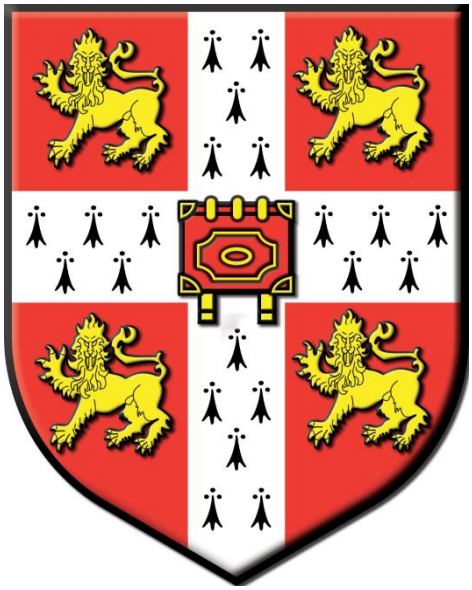
Using Computational Psychology to Profile Unhappy and Happy People



Matthew James Samson

Department of Psychology

Trinity Hall

University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

September 2018

Summary

Using Computational Psychology to Profile Unhappy and Happy People, by Matthew Samson

Social psychology has a long tradition of studying the personality traits associated with subjective well-being (SWB). However, research often depends on a priori but unempirical assumptions about how to (a) measure the constructs, and (b) mitigate confounded associations. These assumptions have caused profligate and often contradictory findings. To remedy, I demonstrate how a computational psychology paradigm—predicated on large online data and iterative analyses—might help isolate more robust personality trait associations.

At the outset, I focussed on univariate measurement. In the first set of studies, I evaluated the extent researchers could measure psychological characteristics at scale from online behaviour. Specifically, I used a combination of simulated and real-world data to determine whether predicted constructs like big five personality were accurate for specific individuals. I found that it was usually more effective to simply assume everyone was average for the characteristic, and that imprecision was not remedied by collapsing predicted scores into buckets (e.g. low, medium, high). Overall, I concluded that predictions were unlikely to yield precise individual-level insights, but could still be used to examine normative group-based tendencies. In the second set of studies, I evaluated the construct validity of a novel SWB scale. Specifically, I repurposed the balanced measure of psychological needs (BMPN), which was originally designed to capture the substrates of intrinsic motivation. I found that the BMPN robustly captured (a) dissociable experiences of suffering and flourishing, (b) more transitive SWB than the existing criterion measure, and (c) unique variation in real-world outcomes. Thus, I used it as my primary outcome.

Then, I focussed on bivariate associations. The third set of studies extracted pairs of participants with similar patterns of covarying personality traits—and differing target traits—to isolate less-confounded SWB correlations. I found my extraction method—an adapted version of propensity score matching—outperformed even advanced machine learning alternatives. The final set of studies isolated the subset of facets that had the most robust associations with SWB. It combined real-world surveys with a total of eight billion simulated participants to find the traits most prevalent in extreme suffering and flourishing. For validation purposes, I first found that depression and cheerfulness—the trait components of SWB—were highly implicated in both suffering and flourishing. Then, I found that self-discipline was the only other trait implicated in both forms of SWB. However, there were also domain-specific effects: anxiety, vulnerability and cooperation were implicated in just suffering; and, assertiveness, altruism and self-efficacy were implicated in just flourishing. These seven traits were most likely to be the definitive, stable, drivers of SWB because their effects were totally consistent across the full range of intrapersonal contexts.

To mum and dad.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

It does not exceed the prescribed word limit (60,000; excluding references, tables, figures, etc.) for the Faculty of Biological Sciences.

Acknowledgements

I enrolled as an undergrad in psychology at Macquarie University after brief but calamitous stints in investment banking and sports retail. Ginger from my transition from orderly comfortable high school, I simply wanted to learn some interesting facts. A decade on, I am somehow submitting a PhD at Cambridge (!!!). In the process, I have studied on four continents, made countless friends and, hopefully, learned as much outside psychology as within. I am extraordinarily lucky that my parents—Conny and Jeff Samson—always put my interests above their own and unconditionally supported my various crazy endeavours. Many other kind people were also hugely influential.

My thanks to Dr Alex Kogan as both my former PhD supervisor, and as a friend and confidante. Dr Kogan agreed to take me on, at short notice, even though I did not have a clearly defined project or any formal training in computational methods. I particularly valued his advice on forming research questions, statistical analyses and communicating research results. He also gave me the autonomy to choose my own research trajectory, learn from the inevitable mistakes, help administer grant funding and circulate research findings. Finally, I also appreciated his optimistic, non-judgmental and problem-focussed approach to the not-so-occasional Lab stressors.

Many other academics and professionals selflessly donated their time to help my PhD. Foremost, my sincerest thanks to Dr Jason Rentfrow for stepping in as my supervisor for the final six months of the PhD. He reviewed *everything* I had written to-that point and guided me through the general introduction and discussion. He also helped clarify the central message in each chapter, was a reservoir for great readings and provided sound methodological advice. His calmness during an especially difficult and upsetting period was essential. Prior to that, Drs Kai Ruggeri, Simona Schnall, Jochen Menges, Joyce Pang, Wayne Warburton, Kay Bussey, Mike Jones and the late Doris McIlwain all patiently indulged my half-cooked academic fancies and, ultimately, inspired the level of confidence I needed to produce independent research. I also enjoyed a particularly fruitful four months computational training with the Advanced Analytics Team at McKinsey in Dusseldorf. A special thanks to Tobias Baer, Frank Jaeger and Myriam Thoemmes for trusting that I could move beyond my comfort zone, formalise novel ideas quickly in code and hit concrete milestones.

Friends have played an equally important role. Thanks to the Cambridge Prosociality and Well-Being lab—Anandhi Vivek, Antonia Sudkaemper, Bryant Hui, Joe Chancellor, Laura Renshaw-Vuillier, Laurie Parma, Moh Yearwood, Rui Sun and Sai Li—for always being there to laugh and help answer loads of empirical (and life) questions. Outside of lab, I'm equally thankful to Alina Guna, Eduardo Machicado, Greg Wilsenach, James Mackovjak, Johanna Lukate, Olly Melville, Shakked Halperin, Somer Greene, Sophie Rosenberg, Tim Rudnicki, Tomas Folke, and many others for the joyous ways they enhanced my PhD experience. Thanks also to the sports teams and student organisations—Cambridge University Cricket Club (CC), Gates Cambridge scholars, Remnants CC, St Giles CC and Trinity Hall MCR (Cohort, Committee, Football Club)—that were so often my favourite escapes.

Finally, I give my humblest thanks to various sources of financial aid. The PhD was only possible with generous support from the Gates Cambridge Trust and the Bill and Melinda Gates Foundation. During the PhD, I also received support from Trinity Hall and the Cambridge Philosophical Society. Prior to that, Trinity Hall, Cambridge Trusts, Endeavour Australia and Macquarie University gave me the opportunities to study and grow in Europe, Asia, North America and South America. I am excited to pay those privileges forward.

Of course, countless others donated their time, gave sage advice and were great friends. Even perceivably innocuous support often had a huge impact. I am extremely lucky. Thank you.

Table of Contents

The Ethical Obtainment of Data 11
1.1. Introduction 11
1.2. Project Overview 11
1.3 Twitter Terms and Conditions 13
1.4. Information and Consent 13
1.4. Mitigating Conflicts of Interest 14
1.5. Data Used in the PhD 15
1.6. Interim Conclusion 15
1.7 AXA Sampling Procedure and Format 15
General Introduction 19
2.1. Abstract 19
2.2. Introduction 19
2.3. Key Constructs 20
2.3.1. Big Five Personality 20
2.3.2. Subjective Well-Being 24
2.4. Bivariate Associations 28
2.4.1. Theoretical Processes 28
2.4.2. Factor Effects 29
2.4.3. Facet Effects 30
2.5. The Problem 31
2.5.1. The Replication Crisis 31
2.5.2. Facet Effect Limitations 33
2.6 Computational Psychology 35
2.7. Present Studies 37
2.7.1 Methodological Decisions 37
2.7.2 Empirical Chapters 39
Can Researchers Predict Psychological Characteristics for Specific Individuals from Their Online Data? 41
3.1. Abstract 41
3.2. Introduction 42
3.2.1. The Problem 43
3.2.2. Everyday Personality Expression 43

3.2.3. Online Personality Predictions 44
3.2.4. The Ecological Fallacy 45
3.2.5. Shifting to Inaccuracy 46
3.2.6. Measuring Accuracy for Specific Individuals 48
3.2.7. Present Studies 49
3.3. Study 1 51
3.3.1. Method 51
3.3.2. Results 52
3.4. Study 2 64
3.4.1. Method 64
3.4.2. Results 65
3.5. Study 3 68
3.5.1. Method 68
3.5.2. Results 72
3.6. Discussion 74
3.6.1. Predicted Big Five Personality 76
3.6.2. Practical Implications 77
3.6.3. Wider Privacy Considerations 78
3.6.4. Limitations and Future Directions 79
3.6.5. Conclusion 80
Rescoring the Balanced Measure of Psychological Needs (BMPN) to Capture Subjective Well-Being 83
4.1. Abstract 83
4.2. Introduction 83
4.2.1. BMPN Development and Validation 84
4.2.2. Alternative Structures in the BMPN 85
4.2.3. Superordinate Psychological Needs Factors 86
4.2.4. Incremental Validity 86
4.2.5. Computational Psychometrics 88
4.2.6. Present Studies 88
4.3. Study 1 89
4.3.1. Method 89
4.3.2. Results 94
4.4. Study 2 97
4.4.1. Method 97
4.4.2. Results 104

4.5. Discussion 109
4.5.1. Implications 110
4.5.2. Limitations and Future Directions 111
4.5.3. Conclusion 112
Propensity Score Matching Increases the Internal Validity of Big Five Facets Effects on SWB 113
5.1. Abstract 113
5.2. Introduction 114
5.2.1. Big Five Personality and SWB 115
5.2.2. Limitations in Existing Research 116
5.2.3. The PSM Solution 117
5.2.4. Combined PSM and Elastic Net 119
5.2.5. The Present Studies 120
5.3. Study 1 121
5.3.1. Method 121
5.3.2. Results 134
5.4. Study 2 141
5.4.1. Method 141
5.4.2. Results 142
5.5. Discussion 145
5.5.1. Big Five Facet Associations with SWB 146
5.5.2. Propensity Score Matching 147
5.5.3. Limitations and Future Directions 148
5.5.4. Conclusion 149
Twenty-Nine Way Interactions? Random Forest Constellations Isolate the Personality Facets that are Prevalent in Extreme SWB 150
6.1. Abstract 150
6.2. Introduction 151
6.2.1. Prevailing Interaction Approaches 152
6.2.2. Problems with the Literature 153
6.2.3. Random Forest 154
6.2.4. Present Studies 155
6.3. Study 1 156
6.3.1. Method 156
6.3.2. Results 159
6.4. Study 2 166
6.4.1. Method 167

6.4.2. Results 168
6.5. Study 3 171
6.5.1. Method 171
6.5.2. Results 172
6.6. Discussion 174
6.6.1. Implications 175
6.6.2. Limitations and Future Directions 177
6.6.3. Conclusion 178
General Discussion 179
7.1. Abstract 179
7.2. Chapter Summaries 179
7.3. Implications 184
7.4. Major Limitations 188
7.5. Future directions 190
7.6. Conclusion 192
Appendix 1.1: Final Approved AXA Ethics Application (PRE.2016.027.V8) 206
Appendix 3.1: 'R' Code for Study 1 and 2 Simulated Correlations 217
Appendix 3.2: Confusion Matrices to Evaluate Bias in Category Assignment 218
Appendix 3.3: 'R' Code for Study 3 Simulated Correlations 221
Appendix 5.1: Comparison Table for Conventional and PSM Models 222

Chapter 1

The Ethical Obtainment of Data

1.1. Introduction

There has recently been increased public scrutiny about the appropriate use of social media data by Dr Kogan—my former PhD supervisor—and his affiliates. To date, the principal focus has been on the large-scale acquisition of Facebook user data without explicit consent, and its subsequent use for commercial purposes. Absolutely none of these data, or the associated self-report data, have contributed to or are reported in my final PhD. Rather, my PhD uses predominantly survey data acquired—to the best of my knowledge—in accordance with a subsequent and fully-approved University ethics application (PRE.2016.027.V8) titled "Cross-National Study of Social Relationships, Prosociality, Well-Being, Health and Political Preferences Using Big Data". Dr Kogan was the primary applicant and I was both a secondary applicant and the corresponding applicant. For full transparency, I describe pertinent details of the ethics application below. The full document is on file at the University of Cambridge School of Biological Sciences. Finally, wider discussion of the AXA project is the ideal context to clarify its sampling procedure and study characteristics, which I do at the end of the chapter.

1.2. Project Overview

The project aimed to use pre-existing archived Twitter behaviour to generate sociodemographic, psychological, health and related survey responses for millions of participants. It was intended for exclusively academic purposes. In the first stage, it proposed that up to 40,000 adult participants from across the world give explicit consent, share their Twitter behaviour (by logging into the Twitter API) and then answer a short battery of pre-approved survey items. Then, different forms of Twitter behaviour—mentions (user references to other Twitter accounts), followed accounts (from which users receive updates), and tweets (user-written content)—were fully anonymised and consolidated into a finite number of abstract dimensions (e.g. 100). These dimensions were used as IVs in statistical models (i.e. algorithms) that predicted users' survey responses. Then algorithms could be exported to predict concomitant survey responses for other users, who did not take the survey battery.

The primary benefit of the project was that it might generate psychological construct scores for a vast proportion of Twitter users from a smaller subset of participants who volunteered both their twitter data and answered my surveys. To this end, Philometrics Inc.—Dr Kogan's private company—donated (totally free of charge) fully anonymous multinational Twitter data for around three hundred million users to the project. As such, algorithms could be exported to an extremely large and heterogenous sample. This could feasibly help overcome common research limitations in social psychology, such as small sample size and unrepresentative participants (Kosinski, Stillwell & Graepel, 2013). As a final step, data were made available for academic research beyond the listed co-applicants—via a pre-approved application procedure—to maximize the potential impact of the project. These data are documented in Figure 1.1.

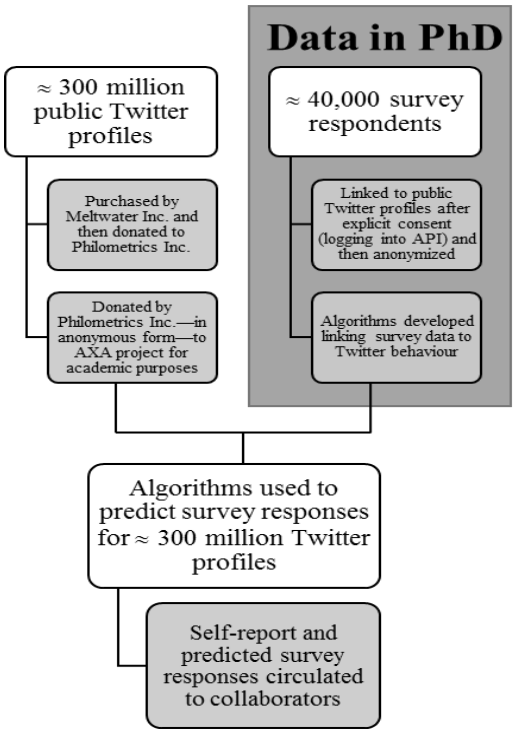


Figure 1.1. Data associated with the approved AXA ethics application and my PhD. The wider project was intended to generate algorithms linking participants' Twitter behaviour to their self-reported psychological construct scores, which could then be exported to users with only logged Twitter behaviour. However, most of my PhD focussed on just the self-report survey data.

1.3 Twitter Terms and Conditions

The project is fully compliant with current Twitter Terms of Service. These were recently made more stringent in the USA and EU, in response to heightened privacy concerns and the introduction of the General Data Protection Regulation respectively (e.g. Henning, 2017; eugdpr.org). At the outset, both emphasise "Twitter is public, and Tweets are immediately viewable and searchable by anyone around the world" (twitter.com/en/tos). Then, they suggest that users can control the extent they publicly disclose contact information, demographics and their existing Twitter behaviour. The US version does not make further provisions, which may indemnify third parties from wrongdoing (iclg.com). The EU version, however, also states

> In addition to providing your public information to the world directly on Twitter, we also use technology like application programming interfaces (APIs) and embeds to make that information available to websites, apps, and others for their use - for example, displaying Tweets on a news website or analysing what people say on Twitter. We generally make this content available in limited quantities for free and charge licensing fees for large-scale access. We have standard terms that govern how this data can be used, and a compliance program to enforce these terms. But these individuals and companies are not affiliated with Twitter, and their offerings may not reflect updates you make on Twitter.

The standard terms are separated into a Developer Agreement and Developer Policy (developer.twitter.com/en/developer-terms). Notably, the former sates that all personal data will be kept "… confidential and secure from unauthorized access…". Then, the latter suggests that direct opt-in user consent is only required when storing non-public Twitter information. To my knowledge, the AXA project kept all Twitter data confidential and used exclusively public information. The EU Twitter policy is also the international default.

1.4. Information and Consent

In the PhD, I obtained fully informed consent for all participants. They gave both their Twitter behaviour and survey responses. In the opening paragraph, the information and consent sheet stated that:

> We are interested in understanding how we can predict people's survey responses from their tweets. Then, I we are also interested in using these predictions to understand the factors that contribute to people's happiness and well-being. To do

> so, we are going to ask you to share your twitter user name with us so that we can look up your tweets and use them in our analyses. (p. 39)

Then it also gave the standard assurances. For example, it stated that data would only be used to generate fully anonymous *group-based* insights for academic research purposes, disclosed other contributing entities (AXA Research Fund, Philometrics), briefly outlined data security protocols and reminded participants they could opt out of the study and/or request to have their data deleted at any time. In addition, it also gave more information about the algorithm generation procedure:

> As part of our project, we will be making use of historic Twitter data that Philometrics Inc. have purchased from Twitter (e.g. tweets, mentions, hashtags). This data will be used in the following way: When your Tweets map onto your survey responses, we will use the information you provide in this study to make algorithms that predict how the users in the purchased historic database of tweets would have responded to these same surveys. (p. 39)

The project did *not* obtain explicit informed consent for participants in the larger Twitter database. The rationale provided in the ethics application was:

> In our view, since the data that is used to generate scores is publicly posted for anyone to see and use, and users of Twitter can be reasonably expected to understand this, it is not necessary to gain [explicit] consent for our application. Furthermore, we minimize any potential harm to the users through (a) anonymization of the data and (b) providing only forecasted scores which, as we describe below, are relatively inaccurate at the individual level. (p. 20)

The technical reasons that individual-level predictions were relatively inaccurate included random errors that emerge when using any inferential statistical approach, and how those errors are compounded with survey measurement imprecision. In Chapter 3, I focus further on the inaccuracy of predicted psychological scores for specific individuals.

1.4. Mitigating Conflicts of Interest

Philometrics donated survey software, Twitter data and data storage for the project. Resources were all provided on an entirely *pro bono* basis. The ethics application stated that listed co-applicants could handle raw Twitter data—principally to extract aggregate logged behaviour

for each user and link it to survey responses—when they were (a) interns at Philometrics; and, (b) working on encrypted and password protected company servers. So far as I know, only aggregated anonymous data was provided to listed co-applicants outside of this context. In accordance with these provisions, I undertook University-sanctioned Leave to Work Away at Philometrics' San Francisco (USA) premises for six weeks in early 2017. My primary tasks were collecting data under the auspices of this ethics application and analysing the results for Chapter 3. To the best of my knowledge, I did not contribute to *any* of Philometrics' commercial activities. All other procedures outlined in the PhD were conducted in a typical University format: I iteratively drafted content that the supervisor then reviewed.

1.5. Data Used in the PhD

My PhD is predominantly based on self-report survey data from the AXA project. Overall, there were approximately 40,000 participants from 33 different countries, who spoke 14 different languages. I describe the full participants at length in Chapters 4 and 5. In addition, I also used consensually disclosed Twitter data—in the above-mentioned formats—from approximately 1,500 participants in the UK, USA and Canada. My PhD does *not* include any data from participants in the larger repository, who did not answer my surveys.

1.6. Interim Conclusion

In summary, all data from the project were obtained with ethics approval for exclusively academic research purposes. All data in my PhD were obtained with *both* prior ethics approval and informed consent. The wider project also obtained archived Twitter data for hundreds of millions of users. These data were originally purchased for commercial purposes—in compliance with Twitter terms of service—obtained by Philometrics and then donated free of charge to the project, along with additional survey software and data storage capabilities.

1.7 AXA Sampling Procedure and Format

The survey data described in this chapter are used throughout the PhD. Thus, I use this final section to give more detailed information about its sampling procedure and format. For full disclosure, Appendix 1.1. contains an *abridged* version of the ethics application. This abridged version excludes cover letters outlining updated revisions—which are already reflected in the attached document—co-applicant details for privacy, and appendices. Appendices were

excluded because they contained mainly standard information and consent protocols that are already quoted at some length in this chapter, and the original-copy grant application that is also explicated in the body of the ethics application. The appendices also contain extensive libraries of scales that were mostly not used in the project. I remind readers that the full, non-abridged, version of this ethics application is available at the Cambridge School of Biological Sciences. Next, I surmise how the project was implemented. Of course, it was fully compliant with the approved ethics application.

The initial primary sampling procedure focussed on using Twitter advertisements. They invited participants to answer a brief survey (e.g. a ten-item personality scale) and then get immediate feedback about their responses. However, Dr Kogan conducted pilot tests in the US and found that the advertisements received very few hits. He instructed me to proceed with the secondary sampling procedure: conventional surveys. There was also a provision of this in the approved ethics application.

I was responsible for the day-to-day development and administration of the surveys. From the outset, the study was intended to be multinational. Initially, Dr Kogan and I identified a small set of geopolitically important countries—China, France, Germany, Japan, Russia, UK and the USA—that it was a priority to assess. Then, I selected as diverse range of other countries that were (a) compatible with collaborator backgrounds, (b) financially feasible with the online survey panel partner (≈ US$2 per completed survey), and (c) projected to yield > 500 participants. Collaborators were responsible for translating and back-translating surveys items into target languages—via the established protocol in psychology (Brislin, 1970)—when there was no existing high-quality translation, per the procedure described in Chapter 4.

Where possible, collaborators were practising research psychologists who had a direct professional or personal (i.e. first-degree) relationship with a member of Dr Kogan's Cambridge Prosociality and Well-Being Lab. On occasion, collaborators had looser connections to the Lab but worked in partnership on their assigned country wave with another first-degree collaborator. Then, I oversaw and managed the work of *all* the collaborators. I determined financial feasibility of each survey based on quotes from our partner: a San Francisco-based online survey panel aggregation company named Cint. Cint offered the cheapest prices I could find across most different countries where survey panels are common. I gained approval from the Cambridge Department of Psychology to use Cint as the exclusive 3rd party survey panel service. Country sample size was projected using a Cint-developed

online tool, which factored in survey length and language requirements. Where there was surplus projected sampling size, I incrementally imposed quotas—for gender, then age, then geographical region—to make the sample as nationally representative as possible. In most cases, surveys were in an official national language. However, there were exceptions. For example, I administered the Finland wave of the survey in English. In such cases, I explicitly asked participants at the outset and in an official national language whether they: (a) were fully fluent in the survey language, and (b) consented to proceeding with the survey.

The default target sample size for each country was N = 1,000. However, due to budget constraints and/or already strong country representation in a world region, such as for some countries in South America, the target was often reduced to N = 500. The only other exception was for the USA, where my target sample size was N = 2,000. That was because it was the first sample, and I wanted to guarantee sufficient power to pilot algorithms linking Twitter behaviour to survey responses. Final sample size in each country differed from its target sample size based on the final number of approved non-acquiescent responses, when precisely the survey link was terminated and, in one case (Bolivia), where it became clear that I was not going to reach the target. Criteria for selecting non-acquiescent cases and final sample size are in Chapter 5: Study 1, which is the first instance in the PhD where I use the entire sample.

Finally, it is important to detail how the survey changed from country-to-country. In all cases, it was designed to be approximately 15-minutes long. I constructed all English-language surveys from the database of questions and scales in the approved ethics application, with oversight from Dr Kogan. It was intended to fulfil grant obligations by primarily assessing participants' personality, health and well-being. I also designated mandatory questions that were administered to every country. They were sex, age, religiosity (no/yes), philanthropy (no/yes), social class, big five personality measured by the 120-item NEO-IPIP scale, at least two convergent measures concerning happiness—one was always satisfaction with life, and the other was usually the balanced measure of psychological needs—subjective health, exercise frequency and fruit and vegetable consumption. Default non-mandatory questions were about emotion experiences, political orientation and an additional convergent measure of big five personality. To incentivize collaborator translations, I invited them to replace non-mandatory questions with their preferred scales—for surveys in their assigned language—provided the scales received prior ethics approval, were compatible with grant objectives, and did not increase the length of the overall survey. Example scales were experiences of gender inequality, coping-style, mental health and right-wing authoritarian beliefs. The majority of items

completed by N > 10,000 participants are described Chapter 4 and/or Chapter 5. The only exceptions are for variables concerning political orientation—e.g. chances of voting in the next national election, right-wing authoritarianism tendencies, attitudes towards women's rights—that were unrelated to my research questions, as well as Twitter behaviour—usage frequency and whether it was participants' primary social media platform—which I ultimately inferred in Chapter 3, directly from usage logs, and were irrelevant thereafter. Full details of the items administered to each country are in the project codebook. It is too large to append to this PhD. Nevertheless, it is available on request to the Department of Psychology, Dr Kogan or myself (matthew.j.samson@gmail.com).

In summary, the AXA sampling procedure, and the survey format, aimed to meet grant obligations and collaborator interests, whilst also pragmatically sampling participants in a wide range of countries and language groups. Where possible, I used quotas to sample nationally representative adult participants. The mandatory section of each survey ensured a high degree of item convergence across countries. While it is unfeasible for the PhD document to disclose the full 300+ page ethics application, or every single survey item that listed in the codebook, both documents are available in full on request.

Chapter 2

General Introduction

2.1. Abstract

What are the personality traits robustly associated with SWB? This guiding empirical focus is predicated on a series of underlying assumptions, which the chapter defends. At the outset, I argue that personality and SWB are both valid and operationalizable psychological constructs. Then, I suggest that they are linked *a priori* by feasible processes. This increases my confidence in the non-spuriousness of existing associations. However, such associations are still often contradictory. To remedy, I suggest that research using high-powered samples and iterative computational analyses can help more definitively isolate the *full set* of personality traits that have robust, internally valid, associations with SWB.

2.2. Introduction

Who is happy? It is perhaps one of the most asked questions in history, occupying the likes of Aristotle, Hume, Freud and the Dalai Lama (White, 2008). It has fostered frameworks advocating everything from virtue, purposefulness and compassion to hedonism, self-enhancement and total pluralism. Ranging perspectives might be galvanized, however, by the simple observation that some people experience happiness more readily than others (Lykken & Tellegen, 1996). This may be captured in personality, which is the default way people interact with the world (Allport, 1937). Personality was documented by the ancient Greeks as the balance of four fluids in the body, and later by psychoanalysts as histrionicism, narcissism and identification with innate human archetypes such as the 'warrior' and 'mother' (McAdams, 1997). Contemporary accounts focus on the universal big five: particularly whether efficiency (conscientiousness) and wanting social harmony (agreeableness) are associated with purposefulness, and negative emotions (neuroticism) and engaging with the outside world (extraversion) are associated with joy (Ryan & Deci, 2001). The preference for novel experiences (openness) is occasionally associated with both purposefulness and joy (e.g. Bardi, Guerra & Ramdeny, 2009). The links between personality and subjectively (self-) appraised well-being (SWB)—the technical equivalent of happiness (Diener, 2000)—are increasingly

important to both policy makers seeking high-fidelity tailored interventions and post-baby boomer generations, who have a particularly strong desire to self-actualize (Diener, Diener & Diener, 2009; Holbrook, 2001).

