of misapplied mathematics, I would not have had my first-ever peer reviewed publication accepted by *American Psychologist* while I was still a first-year Master's student.

As a consequence of those unlikely events, and doubtless a host of others of which I am unaware, in the autumn of 2014 Jim Coyne proposed to me that I should become his first-ever supervisee at Groningen, and also managed to persuade Robbert Sandeman to accept this arrangement. Jim is no longer affiliated with the University of Groningen, but without him very little of this thesis would have been written. I thank Adelita Ranchor for taking over as my main supervisor and Casper Albers for becoming co-supervisor to help gently guide this whole project to a conclusion. Thanks are also due to the members of the assessment committee, professors Chris Chambers, Peter de Jonge, and Mirjam Sprangers. And thanks to my paranymp, Lian Beijers, for her logistical support, and also for her early input into what eventually became Chapter 7.

Next I want to thank my other co-authors on the articles that became the principal chapters of this thesis: Harris Friedman, James Heathers, Doug MacDonald, Julia Rohrer, and Manoj Samanta. I hope to continue working with all of you in the future.

The following people contributed helpful input on one or more chapters: Paul Barrett, Tim Bates, Jim Coan, Nora Egan Demers, Andy Field, Tanya Hart, Hans IJzerman, Toon Kuppens, Daniël Lakens, Rich Lucas, David MacKinnon, Neuroskeptic, Stuart Ritchie, Stefan Schmukle, Sam Schwarzkopf, Rob Summers, and Jeff Walker. My thanks also to a number of anonymous commenters and reviewers, and the members of the Delphi panel that was mentioned in Chapter 7.

For their general help and guidance with questions about psychology, academic practices, research methods, and statistics, I ought to thank a large number of people, many of whom I have only ever “met” online. Some of these people’s names already appear above, but to keep this list to manageable proportions I will mention just four here: Thom Baguley, Hans IJzerman, Carol Nickerson, and Hal Pashler. Also, I wouldn’t be here without the support and encouragement of the teachers from my Master’s course, who encouraged me to be critical in everything I did: Kate Heffernon, Itai Ivtzan, Tim Lomas, Joan Painter, and Nash Popovic.

Martin Brümmer has done a wonderful job with the cover design for the book. I wrote the Dutch summary myself, but since it has been 30 years since I lived in the Netherlands, I enlisted the help of Jan de Ruiter, Tim van der Zee, and Erik-Jan Wagenmakers to correct some of my clumsiness; any remaining linguistic errors are my responsibility. Robin Kok provided some helpful advice about getting the thesis printed.

Although I was told that, as an external PhD candidate, I was not required to take any formal courses (other than one on research integrity) as part of my programme, I decided early on that I would try to follow as many as possible of the classes that “normal” PhD students take. At one point I spent five weeks living in an international student house in Groningen with 40 undergraduates, which was a novel experience for both me and them (but at least they had someone to translate their letters from the bank). I want to thank all of the teachers of these classes for their patience in dealing with my awkward (and sometimes silly) questions, with a particular mention for Eric van Sonderen, who frequently indulged some of my laments about the problems of multiple regression analyses. Thanks also to Annemieke Brouwers and Renate Kroese for their help with the logistics of these courses and my participation in the programme more generally, and Truus van Ittersum for her help in making everything SHARE-compatible.

Finally, I want to acknowledge the love and support of my wife, Siân, and our children, Alex and Joanna. I just wish that my Mum and Dad could be here to see me become Dr. Brown.

Research Institute SHARE

This thesis is published within the **Research Institute SHARE** (Science in Healthy Ageing and healthcaRE) of the University Medical Center Groningen / University of Groningen. Further information regarding the institute and its research can be obtained from our internet site: <http://www.share.umcg.nl/>

More recent theses can be found in the list below.
(supervisors are between brackets)

2019

Dittmar D

Multiple aspects of contact allergy; immunology, patch test methodology and epidemiology
(*dr MLA Schuttelaar, prof PJ Coenraads*)

Beemster T

Quality improvement of vocational rehabilitation in patients with chronic musculoskeletal pain and reduced work participation
(*prof MF Reneman, prof MHW Frings-Dresen, prof CAM van Bennekom, dr JM van Velzen*)

Geerse O

The impact of lung cancer; towards high-quality and patient-centered supportive care
(*prof HAM Kerstjens, prof MY Berger, dr TJN Hiltermann, dr AJ Berendsen*)

Dierselhuis EF

Advances of treatment in atypical cartilaginous tumours
(*prof SK Bulstra, prof AJH Suurmeijer, dr PC Jutte, dr M Stevens*)

Gils A van

Developing e-health applications to promote a patient-centered approach to medically unexplained symptoms
(*prof JGM Rosmalen, prof RA Schoevers*)

Notenbomer A

Frequent sickness absence; a signal to take action
(*prof U Bultmann, prof W van Rhenen, dr CAM Roelen*)

Bishanga DR

Improving access to quality maternal and newborn care in low-resource settings: the case of Tanzania
(*prof J Stekelenburg, dr YM Kim*)

Tura AK

Safe motherhood: severe maternal morbidity and mortality in Eastern Ethiopia
(*prof SA Scherjon, prof J Stekelenburg, dr TH van den Akker*)

Vermeiden CJ

Safe motherhood : maternity waiting homes in Ethiopia to improve women's access to maternity care

(prof J Stekelenburg, dr TH van den Akker)

Schrier E

Psychological aspects in rehabilitation

(prof PU Dijkstra, prof JHB Geertzen)

Malinakova K

Spirituality and health: their associations and measurement problems

(prof SA Reijneveld, prof P Tavel, dr JP van Dijk)

Dijkhuizen A

Physical fitness and performance of daily activities in persons with intellectual disabilities and visual impairment; towards improving conditions for participation

(prof CP van der Schans, dr A Waninge, dr WP Krijnen)

Graaf MW de

The measurement and prediction of physical functioning after trauma

(prof E Heineman, dr IHF Reininga, dr KW Wendt)

Vrijen C

Happy faces and other rewards; different perspectives on a bias away from positive and toward negative information as an underlying mechanism of depression

(prof AJ Oldehinkel, prof CA Hartman, prof P de Jonge)

Moye Holz DD

Access to innovative medicines in a middle-income country; the case of Mexico and cancer medicines

(prof HV Hogerzeil, prof SA Reijneveld, dr JP van Dijk)

Woldendorp KH

Musculoskeletal pain & dysfunction in musicians

(prof MF Reneman, prof JH Arendzen, dr AM Boonstra)

Mooyaart JE

Linkages between family background, family formation and disadvantage in young adulthood

(prof AC Liefbroer, prof F Billari)

Maciel Rabello L

The influence of load on tendons and tendinopathy; studying Achilles and patellar tendons using UTC

(prof J Zwerver, prof RL Diercks, dr I van den Akker-Scheek, dr MS Brink)

For earlier theses visit our website