However, associations between personality and SWB are profligate. Meta-analyses highlight the importance of at least four of the big five, and patterns of effects differ depending on how SWB is operationalized (Lucas & Diener, 2015). Sub-factor (e.g. facet) level analyses are so fragmented—and comprise such interrelated constructs—that it is possible to find evidence for associations between almost every trait and SWB (e.g. Anglim & Grant, 2014). Put simply, research to-date fails to *isolate* a discrete set of robust personality trait associations. This problem is compounded by a possible lack of methodological rigor. The so-called "Replication Crisis" in social psychology (2011-) suggests that existing precedents—e.g. small sample sizes, idiosyncratic study designs and testing multiple *ad hoc* hypothesis—are ill-equipped to detect the often small and fluctuating associations that dominate the sub-field (Shrout & Rodgers, 2018). Computational psychology may help reconcile these limitations. It is predicated partly on automated large-scale data collection and iteratively testing multiple versions of the same research question (Adjerid & Kelley, 2018). In doing so, it offers the potential for more precise effect estimates, fewer methods artefacts and superior statistical control (James, Witten, Hastie & Tibshirani, 2013). My PhD aims to use these emerging computational approaches to isolate the *full but finite* set of personality traits that are implicated in SWB.

2.3. Key Constructs

This section defends the core constructs used throughout the empirical chapters. At the outset, it is important to recognise that I am dealing with fuzzy psychological constructs that might only ever be approximated in data (James, 1890). Thus, it is essential to demonstrate that they are parsimonious—explaining sufficient additional variation in observed behaviour to justify their complexity (Epstein, 1984)—rather than objectively true. I thus show the merits of my preferred personality and SWB frameworks over plausible counterfactuals.

2.3.1. Big Five Personality

There is evidence for personality in written language. In their review, Costa and McCrae (2017) differentiate traits—dispositional styles of interacting with the environment—from characteristic adaptations, which are *context-specific* expressions of the same underlying trait. Both are also separate from abilities, which are learned skills (Goldberg, 1993). According to

the lexical hypothesis—which suggests that all important aspects of personality are encoded in language—researchers can extract traits by examining clusters of related words in a given language (De Raad & Mlačić, 2017). Such approaches typically decompose person descriptors from dictionaries into their grammatical (e.g. nouns, adjectives) and semantic (valence, interpersonal) properties, and then extract clusters that are used interchangeably (e.g. via exploratory factor analysis; EFA). In support, de Raad & Mlačić's (2017) review found that words regularly clustered into extraversion, agreeableness and conscientiousness factors in Germanic, Slavic, wider Indo-European and non-Indo-European languages. These traits were often accompanied by other varying factors, such as neuroticism, openness, honesty/humility and integrity. Although there were inconsistencies, these may have emerged because not all factors are *equally* represented in a given language. This is problematic because techniques like EFA extract a greater number of more granular factors when word frequency increases (Wright, 2017). To illustrate, neuroticism may not have emerged consistently because it is often represented with relatively few common words (e.g. 'stress'), whilst honesty/humility may be an especially well-represented component of agreeableness (Ashton & Lee, 2018). Nevertheless, the lexical hypothesis supports the universality of *at least* three of the big five.

The big five emerged by synthesizing existing personality research. Costa and McCrae's (1976) original big three expanded upon Eysenck and Eysenck's (1975) neuroticism-extraversion framework to also include openness. Openness was, until then, only captured in competing personality scales—notably the 16-PF and the California Q-Sort—through factors like absorption and imaginativeness. Unlike its competitors, however, the big three assumed traits were continuous and normally distributed; they did not necessarily form into either binary categories or discrete trait clusters (Costa & McCrae, 2017). Then, Costa, McRae and Dye (1991) used concurrent findings from the lexical hypothesis to add agreeableness and conscientiousness. The final big five was operationalized in the NEO-PI-R, which asked people to rate the extent that they in general agreed with 240 different pre-validated trait adjectives. Initial support found that the big five structure replicated across multinational populations, there was convergence in self- and informant-reported scores, and scores were mostly stable throughout the adult lifespan (McCrae & Costa, 1992). All these findings have since been extensively replicated in follow up research (for a review, see Allik & Realo, 2017). However, a limitation is that successful replications typically require orthogonal factors. That is, they discount potentially real inter-factor associations (Wright, 2017). In addition, there is no evidence that the big five *comprehensively* accounts for personality. Finally, selectively-

experienced adult milestones (e.g. parenthood) may change putatively stable trait scores (Allik & Realo, 2017). Nevertheless, Goldberg (1990) found that personality scales derived from competing theories all converged on a big five structure, and not alternatives, when using standardized EFA protocols. Thus, the big five may be the least insufficient approximation for true, ecological, personality.

The contemporary big five is hierarchical. Once the big five was established, many researchers relaxed the assumption of orthogonal factors. According to Wright (2017), this led to the discovery of superordinate single- and two-factor personality structures. Using two separate big five scales, Musek (2007) found that up to 50% of the total variation in item responses was attributable to a single meta-factor comprising responses to the more positively valanced pole of each trait. In his meta-analysis, Digman (1997) found that the big five conformed to a two-factor model grouping extraversion and openness, and then the remaining three factors. Musek (2007) continued by finding that one-, two- and (big) five-factor solutions all incrementally increased model explanatory power, albeit with diminishing returns. There are also subordinate facets. Pooling items from multiple operationalisations of the big five, DeYoung, Quilty and Peterson (2007) found that each of the big five had two aspects (e.g. Neuroticism = Volatility & Withdrawal) that could be linked to underlying neuropsychological mechanisms. Conversely, the NEO-PI-R ultimately settled on six theoretically and/or lexical hypothesis derived facets in each factor (Costa & McCrae, 2017). While not always discretely nested in a single superordinate *aspect*, or fully exhaustive, facets still capture a wider plurality of factor components. Recently, Mõttus, McCrae, Allik, and Realo (2014) have also proposed nuances, which are distinguishable *sub-facets*. However, Costa and McCrae (2017) suggest that they still lack a clearly defined taxonomy. Ultimately then, multiple levels of the big five hierarchy may be appropriate to use in associative research.

The recent emergence of aspects has challenged the parsimoniousness of facets. Using an American community sample, DeYoung et al.'s (2007) seminal study found that the 15 different public-domain facets available for each factor collapsed into two correlated aspects. Then, in Canadian university students they found that each of the aspects could be measured using 10 items from the original public domain facet scales. Aspects from the final 100-item scale had common-sense intercorrelations (e.g. assertiveness and industriousness were positively correlated) and at least 8/10 corresponded to the factors that emerge when only heritable aspects of personality are factor analysed (first identified by Jang, Livesley, Angleitner, Riemann & Vernon, 2002). That is, most aspects have plausible genetic bases.

Then, DeYoung, Weisberg, Quilty and Peterson (2013) found that the aspects could explain most of the variation in the competing interpersonal circumplex model of personality, which characterises people on their warmth and dominance. More recently again, DeYoung, Carey, Krueger and Ross (2016) found that the same aspects were also construct valid in abnormal clinical contexts, and could thus be used to differentiate DSM-5 personality disorders. Overall, aspects may be a high-fidelity sub-factor structure of personality, which has plausible corresponding mechanisms and applies to multiple populations.

Nevertheless, facets may still account for noteworthy variation over-and-above aspects. Early support was from Soto, John, Gosling and Potter (2011), who used a large cross-sectional sample of internet panellists to assess age trends in personality scores in over one million participants aged 10-65. They used facets from the big five inventory, which was developed to non-comprehensively but briefly sample two *facets* in each factor. At least 4 facets—depression, anxiety, self-discipline and orderliness—of the 7/10 BFI facets that *directly* corresponded to the NEO-PI-R facets had different developmental trajectories. All seven trajectories were again different when results were also split by gender. Overall, results suggested distinct facet aetiologies and thus dissociable constructs. The big five facets are also supported by research on nuances. Mõttus et al. (2014) found that there was significant cross-observer agreement on individual items after apportioning variance attributable to the superordinate facet. This indicates *non-random* item variation that may capture latent psychological constructs. Most recently, Seeboth & Mõttus (2018) found that such items increased the strength of personality associations—over and above the BFI facets—across 40 sociodemographic and lifestyle outcomes, such as income and sleeping frequency. Together these findings suggest that *real* personality constructs continue to emerge as granularity increases, even to the level of individual items. Facets—the lowest level of the big five with an established taxonomy—may thus explain more variation in target outcomes than aspects.

Facets have wide but not universal empirical support. In McCrae and Terracciano (2005), adults in 50 countries rated the big five using the NEO-PI-R—translated when appropriate—for around 12,000 of their college student peers. After they made scores relative to participants' countrymen, EFA found that the thirty facets collapsed into the orthogonal big five factors—and no additional factors—to parsimoniously explain more response variation than an omnibus general factor. Further, the structure also replicated separately for participants in four different age strata and both genders. Overall, around 95% of the separate country loadings for each factor converged with concomitant, criterion-validity, loadings from an American community

sample. That is, most facets belonged to the same factor across cultures. Exceptions were from some less-industrialized populations. This was corroborated by Zecca et al. (2013), who failed to find any universal big five structure across four regions in Africa. In addition, using both self- and peer-reported big five scales in Bolivian farmer-foragers, Gurven, von Rueden, Massenkoff, Kaplan & Lero Vie (2013) only found support for two superordinate personality factors. Nevertheless, incongruent findings may be caused by the pragmatic need to assess traits through characteristic adaptations, particularly at the granular facet level (LeVine, 2018). Thus, incongruent findings may simply indicate measurement bias rather than culture-specific facets. In addition, a strength of McCrae and Terracciano (2005) is that the factor structure replicated so widely, even when the facets were free to coalesce into alternate structures. Finally, there is still-growing cross-cultural evidence for the NEO-PI-R facet structure. For example, it has been recently documented in both Indonesian and Romanian adults (Wibowo, Yudiana, Reswara & Jatmiko, 2017; Ispas, Iliescu, Ilie & Johnson, 2014). Overall, even universal personality structures will sometimes fail to replicate due to idiosyncratic circumstances. That said, conclusions could be tempered by the apparent increased replication failures in specific non-Western contexts. Thus, the big five facets may commonly, but perhaps not ubiquitously, characterise human personality.

2.3.2. Subjective Well-Being

I associate the granular big five personality facets with general SWB. This is to reconcile the tradeoff between bandwidth and specificity. High bandwidth approaches lead to general conclusions (e.g. Extraverts are happier; Costa & McCrae, 1980), while high specificity approaches lead to more precise conclusions (e.g. Adventurous people are more purposeful; Gavin, Keough, Abravanel, Moudrakovski & Mcbrearty, 2014). They are both important: for charting the parameters of a new field, and isolating associations that may be governed by single discrete mechanisms. Although high-bandwidth approaches are ultimately oversimplifications, there is also risk in prematurely adopting fully high-specificity approaches. According to Rozin (2001), these may neglect the most impactful sub-fields and/or increase the risk that outcomes are confounded by the greater preponderance of proxy variables. For this reason, intermediate investigations—comprising high bandwidth predictors *or* outcomes—can abridge the two approaches. I focus on high-bandwidth SWB for the simple reason that it has less definitive sublevels than personality.

SWB originally concerned human flourishing. It emerged during the wider post-humanistic psychology trend to isolate *discrete psychological constructs* that transcended suffering

(Diener & Ryan, 2009). Diener's (1984) seminal essay introduced both the construct—defined as an introspective (i.e. self-reported) assessment of life quality—and the concomitant tripartite model. The tripartite model suggests that SWB comprises feelings of positive and negative affect, and globally-assessed satisfaction with life. The tripartite model has since become synonymous with the pleasure-oriented, hedonic, approach to well-being (Busseri & Sadava, 2011). In the most recent review, Diener, Lucas and Oishi (2018) noted that there were around 170,000 SWB publications in the past 15 years. They have documented its (a) construct validity both within and across cultures, (b) possible genetic substrates, (c) cross-sectional and longitudinal associations with theoretically linked interpersonal, sociodemographic and other established psychological constructs, (d) environmental contingencies and (e) candidate mechanisms. Overall, SWB and the tripartite model have dominated the well-being literature.

However, there is also countervailing psychological well-being. Ryff (1989) characterized it as a eudemonic approach comprising purposefulness, environmental mastery, positive relationships, growth, autonomy and self-acceptance. It deliberately ignores negative and positive affect. In the most recent meta-analysis, Weiss, Westerhof and Bohlmeijer (2016) documented 27 existing randomized control interventions that aimed to increase aspects of psychological well-being, with outcomes ranging from depressive symptomology to experiences of mindfulness. Results suggested moderate cross-sectional benefits, and small but still significant longitudinal benefits. Overall, researchers continue to distinguish between hedonic and eudemonic well-being. Although the SWB literature may be more substantial,[1] psychological well-being is still relevant in some applied (and likely other) domains.

However, hedonic versus eudemonic well-being may be a false dichotomy. In a recent review, Heintzelman (2018) argued that both may be necessary preconditions for general well-being. Specifically, the author found that the constructs had r = .71 to r = .86 correlations. This is above the common r = .70 threshold for convergent validity, which is used to suggest scales capture aspects of the same underlying construct (Preston & Colman, 2000). In the most recent of these comparisons, Disabato, Goodman, Kashdan, Short and Jarden (2016) evaluated the extent hedonic and eudemonic well-being were differentially associated with various convergent SWB measures such as curiosity, meaning in life and grit. Using over 7,500 participants—who were dispersed across 12 language groups and 7 different world regions—they found that both scales had roughly equal effect sizes for 7/8 outcomes. Failure to find

[1] Ryff (2014) reported only around 350 existing articles on psychological well-being.

discriminant effects suggests that hedonic and eudemonic well-being may both belong to *unitary* SWB.

Indeed, new scales are beginning to reconcile these perspectives. For example, the recently developed Scales of General Well-Being—which comprises 14 facets synopsising the entire SWB literature—has found hedonic feelings of happiness and vitality load positively alongside eudemonic feelings of purpose and connectedness (Longo, Coyne & Joseph, 2017). This has been replicated in the PERMA flourishing scale—comprising positive emotion and accomplishment, among other facets—which is especially popular in applied contexts (e.g. education; Kern, Waters, Adler & White, 2015). Therefore, putatively opposing hedonic and eudemonic theories—and concomitant scales—may capture the same underlying global construct. Thus, they might only differ in the extent they emphasise some (potentially unrepresentative) facets over others.

These findings suggest that the prevailing tripartite model may be competing *directly* with psychological well-being and other emerging frameworks to explain the *same* SWB experience. Importantly, this highlights that the tripartite model is not synonymous with SWB and can thus be evaluated for its parsimony. When doing so, it is important to first classify SWB as a *process* model: it attempts to explain *how* discrete psychological constructs give rise to SWB (Rozin, 2001). Busseri & Sadava (2011) reviewed the existing literature and found there was a lack of consensus on whether its components—negative affect, positive affect and life satisfaction—were three orthogonal outcomes, facets of latent SWB, a pathway where affectivity caused life satisfaction, or a combination of the above. Reconciling using US nationally representative longitudinal data, Busseri (2015) found that sociodemographic variables (e.g. marital status, income) at time one explained the greatest variation in negative affect, positive affect and life satisfaction at time two when they kept the three variables separate, but also apportioned their shared variation into a fourth variable called ‘global SWB’. Jovanovic (2015) found this same structure best accounted for tripartite scale responses in two large cross-sectional samples of young Serbian adults. Together, results suggest that the tripartite model is organised into shared SWB and then the remaining unique variation attributable to its component constructs.

However, this model is only the best of the alternatives tested. Studies all assume affectivity and life satisfaction are initially separate experiences that *subsequently* combine to inform SWB. Parallel streams of research suggest that cognitive appraisals of life satisfaction are

intrinsic to the *formation* of affect, and vice versa (Barrett, 2017). That is, they are not necessarily separate processes from the outset. They are also incomplete. For example, despite including both negatively and positively valanced affect, the tripartite model only includes positively valanced life satisfaction. It fails to account for how cognitive appraisals of life *dissatisfaction* may reduce SWB (e.g. Hoeyberghs et al., 2018). Overall, despite emerging consensus on the structure of the tripartite model, it may still lack ecological validity because it distinguishes between only partial candidate processes.

There are few other established process models. As mentioned above, the Scales of General Well-Being was only recently developed to measure the full range of published SWB *processes*. Importantly, Longo et al. (2017) noted that trends in the literature meant only 1 of the 14 facets (negative affect) measured the absence of suffering, rather than flourishing beyond baseline. Moreover, existing facets may overemphasise the aspects of flourishing that have received greater research attention, other facets may emerge in the future, and the finalised constructs have little demonstrated external validity beyond the original study—either within socio-demographic strata or across cultures. Put simply, comprehensive process models of SWB are still in their infancy.

In the interim, researchers wishing to evaluate SWB as a high-bandwidth outcome may prefer a domain-oriented approach. According to Rozin (2001), domain approaches simply describe the different aspects of life where individuals can experience SWB. For example, they can be either concrete (e.g. sleep, leisure) or motivational (e.g. relationships, achievement). A legacy of humanistic psychology is that the motivation domains are a well-established way of organising wide-ranging behaviours. Seminally, Maslow (1943) expressed them as the hierarchically organised needs for physiological sustenance, safety, love/belonging, esteem and self-actualization. Of course, its strict hierarchy and universality have now been debunked (McLeod, 2007). Nevertheless, Maslow's hierarchy still gave rise to Herzberg's (1966) Motivation-Hygiene Theory, which found that different sets of needs fulfilment protected against job dissatisfaction (e.g. status, comfort) and promoted job satisfaction (e.g. responsibility, growth). It also promoted other general theories of human motivation, such as McClelland's (1965) Acquired Needs Theory, which identified the implicit (i.e. non-conscious) needs for achievement, affiliation and power. These culminated in basic psychological needs theory.

Basic psychological needs theory is the canonical account of SWB motivation domains. It comprises the needs for competence, autonomy and relatedness that manifest in everyday life (Ryan & Deci, 2017). They can be either thwarted or satisfied (Gunnell, Crocker, Wilson, Mack & Zumbo, 2013). A consolidation of the mature needs literature, Ryan and Deci (2001) found other putatively different experiences—such as for self-esteem, meaning and growth—were emergent properties of fulfilling the basic needs. That is, they are comprehensive. The basic needs may also be robustly associated with various measures of SWB both within and across cultures (see Ryan & Deci, 2011), and with a variety of applied outcomes (e.g. van den Broeck, Ferris, Chang & Rosen, 2016). While Ryan & Deci (2001) also suggest the psychological needs foster feelings of intrinsic motivation that ultimately cause SWB, the issue of whether they are separate SWB predictors *or outcomes* is largely trivial. Provided they comprehensively account for the SWB substrates, aggregate psychological needs necessarily capture the overall construct regardless of whether its operationalisations are face valid. Rather, a contemporary area of contention is the extent that basic needs *strength*, which concerns individual differences in the benefits of fulfilling each need (Ryan & Deci, 2000). This was resolved by Chen et al. (2015), who used over 1,500 American, Belgian, Chinese and Peruvian participants to demonstrate that both effects for basic needs satisfaction and needs thwarting on SWB were *not* moderated by individual differences in needs strength. Instead, Rocchi, Pelletier, Cheung, Baxter and Beaudry (2017) found that needs strength was associated with the *sensitivity* that real-world events influenced psychological needs fulfilment. Therefore, needs strength may capture individual differences in *thresholds* for experiencing basic needs thwarting and satisfaction, rather than the phenomenon *per se*.

2.4. Bivariate Associations

The presence of two relatively established psychological phenomena does not guaranteed there will be bivariate associations. As such, I briefly review *a priori* mechanisms linking big five personality to SWB, and then discuss prevailing documented effects. The objective is not to show that effects are definitive, but simply that there are sufficient theoretical and empirical grounds for further associative research.

2.4.1. Theoretical Processes

There are at least three processes that might link personality to SWB. The first is that some traits capture stable sensitivities to directly experience components of SWB, irrespective of context. For example, Schimmack, Oishi, Furr and Funder (2004) found that trait depression

and cheerfulness—the propensities for negative and positive affect—were robustly associated with life satisfaction across cultures. Second, some facets may increase the frequency of extracting the substrates—or nutriments—of SWB from the environment. Evidence for the importance of the environment comes from large national differences in SWB, which tend to have greater variability than even individual differences (Morrison, Tay & Diener, 2011). They highlight that some contexts give drastically more SWB affordances than others. In further support, Prentice, Jayawickreme and Fleeson (2018) recruit whole trait theory to suggest that traits elicit *distributions of states*—via environment selection, selective attention and active interactions—that confer differing autonomy, competence and relatedness experiences. Finally, personality may impact the way people react to various circumstances. Seminally, Boyce and Wood (2011) tracked changes in life satisfaction in German participants from a larger panel who became legally disabled during the four-year sampling period. After a universal decline in SWB immediately post-disability, highly agreeable participants returned to their pre-disability SWB and highly disagreeable participants experienced further decreases. While mechanistic arguments are largely beyond the scope of this PhD, their feasibility still increases my confidence in the non-spuriousness of existing bivariate associations.

2.4.2. Factor Effects

Personality is also linked empirically to SWB. Steel, Schmidt & Shultz (2008) conducted the seminal meta-analysis using the big five. It comprised around 350 different samples and 120,000 participants. They found an up to four-fold increase in effect sizes compared to the previous personality-SWB meta-analysis, which used inconsistent personality operationalizations (see DeNeve & Cooper, 1998). While effects for neuroticism and extraversion were largest, there were also associations across the *entire* big five for job satisfaction, happiness, life satisfaction, affect and life quality. Effects held after controlling for a range of methodological and demographic study differences. More recently, Strickhouser, Zell and Krizan (2017) conducted an empirical *meta-synthesis* of the 36 existing meta-analyses (> 500,000 participants) linking personality to various health outcomes. They found moderate combined effects for the big five ($r \approx .40$) on mental health outcomes—defined as non-physical negatively or positively valanced experiences—in non-clinical adult populations. Effects were largest for conscientiousness, agreeableness and neuroticism. Evidence from prospective longitudinal studies suggested that effects also held over time. Results may have diverged from Steel et al. (2008) because of different SWB definitions.

2.4.3. Facet Effects

Although well-documented, factor associations with SWB are often too broad to be useful. Whilst they may be appropriate in many research contexts. Establishing a scalable framework for more granular constructs may help isolate discrete mechanisms. However, the hierarchical structure of the big five means that there are complex intercorrelations both within and across factors. This makes it difficult to fit controls *a priori* because confounds could differ according to both the facet and outcome in question (and perhaps also the population sampled). Problematically, fitting all 29 facets as controls is usually impossible because it suppresses real effects (Cohen, Cohen, West & Aiken, 2013). Stepwise regression, multiple regression and state-based effects are three documented solutions.

Stepwise regression is when pre-defined superordinate predictors are allocated all the predictor-outcome covariation that they share with subsequent predictors (Cohen et al., 2013). Using stepwise regression, Schimmack et al. (2004) found that the exclusively affective facets—depression from neuroticism and cheerfulness from extraversion—were associated with SWB. Results were consistent using convergent measures of the big five, self- and informant-reported SWB, and heterogeneous student samples. There were no other facet associations across the entire big five. Similarly, Quevedo and Abella (2011) used stepwise regression to evaluate associations across all 30 facets among Spanish students and their friends and family. In partial support, depression and achievement-striving from agreeableness were the only two effects. Overall, stepwise approaches suggest that as few as 2 of the 30 big five personality facets—one being depression—are implicated in SWB. Thus, they may give a particularly narrow account of personality-SWB associations.

Multiple regression is another alternative. Albuquerque, de Lima, Matos & Figueiredo (2012) evaluated the unique effects of each neuroticism, extraversion and conscientiousness facet on SWB after controlling for the other five facets in their factor. Four neuroticism facets, one extraversion facet and three conscientiousness facets emerged. Thus, results suggested that multiple traits impacted SWB. Similarly, Anglim and Grant (2016) examined all facet associations with SWB after controlling for the big five personality *factors*. Also, in partial support of Schimmack et al. (2004), they found effects for depression and self-consciousness from neuroticism, and cheerfulness from extraversion. Thus, results also supported the importance of trait affectivity. Although these studies had divergent results, they did suggest multiple regression yielded *more* facet associations than stepwise regression.

There are also a wide range of possible facet-SWB association in adjacent literatures. For example, according to the emotion regulation perspective, low SWB is rooted in negative phenomenological experiences (Aldao, Nolen-Hoeksema & Schweizer, 2010). As such, facets may be implicated when they cause intense negative emotions—such as in anxiety—or increase the *frequency* of experiencing discomfort, such as in vulnerability (Headey, Kelley, & Wearing, 1993; Steptoe, Hamer & Chida, 2007). Similarly, the inability to mitigate negative emotion elicitors may be captured in both facets associated with conflict management—such as gregariousness and cooperation (Antonioni, 1998)—and a general lack of agency—such as in low self-efficacy and self-discipline (Kim-Cohen et al., 2006). High SWB may be associated with facets that promote goal-oriented behaviour, such as self-efficacy, self-discipline and achievement-striving (Sheldon & Elliot, 1999). Further, friendliness, gregariousness and altruism may help leverage the benefits of social networks (Helliwell, 2006). Therefore, adjacent literatures suggest there may be a wide range of facet associations with SWB.

2.5. The Problem

The simple fact is that existing facet-level associations contradict one another. The only consensus is that depression from neuroticism is implicated, alongside between one and at least seven other facets from across four of the big five. There may be empirical justification to link most of the facets with SWB when also considering research from adjacent literatures. The problem appears to transcend any single study and might thus reflect the wider difficulty of conducting rigorous research in the field. As such, I first contextualise the problem using lessons from the recent 'Replication Crisis' in social psychology. Then I address more specific methodological limitations.

2.5.1. The Replication Crisis

The replication crisis—realization that prevailing social psychology methods may yield unacceptably high error rates—was precipitated by a rapid succession of tenuous findings. From 2010 to 2012, prominent researchers were caught fabricating data (e.g. Diederik Stapel; Levelt, Drenth, & Noort, 2012); there was a putatively gold-standard proof that humans could see into the future (Bem, 2011); and, a famous unconscious priming effect—subliminal elderly primes slowed walking speed—was debunked because it was highly contingent on the original, idiosyncratic, study context (Doyen, Klein, Pichon & Cleeremans, 2012). These events activated dormant concerns about artificially dichotomous p-values, underpowered studies and publication bias (e.g. Cumming & Calin-Jageman, 2016). They also ushered a period of intense

methodological scrutiny. New criticisms emerged about the pervasive multitude of ways that researchers (a) failed to weight results by their *a priori* plausibility (Bayesian inference; Wagenmakers, Wetzels, Borsboom & Van Der Maas, 2011), (b) artificially inflated both effect precision and magnitude (Szucs & Ioannidis, 2017), and (c) failed to corroborate existing findings (Everett & Earp, 2015). Early findings indicted the entire discipline.

Then, large-scale collaborative studies investigated the true extent of replication failure. Prominently, Klein et al. (2014) attempted to replicate 13 published effects in psychology using over 6,000 participants in 33 different research labs (majority American). Ten of the findings replicated consistently, but there was substantial variability in even their effects across labs. Then, the Open Science Collaboration (2015) replicated 100 studies published in three prestigious psychology journals during 2008. Despite similar or larger samples than most of the original studies, there were only 36% successful replications. Aggregate observed effect magnitude was halved. It is difficult to determine the extent findings generalise to the entire field because replications were unrepresentative and conceptual, rather than direct, and/or used different populations. Nevertheless, findings did highlight that even seemingly robust psychological research often fails to clearly isolate target phenomena.

Recent review articles have attempted to consolidate lessons from the replication crisis. Nelson, Simmons & Simonsohn (2018) highlight the fallibility of p-values, suggesting that statistical power is compromised every time researchers make design and/or analysis decisions based on partial data. In addition to full disclosure and open materials, they suggest study preregistration to commit researchers to their *a priori* defined protocols. Shrout and Rodgers (2018) also highlight the importance of distinguishing between exploratory and confirmatory research, evaluating the conservative bound of CIs, and fully accounting for statistical power and design features (e.g. self-report vs behavioural outcomes) in meta-analyses. Ideally, they suggest that effects are triangulated across *populations* of high-powered replication studies. However, these recommendations also implicitly recognise that genuine psychological effects are fickle. De Boeck and Jeon (2018) review existing meta-analyses to suggest that most effects explain 10-25% predictor-outcome covariation. This is almost equivalent to the variation caused by both differences in within-population study designs and between-population moderators. Overall, I thus conclude that methods must improve to re-enforce the marginal superiority of real but fluctuating effects over their contingencies.

How does this lesson apply to facet-SWB associations? The multitude of facet IVs reduce the effect magnitude of any single association. Then, facet specificity introduces further measurement error because, by necessity, it assesses more discrete behaviours that are prone to being *culture-specific* characteristic adaptations (Costa & McCrae, 2017). Further, there are a lack of attempted replications. This limits the ability to explain inconsistent findings through discrete population-level moderators (e.g. collectivism). Although studies may have been sufficiently powered, their lower-bound CIs are often so negligibly beyond zero that effects could have disappeared in even slightly different testing conditions. Ultimately, these contingencies all jeopardise the already-small margin between genuine but changing population effects, and methodological artefacts. There are also more specific limitations.

2.5.2. Facet Effect Limitations

Existing facet-level research tends to use either stepwise or multiple regression to control for potentially confounding personality variables. In doing so, they make *a priori* decisions about whether personality facet covariation should be allocated to one facet over another, or completely apportioned from the analysis. This is reasonable provided: (a) there are a relatively small number of covarying facets that can be organised by theoretical precedence for stepwise regression, and/or (b) facets form a comprehensive set of plausible multiple regression controls that are not so exhaustive that they compromise the construct validity of the target. Problematically, these conditions are not met for the big five facets. Their complex structure means that any of the 29 non-target facets could confound associations with SWB.

The safest strategy is to control for *every* facet. However, this grossly limits the extent that any facet can share a *unique* association with the outcome—over and above the other facets—due to multicollinearity. The alternative is to relax controls. However, this increases the likelihood of confounding. Existing stepwise approaches adhere to this tradeoff. Stepwise regression may be adept at finding (confounded) support for the variables with the strongest theoretical associations because they are given *all* the shared facet covariation (Thompson, 1995). However, other facets must then compete for proportionally little variance in the outcome. As this process is repeated across multiple steps, the burden of proof may become untenably high. Results become self-fulfilling: Variables arbitrarily assigned to the first step have the highest chance of emerging as significant. Had Schimmack et al. (2004) cited the importance of (e.g.) self-efficacy and self-discipline in SWB—which is certainly defensible (Duckworth, Peterson, Matthews & Kelly, 2007)— it is unclear whether depression and cheerfulness would have even emerged as robust predictors. Further, assuming all facets in the subsequent step have equal

predictive power, the ones least associated with the superordinate facet will be most strongly associated with the outcome. This may cause *cascading* arbitrary effects. Therefore, at best stepwise regression may artificially diminish the full range of facet predictors. At worst, it may mean that emergent facet-SWB associations are completely spurious.

Multiple regression approaches are equally fraught. Controlling for just intra-factor facets does not mitigate the likelihood of confounding from facets in other factors. Alternatively, controlling for the big five factors completely removes any covariance the facet shares with its *own* superordinate factor. This is especially problematic because the factor and facet are, by definition, intended to measure overlapping aspects of the same global personality trait. Further, circumplex approaches suggest that any single facet can be conceptualised as a constellation of other facets (Saucier & Ostendorf, 1999). For example, friendliness might comprise high trust and gregariousness. Thus, fitting even a moderate number of controls can remove components of the target facet that are intrinsic to the very underlying construct it is intended to measure. Put simply, multiple regression compromises facet construct validity. This could be another reason for the fluctuating associations reported to date.

Contrastingly, state-based effects are so fragmented that there is little cross-study equivalence. Put another way, studies comprise largely different samples, variables and variable operationalizations. Thus, it is difficult to disambiguate real and artefactual effects. For example, findings could (a) be obfuscated by legitimate population contingencies; (b) proxy for other facet effects that are either uncontrolled or unaccounted for in the manipulation check; (c) be specific to partial operationalizations of SWB; and/or, (d) use scales/manipulations that have different intensities, thus compromising effect size estimates. Indeed, the extent of these inconsistencies means that it is feasible to find '*a priori'* evidence linking almost every facet to SWB. Thus, results simply affirm factor-level findings that implicate most traits. Over time, cross-study equivalence might be increased through meta-analyses, which fit many of these design contingencies as moderators. However, this relies on the kind of critical mass of studies that only accumulates over a long period of research and with sufficient expenditure. Even then, it is unclear whether states capture the omnibus facet. They may instead capture specific nuances, and thus have limited applicability to any current, definitive, trait taxonomy. Overall, research on states suggests wider ranging facet effects, but findings are only ever preliminary.

2.6 Computational Psychology

Emergent computational methods may offer a solution. Although replication crisis reviews focus on increasing power, pre-planning analysis and conducting more replications, they also briefly examine the potential for new technologies. For example, Nelson et al., (2018) suggest making comprehensive study materials open access online. Shrout and Rodgers (2018) highlight that iteratively analysing sub-samples of participants can help stabilize effect estimates. However, there may be wider-ranging benefits. Lazer et al. (2010) introduced Computational Social Science as "The capacity to collect and analyse massive amounts of data" (p. 721). They highlighted the potentials of live GPS-based person tracking, automatically imposing structure on complex sources of information (e.g. videos, natural language) and making inferences from online behaviour. Of these, sampling and iterative testing strategies may be particularly adept at reconciling facet associations with SWB.

One noteworthy benefit of the computational paradigm is cheap large samples. This is highlighted by both online panel and algorithmic sampling strategies. Publicly available online survey panels have dramatically reduced the cost of collecting data. Buhrmester, Kwang and Gosling (2011) evaluated data quality on Amazon's pioneering Mechanical Turk platform. They found scale internal consistency and intercorrelations across multiple testing occasions were comparable to offline methods, and that compensation rates as low as US$0.02 only marginally reduced data quality. Results were recently questioned by Matherly (2018), who found that panel members with the same 'reputation'—the metric used to determine their level of pay—produced highly-variable data quality. However, this can usually be remedied by post-test filtering for atypical response patterns (e.g. acquiescence, manipulation check items; Cohen et al., 2013). Moreover, Walter, Seibert, Goering and O'Boyle (2018) conducted a meta-analysis of online panel internal consistencies and effect sizes, which comprised 90 different samples and over 30,000 participants. Internal consistencies were all within the credibility bounds of their offline equivalents—which were taken from existing meta-analyses—most effect-size estimates overlapped, and there were no differences in aggregated overall effects. Therefore, it may be possible for *individual research teams* to collect large survey data, even after adjusting results from the above studies for price inflation.

The algorithmic approach was popularized using social media data. Kosinski et al. (2013) used Facebook page likes for over 58,000 participants to predict their self-reported sociodemographic (e.g. sexuality, religion) characteristics and big five personality. This was consolidated by Youyou, Stillwell and Kosinski (2015), who found that predictions were more

accurate than most ‘classical’ informant reports. Recent research has even found that single Facebook Likes and single online dating profile images can predict personality and sexual orientation respectively (Wang & Kosinski, 2017; Matz, Kosinski, Nave & Stillwell, 2017). The implication is that algorithms can also generate predictions for people who do not also give self-reports. It opens the possibility that individual research teams could either borrow or create algorithms developed using a relatively small sample of participants and apply them to massive databases of online behaviour. Online panel and algorithmic approaches may render p-values and even confidence intervals—which are also predicated on the SE and thus sample size (Cohen et al., 2013)—redundant. That is, they might isolate extremely precise effect patterns in the population sampled, at least presuming the scales measured are construct valid.

Automated iterative analyses—also referred to as machine learning—are characterized by resampling and parameter tuning. Resampling involves evaluating the same research question in multiple subsamples (James et al., 2013). Two prominent examples are bootstrapping and k-fold cross-validation. Both involve aggregating results from an entire population of models to increase robustness. I also extend resampling to include any other analysis that iteratively uses different participant and/or variable subsamples, as well as simulations (Adjerid & Kelley, 2018). For example, different subsamples could ask the same research question across varying sociodemographic strata or across multiple construct operationalizations. Simulations are when each constituent subsample is *generated* according to a different (defensible) permutation of assumptions (James et al., 2013). For example, I might assume that two variables are normally distributed and have a moderate (e.g. r = .30) correlation. The primary benefits of resampling are triangulating results across multiple variable operationalizations, and increasing the information about the conditions that give rise to especially weak and strong effects.

Parameter tuning varies the statistical analyses. Ridge and LASSO regression are perhaps the most prominent examples in psychology. They both systematically vary individual effect estimates to maximize the explanatory power of the entire model (Zou & Hastie, 2005). Another example is participant weights, which might be used to test models on subsets of participants with different characteristics. Weights could be used to select participants with the same covariates, or to test the extent results generalise to different populations. Overall, iterative approaches are thus concerned with isolating particularly robust effect estimates.

These computational approaches can mitigate fickle facet effects. For example, large sample size gives unprecedented power to evaluate patterns of effect sizes without conflating Type 1

error. They are so economical that it is often feasible to increase survey length, which allows researchers to evaluate the same hypothesis with multiple converging IVs and DVs. An auxiliary benefit is that both online panels and logged internet behaviour give additional access to non-WEIRD (western, educated, industrialised, rich, democratic) participants with little extra effort and often at reduced cost (Henrich, Heine & Norenzayan, 2010). During resampling, relatively large effects may indicate that certain variable operationalizations capture extremely robust effects. They may also show that results are not simply artefacts of idiosyncratic, singular, measurement choices. Averaging results across models with iteratively sampled *partial* controls may increase internal validity without causing multicollinear effects. In many cases, parameter tuning can help extract more ecologically valid associations between variables, ensuring the statistics are optimized to account for true real-world phenomena (James et al., 2013). Overall, computational psychology offers a battery of possible solutions.

2.7. Present Studies

The empirical chapters revisit bivariate associations between big five personality and SWB using a computational psychology paradigm. Throughout the General Introduction, I argued that the big five facets are the most granular—robustly-supported—level of personality. Then I suggested that it is premature to define SWB as a series of discrete psychological processes. Instead, it can be viewed as the combined feelings of autonomy, competence and relatedness. The existence of plausible processes linking personality to SWB is corroborated by existing bivariate associations. However, isolating specific effects, and effect patterns, is more difficult. The replication crisis in social psychology highlighted that effects may be contingent on unergeralizable sampling decisions, study designs and analyses. Computational psychology—which leverages huge samples and iterative analyses—mitigates these limitations by increasing sample heterogeneity and power, systematically accounting for methodological artefacts and increasing statistical control. Next, I document my general methods decisions and then briefly surmise the empirical chapters.

2.7.1 Methodological Decisions

There are some methods decisions that span the empirical chapters. I focussed on personality and SWB because they were relatively incontrovertible psychological constructs that I assessed in most country waves of the AXA study. They have widely established cross-cultural validity. They were also expediently translatable into multiple languages—when an existing translation did not exist—and could be assessed using self-report. Chapter 1 contains transparent

disclosure of the AXA study sampling procedure and the concomitant survey format. Specific sample characteristics and key variables are described in Chapter 5, which is the first instance that I use the full AXA study data. Personality-SWB associations—whilst interesting in their own right—are also a use-case for my documented battery of methods interventions. The primary data to hand is a cross-sectional dataset of ≈ 36,000 participants (≈ 40,000 prior to removing acquiescent responses) spanning all 6 permanently inhabited continents, 14 different language groups and 33 countries. Psychological construct scores were relative to participants' countrymen. This meant I controlled for *all* country-level effect contingencies (Aguinis, Gottfredson & Culpepper, 2013). Although I often had intuitions about effects, I refrained from proposing specific directional hypotheses because the literature often supports multiple conflicting perspectives.

Of course, my computational approaches involve performing far more statistical models than conventional alternatives. However, they may not lead to excess Type 1 error because they avoid using parameter estimates. For instance, they ignore t- and p-values. Instead, they take either the overall model explanatory power and/or effect estimate for that *specific sample* without making normative inferences (James et al., 2013). Such inferences are instead made at a second-step, using the bootstrapped *population of results* from across changing models. In most instances, multiple effects can thus be consolidated into single parametric inferences.

It was appropriate to use the prevailing power analysis method in social psychology. Throughout the PhD, I evaluate associations *within* the big five factors and facets, and also within convergent measures of SWB. Then, I of course evaluate associations between the thirty big five facets and my two primary SWB outcomes, which capture suffering and flourishing. I use a power analysis predicted on the very dichotomous p-values I criticized above (Cohen, 1992). To explain, such power analysis is advocated in widely influential instructional texts (most notably "The New Statistics"; Cumming, 2013), alongside increased focus on confidence intervals. Further, large sample sizes yield precise effect estimates because they are predicated on the standard error. That means the bounds of even extremely conservative confidence intervals often converge with the point estimate. The putatively dichotomous power calculation may, in the present study context, apply almost equally to the entire *range* of plausible effects. Finally, it is only a provisional metric. I further mitigate the likelihood of spurious findings throughout the PhD by only evaluating (a) effects with noteworthy magnitudes, and/or (b) the consistency of *whole patterns* of effects, which are far less likely to be caused by Type 1 error (Murayama, Pekrun & Fielder, 2014). Overall, such a power analysis was appropriate because

of existing precedents (and thus also familiarity in the field), the large sample size and its use as a provisional tool alongside other approaches that were intended to mitigate spuriousness.

Indeed, confidence intervals are often calculated from the standard error or point-estimate effects from 500+ bootstrapped resamples (Preacher & Hayes, 2004). Both have high levels of convergence with point estimates when there is extremely large survey data, such as in the present PhD (Cohn & Becker, 2003). Thus, assessing the probability that effects of a pre-specified magnitude truly exist in the population is almost the equivalent of assessing the probability that the same target effect adjusted for its conservative plausible magnitude (i.e. the CI bound nearest zero) also exists the population. Finally, power analyses focus on effects that are large enough to be theoretically important (e.g. r = .10, which is the magnitude that often corresponds to a 'small' effect in social psychology; Cohen, 1992). Thus, I replace the absolute null hypothesis (i.e. no effects whatsoever) with a more parsimonious threshold.

There was ample study power. I targeted at least 99% chance of detecting real effects in the population. Even assuming I perform 1,000 inferential statistical tests using the conservative Bonferroni-corrected p-value significance threshold of .001 (i.e. .001/1,000), I have sufficient power to detect correlations as small as r = .04. That is, given 1,000 tests of effects around r = .04, there is 99% probability that at most 1 is a false positive. There is a virtually perfect chance I detect all exclusively real effects r > .10. Standard error is inversely proportional to sample size, and thus point estimates and confidence intervals often converge. This means I can isolate effects very precisely and contrast their magnitudes. Importantly, there is also still surplus power for the planned analysis, which can offset any additional sources of statistical error.

2.7.2 Empirical Chapters

The following chapters use computational psychology approaches to sequentially reappraise components of the research process, as they pertain to the cross-sectional study of personality associations with SWB. Full comprehensiveness is beyond the scope of any single PhD, and I thus focus on especially topical and/or empirically relevant issues. Chapters 3 and 4 are on univariate measurement, and then Chapters 5 and 6 are on bivariate associations.

In univariate measurement, I focus on variable operationalization. Chapter 3 begins by evaluating the feasibility of examining psychological characteristics online. Emergent research suggests that logged online behaviour can be used to predict personality. I evaluated the veracity of this claim by quantifying the extent prediction algorithms—which are normative—apply to specific individuals. I found individual predictions are negligibly above chance for

even hypothetically accurate algorithms that are beyond the scope of current technologies. This suggests that predicted personality may be safe to use for normative analyses without compromising individual privacy. Chapter 4 continued with the dependent variable group: SWB. I evaluated the extent a prominent self-report operationalization of the basic psychological needs could be repurposed to measure global SWB. I found consistent evidence for separate needs thwarting and satisfaction factors. They were disproportionately associated with subjective ill- and well-being respectively, captured more transitive SWB than life satisfaction, and additional explained unique variation in real world outcomes. Thus, needs thwarting and needs satisfaction were the primary outcomes for the subsequent two chapters.

In bivariate associations, I focussed on internal validity. Chapter 5 evaluated bivariate associations between all 30 big five facets and SWB. I increased internal validity by iteratively extracting pairs of participants who had similar covarying facets but differed on the target facet. Observed effect patterns were *more accurate* than both conventional multiple regression and even advanced machine learning alternatives. Thus, they may help better identify the most promising effects for further analysis. Finally, Chapter 6 found the most robust, internally valid, personality-SWB associations. Specifically, I evaluated the facets that were consistently associated with extreme SWB across the full range of changing intrapersonal trait contexts. I initially found that plausibly *real-world* individuals with extreme SWB had personality profiles that (differentially) deviated from the population mean on most facets. Then, I found that *hypothetical* simulated individuals—who were free to have both likely and unlikely patterns of facet scores—deviated from the population mean on a total of only 9/30 facets. These were the best-candidates to be internally valid effects because prevalences were robust regardless of individuals' wider personality. That is, they were not simply piggybacking on the covarying facet effects most prevalent in the population. Overall, I found first evidence for the comprehensive, yet still discrete, range of personality facets implicated in SWB.

Chapter 3

Can Researchers Predict Psychological Characteristics for Specific Individuals from Their Online Data?

3.1. Abstract

A recent wave of research has found that online social media behaviour can be used to generate prediction algorithms about peoples' psychological characteristics from their online behaviour. Findings are robust at the group level: there are 'statistically significant' associations between Facebook, Instagram and Twitter logs, and self-reported psychological construct scores. Effects also regularly exceed r = .30, which is *large* by social psychology standards. That means they explain more than the typically observed amount of predictor-outcome covariation. As such, existing research might reasonably conclude that patterns of online behaviour reflect psychological characteristics, on average and in the population/s sampled. However, many of these studies extrapolate from their group-based insights to claim that algorithm-predicted scores are also highly accurate for specific individuals. This claim does not necessarily follow from the existing data. The ecologically fallacy suggests that sample-level trends only weakly manifest at more granular levels, such as for specific individuals. The problem may be exacerbated for predicted psychological characteristics, because sample-level performance metrics are relative and not absolute. To reconcile, I directly evaluated the veracity of these extrapolations. I predicted big five personality using a combination of simulated (N = 10,000) and real-world survey data (N = 3,132,610), and machine learning with multifaceted Twitter data (N = 1,471) at realistic and hypothetical future accuracies. I found that scores were usually too imprecise to capture specific individuals' self-reported personality, differentiate between individuals with varying levels of each trait, or correctly assign individuals to low, medium, and high categories. Results confirm that even highly robust and *relatively* large group trends are only marginally prevalent in specific individuals. Overall, I conclude that predicted psychological characteristics can be used for normative cross-sectional research—of the kind featured throughout my PhD—without violating individual privacy. It is highly unlikely, however, that they can be used to make definitive inferences about specific individuals.

3.2. Introduction

Academics, policy makers, the private sector, and the wider public are all interested in the potential of emerging large-scale data and computational approaches to improve daily life. For example, entire genomes can be processed to isolate hereditary illnesses (Adams, 2015), Google searches can track the spread of civil unrest (Manrique, Morgenstern, Velasquez & Johnson, 2013), and past Amazon purchases can improve future product recommendations (Chen, Chiang & Storey, 2012). Such technologies might also completely change the ways humans interact with the physical world. For example, 3D printing can encode and then recreate the structure of even complex objects (Rengier et al., 2010), self-driving cars operate on live-stream real world data (Yang & Coughlin, 2014), and virtual reality utilizes multifaceted biomechanical feedback (Burdea & Coiffet, 2003). There are corresponding advancements in social psychology. Prominently, social media information may be used to generate unique psychological characteristic profiles for individual users (Golbeck, Robles, Edmondson & Turner, 2011; Kosinski et al., 2013). The technology may outperform predictions made by work colleagues, close friends, and even family (Youyou et al., 2015). Researchers have even set up *one-click* tests that claim to give accurate profiles (Cambridge Psychometrics Centre; applymagicsauce.com). Overall, proponents claim that online predicted psychological characteristics may usher in a new era of exceptionally precise research.

However, such research may be derailed by concerns that predicted psychological characteristics, such as personality, violate individual privacy. Existing research invites this concern, by claiming that the technology is highly accurate for specific individuals (e.g. see the 'Discussion' sections for Kosinski et al., 2013 & Youyou et al., 2015). Subsequently, there was public outcry that social media-based predictions were used to target individual voting behaviour in the 2016 US presidential election (e.g. Davies, 2015; Grassegger & Krogerus, 2017; Lapowsky, 2017). Prominent figures with access to social media data—such as Facebook founder and CEO Mark Zuckerberg and Cambridge psychologist Dr Aleksandr Kogan—then both testified about their practises in front of US congress and UK parliament (Watson, 2018; Lomas, 2018). Adjacently, there were landmark legal cases where the plaintiff won the right to be forgotten online (Grierson & Quinn, 2018; Mantelero, 2013). In May 2018, the EU's General Data Protection Regulation (GDPR) also came into effect. It drastically increases individual privacy safeguards (Burgess, 2018). Overall, researchers, policy makers and the

public all assume that social-media predicted psychological characteristics apply to specific people. This may have spurred the broad global trend towards protecting privacy rights online.

3.2.1. The Problem

The algorithm technology in question can, feasibly, predict any psychological characteristic that is stable enough to manifest in logged online behaviour. Concerns are not specific to personality. Nevertheless, personality has to date been the focus of both proof-of-concept research and public outcry. There are no rigorous tests of the extent psychological characteristics—as demonstrated through personality—apply to *specific* individuals. Investigating this will help clarify the privacy implications of using algorithm-predicted construct scores throughout the remainder of my PhD.

I contend that it is inappropriate to use any predicted psychological information that is highly accurate for specific individuals without their *explicit prior consent*. Thus, I am in the unusual position of suggesting that prediction algorithms are only useful for psychological research if they are sufficiently *inaccurate*. There are *a priori* reasons that this might be the case. The ecological fallacy suggests that aggregate-level trends—for example, that personality is nested in online behaviour—do not necessarily manifest in sub-levels of the data, such as for specific individuals (Robinson, 1950). Further, prediction accuracy has to date been evaluated using correlations (e.g. Pearson's R) and variance explained (e.g. R^2). These metrics are relative and not absolute, which means that effect sizes are interpreted with reference to arbitrary yardsticks (e.g. r = .30—or 9% explained variation in the DV—may be a 'medium' effect). Thus, they may further mask the true extent of the ecological fallacy. To remedy, I switch to absolute prediction error, which is a person-specific measure of *inaccuracy* (Kelley, 2007). It is less common in psychology because it obfuscates the general population trend. It is more common in applied disciplines because it gives practicable estimates about how much the overall effect manifests in specific cases. With it, I aim to evaluate the full extent that psychological characteristics predicted from existing logged (e.g. online social media) behaviour, and expressed in terms of personality, apply to specific individuals.

3.2.2. Everyday Personality Expression

Personality is the constellation of stable ways that people interact with the world. The most robust conception is the big five, which comprises neuroticism, extraversion, openness, agreeableness, and conscientiousness (John & Srivastava, 1999). Although exact

manifestations of personality may be context dependent, they may also have some underlying commonalities (Allik, 2002). For example, extraverts may almost always prefer socialising with a wider range of people than introverts. This opens the possibility that people may leave constant and thus *recognizable* traces of their behaviour. Gosling, Ko, Mannarelli and Morris (2002) suggested that such traces can be reverse-engineered to reconstruct aspects of individuals' personality. For example, they suggested photos of foreign places might convey high openness, and an ordered desk might be the residue of high conscientiousness. Then, they also found evidence that there were convergent observer ratings of personality from office spaces and bedrooms, and that ratings were positively associated with self- and peer-reported personality. Personality may also leave *constant*, interpretable, physical traces.

Personality expression may also behave similarly in non-physical mediums. For example, Rentfrow and Gosling (2003) found four music preference dimensions that were differentially related to participants' personality profiles. Then, Bonneville-Roussy, Rentfrow, Xu, and Potter (2013) found that music preference expression was constant across ethnicities, ages and countries. Convergently, Gerber, Huber, Doherty, & Dowling (2011) found that researchers could infer the big five from people's political expressions. Then, Hirsh, DeYoung, Xu, & Peterson (2010) found universal associations between political conservatism-liberalism and the big five. In the context of computer-modulated personality, Nass and Lee (2001) found that participants could differentiate between standardized digital voices that had similar and dissimilar personality profiles. Finally, Alam and Riccardi (2014) used machine learning to infer personality from only traces of non-descript spoken conversations. Therefore, personality might also be inferred from immaterial expressions and even computerized speech.

3.2.3. Online Personality Predictions

Social media may be another medium that leaves reliable traces of personality. Gosling, Gaddis, and Vazire (2007) found convergent observer ratings of participants' personality from their Facebook and MySpace profiles, which were then also positively associated with both existing self- and peer-reported personality. Then, Gosling, Augustine, Vazire, Holtzman, and Gaddis (2011) found that the underlying properties of personality expression were constant across online and offline contexts. For example, extraverts engaged in more social interactions on both mediums. Across 89 different countries, Sumner, Byers, Boochever and Park (2012) found that psychopathic, Machiavellian, and narcissistic personality traits were expressed in

approximately the same way offline and on Twitter. Overall, personality may thus be expressed consistently in offline and online contexts, and in similar ways across cultures.

Personality may also be *automatically* inferred from social media data. Golbeck et al. (2011) found that individuals' entire Twitter profiles explained 11% to 18% of the variation in self-reported personality. Then, Kosinski et al. (2013) found that just Facebook Likes predicted 8% to 16% variation in self-reported personality. In a follow up study, Youyou et al. (2015) found that computer personality predictions outperformed friends when participants had just 65 Likes, and outperformed family when participants had 125 Likes. Predictions were forecast to outperform even romantic partner ratings when there were around 275 Likes. Thus, online personality predictions may outperform even human raters as the volume of data increases.

Personality predictions may also improve with increased data *heterogeneity*. Heterogeneity is important because personality is the *average* way people interact with the world across a variety of different contexts (Paunonen & Ashton, 2001). In support, Skowron, Tkalčič, Ferwerda & Schedl (2016) found that user data from both Twitter and Instagram produced more accurate personality predictions than data from either platform in isolation. Ongoing technological advancements promise even more heterogeneity. For example, Volkova, Bachrach, Armstrong and Sharma (2015) predicted personality from natural language in participants' Tweets. Wang and Kosinki (2017) demonstrated that big data technologies could even infer psychodemographic information—homosexuality—exclusively from users' online dating profile pictures. Overall, personality predictions may thus become even more accurate as user data is continuously logged, quantified and then linked across different online mediums—for example, with the assistance of singular user IP addresses and universal logins.

3.2.4. The Ecological Fallacy

However, the ecological fallacy suggests there may be a ceiling accuracy of these predictions that precludes making inferences about specific individuals. It is the faulty assumption that model predictions apply equally to all cases in a population (Brewer & Venaik, 2014). It was first demonstrated by Robinson (1950), who found that nationwide difference in literacy between African Americans and Caucasian Americans were the result of *varying* regional differences. Moreover, a range of *different* regional effects could have produced the same aggregate results. Then, Freedman (1999) demonstrated that the plausible bounds—or confidence intervals—of specific region effects were generally too broad to yield conclusions that applied specifically to those regions. Gerhart (2009) gave a contextualised example:

despite the well-established differences in individualism-collectivism *between countries*, institutional individual-collectivism varied far more *within* than between countries. Overall, the ecological fallacy is thus the failure to recognise that sub-level effects are free to vary in a wide range of ways that deviate from the aggregate effect.

Psychological research also suffers from the ecological fallacy. Like regions within a country, individuals in a sample can show different patterns of covariance that produce the same aggregate result (Eisenhauer, 2008). This is supported by the ubiquity of residuals across typical psychological models, such as linear regression. To elaborate, even highly accurate effects can tolerate discrepant and unaccounted for sub-level patterns of behaviour, which manifest as unexplained variation in the DV (James et al., 2013). These may have negligible impact on the magnitude or robustness of the overall trend, especially when (a) they are cancelled out by discrepant patterns of behaviour in the opposite direction (almost guaranteed when there is sufficient power, because of central limit theorem), and (b) sample size is large enough to offset any increases in model uncertainty that are introduced by especially large or frequent residuals. Overall, the ecological fallacy suggests that research to-date *only* speaks to the group-level effectiveness of using logged online behaviour to predict personality.

3.2.5. Shifting to Inaccuracy

The ecological fallacy may also be *exacerbated* by relative and not absolute effect estimates. Researchers use Pearson's R correlations and R^2 to conclude that are non-trivial relationships—e.g. between internet behaviour and personality—and thus evidence the phenomenon exists in the population. However, precise definitions of non-triviality differ. In social psychology, common heuristics for small ($r = .1$; $R^2 = 1\%$), medium ($r = .3$; $R^2 = 9\%$) and large ($r = .5$; $R^2 = 25\%$) effects allow for at least 75% of covariation between the predictors and the outcome to remain unexplained (Cohen et al., 2013). These benchmarks are sensible when studying *population* trends: most social phenomena have multiple underlying factors working together to shape behaviour, and thus no one factor can account for a substantial proportion of the total variance in the outcome. Social psychology effects may dilute whatever remaining correspondence there is between the aggregate effect and its sub-level effects—in objectively high-accuracy models from other fields—because they often fail to explain the absolute majority of variation in the DV. Our most common interpretation heuristics obfuscate this fact.

There is an alternative way to interpret predicted scores. Any single person's predicted personality may be best seen to exist in a normal *distribution of feasible personalities* (this idea

was first introduced by Fleiss, 1971). To clarify, any predicted score that uses a partial sample of online behaviour is a potentially-biased point estimate. Researchers may only know that the actual score lies within a wider range of plausible values surrounding the estimate (i.e. a confidence interval). I illustrate with 1000 hypothetical people who have all been predicted to be a 7 out of 10 on a measure of extraversion. Central limit theorem suggests a histogram of their true extraversion scores would be normally distributed with a mean of seven, and with a standard deviation that is inversely proportional to the prediction accuracy (James et al., 2013). That is, as accuracy increases the standard deviation decreases and there is greater convergence between true and predicted scores. In the absence of any additional information, I must assume every individual true score lies somewhere in the distribution, but not necessarily on seven. To date, researchers have not evaluated the extent these distributions yield predicted scores that on average differ, in qualitatively meaningful ways, from individuals' true scores. My intuition is that prediction errors for even 'large' social psychology effect sizes ($r > .5$; unexplained variation < 75%) are substantial. If true, I may have only marginal confidence in predicted scores for any specific individual.

These limitations may also apply to the future hypothetical limits of big data predictions. Indeed, increasingly comprehensive data on specific individuals has diminishing returns. When using Facebook to predict personality, Kosinski et al. (2013) found accuracies plateaued at around 300 Likes. Of course, combining multiple sources of big data might prolong the onset of diminishing returns. However, this is predicated on every new source of data (a) being at least partially distinct from all other existing sources, and (b) creating models that generalize to an entire population of people and do not simply reflect the idiosyncratic usage patterns of the sub-group sampled (i.e. the problem of overfitting; James et al., 2013). Even then, any degree of non-distinctiveness will still result in at least some diminishing returns. For example, Skowron et al. (2016) explained 32% of variation in personality predictions using Twitter data, but only an additional 14% using Twitter and Instagram data combined—despite both sources containing wide-ranging and comprehensive user behaviour logs. Thus, even increasingly comprehensive and heterogeneous data might reach an imperfect ceiling level of accuracy.

These problems are not negated by simply using more participants. All else being equal, large sample size increases the certainty that even negligibly explanatory variable effects are non-spurious (i.e. p-values are contingent on sample size; Loken & Gelman, 2017). However, increased sample size has no bearing on *effect size* because predictors still explain the same amount of total variation in the outcome (Fraley & Vazire, 2014). Thus, high powered studies

might only help researchers evaluate subtle phenomena or the same phenomenon in multiple different sub-populations. This constraint also applies to cutting-edge machine learning. The most complex approaches—such as elastic nets, random forests and deep neural networks—certainly do benefit from more participants. However, that is only because it improves the participant-to-predictor ratio, which allows models to capture more complex non-linear relationships (James et al., 2013). They—like all forms of machine learning—suffer equally from the problems of diminishing returns and overfitting because they also depend on the richness and generalizability of the original data.

3.2.6. Measuring Accuracy for Specific Individuals

The accuracy of predicted personality scores may be overstated because it is typically measured using correlations (r) and variance explained (R^2). R famously does not have a clear interpretation beyond merely suggesting negative or positive covariance (Bosco, Aguinis, Singh, Field & Pierce, 2015). Classically, Ozer (1985) argued that R^2 differs depending on whether shared covariation is attributed to just the nominated predictor or outcome, whether it accounts for overlapping variance from multiple predictors, and whether it is calculated as a percentage of total outcome (or predictor) variation versus partial unexplained variation. Put simply, although highly useful because of their sometimes-standardized nature, these metrics still lack unambiguous natural units that people can reference to daily experiences. Thus, researchers, practitioners and the general public may be left to accept that a 'large effect' (r = .5) is indeed objectively large, and that it signifies a highly accurate prediction. Fortunately, there are other methods for measuring accuracy at the individual level, which are more intuitive and lend themselves to idiographic interpretations. I use three in this chapter:

3.2.6.1. Mean Absolute Error

Mean absolute error (MAE) is one of the simplest possible measures of accuracy. It captures the average difference between predicted and actual scores for each participant in the original units of the measurement scale. For example, if I predicted Joe is 27 years old when he is 30, then the error is 3 years. Then, if I predict Sally is 36 years old when she is 31, the error is 5 years. In this case, MAE for the sample (Joe and Sally) is 4 years.

3.2.6.2. Classification Accuracy by Category Assignment

Alternatively, I can assign true and predicted scores to categories and evaluate how often they match. The simplest approach is to create 'low' and 'high' categories via a median split.

However, three categories may be minimally viable because they mitigate the limitations of imposing a false dichotomy on the data by also accounting for relatively neutral cases (Maxwell & Delaney, 1993). Thus, I might assign each person into low, medium, and high (e.g.) extraversion categories by dividing the continuous range of possible extraversion scores into thirds. The three minimally viable categories increase the likelihood of accurate predictions—compared to using additional categories—because (a) prediction errors are only possible in one direction for two of the three categories, and (b) there are no within-category prediction errors. Then, accuracy is simply how often the true and predicted categories match. Given even categories, pure chance is 50% for median splits, 33% for three categories, 25% for four categories, and so forth.

3.2.6.3. Correctly Ranking Pairs of Cases

Finally, I can evaluate the extent predicted scores correctly rank randomly drawn pairs of participants. This is an analogue to area under the curve (AUC), which evaluates prediction accuracy in experiments with two conditions—control and treatment—and is also commonly used in other social media prediction studies (e.g. Wang & Kosinski, 2017). AUC is the percentage of participants—one randomly sampled from the control, the other from the treatment—who are then correctly classified back into their respective conditions based on their predicted scores (Hanley & McNeil, 1982). The interpretation changes slightly when using continuous variables: large discrepancies between true and predicted scores decrease the probability of obtaining matching rank orders. This may mean correct rankings are only marginally above 50/50 chance, at least until prediction accuracies are very high.

3.2.7. Present Studies

My primary aim in this chapter was to demonstrate the accuracy of predicted psychological characteristic scores for *specific individuals* using algorithms that are statistically robust, and highly accurate, at the group level. I did this through the prism of personality (1) because of its topicality in the public domain, (2) to converge with the exiting literature on online psychological characteristic predictions, and (3) because I was directly interested in the privacy implications of using predicted personality throughout the remainder of my PhD. Nevertheless, results can be translated so that they apply equally to any normally distributed variable, via a simple linear transformation.

The three approaches—MAE, classification accuracy, and correctly ranking pairs of cases—yielded at least seven unique demonstrations: the extent (1) aggregate-level correlations

translate into individual prediction errors; (2) predicted scores for individuals with different true scores overlap; (3) predicted score MAEs for people with middling true scores are more accurate than those for people with extreme true scores; (4) prediction errors increase when they are corrected to have realistic and not artificially condensed model SDs; (5) outlier predictions are *classified* more accurately than middling predictions; (6) individuals are correctly classified across multiple characteristics; and, (7) pairs of cases are correctly ranked based on their predicted scores.

To achieve these aims, I used combinations of simulated and real-world data. Monte carlo simulations—which are predicted on randomly generated and normally distributed numbers—helped to evaluate even extreme best-case prediction accuracies that are largely beyond the scope of current technologies. They meant I could focus on three prediction accuracy benchmarks—'Best-case' (r = .90), 'Demographic' (r = .60) and 'Personality' (r = .30)—that were, to varying extents, grounded in combinations of real-world and simulated data. They were deliberately larger than typical social psychological benchmarks (e.g. r = .50, .30, .10; Cohen et al., 2013) to offer a more rigorous test of the overarching question. Best-case is a hypothetical upper-bound of future prediction technologies, of the kind that integrates data across platforms and uses advanced machine learning. Whilst difficult to accurately project, I settled on this imperfect upper bound because of diminishing returns from new forms of data and the likelihood that at least a small portion human behaviour (i.e. ≈ 19%) will never be logged. Further, it also meant that Pearson's R differences were kept constant between benchmarks. Demographic is the prediction accuracy researchers might expect using a single, reliable, source of online data to predict sociodemographic characteristics such as age, education level and political orientation (Youyou et al., 2015). It might be considered the plausible upper limit of current technologies. Finally, Personality is the prediction accuracy researchers might expect using a similar source data to predict any of the big five personality traits (e.g. Kosinski et al., 2013). It also coincides with a 'medium' effect in social psychology.

As I mentioned above, findings apply equally any normally distributed trait. As such, we interpret omnibus psychological characteristics through the prism of personality. It is possible to use this same logic to make results *even more intuitive*. Personality traits are often scored on unintuitive ordinal scales (e.g. from one to five). Ordinal scales are those where a value of zero only has meaning relative other scores in the sample. For example, I can only conclude that someone scoring 3/5 on extraversion has more of the trait than someone scoring 2/5, but not that they are neutral or extraverted *per se*. It is difficult to understand how these individual

scores translate into different real-world behaviours. Ordinal scales can be contrasted with ratio scales, where zero has an intrinsic meaning. For example, something with a height of zero is entirely flat and a person aged zero is a new-born. They are far easier to interpret. Thus, I linearly transformed (i.e. multiplied and added a constant) ordinal true and predicted personality so that scores could be interpreted as ratio age. I settled on age because it is perhaps the most consistently understood and interpreted ratio scale across cultures. I stress that results could be *transformed back* to personality, or indeed any other continuous psychological characteristic, at any time. Thus, I fully preserved the integrity of the original scores whilst also making them more digestible to a wide range of researchers and practitioners.

3.3. Study 1

In Study 1, I applied the seven approaches to fully simulated true and predicted personality. Thus, I evaluated effects independent of idiosyncratic variable distributions and unique statistical/machine learning models. It meant that errors were *completely random*—which is a precondition of all *unbiased* prediction algorithms—and that true and predicted scores were both normally distributed. I thus simulated predictions that met general linear model (and other model) assumptions for extremely high robustness.

3.3.1. Method

3.3.1.1. Simulations

There were 10,000 simulated cases with normally-distributed personality (M = 3; SD = 1), which I then linearly transformed into age (M = 35; SD = 18). Age converged with what researchers might obtain from a typical community sample of US residents (e.g. in the American Community Survey). I called these values ‘true’ because they were designed to proxy for original self-report scores. I truncated true ages outside 0-80 to fit this range.

3.3.1.2. Procedure

Predicted scores were derived from true scores. Specifically, I generated predicted scores by iteratively adding random noise sampled from a normal distribution (M = 0; SD = 0.05 X SD of true age) to true age to down-regulate their convergence, until I reached their target correlation. Then, I corrected predictions so that they had the same mean, minimum and maximum as trues scores. This code to generate predicted scores is in Appendix 3.1. Importantly, as normative statistical model accuracy decreases, it relies more on the predictive

power of the intercept term—which assumes everyone has the same baseline score—over the coefficients (James et al., 2013). As such, I corrected the predicted score SDs, so that they were proportional to the target correlation. For example, the target r = .60 yielded predicted scores with 60% of the true age SD. Finally, I repeated this entire simulation protocol ten times to increase the stability of my estimates. Then, I created MAE by taking the average of the absolute difference between true and predicted scores for each case. MAE is a more liberal estimate of model accuracy than either RMSE or R-squared—other popular contextualized metrics of inaccuracy—giving the ecological fallacy, and thus my intuitions, the maximum chance of failing.

3.3.2. Results

I demonstrated all proofs statistically using prediction accuracies from r = .01 to r = .99, increasing in increments of r = .01 (i.e. N = 99). Then, I focussed on the three prediction accuracy benchmarks—best-case (r = .90), demographic (r = .60), and personality (r = .30)—where I concretely interpreted results in terms of analogous age. For every accuracy, I aggregated MAE for each of the ten monte carlo simulations. Despite large statistical power, I still used 95% CIs—which I defined as 2.5^{th} and 97.5^{th} percentile absolute errors—because they offered narrower distributions of prediction errors than more conservative CIs, thus again increasing the chances that my intuitions were wrong.

3.3.2.1. MAE by Prediction Accuracy

Approach one evaluated whether predicted scores were too imprecise to make inferences about specific individuals, even at high training accuracies. There was a negative quadratic (inverted u-shaped) relationship between prediction accuracy and MAE (B = -10.78, CI = (-11.51, -10.06), t(96) = -29.61, p < .001). Results are in Figure 3.1. Prediction errors only decreased rapidly for very extreme accuracies. However, even at the best-case benchmark they may have already been too big to be useful (MAE = 6.15; CI = 0.24, 17.14). They on average mistook a 17-year-old for an 11-year-old middle schooler or a 23-year-old. At the demographic benchmark, MAE = 11.28 (CI = 0.45, 30.91). They on average mistook the 17-year-old for a 6-year-old infant or a 28-year-old. At the personality benchmark, MAE = 13.50 (CI = 0.54, 36.41). They on average mistook the 17-year-old for a 3-year-old toddler or a 31-year-old. At the personality benchmark, MAE was negligibly better than simply assuming everyone was average (MAE = 14.23; CI = 0.51, 35.13). Thus, in my view, such predictions were too

imprecise to make inferences about specific individuals at every benchmark, and negligibly better than assuming everyone was average at the personality benchmark.

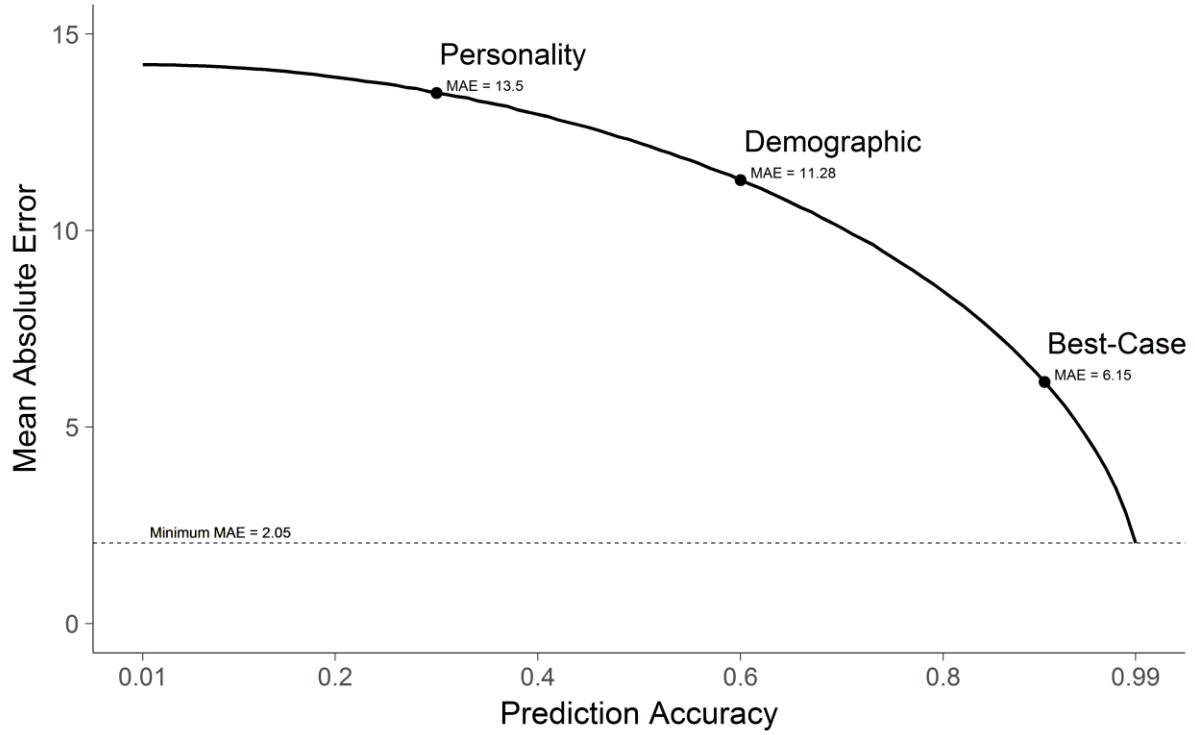


Figure 3.1. MAE as a function of training accuracy for psychological characteristics. Prediction accuracy is the Pearson's R correlation between simulated true and predicted scores. True scores were simulated normally distributed variables that proxied for self-reported personality but reported in terms of age. I generated predicted scores by incrementally adding random noise to true scores until I reached the target correlation (R = .01 to .99, by R = .01). Mean absolute error (MAE) was the mean of the absolute difference between true and predicted scores for each case. 'Personality', 'Demographic' and 'Best Case' reflect R = .30, R = .60 and R = .90 true-predicted score correlation benchmarks, respectively. The dashed line is MAE when R = .99.

3.3.2.2. Overlapping Predictions for Divergent True Scores

Approach two evaluated the percentage of predicted scores that were shared between extreme cases. To this end, I retained the bottom and top 20% of true ages. Descriptive statistics are in Table 3.1. Then, I evaluated the extent that their predicted score distributions overlapped when accuracy increased from r = .01 to r = .99. Training accuracy positively moderated the mean difference between bottom vs top 20% predicted scores (B = 49.71, CI = (49.56, 49.86), t(395996) = 669.72, p < .001). Put another way, both groups' mean age converged on the full sample mean as training accuracy decreased. While group differences may have remained significant for each benchmark, the average difference in true ages (49.58 years; SD = 9.91) was only 40.35 years (SD = 11.38) for best-case, 17.79 (SD = 10.91) years for demographic,

and 4.43 (SD = 6.92) years for personality benchmarks. At the personality benchmark, I on average predicted 60-year-olds were only 5 years older than 10-year-olds.

Table 3.1.

Mean predicted age for the bottom 20% and top 20% true ages

Prediction accuracy	M (SD)		M-DIFF (SD)
	Bottom 20%	Top 20%	
Best-case (r = .90)	15.34 (7.07)	55.7 (8.97)	40.35 (11.38)
Demographic (r = .60)	26.66 (6.92)	44.46 (8.44)	17.79 (10.91)
Personality (r = .30)	32.98 (4.63)	37.41 (5.1)	4.43 (6.92)
True	10.48 (6.34)	60.05 (7.77)	49.58 (9.91)

Notes. M-DIFF = Mean difference. Cases were allocated to bottom and top 20% based on their true age.

Then, I evaluated the overlap between predicted score distributions from the bottom and top 20% cases. They are in Figure 3.2. Predicted score overlap was 2% for best-case, 28% for demographic, and 63% for personality benchmarks. At the personality benchmark, there was a 2/3 chance that any given 10-year-old's predicted score was drawn from the predicted-score distribution of a 60-year-old, and vice versa. Therefore, I found that even when predicted scores differentiated between extreme cases, they (a) underestimated the true magnitude of the differences, and (b) yielded predictions that could feasibly belong to either very young or very old cases, especially at the more realistic prediction benchmarks.

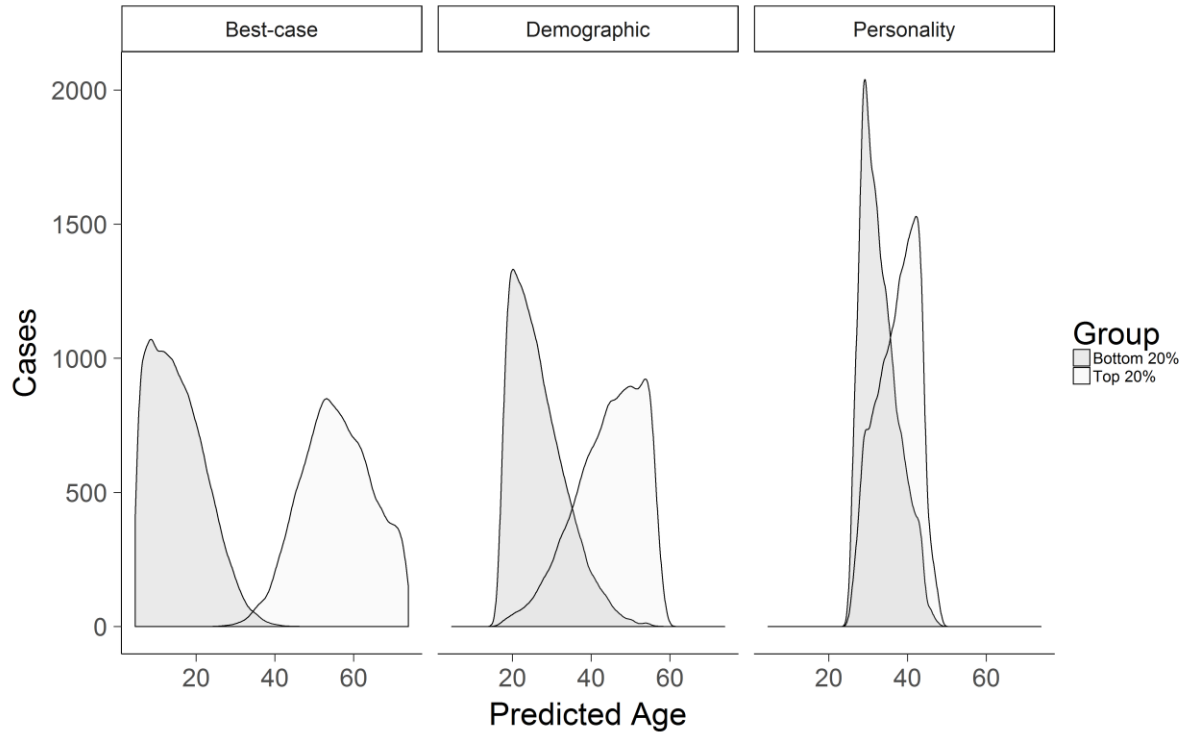


Figure 3.2. Predicted score distributions for participants with bottom 20% and top 20% age. Age in the graph can proxy for any normally distributed variable (via a linear transformation); in this case it proxies for personality. Predicted age was generated from simulated true age by adding random error. There were three true-predicted score correlation benchmarks: Best-Case (R = .90), Demographic (R = .60) and Personality (R = .30). The shaded regions reflect the number of cases for each value of predicted age (i.e. the density), at each correlation benchmark. The darker shading reflects cases with bottom 20% true ages, and lighter shading reflects cases with top 20% true ages. Where shading overlaps, there is shared predicted score density between bottom and top 20% true age. This shared density was 2% for Best-Case, 28% for Demographic and 63% for personality.

3.3.2.3. MAE at Different Values of Original Scores

Approach three explored how prediction accuracy changed according to the magnitude of true scores. I evaluated MAE by true score decile for r = .01 to r = .99. MAE for the 1st, 5th, 6th and 10th deciles are in Table 3.2. The quadratic effect for decile on MAE became less positive as accuracy increased (B = -365.43, CI = (-376.01, -354.85), t(984) = -67.79, p < .001). Results are in Figure 3.3. That is, at low accuracies, errors were disproportionately large for the extreme-most deciles (i.e. 1st and 10th, then 2nd and 9th, etc.). Then as accuracy improved, these extreme deciles also had the largest decreases in MAE. There were also small accompanying *fluctuations* in MAE for the middle deciles, as prediction accuracy increased. This may have been because prediction models with lower accuracies relied more on the assumption that everyone was average (i.e. they relied on the intercept term). Thus, they were disproportionately accurate for those cases with true ages closest to the average, and disproportionately inaccurate for cases with other true ages.

Table 3.2.

Mean predicted age for the bottom 20% and top 20% of true ages at different prediction accuracies

Decile	1st	5th	6th	10th
Personality (r = .30)	27.73	4.52	4.78	28.31
Demographic (r = .60)	20.64	7.47	7.83	19.78
Best-case (r = .90)	7.56	5.68	5.65	7.21
Assuming average	29.93	2.51	1.94	31.18

Note. Decile = True age rank.

For the 1st and 10th deciles, MAE = 7.38 years (CI = 0.31, 19.12) for best-case, MAE = 20.21 years (CI= 6.50, 37.20) for demographic, and MAE = 28.02 years (CI = 16,33, 42.94) for personality benchmarks. For the personality benchmark, this was the equivalent of mistaking a 66-year-old for a 38-year-old or a 94-year-old. Predictions were more precise for 5th and 6th deciles across all benchmarks, fluctuating between MAE = 5.66 (CI = 0.23, 15.90) for best-case, MAE = 7.65 (CI = 0.33, 19.33) for demographic and MAE = 4.65 (CI = 0.22, 11.60) for personality. However, in every instance, it was at least twice as accurate to simply assume *all* cases in the 5th and 6th deciles were the average (MAE = 2.22, CI = 0.20, 4.42). Thus, predictions for extreme cases were impractically large at every benchmark, while for middling cases it was always far more accurate to simply rely on the mean.

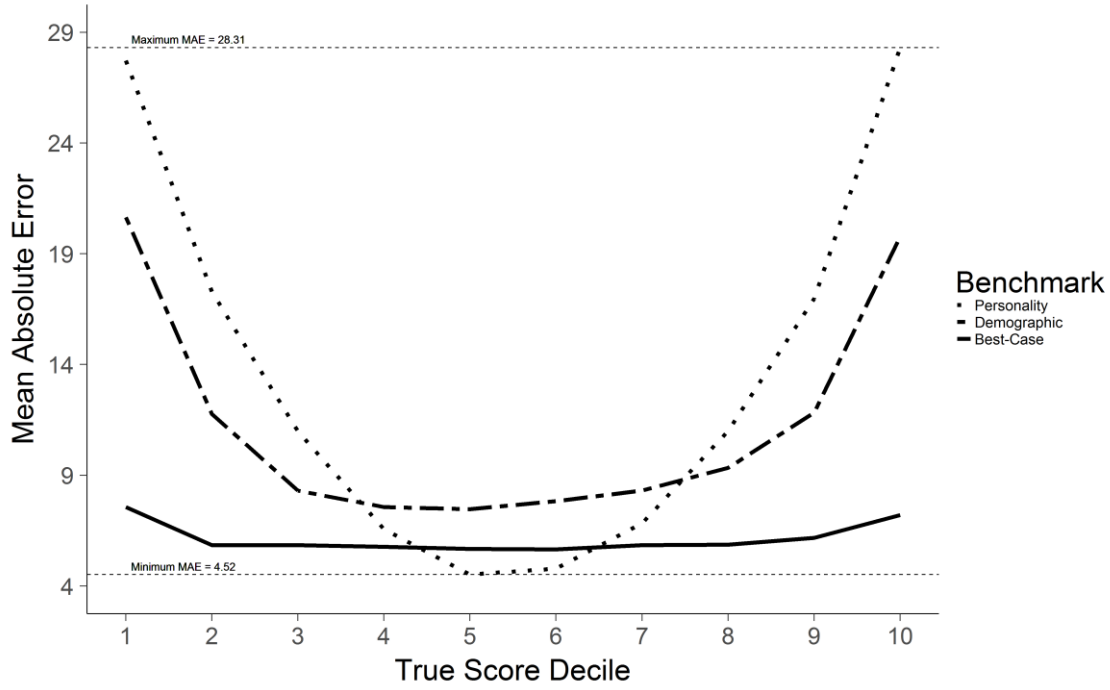


Figure 3.3. Predicted score MAE by true age decile at different training accuracies. True score decile is the rank-order magnitude of the true scores. True scores proxy for self-report survey responses. Mean absolute error (MAE) is the mean of the absolute difference between true and predicted scores for each case. MAE is reported here in terms of age. It can be linearly transformed to reflect any continuous variable. 'Personality', 'Demographic' and 'Best Case' benchmark lines reflect MAE by decile at R = .30, R = .60 and R = .90 true-predicted score correlations respectively.

3.3.2.4. Adjusting for Realistic SDs

Approach four evaluated how MAE changed as I adjusted the SDs for predicted scores to realistic magnitudes. Although there were 18% children (< 18 years) in true age, predicted score SD shrinkage meant there were 16% children at best-case, only 2% at demographic and < 1% at personality benchmarks. In real world use cases, such shrinkage would not be practicable. Thus, I corrected for SD shrinkage by artificially spreading out predicted scores until they had the same SD as the true scores, for prediction accuracies from r = .01 to r = .99. There was a negative quadratic association between prediction accuracy and MAE (B = -7.85, CI = (-8.56, -7.15), t(96) = -22.02, p < .001). Results are in Figure 3.4. While SD-corrected prediction errors for best-case were relatively stable (MAE = 6.29; CI = 0.25, 17.63), they increased rapidly for demographic (MAE = 12.58; CI = 0.49, 35.07) and personality (MAE = 16.66, CI = 0.66, 46.41) benchmarks. In fact, MAE for personality was *larger* than MAE for simply assuming everyone was average (MAE = 14.23). It mistook a 17-year-old for a new-born or a 34-year-old. The minimum prediction accuracy where a full SD correction did not

push MAE beyond this threshold was r = .49. This threshold is above the vast proportion of documented personality predictions to date.

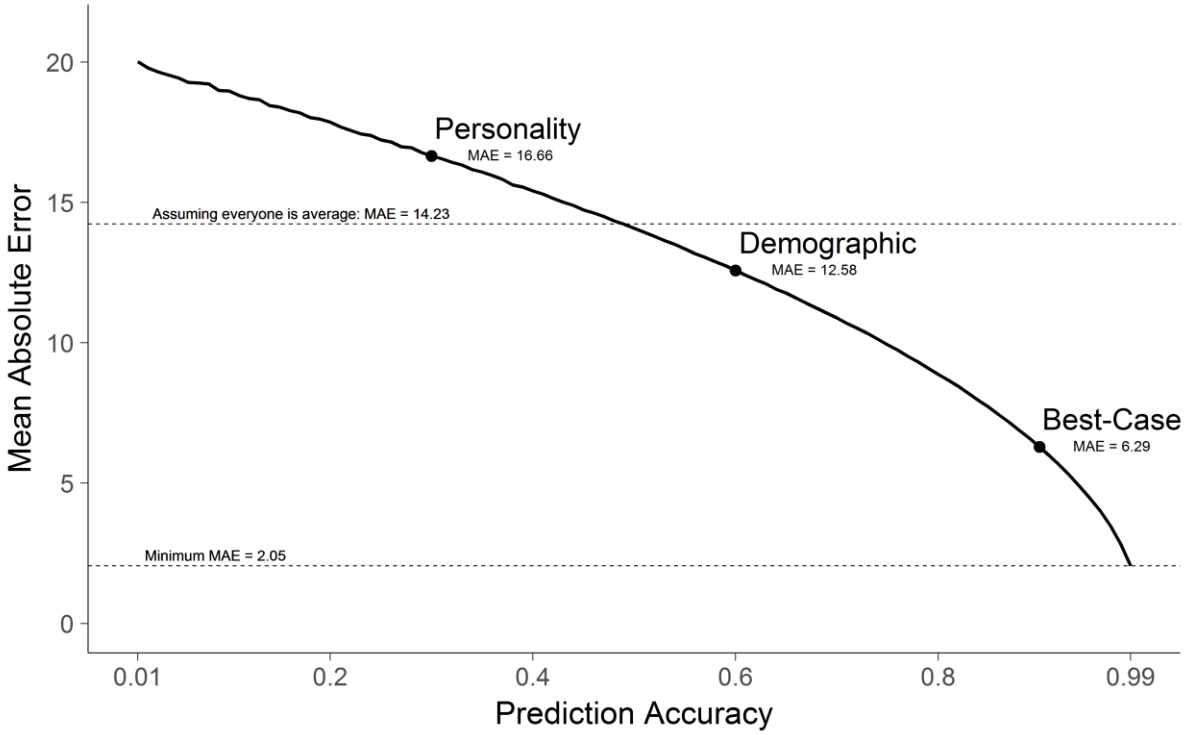


Figure 3.4. MAE as a function of training accuracy for psychological characteristics, when predicted score SDs are corrected to match true score SDs. Correction is necessary because predicted score SDs are proportional to training accuracy. This reflects the increasing importance of the constant intercept term at low accuracies. Prediction accuracy is the Pearson's R correlation between simulated true and predicted scores. True scores were simulated normally distributed variables that proxied for self-reported personality but reported in terms of age. I generated predicted scores by incrementally adding random noise to true scores until I reached the target correlation (R = .01 to .99, by R = .01). Mean absolute error (MAE) was the mean of absolute difference between true and predicted scores for each case. 'Personality', 'Demographic' and 'Best Case' reflect R = .30, R = .60 and R = .90 true-predicted score correlation benchmarks respectively. The dashed line in MAE is when R = .99.

Inflating SDs exposed the tendency for predicted scores to cluster randomly around the sample mean. This was reflected in large discrepancies in the rank order of participants' predicted scores. Ranking error for the 10,000 cases was MAE = 989.61 places (CI = 29.49, 3132.76) for best-case, MAE = 2014.57 places (CI = 65.00, 5915.07) for demographic and MAE = 2723.15 (CI = 95.09, 7588.49) places for personality benchmarks. For comparison, across 100 iterations of randomly assigning 'predicted' values of true age, chance ranking error was MAE = 3333.82 (CI = 125.56, 8419.65). Therefore, as prediction accuracy decreased, SD corrections may have inflated MAE because the rank order of participants increasingly converged with completely random predicted scores.

3.3.2.5. Classification Accuracy by Category Assignment

Approach five was whether I classified extreme predicted scores into low, medium and high categories more accurately than I classified middling predicted scores. As a preliminary step, I evaluated classification accuracy for all 10,000 cases. Correct classification for best-case (75%), demographic (55%) and personality (43%) benchmarks moved progressively towards baseline chance (33%). Then, I also used confusion matrices to evaluate whether there were any biases in category assignment (e.g. middle third true scores were disproportionately misclassified into the bottom compared to top third). Results are in Appendix 3.2. Overall, results suggested that classifications were equally accurate—across all four performance metrics that are commonly calculated from the confusion matrix—for bottom and top third true scores at each benchmark. Classification accuracy was consistently lower for middle third true scores, because errors could occur in two directions. Overall, there was no evidence of classification bias, meaning I could evaluate classification accuracy for bottom and top percentile cases together in the same models.

Next, I evaluated whether classification accuracies varied as predicted score values changed. To this end, every case was assigned a rating based on the extremeness—or deviation from the mean—of their predicted scores (1 = "least extreme", 100 = "most extreme"). Thus, the least extreme cases had 49.50^{th} to 50.50^{th} percentile predicted scores, and the most extreme cases had the bottom and top 0.50^{th} percentile predicted scores. Then, I evaluated classification accuracy—which was the percentage of time true scores were classified into the same third as predicted scores—as a function of extremeness. I focused on the cases where there was a more than 50/50 chance of obtaining correct classifications.

I evaluated classification accuracy as a function of predicted score extremeness for training accuracies ranging from r = .01 to r = .99. The quadratic effect for extremeness on correct classification was positively moderated by training accuracy (b = 8.79; CI = 8.39, 9.19; t(9894) = 42.81; $p < .001$). Results are in Figure 3.5. Put more simply, as training accuracy progressed from extreme hypothetical to more realistic, only increasingly extreme cases were classified at above 50/50 chance. To illustrate, best-case classification accuracy was mostly above 50% across extremeness scores. However, it was still only 45% at the threshold of two different predicted score categories (i.e. extremeness = 34). That is, best-case accuracy failed to reliably differentiate between the edge cases from two different categories. For the demographic benchmark, just under half the cases—those with the most middling predicted scores

(extremeness < 45)—were *incorrectly* classified more than 50% of the time. For the personality benchmark, more than the 2/3 of cases with the least extreme predicted scores (extremeness < 68) were *incorrectly* classified over 50% of the time. Therefore, at realistic accuracies only *extreme*—or increasingly outlier—predicted scores could be classified into the correct age third at more than 50/50 chance. For the personality benchmark, this was the equivalent of only classifying participants aged below 16.50 and above 52.82 correctly more than 50% of the time. Thus, attempts to mitigate MAE by classifying participants into one of three age buckets were ultimately ineffective for a large proportion of cases with middling predicted scores.

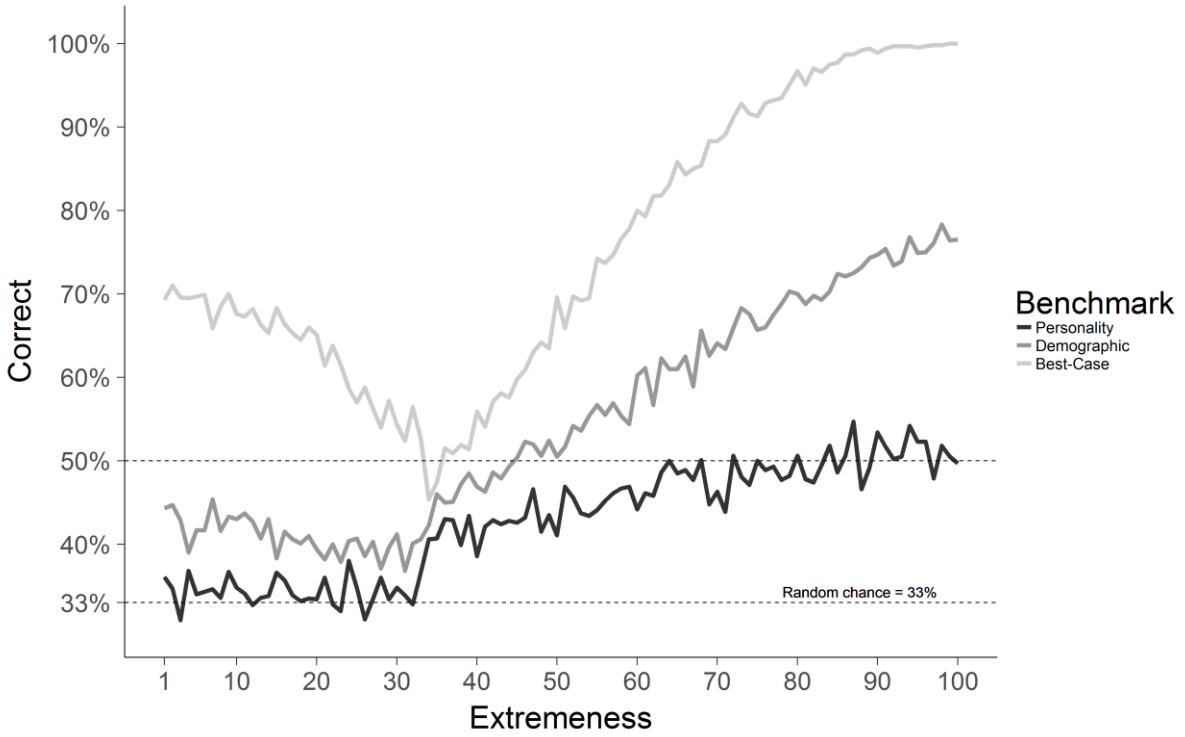


Figure 3.5. Percentage of cases correctly classified by the extremeness of their predicted scores. Extremeness was calculated by percentile. The 1st percentile were the 1% absolute predicted scores closest to the mean. The 100th percentile were the 1% absolute predicted scores furthest from the mean. 'Correct' is the percentage of cases correctly bucketed into equal-sized 'low', 'medium' and 'high' thirds for cases in each percentile. 'Personality', 'Demographic' and 'Best Case' benchmark lines reflect classification accuracy as a function of extremeness at R = .30, R = .60 and R = .90 true-predicted score correlations respectively. The dashed line reflects the 33% random chance of bucketing each case correctly.

However, even classification accuracy for cases with extreme scores—putatively the *easiest* to put into buckets because classification errors could only occur in one direction—had ceiling effects that converged with 50/50 chance as they progressed to realistic training accuracies. For extremeness > 80—when cases had predicted scores that were in the bottom and top 10%—

classification accuracy was 99% for best-case and 74% for demographic benchmarks, but *still only* 51% for the personality benchmark. These classification ceilings further diminished when I increased the number of buckets to create more homogenous—and thus practicable—groups of predicted ages. When there were four age buckets, correct classification for extremeness > 80 was 96% for the best-case benchmark, but only 62% for demographic and 40% for personality benchmarks. With five buckets—which is the minimum number needed to create a separate category comprising mostly children (> 18)—correct classification was 92% for the best-case benchmark, but only 54% for demographic and 33% for personality benchmarks. Therefore, progression from best-case to realistic training accuracies meant I increasingly misclassified even the most extreme cases. Indeed, classification accuracy was so poor for the personality benchmark—despite the relative ease of classifying these extreme scores—that even it was still only around 50% when using three categories. Classification accuracy also got progressively worse—falling below 50/50 chance—when attempting to put cases into four or five buckets at realistic training accuracies.

3.3.2.6. Multiple Classifications

Approach six investigated the extent errors were compounded when I attempted to classify the *entire* big five into low, medium, and high buckets. I simulated predicted scores from r = .01 to r = .99 for five different *orthogonal* sets of age—which each proxied for a big five factor. Correct classification for *at least* 1/5 factors was > 99% for best-case, 98% for demographic and 94% for personality. The quadratic association between training accuracy and percentage correct became more positive as the number of correct classifications increased (b = .54; CI = .41, .67; t(364) = 8.39; p < .001). Results are in Figure 3.6. Put another way, as prediction accuracy decreased, classification accuracy for multiple traits also decreased. Across the entire big five, there was > 50% chance of correctly classifying 4/5 factors at best-case, 3/5 factors at demographic and 2/5 factors at personality benchmarks. Thus, at the best-case benchmark I generated mostly but not fully correct personality profiles at above 50/50 chance. As I progressed to the realistic benchmarks, it was increasingly likely to categorize more than half the factors incorrectly, rather than correctly.

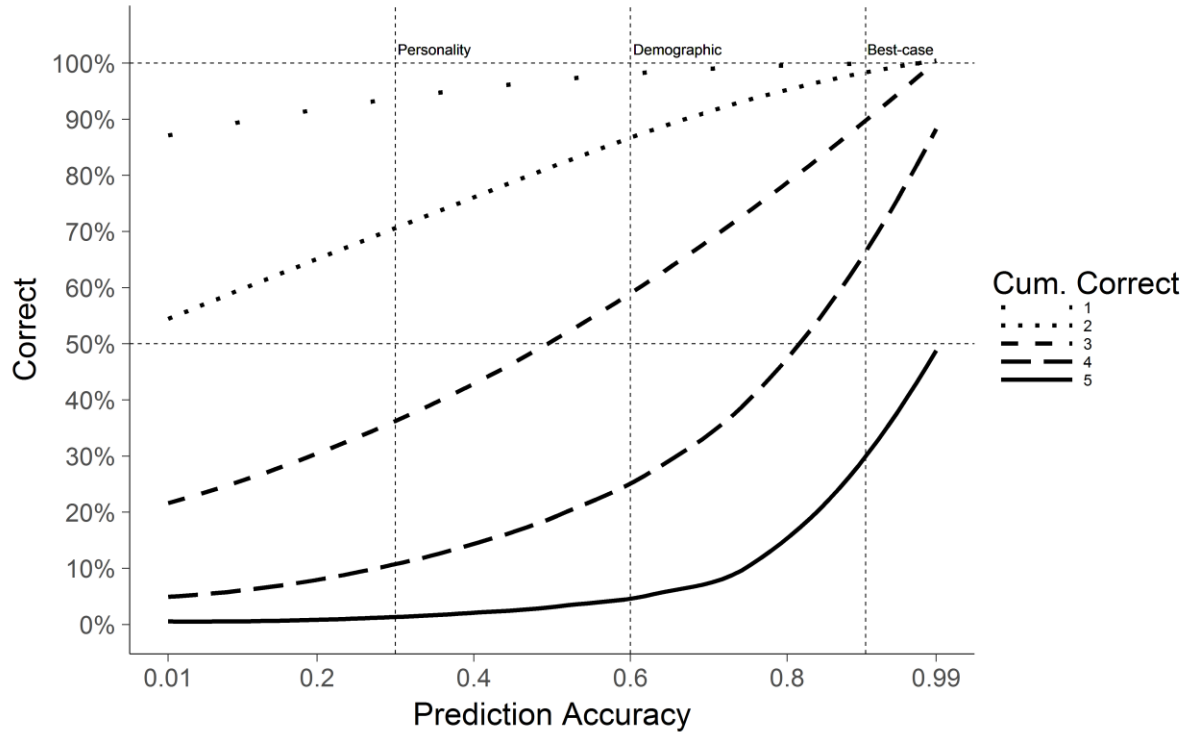


Figure 3.6. Cumulative percentage of correctly classifying one to all five big five personality traits. Prediction accuracy is the Pearson's R correlation between simulated true and predicted scores, which I generated by incrementally adding random noise to true scores until I reached the target correlation (R = .01 to .99, by R = .01). The dashed vertical lines reflect 'Personality' (r = .30), 'Demographic' (r = .60) and 'Best Case' (r = .90) benchmark correlations. Correct is the percentage of cases correctly classified into their true score bucket at each prediction accuracy. Cum. Correct lines are the cumulative percentage correct when bucketing 1/5, 2/5, 3/5, 4/5 and 5/5 psychological characteristics. The horizontal dashed line is 50/50 chance of classifying the target number of characteristics correctly.

Then, I evaluated the extent simulations classified big five traits 100% correctly compared to 100% incorrectly. For best-case accuracy, 24% of cases were totally correct and < 1% were totally incorrect. Despite this large ratio, approximately 3/4 of cases were wrongly classified on at least one trait. At demographic accuracy, 5% of cases were totally correct, and 2% of cases were totally incorrect. At personality accuracy, 1% of cases were totally correct, and 6% of cases were totally incorrect. Thus, as simulations progressed to the personality benchmark, it became more than four times as parsimonious to assume classifications were totally incorrect rather than totally correct. The minimum accuracy where predictions were more likely to be totally correct than totally incorrect was r = .51—which is beyond the limits of most current social media personality predictions. Finally, it was unfeasible to create a *comprehensive* big five personality profile at greater than 50/50 chance for any training accuracy.

3.3.2.7. Correctly Ranking Participant Dyads

Finally, approach seven evaluated the extent predicted scores correctly ranked random pairs of participants. For every training accuracy from r = .01 to r = .99, I randomly split participants into halves 100 times. Thus, each random split created 5,000 pairs. The mean true age difference in pairings was M = 20.00 years (SD = 14.73). Then, I determined the extent that true score rank matched predicted score rank. There was a positive linear association between training accuracy and the percentage of pairs that were ranked correctly (b = .41; CI = .40, .42; t(97) = 72.67; p < .001). Results are in Figure 3.7. Correct ranking was 86% for best-case, 71% for demographic and 60% for personality benchmarks (random guessing would have resulted in 50% accuracy). Despite these putatively large percentages, correct rankings for demographic and personality benchmarks were still nearer to chance than being 100% correct. The first training accuracy where correct classification was > 75% was r = .70. Thus, even when I progressed from the best-case to the demographic benchmark—the top end of realistic prediction accuracies—it was *already* more parsimonious to assume predictions could not differentiate between participants at all above chance, than to assume it differentiated between them perfectly.

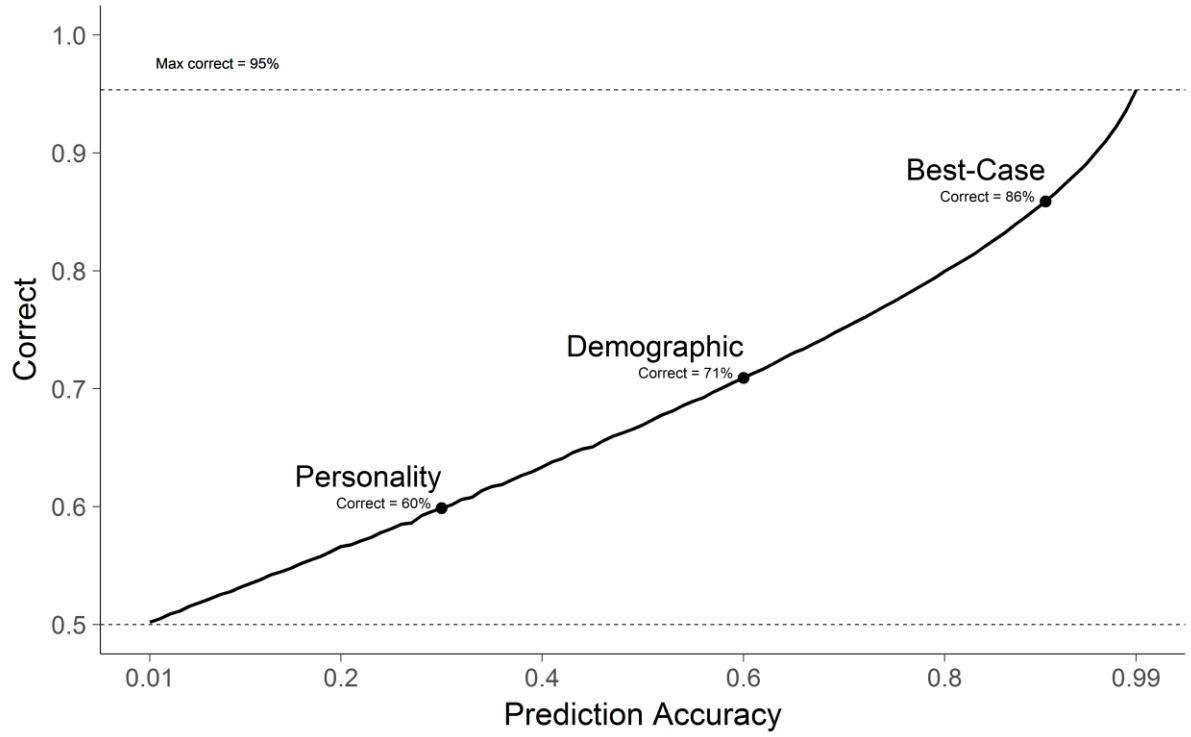


Figure 3.7. Percentage of random pairs of cases that were correctly ranked using their predicted scores. Prediction accuracy is the Pearson's R correlation between simulated true and predicted scores, from r = .01 to r = .99 and increasing in increments of r = .01. Correct is the proportion of pairs where true score order matched predicted score order. 'Personality', 'Demographic' and 'Best Case' reflect R = .30, R = .60 and R = .90 true-predicted score correlation benchmarks, respectively. The bottom dashed line is 50% correct, which is the accuracy obtained by random chance. The top dashed line is the proportion correct when the correlation between true and predicted scores is R = .99.

3.4. Study 2

Study 2 aimed to confirm whether simulated results also applied to real world self-reported age for more than three million participants. Specifically, I evaluated (a) whether results held with true scores that were not perfectly normally distributed, and (b) the extent very big data changed the conclusions from Study 1. Again, I created predicted scores by adding random noise to real existing true scores. As per Study 1, simulated errors were normally distributed and random. Thus, I again evaluated prediction errors from general linear models that met all assumptions for high robustness. I focused exclusively on simulated prediction accuracies for best-case (r = .90), demographic (r = .60) and personality (r = .30) benchmarks.

3.4.1. Method

3.4.1.1. Participants and Procedure

I used publicly available demographic data from 2014 American Community Survey (N = 3,132,610; United States Census Bureau, 2016). Mean age was 40.82 (SD = 23.55) and there

were 51% women. I repeated the simulation procedure from Study 1 (the simulation code is in Appendix 3.1). Thus, I simulated predicted scores by iteratively adding random noise sampled from a normal distribution (M = 0; SD = 0.05 X SD of age) until I reached the benchmark accuracies. I again truncated predicted ages outside the true age range (0-96), equalised true and predicted score means and adjusted predicted score standard deviations to have realistic shrinkage. I also repeated each simulation 10 times to increase effect stability. Throughout the results, I again used aggregate means and 2.5 and 97.5 percentile CIs. Table 3.3 compares key results across studies.

Table 3.3.

Key results for all seven approaches across the three studies

		S1 (M= 35, SD = 18)			S2 (M = 41, SD = 24)			S3 (M= 35, SD = 18)		
Benchmark (r =)		**.90**	**.60**	**.30**	**.90**	**.60**	**.30**	**.90**	**.60**	**.30**
MAE	**A1**	6.15	11.28	13.50	8.21	15.39	18.94	6.09	11.30	13.56
Overlap extreme quintiles	**A2**	2%	28%	63%	1%	28%	64%	2%	28%	62%
MAE Extreme deciles	**A3**	7.38	20.21	28.02	9.26	25.19	34.77	8.79	20.81	27.97
SD Corrected MAE	**A4**	6.29	12.58	16.66	8.39	16.69	22.24	5.96	12.11	16.10
Buckets – 1 trait	**A5**	75%	55%	43%	77%	55%	43%	76%	55%	42%
Buckets – 5 traits	**A6**	24%	5%	2%	27%	5%	1%	24%	4%	1%
AUC	**A7**	86%	71%	60%	85%	70%	59%	86%	71%	60%

Notes. A1-7 = Approaches 1-7. S1-3 = Studies 1-3 (M = Mean age in sample, SD = Age SD in sample). S1 = Full simulations. S2 = Mixed real-world data and simulations. S3 = Real world online data and machine learning. MAE = Mean absolute error. Buckets = Correct classification into low, medium and high categories. AUC = Area under curve. Only point estimate descriptives provided for expedience.

3.4.2. Results

Approach 1 evaluated whether predicted scores were too imprecise to make inferences about specific individuals, even at high training accuracies. Best-case MAE = 8.21 (CI = 0.32, 23.11), demographic MAE = 15.39 (CI = 0.65, 40.36), and personality MAE = 18.94 (CI = 0.91, 44.44). For the personality benchmark, predictions were again negligibly better than assuming everyone was average (MAE = 20.20, CI = 1.18, 43.18). At all benchmarks, predictions were in my view too imprecise to make inferences about specific individuals.

Approach 2 evaluated the percentage of plausible predicted scores that were shared between the bottom and top 20% of cases. Mean age was 8.42 (SD = 4.89) for the bottom 20% and 73.57 (SD = 8.04) for the top 20%. While the mean difference in true scores was thus 65.16 years (SD = 9.14), it was only 52.42 (SD = 13.44) for best-case, 23.00 (SD = 14.48) for demographic and 5.75 (SD = 9.27) for personality benchmarks. This corresponded to overlap in the predicted score distributions between the bottom and top 20% of 1% for best-case, 28%

for demographic and 64% for personality benchmarks. For personality, the average 74-year-old was predicted to be only *6 years* older than the average 8-year-old, and there was approximately 2/3 chance that predicted scores from one group were drawn from the predicted scores of the other group. Therefore, at realistic training accuracies it was unlikely to adequately differentiate even opposing extreme cases based on their predicted scores.

Approach 3 explored how prediction accuracy changed according to true score decile. For true age, the 1st decile mostly comprised infants (M = 4.19; SD = 2.59), the 5th (M = 37.25; SD = 2.55) and 6th (M = 45.60; SD = 2.26) deciles mostly comprised middle-aged adults, and the 10th decile mostly retirees (M = 80.09, SD = 6.15). For the 1st and 10th deciles, MAE = 9.26 (CI = 0.38, 25.20) for best-case, MAE = 25.19 (CI = 9.18, 48.25) for demographic, and MAE = 34.77 (CI = 21.78, 51.27) for personality benchmarks. Although best-case accuracy was superior to the other two benchmarks, even it was still too imprecise to predict age for specific individuals. Errors again fluctuated for the 5th and 6th deciles, such that MAE = 7.86 (CI = 0.31, 22.06) for best-case, MAE = 10.57 (CI = 0.45, 26.37) for demographic, and MAE = 6.71 (CI = 0.28, 17.08) for personality benchmarks. Although relatively precise, it was still more than 1.5 times as accurate to simply assume all these participants were average (MAE = 4.21, CI = 0.18, 8.18). Thus, training accuracy may not have improved enough to make best-case predictions for the extreme deciles practicable, and it was again more accurate to simply assume all middling cases were average regardless of the benchmark.

Approach 4 evaluated how MAE changed as I adjusted the SDs for predicted scores to realistic magnitudes. Although there were 21% children in true age, there were 18% children at best-case, only 1% at demographic and < 1% at personality benchmarks. When predicted age SDs were corrected to match true age, MAE = 8.39 (CI = 0.32, 23.87) for best-case, MAE = 16.69 (CI = 0.63, 47.81) for demographic, and MAE = 22.24 (CI = 0.85, 61.88) for personality benchmarks. Thus, correcting for realistic SDs made predictions increasingly redundant. At the personality benchmark they were again worse than assuming everyone was average.

Approach 5 evaluated whether classification accuracy—into low, medium, and high buckets—differed by predicted score extremeness. Overall, correct classification was 77% for best-case, 55% for demographic, and 43% for personality benchmarks (33% is chance). Then, I scored predicted age by extremeness (1 = "least extreme", 100 = "most extreme"). Best-case classifications were mostly above 50% across extremeness scores, but were again at or below 50% for edge cases (extremeness = 33 to 34). Prediction accuracy was only > 50% for

increasingly extreme cases at demographic (extremeness = 41) and personality (extremeness = 87) benchmarks. For the personality benchmark, this was the equivalent of only classifying infants aged under 2.25 and the elderly aged over 76.00 more than 50% correctly. When extremeness > 80, classification accuracy was 98% for best-case and 72% for demographic benchmarks, but still only 51% for the personality benchmark. When there were four categories, classification accuracy when extremeness > 80 was 92% for best-case, 59% for demographic and only 40% for personality. When there were five categories, it was 85% for best-case, 50% for demographic and only 33% for personality. Therefore, as simulations progressed to realistic accuracies, and/or I increased the number of categories, even extreme predictions were incorrect for more than 50% of cases.

Approach 6 investigated the extent classification errors were compounded when I attempted to put the entire big five into buckets. Simulations classified at least one of the five factors correctly for > 99% of cases at best-case, 98% of cases at demographic, and 94% of cases at personality benchmarks. To find cumulative accuracy, I raised the percent correct for at least 1/5 big five traits to the powers of two, three, four and five. Correct classification was again > 50% for only a maximum 4/5 factors for best-case, 3/5 for demographic, and 2/5 for personality benchmarks. The likelihood of classifying all 5 factors correctly was 27% for best-case, 5% for demographic and 1% for personality benchmarks. By contrast, the likelihood of classifying all cases incorrectly was < 1% for best-case, 2% for demographic and 6% for personality. Thus, it was unlikely that simulations classified the entire big five correctly at any benchmark. As simulations progressed to more realistic accuracies it was more likely that they classified the entire big five incorrectly, rather than correctly.

Finally, approach 7 evaluated the extent that simulations correctly ranked random pairs of participants based on their predicted scores. At every benchmark, I evaluated the extent true and predicted score ranks matched for 100 random split-half pairings. The mean true age difference in pairings was M = 27.11 years (SD = 19.34). Correct classification was 85% for best-case, 70% for demographic, and 59% for personality benchmarks. For demographic and personality benchmarks, correct classification was nearer to 50/50 chance than 100% correct. Thus, it was more parsimonious to conclude prediction models could not differentiate between participants *at all* despite large mean age differences, than to conclude they did so perfectly.

3.5. Study 3

In Study 3, I evaluated whether results also applied to real-world machine learning personality predictions from social media data. Real-world prediction model accuracy depends partly on sample size, sampling variance, the precise forms of social media data available, and the choice of machine learning models. Thus, to increase the generalisability of results, I iteratively *removed random error* from predicted scores until they reached best-case (r = .90), demographic (r = .60) and personality (r = .30) benchmarks. Randomly removing errors preserved original model robustness and allowed me to evaluate whether predicted score accuracies converged across all three studies.

3.5.1. Method

3.5.1.1. Participants

I recruited 1,471 American, British, and Canadian participants. They were retained from an original sample of 3,579 participants—who all took part in the *first wave* of the wider AXA study—because they volunteered their Twitter data. This first wave was a convenience sample of online survey panel participants from the most easily accessible English-speaking countries. It was indented to (a) give preliminary indication of differences in survey response styles between even countries with relatively similar cultures; and, (b) be large enough to build robust algorithms liking participants' Twitter behaviour to their survey responses. The full survey procedure is in Chapter 1. There were 74% women and the mean age was 36.17 (SD = 13.72).

3.5.1.2. Materials

I assessed big five personality using the 120-item IPIP-NEO personality inventory (Johnson, 2014), which was administered to this and all subsequent waves of AXA participants. Each factor was the average of 24 items rated from 1 = "strongly disagree" to 5 = "strongly agree". I used Twitter (a) mentions of other user accounts, and (b) content words (bag of words) and (c) phrases (word vector index) from tweets as the model inputs. For each type of Twitter data, I created a matrix where each unique behaviour—e.g. a specific account mentioned—had a separate column. Then, every participant scored 1 if they engaged in that behaviour and 0 if they did not. To mitigate sparsity, I collapsed each form of data into 100 factors using principal components analysis. This yielded a total of 300 factors. Then, I converted all factors to z-scores (M = 0, SD = 1).

3.5.1.3. Procedure

Many prevailing machine learning approaches use either LASSO or ridge regression to mitigate spurious coefficients (e.g. Kosinski et al., 2013). Thus, I used elastic net regression—which combines the two—to optimize predictions. It is at least as accurate as either approach in isolation (Zou & Hastie, 2005). I evaluated model accuracy using 25 different combinations of elastic net regression parameters. Then, I repeated this process using 10-fold cross validation, which (1) randomly assigns participants to ten groups, (2) creates an exhaustive set of models using 9/10 of the groups, and (3) evaluates accuracy by correlating true and predicted scores in the excluded group. Overall, this procedure means model accuracy is never artificially inflated by using the same participants in model development and evaluation. Each final model was the combination of elastic net parameters that had the highest average accuracy across 10-fold cross validation.

3.5.1.4. Big Wide Data Saturation

First, I evaluated the extent the Twitter variables were sufficiently comprehensive to maximize raw prediction accuracy. To this end, I used the full sample to generate 10-fold cross validated elastic net predictions with an increasing number of Twitter variables. I iteratively generated models using just the 1st to all 100 factors from each type of data. Thus, the number of predictors increased from 3 to 300, in increments of 3. Total prediction accuracy—with all 300 predictors—was r = .23 for neuroticism, r = .20 for extraversion, r = .35 for openness, r = .26 for agreeableness, and r = .21 for conscientiousness. Then, I focussed on average prediction accuracy across the big five (r = .25). There was a positive logarithmic association between the number of twitter variables and prediction accuracy (b = .05; CI = .04, .05; t(97) = 20.22, p < .001). Results are in Figure 3.8. Put another way, accuracy increased rapidly but then plateaued at around 80 variables. Thus, adding additional qualitatively similar Twitter variables was unlikely to further improve model accuracy.

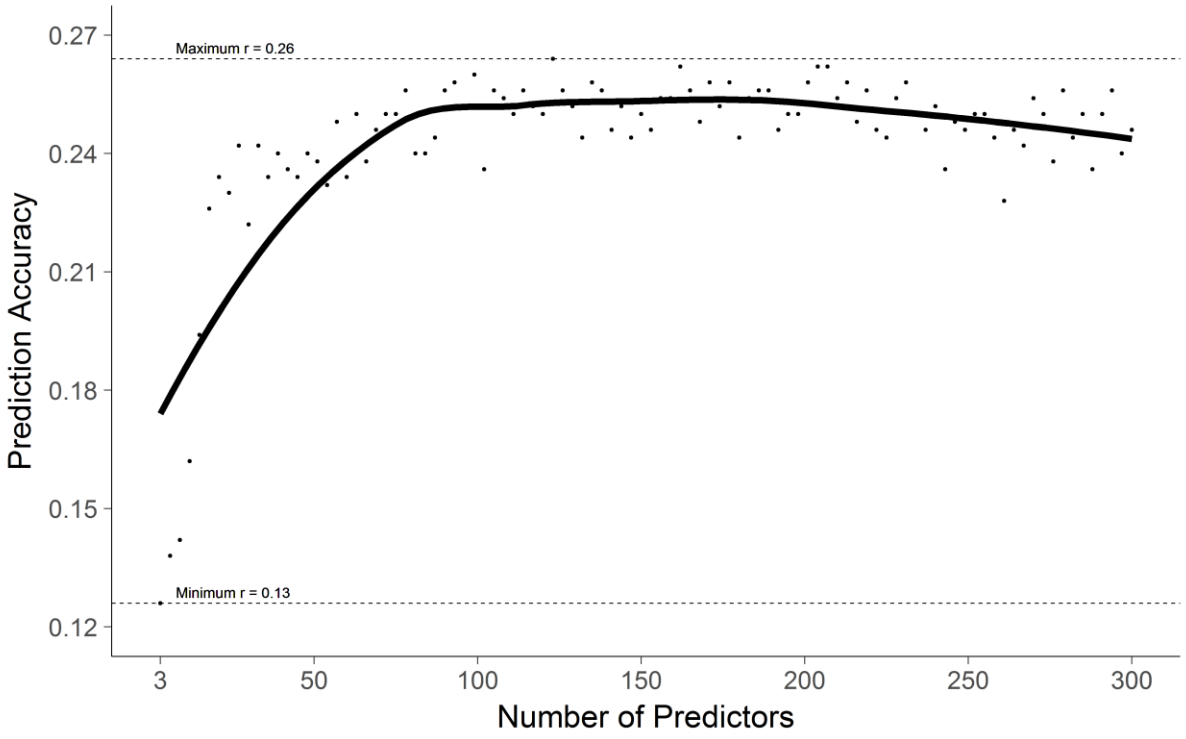


Figure 3.8. Big five personality prediction accuracy with increasing predictors. Number of predictors is the number of different Twitter variables used. There were three categories: mentions of other user accounts, word vectors and bags of words. Each category had behaviours that were collapsed into 100 variables using principle components analysis. The first variable explained the most total variation in observed behaviours, and so forth. I incrementally added variables in threes—one at a time from each category—using the most explanatory variable still available. Prediction accuracy was the correlation between true and predicted scores for models with incrementally increasing predictors. I averaged accuracy across the entire big five. There was some variability in accuracy between adjacent predictor numbers, and thus I show both individual points and the loess line.

3.5.1.5. Big Long Data Saturation

Then, I evaluated the extent that the sample size was large enough to maximize predication accuracy. To this end, I again generated 10-fold cross validated elastic net regression models, this time with an increasing number of randomly sampled participants for each of the big five. I began with models generated from 350 randomly selected participants, and then incrementally increased sample size by 10 until I utilized the entire sample. I repeated this entire procedure 10 times to increase the stability of the estimates at each increment. There was a positive linear association between sample size and prediction accuracy ($b < .01$; CI = < .01, <.01; $t(682) = 7.86$, $p < .01$). Results are in Figure 3.9. However, inspection of the loess line suggested there was also preliminary evidence for a positive logarithmic trend, where accuracy increased steeply and then stabilized. However, overall sample size was too small to show this trend in full. Thus, sample size was large enough to find reliable evidence that Twitter data contained personality information, but it did not maximise the amount of information extracted.

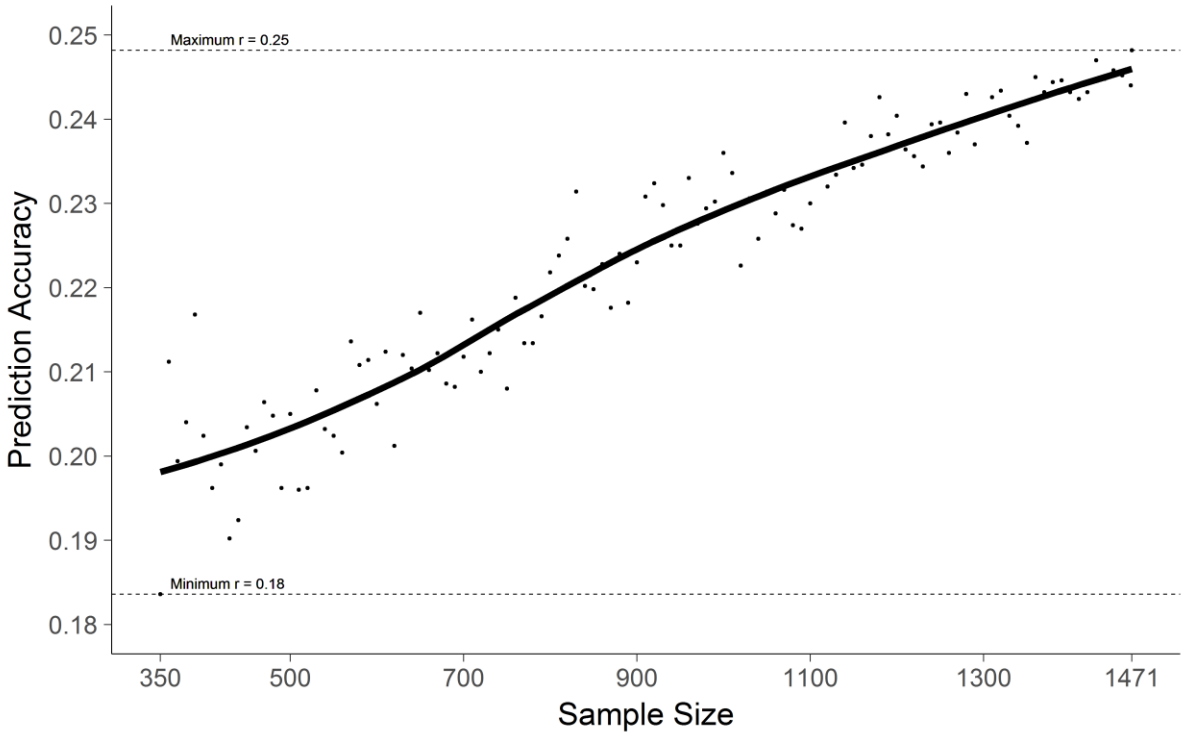


Figure 3.9. Big five personality prediction accuracy as a function of sample size. Sample size was from 350 to 1,471, increasing in increments of 10. Prediction accuracy was the Pearson's R correlation between true and predicted scores—generated from all 300 Twitter variables—at each sample size, which was then averaged across (a) the entire big five and (b) 10 iterations of each model, which each had different randomly selected training participants. There was some variability in accuracy between adjacent sample sizes, and thus I show both individual points, and the overall loess line trend. The bottom and top lines reflect the minimum and maximum average training accuracies observed for any single sample size.

3.5.1.6. Final Predictions

Finally, I corrected the predictions upwards so that they reached best-case (r = .90), demographic (r = .60) and personality (r = .30) benchmarks. To this end, I iteratively *removed* random noise from predicted scores. To this end, I first combined predictions for the entire big five. Then, to avoid inflated correlations, I adjusted true score means and SDs for each of the big five to match the average (M = 3.37; SD = 0.64). I repeated this process for predicted scores (M = 3.37; SD = 0.31). Then, I found the residuals from the correlation between true and predicted scores. I randomly removed between 0% and 1% of each residual from its corresponding predicted score. I repeated this process until I reached each of the benchmarks, and then repeated the entire process 10 times to increase the stability of the estimates. The exception was openness for the personality benchmark, which I kept at its raw training accuracy (r = .35). To preserve ecological validity, I also *increased* predicted score SDs so that they remained proportional to the magnitude of the upward prediction accuracy adjustment. Thus,

they realistically converged with true score SDs. This simulation code is in Appendix 3.3. The ratio of predicted to true score SDs was .80 for best-case, .52 for demographic, and .25 for personality benchmarks. Thus, they converged with the ratios used in Study 1 and Study 2. Finally, I linearly transformed true personality to match age from Study 1 (M = 35; SD = 18), and transformed predicted personality using the coefficients from the true score transformation, to aid interpretability. I truncated all true and predicted scores so the range was again 0-80.

3.5.2. Results

To preserve original true and predicted score distributions, I generated evidence for Approaches 1-7 separately for each big five factor, training accuracy, and iteration. Then, I aggregated their MAEs, confidence intervals, and classification accuracies. Thus, I report mean MAE, and mean 2.5^{th} and 97.5^{th} percentile confidence intervals for the absolute errors. Classification accuracy was the mean percentage correct. Throughout the results, I continue to use data linearly transformed to age to enhance interpretability. Nevertheless, the underlying data were predicted personality, and thus conclusions apply equally to the big five. Conclusion also apply equally to any normally distributed psychology characteristic, with the appropriate linear transformation. Table 3.3 (in the Study 2 methods) compares results to studies 1 and 2.

Approach 1 evaluated whether predicted scores were too imprecise to make inferences about specific individuals, even at high training accuracies. Supporting Studies 1 and 2, MAE = 6.09 (CI = .18, 17.70) for best-case, MAE = 11.30 (CI = .50, 30.86) for demographic, and MAE = 13.56 (CI = .53, 35.75) for personality benchmarks. For the personality benchmark, I on average mistook a 17-year old for a 3-year-old toddler or a 31-year-old. As models progressed to realistic accuracies, predictions were again negligibly better than assuming everyone was average (MAE = 14.29, CI = 0.63, 35.32). Overall, they were again too imprecise to make inferences about specific individuals, at every benchmark.

Approach 2 evaluated the percentage of plausible predicted scores that were shared between the bottom and top 20% of cases. The mean difference in true age between these groups was 49.42 (SD = 9.60). However, the mean difference was 35.74 (SD = 11.38) for best-case, 15.56 (SD = 11.05) for demographic and 3.82 (SD = 6.13) for personality benchmarks. Thus, differences shrunk when progressing to realistic benchmarks. This corresponded to 2% predicted score overlap for best-case, 28% for demographic and 62% for personality benchmarks. Overall, at realistic benchmarks the difference in predicted scores between extreme cases was negligible, and distributions increasingly overlapped.

Approach 3 explored how prediction accuracy changed according to the decile of the original score. For the 1st and 10th deciles, MAE = 8.79 (CI = 0.00, 20.27) for best-case, MAE = 20.81 (CI = 4.83, 36.26) for demographic, and MAE = 27.97 (CI = 16.51, 40.60) for personality benchmarks. Although best-case accuracy was superior to the other benchmarks, even it may have still been too imprecise to accurately predict specific ages. For the 5th and 6th deciles, MAE = 4.69 (CI = 0.20, 14.65) for best-case, MAE = 5.73 (CI = 0.21, 17.65) for demographic, and MAE = 3.74 (CI = 0.15, 10.97) for personality benchmarks. Although these were relatively precise, it was still at least 1.5 times more accurate to simply assume all participants were average. Thus, training accuracy may not have improved enough to make best-case predictions for the extreme deciles practical, and it was again more accurate to assume that all middling cases were simply the average, regardless of benchmark.

Approach 4 evaluated how MAE changed as I adjusted the SDs for predicted scores to realistic magnitudes. Although there were 18% children in true age, predicted score SD shrinkage meant there were 12% children at best-case, 3% at demographic and < 1% at personality benchmarks. When predicted age SDs were corrected to match true age, MAE = 5.96 (CI = 0.21, 18.21) for best-case, MAE = 12.11 (CI = 0.46, 36.18) for demographic, and MAE = 16.10 (CI = 0.63, 47.05) for personality benchmarks. For personality, MAE was again worse than assuming everyone was average (MAE = 14.29). Thus, correcting for realistic SDs made predictions increasingly redundant.

Approach 5 evaluated whether classification accuracy—into low, medium and high buckets—differed by predicted score extremeness. Overall, correct classification was 76% for best-case, 55% for demographic and 42% for personality benchmarks. Then, I ranked predicted scores by their extremeness (1 = "least extreme", 100 = "most extreme"). Best-case classifications were mostly above 50/50 chance across extremeness scores, but were again at or below 50% for edge cases (extremeness = 33 to 36). Prediction accuracy was only > 50% for increasingly extreme cases at demographic (extremeness = 52) and personality (extremeness = 62) benchmarks. For personality, this was the equivalent of only classifying cases under 18.54 and over 51.15 years more than 50% correctly. When extremeness > 80, classification accuracy was 97% for best-case and 76% for demographic benchmarks, but still only 54% for the personality benchmark. When there were four categories, it was 95% for best-case, 68% for demographic and 45% for personality. When there were five categories it was 91% for best-case, 60% for demographic and only 38% for personality. Thus, as models progressed to

realistic accuracies, even extreme predictions were increasingly difficult to classify correctly, especially into more than three categories.

Approach 6 investigated the extent classification errors were compounded when I attempted to classify the entire big five into low, medium and high buckets. For this approach, I treated each of the big five from my dataset separately and *did not* transform them into age. I classified at least one of the five factors correctly for > 99% of cases at best-case, 98% at demographic, and 94% at personality. However, again I only classified > 50% of cases correctly for 4/5 factors at best-case, 3/5 at demographic, and 2/5 at personality benchmarks. The likelihood of classifying all 5 factors correctly was 24% for best-case, 4% for demographic and 1% for personality benchmarks. The likelihood of classifying all cases incorrectly was < 1% for best-case, 2% for demographic and 6% for personality. Thus, it was unlikely that models classified the entire big five correctly at any benchmark, and it was again more likely that they classified the big five entirely incorrectly than entirely correctly at the personality benchmark.

Finally, approach 7 evaluated the extent I correctly ranked random pairs of participants based on their predicted scores. At every benchmark, models evaluated the extent true score ranks matched predicted score ranks for 100 random split-half pairings. The mean true age difference in pairings was M = 19.98 years (SD = 14.67). Correct classification was 86% for best-case, 71% for demographic, and 60% for personality benchmarks. For demographic and personality benchmarks, correct classification was again nearer to chance than being 100% correct. Thus, it was more parsimonious to conclude prediction models could not differentiate between participants, despite their large mean age difference, than conclude they did so perfectly.

3.6. Discussion

Until now, scientists, practitioners and the broader public have assumed that predicted psychological characteristics, especially those derived from algorithms using online behaviour, are highly accurate for specific individuals (e.g. Golbeck et al., 2011; Kosinski et al., 2013; Youyou et a., 2015; Grassegger & Krogerus, 2017). My aim was to directly evaluate the veracity of this claim. Although results apply equally to any normally distributed variable, interpretations were through the prism of personality because of its topicality and focus in the research literature to date, and to evaluate the privacy implications of using predicted personality throughout the remainder of this PhD. In three studies, I found fully convergent evidence that, at realistic accuracies, individual predictions are only marginally better than

chance at capturing specific individuals' big five personality, differentiating between cases with different true scores and correctly bucketing cases into low, medium and high thirds.

In approach one, I iteratively simulated predicted scores that had almost zero to almost perfect associations with true scores. At the best-case hypothetical prediction accuracy—of the kind researchers and practitioners might reach sometime in the future—models on average mistook a high school senior for a middle schooler or a late college graduate. At the upper limit of current demographic predictions, models mistook them for an elementary schooler or a 28-year-old. At realistic accuracy for current personality predictions, models on average mistook them for an infant or a thirty-something. Thus, predictions at all benchmarks may have had average errors that were too large to meaningfully correspond to individuals' true scores.

In approaches two and three, I found that predicted scores artificially converged on the full sample mean. In approach two, I evaluated the distributions of predicted scores for cases with the bottom and top 20% of true scores. At the best-case benchmark, there was only 2% overlap between predicted scores across the two groups. However, there was greater overlap at demographic and personality benchmarks. This resulted in up to a 2/3 chance that predicted scores drawn from the bottom 20% also belonged to the distribution of predicted scores from the top 20%. Put another way, at realistic accuracies, mean difference between predicted scores decreased to the extent that the average retiree was predicted to be only around 5 years older than the average primary schooler. In approach three, I evaluated prediction accuracy by true score decile. At the best-case benchmark, prediction errors were uniformly inaccurate across deciles, by around six years. As accuracy decreased, predictions remained equally inaccurate for true scores around the mean. Thus, regardless of benchmark it was at least 1.5 times more effective to simply assume all middling cases were the average. However, errors were also more inaccurate at the extreme deciles. At demographic and personality benchmarks, MAE was 20 and 28 years respectively. Thus, accuracy was driven by reductions in extreme prediction errors that—while noteworthy—were again too large to be practicable.

In approach four, I found that attempts to make predicted scores more realistic also increased prediction errors. Imperfect predictions increasingly cluster around the mean, thus undermining their capacity to identify real world thresholds—e.g. between children and adults, or non-extraverts and extraverts. To restore these thresholds, I increased predicted score variation to match true score variation. However, this caused error inflation to the point—for the personality benchmark—that models on average mistook a 17-year-old for a new borne. At

realistic prediction accuracies, corrected scores were in fact *worse* than simply assuming everyone was average. This exposed the tendency for predictions to cluster randomly around the mean, which thus obfuscated participants' true rank order and the subsequent corrections.

Considering these constraints, researchers might attempt to retroactively categorize predicted scores. However, in approaches five and six I found that classifications were usually nearer 50/50 chance than perfectly correct. In approach five, I found that at realistic accuracies only cases with extreme predicted scores were correctly categorized into thirds more than 50% of the time. Models also almost perfectly classified the 20% most extreme cases, regardless of whether there were three, four or five buckets. However, as they progressed to the personality benchmark, there was less than 50% chance of classifying cases between adolescence and late middle age—i.e. cases with the middle 2/3 of predicted scores—into their correct thirds. Moreover, correct classifications at the personality benchmark plateaued for even the extreme cases—which were the easiest to classify because models could only make classification errors in one direction—such that the ceiling accuracy was still only marginally better than 50/50 chance. Correct classifications dropped even more sharply when attempting to use more than the minimally viable three categories. Then, approach six found that classification errors were compounded across multiple predicted traits. Even for the best-case benchmark, there was only a 1/4 chance that models classified the entire big five correctly. Chance fell to 1/20 for the demographic benchmark. At the personality benchmark, models were up to 6 times more likely to classify individuals' big five 100% incorrectly than 100% correctly.

Finally, in approach 7, I found that random pairs' predicted score ranks were more likely to be caused by random chance than perfectly accurate. Although models were nearer perfect rank-ordering (i.e. > 75% correct classifications) for best-case accuracy, this pattern was inverted at demographic and personality benchmarks. At the realistic accuracies, predicted scores regularly failed to differentiate between cases that were actually different by 20 years, on average. At these accuracies, it was usually more reasonable to assume that predicted score rankings were no better than chance than to assume they were perfectly accurate.

3.6.1. Predicted Big Five Personality

Best-case predictions—the kind researchers and practitioners might hypothetically achieve in the future—(a) reliably differentiate between opposite extreme cases, (b) are almost equally efficacious for participants across the entire spectrum of true scores, (c) can be corrected to capture some real-world thresholds, (d) can be accurately categorized into thirds on a single

variable, especially at extreme values, and (e) are likely to correctly rank randomly drawn pairs of participants. Even so, they may still be incorrect to the extent that they miss important thresholds, which coincide with qualitative behaviour changes, and cannot be used to reliably classify cases across multiple traits. Thus, even this extreme benchmark may only allow some restricted conclusions about specific individuals, and only then on a small subset of traits rather than their entire psychological profiles. At demographic and personality prediction accuracy benchmarks—the best researchers might expect given emergent and current technologies—scores may have been too imprecise to be practicable, unable to be transformed to capture realistic thresholds, and unable to be classified or ranked meaningfully above chance for a large portion of cases. At the personality benchmark—which coincides with a 'medium' effect in social psychology—it was often more parsimonious to conclude predictions were no better than assuming everyone was average, rather than to conclude they were totally correct.

Results highlight that theoretically significant trends—such as that humans leave traces of their psychological characteristics like personality online (Golbeck et al., 2011)—are exclusively normative. By switching from correlations to absolute prediction error, I found that even putatively 'accurate' models may not yield informative predictions for specific individuals. This may, counter-intuitively, benefit researchers in social psychology. They might use the small kernels of accuracy in predictions to evaluate group-level effects without jeopardising individual privacy, provided prediction errors are both fully random and offset by increased sample power.

3.6.2. Practical Implications

This chapter explored the extent online predicted psychological characteristics apply to specific individuals, at various feasible and hypothetical future accuracies. I intended for it to help make a more informed decision about the appropriateness of using non-consensually obtained predicted psychological scores for the upcoming chapters. Overall, my findings suggested that they were not intrusive, and might thus be appropriate for further research without obtaining explicit consent. Such predicted scores may be especially useful for generating such large data that I could (a) test extremely subtle effects for personality on SWB, and (b) quantify the *average* personality of participants' social networks, local regions, states and countries—to examine how social contexts change individual-level effects. However, my PhD evolved as I developed this chapter. My focus was increasingly on the more fundamental and still largely unanswered question 'what facet-level personality traits are robustly associated with SWB?'. It was possible to begin addressing this question with exclusively AXA self-report data. A

benefit was that, unlike algorithm-predicted scores, self-report data only suffers from measurement error and not compounded measurement and prediction errors. Nevertheless, findings suggest that future research could use online-predicted personality to replicate and then extend results throughout the remainder of my PhD, without serious privacy ramifications.

3.6.3. Wider Privacy Considerations

Although results ultimately did not inform the direction of subsequent chapters, topics raised herein may still have practical implications outside the PhD. I find it important to broach some of these implications here—even when they are not the primary focus of the results section—because of both the topicality of the research area, and the allegations raised against my former PhD supervisor.

First, internet users do not normally have the time or expertise to understand the full implications of their consent (Tam et al., 2015). Thus, they must trust that terms and conditions converge with their own *privacy expectations*. Regardless of algorithm precision, a prevailing expectation may be that it is *only* appropriate to use individual-level predictions to market consumer products (Custers, van der Hof & Schermer, 2014). Domains like political decision making may be sacrosanct. It is imperative to educate the public on the true informativeness of predicted scores, and then also obtain explicit informed consent when there is any possible real *or perceived* breach of privacy expectations.

I must also distinguish between openly disclosed social media information and predicted private traits. The former comprises everything from (e.g.) indicating gender on Facebook, to buying a blender on Amazon or rating a TV series on Netflix. This information is used by recommender systems to suggest either similar products and services—or products and services used by similar people—that individuals have a higher-than-chance probability of liking (Gomez-Uribe & Hunt, 2016). In some circumstances, these recommendations may be extremely accurate (Koutrika, 2018). However, they typically bypass predicted psychological characteristics—like personality—which introduce another level of inferential analysis and thus cause compounded statistical errors. Instead, to-date the most prominent use of predicted private traits is to generate *bespoke advertisements* for existing recommendations (Sun, Li & Zha, 2017). Considering results from the present studies, *correct* bespoke advertisements may occur at slightly better than chance across one or a small number of traits. Notably, any marginal benefit must also be offset by the potentially asymmetrical consequences of incorrectly targeted ads.

Importantly, psychological characteristic predictions may also have optimum fidelity when they use relatively few, maximally informative, sources of information. This speaks to the phenomenon of overfitting, which is when predictions are idiosyncratic to the specific sample and not generalisable to the entire population (James et al., 2013). Relatively uninformative social media behaviour may only help predict private traits when it adds *more* new information—over and above other consensually disclosed (e.g.) concrete demographic information—than it adds superfluous model complexity. One putative remedy is elastic net. It is the prevailing machine learning approach that mitigates overfitting because it extracts clusters of related predictors (Zou & Hastie, 2005). However, this means stronger and weaker predictors are clustered together, which compromises the fidelity of the stronger predictors (Hu, Singh & Scalettar, 2017). Put more simply, it may be preferable to fit *only* the stronger predictors at the outset. In online contexts, the stronger predictors may be openly disclosed and face valid demographic variables (e.g. checking a box that one is 'female'), rather than comparatively transitive and ambiguous (e.g.) Facebook Likes. Overall, even when online behaviour contains traces of true personality, it may still be useless during modelling.

Group-level privacy is at least as pressing a concern. In some cases, social media information might be used to improve the *chances* that individuals respond favourably to bespoke advertisements for individual recommendations. This is especially relevant in at least two cases: sensitive domains in the general adult population, and at-risk groups. A prominent sensitive domain is partisan political decision making. Political campaigns will inevitably attempt to persuade voters using battery of strategies that transcend officially-stated policy. However, there may be consensus that certain types of attempted persuasion are, functionally, indistinct from widely rebuked malpractices—such as defamation or coercion—and thus unethical. Prominent at-risk groups include children and various adult populations. Such groups may have compromised decision-making capacity and thus heightened susceptibility to persuasion interventions. In such cases, such targeted interventions may *always* be unethical.

3.6.4. Limitations and Future Directions

There are at least five limitations in the present chapter. First, prediction error magnitude will change from study to study depending on the variables/transformations (e.g. age vs income). Researchers might reduce absolute error by increasing sample homogeneity. While this may cause restriction of range and thus compromise external validity, it may also create some instances where numeric predictions are more practicable. Second, results were based on the kind of normally distributed prediction errors typical in general linear models. Although I

expect at least some universal support for my conclusions, some findings—e.g. relatively large prediction inaccuracy for extreme cases—may be less generalisable to other error distributions. The third limitation concerns the analogical use of age to illustrate prediction inaccuracy. Meaningful age thresholds (e.g. between children and adults) may not correspond to meaningful personality thresholds (e.g. between extraverts and non-extraverts). As such, it may be appropriate triangulate across multiple concrete ratio scales. Fourth, allowing unequal bucket sizes may have improved categorisation accuracy. For example, disproportionately allocating participants to middling buckets may have reduced the likelihood of extreme bucketing errors. Finally, prediction accuracies using our Twitter data were lower than those observed with other social media data (e.g. Youyou et al., 2015). This may be because of the relatively small sample size, sampling variability, or insufficient data heterogeneity. Thus, I corrected predictions upwards by removing random noise. Although corrections preserved ecological validity because they did not introduce any additional sources of bias, they only replicated the results from the other studies using *partially* real-world data.

Future research could extend findings to other empirical domains where the individual is the unit of analysis. While I demonstrated model inaccuracy for online personality predictions, it ought to apply (a) to any normally distributed variable, and (b) in any domain that uses normative statistics. Results might hold *even* when those statistics are used to make inferences about internal processes, such as in cognitive neuroscience. It may be especially usefully to apply the seven approaches to practical domains, where researchers are interested in both validating theory and evaluating *practical implications*. For example, they could help better understand the extent a specific form of intelligence impacts job performance, and the effectiveness of targeted advertising on actual voting behaviour. Where the outcome refers to an intangible psychological construct—or whenever else a score of zero has no intrinsic meaning—researchers might improve comprehension by using concrete and well-known analogies, such as age, height, weight, and income. Finally, there could be increased attempts to quantify *all* sources of prediction inaccuracy. For example, measurement imprecision, self-report bias and legitimate changes in both predictors and outcomes over time may all further undermine predicted score accuracies. They might combine with algorithm predictions to create multiple, propagating, errors.

3.6.5. Conclusion

Emerging computational approaches have the potential to help us live longer, more peacefully, and in greater abundance. They can also help augment the ways people interact with the world,

potentially enriching their social and intellectual lives. However, realization of these potentials might depend on the extent that big data intrudes on individual privacy. Across three studies, I found consistent support that predicting psychological characteristics from online behaviour at realistic and even best-case future hypothetical accuracies yields insights that are usually too imprecise to apply to specific individuals. Results exposed the discrepancy between theoretically relevant and individually practicable research. Put another way, people's psychological characteristics may indeed manifest in their quantifiable internet behaviour, whilst they simultaneously remain largely unknowable.

Short of actively disclosing private information (e.g. Facebook Liking the US Republican Party or registering for a homosexual dating website)—researchers might only ever extract fuzzy and imprecise psychological information that is, at best, marginally better than chance guessing. If one was to make a binary assessment about the fidelity of social media algorithm predictions using current technology, they would likely conclude it yields zero new psychological insights for specific individuals rather than totally accurate insights. Thus, researchers and practitioners might place renewed focus on the appropriateness of using predicted variables to evaluate *group trends*. This opens the possibility of building algorithms that unlock the psychological profiles of entire databases, for exclusively normative analyses.

Predicted psychological characteristic scores may still be especially useful during the exploratory research phase. Multiple algorithms can be used to construct comprehensive private trait profiles for individuals in a database. It is still possible to use error-prone individual predicted psychological characteristic scores to find effects between clusters of related variables (e.g. via structural equation modelling). Using millions of people from social media also means there is likely sufficient statistical power to evaluate very subtle and/or complicated effects. Finally, variables can also be averaged upwards to find context-level psychological characteristics, where individual prediction errors would cancel each other out. This opens the possibility of investigating the extent individual-level effects change across contexts.

Ultimately, however, I did not proceed with online-predicted psychological scores in the remaining PhD. As research for this chapter progressed, I became more interested in addressing the fundamental question 'what personality facets are robustly associated with SWB?'. The questions I was beginning to ask could all be addressed with the large self-report database of personality scores described in Chapter 1, at least in the first instance. These scores only suffered from measurement error, and not combined algorithm prediction and measurement

error. Thus, they were better approximations for the target underlying psychological constructs, which meant higher fidelity effect size estimates and less risk of confounding due to added spurious score variation that, if non-random, could have captured other unwanted phenomena.

Chapter 4

Rescoring the Balanced Measure of Psychological Needs (BMPN) to Capture Subjective Well-Being

4.1. Abstract

In the present chapter, I evaluated whether the 18-item balanced measure of psychological needs (BMPN)—originally intended to measure the separate feelings of autonomy, competence and relatedness—could be rescored to measure overall SWB. To this end, I assessed the BMPN, other SWB scales and sociodemographic variables in 28,000+ adult participants from between 28 and 33 countries. Using EFA (Study 1), I found consistent evidence for two separate factors, which comprised needs satisfaction and needs thwarting. In Study 2, I then found that (a) both factor scores converged with other measures of SWB, (b) needs satisfaction and needs thwarting disproportionately measured positively and negatively valanced SWB respectively, and (c) results replicated across different countries and demographic strata (e.g. just women). Then, I found that both needs satisfaction and thwarting (d) captured more transitive SWB than the criterion satisfaction with life scale, and (e) explained additional unique variation in lifestyle outcomes such as physical health and relationship status. Results were not artefacts of response bias or differently worded BMPN items. Therefore, I concluded that the BMPN successfully collapsed into needs satisfaction and thwarting factors measuring overall SWB.

4.2. Introduction

The previous chapter focussed on the feasibility of using online predictions to measure the predictor-group—big five personality—which I focus on throughout the remainder of the PhD. The present chapter continues this univariate focus by switching to SWB, which is the outcome-group. SWB involves self-appraised feelings of suffering and flourishing (Diener, 1984). To date, its operationalisations include emotion experiences, cognitive reflections and the sense of purposefulness (Diener et al., 2017). SWB is a meaningful outcome: there is convergent evidence, from such measures and others, that individuals often prioritize

maximizing their SWB over their objective well-being (Ryan & Deci, 2001). They may also use their SWB to inform contemporaneous religious, purchasing and political decisions (Ellison, 1991; Baumeister, 2002; Bok, 2010). Further, SWB may partly cause other important individual outcomes, such as professional success, strong relationships and longevity (Csikszentmihalyi, 1997; Daley, Burge & Hammen, 2000; Diener & Chan, 2011). At the aggregate level, both government and private institutions are increasingly monitoring SWB—as they do GDP and education attainment—and attempting to both mitigate low SWB in specific populations and correct for forecasted aggregate-level changes (e.g. The World Happiness Report; Helliwell, Layard & Sachs, 2018). Appropriate measurement is essential for assessing personality facet associations with an unbiased and comprehensive measure of SWB.

As I outlined in the General Introduction, however, there may be a lack of scales that unbiasedly capture overall SWB. For example, the prevailing tripartite model—comprising negative affect, positive affect and life satisfaction—captures an arbitrary subset of only three SWB facets. Contrastingly, Ryff's (1989) psychological well-being framework deliberately ignores affect. While more comprehensive scales are emerging—such as the Scales of General Well-Being (Longo et al., 2017)—these are not yet widely replicated, may overweight facets from established sub-fields and disproportionately focus on positively valanced SWB. An alternative is to appropriate scales from the mature field of human motivation (Robbins, 2008). Specifically, basic psychological needs theory—a component of the hugely influential self-determination theory (SDT; Ryan & Deci, 2000)—suggests that the combined feelings of autonomy, competence and relatedness needs exhaustively cause omnibus SWB. Measuring these domains may negate the need to use incomplete, process-oriented, measures. Thus, I aim to evaluate whether one prominent operationalization of basic psychological needs theory—the balanced measure of psychological needs (BMPN)—captures overall SWB.

4.2.1. BMPN Development and Validation

The BMPN measures the SDT concept of psychological needs satisfaction. According to SDT, *all three* psychological needs—feeling autonomous, competent and related—are necessary for SWB (Ryan & La Guardia, 2000). As such, the psychological needs are most often conceptualized as separate mechanisms that *explain* the benefits of different types of goal satisfaction. The 18-item BMPN was originally developed to supersede the existing domain-general (rather than context-specific; e.g. workplace) measure of psychological needs satisfaction (Sheldon & Hilpert, 2012). There are six unambiguously worded items—three positively and three negatively valanced—measuring each need. Using confirmatory factor

analysis, Sheldon and Hilpert (2012) found evidence for three separate need factors, which each had nested facets for feelings of thwarting and satisfaction. Then, they found the three factors were positively associated with a composite measure of tripartite SWB. Therefore, there was preliminary support that the BMPN captured the three different psychological needs.

The original BMPN factor structure has since been replicated. Using German university and young adult community samples, Neubauer and Voss (2016a, 2016b) used confirmatory structural equation modelling to replicate the prevailing BMPN factor and facet structure. Then, they also found that its scales were positively associated with life satisfaction and negatively associated with depression. Convergently, Chen et al. (2015) administered a 24-item adaptation of the BMPN—the balanced psychological needs satisfaction and frustration scale—to university students from four separate countries, which each spoke a different language. Using facet-level analysis, they found that the each of the three needs satisfaction facets had positive associations with life satisfaction and vitality, and that each of the three needs thwarting facets were positively associated with depression. Results applied equally to each country sample. Finally, there is consistent evidence for positive associations between the BMPN needs factors and SWB in American, UK and Chilean community populations, as well as in heterogeneous multinational students (Martela & Ryan, 2016; Unanue, Dittmar, Vignoles & Vansteenkiste, 2014; Yang, Zhang & Sheldon, 2017). Thus, there is further support for the robustness of the BMPN across a variety of different research and cultural contexts.

4.2.2. Alternative Structures in the BMPN

However, the BMPN may have multiple construct valid structures. The above studies used confirmatory factor analytic approaches to find evidence for the superior explanatory power of the three-factor solution. However, *most* of the explained variation in item responses may still be accounted for by one or two latent superordinate factors. It is unclear whether the tradeoff between increased explanatory power and the increased complexity of their factor structure is always worthwhile, especially in research contexts that do not explicitly differentiate between two or all three of the psychological needs. Further, the same convergent SWB scales were typically used to demonstrate the validity of all three BMPN factors. This may suggest *either* that each scale captures separate aspects of latent SWB, or *alternatively* that variance shared between scales captures a single shared SWB construct. Finally, the main validation studies—in America and Germany—used young adults. Young adults may experience more asymmetric needs satisfaction because they tend to be financially dependent on their family, yoked to tertiary training programs or disproportionately motivated to fulfil relatedness and/or autonomy

needs (Schulenberg, Sameroff & Cicchetti, 2004). Thus, they may experience greater affordances to fulfil just one or two needs, which exaggerates the multi-factor BMPN structure.

4.2.3. Superordinate Psychological Needs Factors

SDT provides rationale for a one-factor BMPN solution. It suggests that people are intrinsically motivated to fulfil all three of their psychological needs (Ryan & Deci, 2000). While there may be differences in need *strength*—that is, some needs have a higher threshold for fulfilment than others—*all three* are still prerequisites for SWB (Sheldon, Elliot, Kim & Kasser, 2001). For example, Reis, Sheldon, Gable, Roscoe and Ryan (2000) found that the *combined* feelings of autonomy, competence and relatedness explained daily fluctuations in positive and negative affect, subjective vitality and signs of physical illness. Similarly, Newman, Tay and Diener (2014) found that leisure activities enhanced SWB when they activated all the needs pathways, alongside feelings of escape and meaning. Even though the psychological needs are qualitatively distinct, their satisfaction may still be relatively constant.

Alternatively, the BMPN may comprise two separate SWB factors. Negatively coded BMPN items focus on needs thwarting rather than the mere absence of needs satisfaction. For example, relatedness needs thwarting might involve *feeling lonely* rather than simply not feeling connected with close others. Such a two-factor structure—with additional feelings of suffering—aligns with the founding premise of positive psychology: that subjective ill- and well-being are qualitatively distinct phenomena, rather than simply opposite ends of a single continuum (Seligman & Csikszentmihalyi, 2014). This was support by Ryff et al. (2006), who examined the biological correlates of subjective ill-being and SWB. Averaging across multiple measures—e.g. depressive symptoms and trait anxiety, and eudemonia and hedonia—they found differences in the extent SWB variables were associated with salivary cortisol, systolic blood pressure and cholesterol. Most recently, Vanhove-Meriaux, Martinent and Ferrand (2018) evaluated this distinction directly in geriatric participants. They found that omnibus needs thwarting was uniquely related to negative affect, and omnibus needs satisfaction was uniquely related to positive affect, eudemonia and a sense of vitality. Thus, the basic psychological needs may, alternatively, collapse into two superordinate factors.

4.2.4. Incremental Validity

To have practical utility, the BMPN must outperform the criterion satisfaction with life scale (SLS), which measures general cognitive appraisals of SWB. In the General Introduction, I argued that SLS has portions of variation that are both facet-specific and shared with feelings

of negative and positive affect. However, Busseri & Sadava's (2011) review also argues that an alternate, prevailing, structure suggests negative and positive affect *cause* changes in SLS. Thus, SLS may be the most central (i.e. mediational) of the three tripartite constructs to overall SWB. There are also practical reasons for viewing SLS as the single criterion measure. Life satisfaction and affectivity are measured with different surveys. For example, SLS involves rating global abstract statements (e.g. "I am satisfied with my life"; Diener, Emmons, Larsen & Griffin, 1985). Contrastingly, positive and negative affect may be measured using responses about the frequency of emotion experiences (e.g. "Excited") in specific windows of time (e.g. "… the last few weeks"; Watson, Clark & Tellegen, 1988). Thus, aggregated SWB scores may be confounded by differential scale instructions, item wording and/or response options.

Lucas and Diener's (2015) review gives further support for the precedence of SLS over positive and negative affect. First, they concede a weakness in global retrospective measures of life satisfaction: they are either assessed heuristically—quickly, using implicit rules of thumb—or require extremely onerous cognitive appraisals about multiple life domains. Nevertheless, they also suggest a lack of definitive evidence for moderating variables. They illustrate by citing Schwarz & Clore's (1983) famous finding that weather (sunshine vs rain) changed self-reported SWB, and that moderation effects were *marginal* and only replicated in certain contexts. Thus, Lucas and Diener (2015) conclude that any more transient effects for affect on SWB would manifest as *random error* rather than systematic confounds. Finally, they suggest measures of affect may be more fraught because they are partly determined by extraneous factors—many of which are only weakly related to objective life circumstances (e.g. diet, current medicines)—multiple reports are needed to establish the central tendency of affect experiences and new responses may be especially biased by previous response patterns. Overall, SLS may thus be the least imperfect of the tripartite model SWB scales.

The BMPN may outperform criterion SLS in many ways. It may more unbiasedly capture entire SWB because it comprises an equal balance of negatively and positively valanced items. Individual items also comprise a more ecologically valid combination of cognitive and affective appraisals, rather than artificially delineating the constructs. In addition, BMPN may also capture more transitive SWB. The BMPN comprises items that reference contemporary experiences (e.g. "… *successfully* completing difficult tasks and projects"). Thus, any extraneous variable must covary with both the potentially stable predictor and the more transitive outcome to offer a plausible alternative explanation. This reduces the absolute number of possible confounds, thus increasing our confidence in non-spurious effects. By

contrast, SLS may be largely determined by features that are relatively stable. In support, Diener et al. (2017) found large structural differences in average life satisfaction between countries, ethnicities and social classes. Further, Schimmack et al. (2004) found that people who are predisposed to experience high cheerfulness (a facet of extraversion) and low depression (a facet of neuroticism) may experience particularly high life satisfaction, even after accounting for a range of demographic factors. Thus, life satisfaction may itself be a kind of stable character trait, at least during adulthood. This makes it especially vulnerable to confounding when used as an outcome variable; there are likely a range of covarying stable variables—like socio-economic status, education attainment and personality—that cannot all be controlled because they would *explain away* most of the variation in life satisfaction (Heller, Judge & Watson, 2002). Thus, they can act as confounders. Overall, such features suggest that BMPN may have predictive utility over-and-above SLS, in a variety of research contexts.

4.2.5. Computational Psychometrics

For simplicity, I define 'computational' as any approach that uses large-scale data to perform analyses using automatically changing parameters, which thus yields full *patterns* of results. Most relevant to the present chapter, this includes bootstrapping—i.e. repeatedly generating statistical models using random sub-samples—and evaluating relative effect magnitudes for an exhaustive combination of variable pairs. Iterative methods, in particular, can enhance existing psychometrics. Psychometrics examines whether scales measure their intended psychological constructs. At the outset, performing bootstrapped EFA—which evaluates the items that covary most together—allows researchers to quantify the extent observed patterns are due to sampling error, and adjust confidence in the results accordingly (Hox, Moerbeek & van de Schoot, 2017). Iterative approaches may also be especially useful when evaluating discriminant validity, which is when theoretically unrelated variables have *weak* associations. Importantly, weak associations may still be significant due to high statistical power, common method variance and/or positive manifold (Murayama et al., 2014). Thus, it may be useful to iteratively evaluate the *relative* magnitude of exhaustive convergent (theoretically related) vs discriminant associations. Overall, computational psychometric approaches may thus help increase confidence in the construct validity of the rescored BMPN.

4.2.6. Present Studies

My overall aim was to determine whether BMPN sub-scales could be combined into one or two superordinate factors that captured overall SWB. Thus, I used EFA to determine whether

there were superordinate factor structures in the BMPN. Then, I evaluated the psychometric properties of the emergent factor/s. Specifically, (a) convergent validity was the extent of associations with other measures of SWB, in the intuitive directions; (b) discriminant validity was the extent associations were *larger* when variables *both* captured flourishing or *both* captured suffering; (c) external validity was the extent (a) and (b) held when using deliberately unrepresentative sub-samples; (d) pragmatic validity was the extent that BMPN factor/s were more associated with transitive convergent measures of SWB than SLS; and, (e) incremental validity was when the emergent factors were robustly associated with real-world outcomes after controlling for sociodemographic factors, response bias and SLS.

4.3. Study 1

Study 1 evaluated the BMPN structure that explained the most variation in individual item responses. I theorized that items would collapse into either one factor measuring overall needs satisfaction, or two factors measuring needs satisfaction and thwarting. I evaluated all BMPN effects relative only to participants' countrymen. This was because I was interested in evaluating BMPN independent of the structural factors that cause country-level differences. I used EFA, which meant I could find the endemic structure of BMPN items rather than having to pick from a pre-defined structure (as in confirmatory factor analysis). To increase objectivity, I used three *a priori* triangulated criteria to decide how many factors to retain. A robust solution was when the metrics converged to suggest the same number of factors.

4.3.1. Method

4.3.1.1. Participants

Participants were the internet panellists who took part in my large multinational AXA survey project. The full survey procedure is in Chapter 1. The BMPN was administered in 28/33 countries. This yielded a subsample of 28,952 participants from the total retained N = 36,498. Participants were retained (89%; $SD_{Country}$ = 5%) because they took > 5 minutes to complete the survey, had > 70% non-missing responses and had > .25 response variance in the ubiquitously administered NEO-IPIP-120 (Johnson, 2014). Overall, there were 53% men and the mean age was 34.42 (SD = 11.70). Participant demographics by country are in Table 4.1. There were between 164 and 1,438 participants in each country, of which between 35% and 72% were men and the mean age ranged from 30 to 42.

4.3.1.2. Materials

All scales—including the BMPN—were originally developed in English. I used established scale translations where possible. The remainder were translated and then back translated to English by two expert-language speakers, using the established protocol in cross-cultural psychology (Brislin, 1970). When the back-translation failed to converge with the original English version, the translators reached consensus on the final wording. Then, a trained social psychologist reviewed and approved each final back-translated scale.

Table 4.1

Participant demographics in countries that were administered the BMPN

ISO	Country	Language	N	% Retained	% Male	Age (SD)
ARG	Argentina	Spanish	1,106	88%	50%	36.24 (12.27)
AUS	Australia	English	1,149	90%	47%	40.19 (12.77)
AUT	Austria	German	1,240	93%	45%	39.89 (12.62)
BOL	Bolivia	Spanish	164	79%	55%	33.19 (12.06)
CAN	Canada	English	1,295	91%	38%	35.98 (13.85)
CHL	Chile	Spanish	1,121	89%	50%	33.58 (11.14)
CHN	China	Mandarin	960	86%	57%	32.93 (8.78)
COL	Colombia	Spanish	1,083	94%	66%	30.55 (9.43)
DEU	Germany	German	1,128	94%	55%	37.53 (13.42)
ECU	Ecuador	Spanish	1,148	84%	47%	34.17 (11.8)
ESP	Spain	Spanish	1,014	94%	63%	33.82 (9.7)
FIN	Finland	English	1,028	91%	48%	38.36 (12.46)
GBR	United Kingdom	English	1,438	92%	35%	35.08 (12.93)
IND	India	English	980	91%	78%	30.57 (9.4)
ITA	Italy	Italian	1,108	92%	49%	34.68 (10.78)
JPN	Japan	Japanese	458	83%	48%	42.11 (12.17)
KOR	South Korea	Korean	493	91%	48%	36.92 (11.11)
MEX	Mexico	Spanish	1,185	96%	59%	30.09 (9.13)
PER	Peru	Spanish	1,078	89%	61%	29.72 (9.64)
POL	Poland	Polish	970	89%	66%	30.9 (11.23)
PRY	Paraguay	Spanish	980	80%	47%	29.86 (9.04)
RUS	Russia	Russian	1,149	94%	46%	36.85 (11.86)
THA	Thailand	Thai	1,079	89%	49%	33.95 (9.14)
TUR	Turkey	Turkish	1,106	81%	72%	30.56 (9.48)
TWN	Taiwan	Mandarin	1,025	84%	50%	34.22 (10.62)
URY	Uruguay	Spanish	1,214	85%	45%	35.89 (12.27)
VEN	Venezuela	Spanish	1,092	94%	61%	32.41 (10.95)
ZAF	South Africa	English	1,161	90%	47%	35.35 (11.44)
Total			28,952	89%	53%	34.42 (11.70)

Notes. N = Total number of participants who were retained in each country.

4.3.1.2.1. BMPN

Of the eligible participants, I retained the 98% who completed every BMPN item. I did not impute missing item scores because this may have artificially inflated any emergent factor

structure (Siddique, de Chavez, Howe, Cruden, & Brown, 2018). Thus, the final sample size was 28,372. There were six items for each psychological need (autonomy, competence, relatedness), which were each rated on a 5-point Likert scale (1 = "Strongly disagree", 5 = "Strongly agree"). Three items were reverse coded in each subscale. For the present study, I converted items from their original past-tense to present tense, so that they captured *current* appraisals of SWB. Descriptive statistics, scale intercorrelations and associations with life satisfaction are in Table 4.2. This table also includes the intra-class correlation coefficient (ICC), which is the proportion of score variation attributable to participants' country (Aguinis et al., 2013). All three original BMPN factor scores were normally distributed, there was no evidence for floor or ceiling effects and varied mostly at the individual level. However, internal consistency was also below the conventional threshold ($\alpha = .70$) for inferring that each subscale measured a single construct. This was first evidence that an alternate structure may be more appropriate. Bivariate associations with satisfaction with life—the established criterion measure of SWB in the data—were all positive and approximately the same magnitude. Finally, subscale intercorrelations were all also positive and approximately equal magnitude. Notably, they were also approximately as large as the internal consistencies. This suggested equal magnitude correlations within and *between* the needs factors. All BMPN items are reported in Table 4.3, which is in the Results section.

Table 4.2

Descriptive statistics for the BMPN relatedness, competence and autonomy subscales

Subscale	Mean (SD)	α	ICC	R_{SLS} (99% CI)	R_{REL} (99% CI)	R_{COM} (99% CI)
Relatedness	3.61 (0.69)	.62	5%	0.42 (0.41, 0.44)	-	-
Competence	3.44 (0.69)	.65	7%	0.43 (0.42, 0.44)	0.56 (0.55, 0.57)	-
Autonomy	3.45 (0.69)	.64	6%	0.46 (0.45, 0.47)	0.58 (0.57, 0.59)	0.59 (0.58, 0.6)

Notes. SLS = Satisfaction with life scale, which was the prevailing criterion measure of SWB. REL = Relatedness subscale. COM = Competence subscale.

4.3.1.3. Procedure

4.3.1.3.1. Variable Transformations

I performed three *a priori* variable transformations prior to the analysis. Within each country, I log transformed those continuous variables where skew/SE(skew) > |1|. This helped ensure that the regression assumption of normally distributed residuals was met for each separate country, and by extension the sample at-large. Also within countries, I converted all BMPN items to z-scores (M = 0; SD = 1). This meant that scores were relative to participants' countrymen, rather than absolute. An advantage of z-scores is that they eliminate the need to control for country-level main effects because there is no actual variation in country means for each BMPN item.[2] In support, after z-scoring all ICCs were approximately zero. Finally, another benefit was that z-scores held item effect interpretations constant. A unit change always equated to an SD change, relative to participants' countrymen. As a final step, I then also z-scored BMPN items again, this time across countries. This ensured that they all had exactly equal influence when they were summed to create the emergent factor scores. This was appropriate because BMPN items were designed to capture different sub-components of SWB. If some items had more variability than others, they would have been *overrepresented* in any emergent, aggregate, scores.

4.3.1.3.2. Exploratory Factor Analysis

I also selected the EFA approach *a priori*. It was intended to minimise residuals—rather than prioritize fit for certain subgroups of items—because I was interested in the solution that explained the *maximum variation* in the BMPN. I optimized the factor solution using oblique rotation because it allows emergent factors to be correlated.[3] It is convention to use principal components analysis (PCA) to determine the optimum number of factors, and then switch to the chosen factor analytic method for the primary analysis. However, PCA factors are always orthogonal and may thus fail to converge with a non-orthogonal factor analytic solution (Hox et al., 2017). I thus used the specified EFA to both determine the optimum number of factors, and to evaluate the final solution.

[2] While it was possible that individual*country level interactions still had effects on SWB, I discounted them because (a) they were likely very small, and (b) I was primarily focussed on the individual-level phenomena.
[3] If the factors are uncorrelated, oblique rotation is identical to 'Varimax' rotation.

4.3.1.3.3. Optimum Factor Structure

There is no gold-standard method to evaluate the optimum number of BMPN factors. Thus, I triangulated Kaiser's eigenvalue criterion, Cattell's scree plot criterion and Velicer's minimum average partial (MAP) test (Cohen et al., 2013).[4] Kaiser's and Cattell's criteria are predicated on eigenvalues, which are the extent that using emergent factor weights to aggregate item responses explains total variation in their scores, proportional to unweighted aggregation. Thus, eigenvalues > 1 explain more item covariation than a factor where all items have weights of |1|. The first factor has the highest eigenvalue, followed by the second factor, and so forth.

The three techniques applied different rules to determine the optimum number of factors. Kaiser's criterion is to simply select the factors with eigenvalues > 1. Thus, it only introduces added complexity when there is also added explanatory power. Cattell's criterion for inspecting the scree plot is to select only factors to the left of the inflection point—or the point where the trend between factor number (IV) and eigenvalue (DV) changes from steeply negative to shallowly negative. Cattell's criterion thus identifies the superordinate factors that explain much more item covariation than the remaining factors. For the present study, I defined the inflection point as the first factor where the upper bound 99% CI eigenvalue was < 20% of the *lower bound* 99% CI of the eigenvalue from the first factor. Finally, I also used Velicer's MAP test. It fits separate models for each possible factor solution. Then, it calculates all bivariate item correlations after removing variance explained by the emergent factors. It selects the factor solution that yields the smallest average squared value of these correlations. As such, it suggests the solution where the residuals are most unrelated to one another, and thus unlikely to belong to another unaccounted-for latent factor. Overall, all three techniques used different criteria to find the most parsimonious emergent factor structure in the BMPN.

4.3.1.3.4. Bootstrapped Factor Solution

Like other parametric statistical approaches, EFA is also subject to sampling error. To quantify the error around each of my eigenvalue and MAP estimates, as well as my final factor loadings, I evaluated the optimum factor structure in 1,000 bootstrapped samples of participants. In each iteration, participants were randomly sampled with replacement to match the total sample size.

[4] I opted not to perform parallel analysis—another prevailing method to determine the optimal EFA solution—because it defined non-spurious factors as those that had larger eigenvalues than the upper-bound CI of the eigenvalue for the corresponding factor in a dataset comprising random simulated and uncorrelated item responses. My sample size was so large that the upper bound CI and point estimate random eigenvalues converged even when using even stringent 99% CIs. Thus, parallel analysis was likely to yield spurious BMPN factors because even factors with marginal eigenvalues were retained.

Bootstrapping has been shown to generate accurate error margins across a variety of factor analytic and multiple regression approaches (Larsen & Warne, 2010). One thousand samples is sufficient to activate central limit theorem, which almost guarantees that results are normally distributed and can thus be summarized by the mean (Abranovic, 1997). My 99% CIs were 0.5^{th} and 99.5^{th} percentile bootstrapped factor loadings. As such, bootstrapping meant I could evaluate the plausible range of true population estimates.

4.3.2. Results

I used factor analysis to establish the emergent structure of the BMPN. When evaluating scales derived from a pre-existing theory—such as the BMPN from SDT—researchers often use confirmatory factor analysis because it seeks evidence for a predetermined structure over and above plausible alternatives. However, I switched to EFA to evaluate the *emergent structure* of the BMPN. This meant items were free to cluster together into one, two or other alternative superordinate SWB factors. I used a triangulated approach to determine the optimum number of factors, and then evaluated factor loadings. Overall, I determined the factor solution that accounted for the largest possible variation in items that did not also introduce additional, unnecessary, factors.

4.3.2.1. Optimum Number of Factors

I evaluated the optimum number of factors using Kaiser's Criterion, Cattell's criterion and Velicer's MAP test. They were all derived from bootstrapped EFA that minimized residuals and allowed factors to be correlated. Plots of the eigenvalues and MAP values are in Figures 4.1a and 4.1b respectively. According to Kaiser's criterion, only the eigenvalues for factors one (M = 4.00; CI = 3.92, 4.09) and two (M = 2.14; CI = 2.07, 2.21) were greater than the threshold of one. According to Cattell's criterion, only the variation explained by the second factor as a proportion of the first factor (59%) was greater than 1/5. Variation explained by the third (17%), fourth (11%), fifth (6%) and remaining factors decreased progressively, below the 20% threshold. Finally, Velicer's MAP test also yielded a two-factor solution. Both the point estimate MAP correlation and its CI (MAP = .012; CI = .012, .012) were smaller than the next-best, three factor, solution (MAP = .014; CI = .014, .014). As such, the triangulated results all converged to suggest that the two-factor solution (a) explained more item covariation than a default single factor, (b) explained substantially more item covariation than subsequent factors, and (c) minimized the chances that item residuals collapsed into an unaccounted-for, factor.

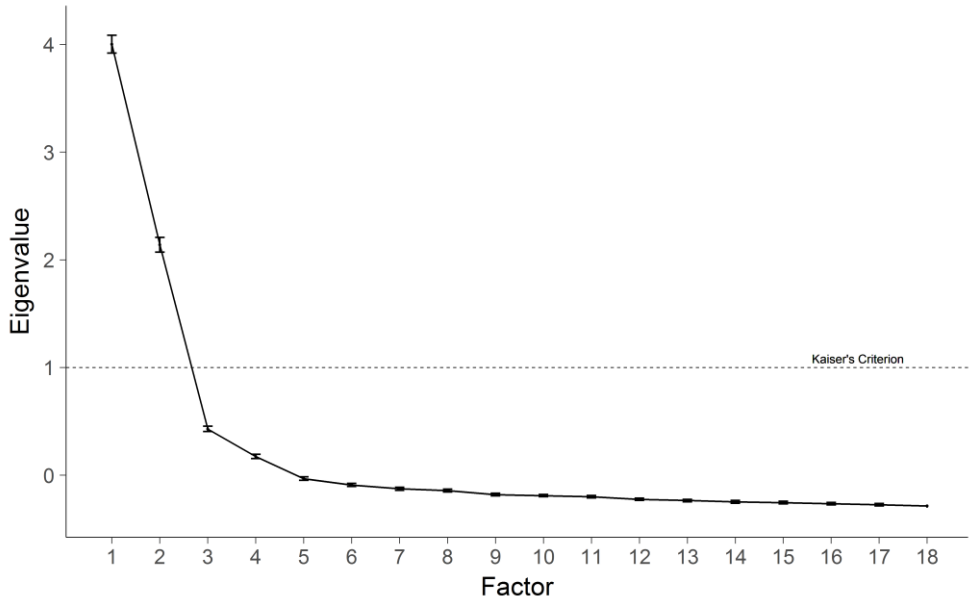


Figure 4.1a. Eigenvalues for each exploratory BMPN factor. Factor is the number of designated factors using oblimin factor analysis—which allows for correlated factors—with minimized residuals. Eigenvlaues are the added explanatory power of using the weights from each pre-defined factor number solution, proportional to the unweighted solution. The dashed line is Kaiser's Criterion, suggesting factors are only retained when values > 1.

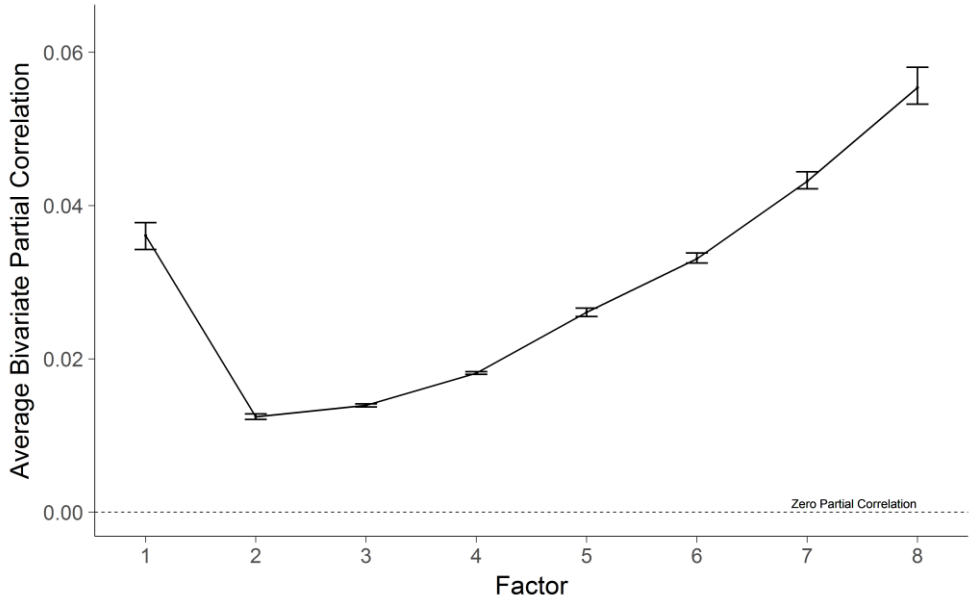


Figure 4.1b. Velicer's MAP test to determine the optimum number of BMPN factors. Factor is the number of designated factors using oblimin factor analysis—which allows for correlated factors—with minimized residuals. Average squared bivariate partial correlations are the absolute R-values after removing all item variation attributable to the designated factors. The dashed line reflects zero partial item correlations, which meant that items were completely unrelated after taking the superordinate factors into account.

4.3.2.2. Exploratory Factor Analysis Solution

In the two-factor solution, I theorized that BMPN items would cluster together to measure psychological needs satisfaction and thwarting. To this end, I performed the final bootstrapped EFA again, specifying two factors. Results are in Table 4.3. All items that measured the presence of needs satisfaction loaded positively onto the first factor and negatively onto the second factor. Each factor one loading was larger—in absolute terms—than the corresponding factor two loading. Eight of the nine items that measured the presence of needs thwarting loaded negatively onto the first factor and positively onto the second factor. The only exception was "I am struggling doing something I should be good at", which had a negligible positive loading on the first factor and a strong positive loading on the second factor. Thus, it still unambiguously loaded onto only the second factor. Each needs thwarting factor two item loading was larger—in absolute terms—than the corresponding factor one loading. Thus, they showed the exact inverse pattern of loadings. When I set the threshold loading for item retention to |0.40|, all items measuring needs satisfaction were retained in factor 1 and all items measuring needs thwarting were retained in factor 2. Overall, item loadings thus suggested that the first factor exclusively captured needs satisfaction, and the second exclusively captured needs thwarting. Differences in loading magnitude suggested that each item measured only one of the constructs, rather than both.

Table 4.3

EFA loadings for each BMPN item

BMPN Item	F1: Satisfaction	F2: Thwarting
I feel a sense of contact with people who care for me…	**0.58 (0.56, 0.59)**	-0.16 (-0.17, -0.14)
I am lonely.	-0.3 (-0.32, -0.29)	**0.62 (0.6, 0.63)**
I feel close … with other people who are important to me.	**0.6 (0.59, 0.62)**	-0.2 (-0.22, -0.18)
I feel unappreciated by one or more important people.	-0.18 (-0.19, -0.16)	**0.62 (0.6, 0.63)**
I feel … intimacy with the people I spend time with.	**0.44 (0.42, 0.46)**	-0.01 (-0.03, 0.01)
I have … conflicts with people I usually get along with.	-0.06 (-0.08, -0.05)	**0.51 (0.49, 0.52)**
I am successfully completing difficult tasks and projects.	**0.69 (0.67, 0.7)**	-0.22 (-0.24, -0.2)
I am experiencing … failure … at something.	-0.28 (-0.3, -0.27)	**0.68 (0.67, 0.69)**
I take on and master hard challenges.	**0.69 (0.67, 0.7)**	-0.16 (-0.18, -0.15)
I do some stupid things that make me feel incompetent.	-0.22 (-0.24, -0.21)	**0.63 (0.62, 0.65)**
I do well even at hard things.	**0.67 (0.65, 0.68)**	-0.18 (-0.2, -0.16)
I am struggling doing something I should be good at.	0.04 (0.02, 0.06)	**0.42 (0.4, 0.43)**
I am free to do things my own way.	**0.54 (0.53, 0.56)**	-0.23 (-0.24, -0.21)
I have a lot of pressures I could do without.	-0.1 (-0.12, -0.08)	**0.58 (0.57, 0.59)**
My choices express my "true self".	**0.59 (0.57, 0.6)**	-0.16 (-0.17, -0.14)
There are people telling me what I have to do.	-0.1 (-0.12, -0.09)	**0.54 (0.52, 0.55)**
I am really doing what interests me.	**0.6 (0.58, 0.61)**	-0.25 (-0.26, -0.23)
I have to do things against my will.	-0.17 (-0.19, -0.15)	**0.59 (0.58, 0.61)**

Notes. BMPN = Balanced measure of psychological needs

4.4. Study 2

In Study 1, I found support for two separate BMPN factors comprising psychological needs satisfaction and thwarting. In Study 2, I evaluated their construct validity. At the outset, I was interested in the simple direction of linear regression effects. Then, I switched to examine the *differential magnitudes* of associations between the emergent BMPN factors and various other SWB outcomes. Doing so better isolated the different underlying SWB components that needs satisfaction and thwarting captured, such as valence and transitiveness. An additional benefit was that I could also rule out the possibility that needs thwarting—which fully comprised items originally intended to be reverse scored—was not simply a methodological artefact.

4.4.1. Method

Construct validity was evaluated using the same participants as Study 1. Participant characteristics—other than sex and age—relevant to the present study were binary variables about ethnic minority (N = 17,452; 18%; ICC = 11%) and heterosexuality (N = 27,982; 87%; ICC = 2%). I used sex, age, ethnicity and sexuality to evaluate the construct validity of needs satisfaction and thwarting in specific biased sub-populations, which were taken from the larger sample. In addition to the BMPN, I used SLS and all other partial SWB variables completed by at least 10,000 participants. In some cases, variables may have been administered to participants from a subsample of countries that were unrepresentative of the entire population sampled. Nevertheless, I mitigated the risk of confounding by evaluating convergent effects from variables in different subsamples.

4.4.1.1. Materials

There were six categories of scales: BMPN, life satisfaction, other SWB scales, structural markers, lifestyle markers and response bias. Descriptive statistics and associations with criterion SLS are in Table 4.4. All continuous scores were normally distributed and there was no evidence for floor or ceiling effects. Internal consistency was always above $\alpha = .70$. Both binary variables—religiosity and relationship status—had at least 38% of cases in the smallest category. All continuous and binary variables also mostly varied at the individual level. Thus, it was appropriate to use them in linear regression.

Table 4.4.

Descriptive statistics for SLS, other SWB scales, structural markers, lifestyle markers and response bias

Category	Variable	C	N	Mean (SD)	α	ICC	R_{SLS} (99% CI)
Life Satisfaction	Satisfaction with life	28	28,089	4.39 (1.43)	.87	9%	-
Convergent Scales	Positive affect	14	13,406	3.41 (0.78)	.91	12%	0.37 (0.35, 0.39)
	Negative affect	14	13,406	2.33 (0.86)	.90	6%	-0.21 (-0.23, -0.19)
	Cheerfulness	28	28,372	5.22 (1.19)	.79	15%	0.54 (0.53, 0.55)
	Depression	28	28,372	3.14 (1.35)	.76	8%	-0.47 (-0.48, -0.45)
	Happiness	12	11,138	4.69 (1.25)	.75	9%	0.61 (0.59, 0.63)
Structural Marker	Social Class	28	27,743	41.52 (22.6)	-	10%	0.29 (0.27, 0.3)
	Religiosity	28	27,694	40%	-	15%	0.15 (0.14, 0.17)
Varying Marker	Physical Health	28	27,913	63.55 (23.8)	-	7%	0.34 (0.33, 0.36)
	Fruit and Veggies	28	27,864	67.71 (23.97)	-	4%	0.18 (0.17, 0.2)
	Exercise Frequency	28	27,860	48.79 (31.11)	-	6%	0.22 (0.2, 0.23)
	Household Income	17	17,061	39.7 (23.42)	-	2%	0.28 (0.26, 0.3)
	Relationship Status	28	27,498	62%	-	2%	0.16 (0.15, 0.18)
Response Bias		28	28,372	4.18 (0.47)	-	7%	0.18 (0.17, 0.2)

Notes. C = Number of Countries. ICC = Intraclass Correlation Coefficient. SLS = Satisfaction with life scale. Fruit and veggies = Fruit and vegetable consumption.

4.4.1.1.1. Needs Satisfaction and Thwarting

I summed the nine items from the BMPN that were retained in each of the emergent factors. Both needs satisfaction (α = .83) and needs deprivation (α = .82) factors had good internal consistency. For comparison, I used the Spearman-Brown correction to evaluate the projected internal consistency when the three originally six-item BMPN subscales were expanded to have nine items. Corrected internal consistency was still markedly lower for autonomy (α = .73), competence (α = .74) and relatedness (α = .71). Thus, newly-obtained needs satisfaction and thwarting from Study 1 may have more consistently captured their respective underlying SWB factors, compared to the existing factor structure. I further evaluate this claim in Chapter 5, using multigroup confirmatory factor analysis. Both needs satisfaction (M = 0; SD = 0.66; ICC < .01) and thwarting (M = 0; SD = 0.64; ICC < .01) were normally distributed, showed no evidence of floor or ceiling effects and varied exclusively at the individual level.[5] The two factors had a robust small to moderate negative association (R = -.22; CI = -.24, -.21; T(28,370) = -38.89; p < .001). This suggested that people experiencing more needs satisfaction also experience less needs thwarting on average, and vice versa. That said, there was only 5% shared variation. This suggested that the factors also captured largely distinct aspects of SWB.

[5] Final scores comprised items that had already been transformed to have equal influence and capture participants' scores relative to their countrymen. Thus, they all had mean ≈ 0, and SD ≈ 1.

4.4.1.1.2. Life Satisfaction

I measured life satisfaction with SLS (Diener et al., 1985). It has demonstrated construct validity across a variety of different cultures and language groups. It is also associated with a plurality of sociological, health and psychological outcomes (Diener et al., 2018). SLS comprises five items, which are each rated on a 7-point Likert scale (1 = "Strongly disagree", 7 = "Strongly agree"). An example item is "I am satisfied with my life". For the present study, I conceptualized it as the criterion—or established gold-standard—measure of SWB.

4.4.1.1.3. Other Scales

The other scales—positive affect, negative affect, cheerfulness, depression and happiness—all measured predominantly affective components of SWB. Positive affect and negative affect were each measured by rating 10 emotion words from the PANAS scale (Watson et al., 1988). Participants were asked how frequently they had experienced each emotion "in the past few weeks" on a five-point Likert scale (1 = "Very slightly or not at all"; 5 = "Extremely"). An example item for positive affect is "Excited". An example item for negative affect is "Scared". Cheerfulness and depression were constituent facets of NEO-IPIP-120 extraversion and neuroticism factors, respectively (Johnson, 2014). In the present study, I measured both cheerfulness and depression using their four item sub-scales (both rated on 7-point Likert scales;[6] 1 = "strongly disagree", 7 = "strongly agree"). Although they are big five facets, I used them here because they also capture the direct sensitivity to experience the affective components of SWB (Schimmack et al., 2004). An example item for cheerfulness is "Radiate joy". An example item for depression is "Dislike myself". Happiness was measured with the first two items from the four-item subjective happiness scale (Lyubomirsky & Lepper, 1999).[7] They were both rated on seven-point Likert scales (1 = "Not a very happy person"; 7 = "A very happy person"). The items were "In general, I consider myself…" and "Compared with most of my peers, I consider myself…"). The positively valanced convergent measures—positive affect, cheerfulness and happiness—were all positively associated with SLS. The negatively valanced convergent measures—negative affect and depression—were both negatively

[6] In Chapter 3, the NEO-PI-R was measured using a 5-point Likert scale. However, in subsequent country waves I used a seven-point scale. When using this more comprehensive data, I thus transformed the minimum and maximum item scores from the first wave to one and seven respectively when evaluating raw descriptive statistics. It was unlikely that different Likert scale lengths impacted actual results because personality was z-scored separately in each country for all the analyses.

[7] Translations failed for the remaining two items, likely because they had more complex sentence structures.

associated with SLS. Thus, there was evidence that all scores converged, in the intuitive directions, to measure latent aspects of SWB.

4.4.1.1.4. Structural Markers

The structural markers—social class and religiosity—were both sociodemographic variables that were (a) relatively stable during adulthood, and (b) have established positive associations with SWB, both within and across cultures (Verdugo, 2002; Heiphetz, Spelke & Banaji, 2013). Social class was measured with the single item "Where do you place yourself on the following spectrum of social class?" on a 100-point sliding scale (1 = "Working class"; 100 = "Upper class). Religiosity was measured with the single binary item "Do you currently practise a religion? (e.g. Pray, attend regular services)" (1 = "Yes"). Both structural markers were positively associated with SLS.

4.4.1.1.5. Lifestyle Markers

The lifestyle markers—physical health, fruit and vegetable consumption, exercise frequency and relationship status—were all sociodemographic variables that also had established positive associations with SWB (Walsh, 2011). They were all measured with single items. Physical health was "How do you rate your health in the past 12 months?" and was rated on a 100-point sliding scale (1 = "Very poor"; 100 = "Very good"). Fruit and vegetable consumption was "How often do you eat fruit and vegetables?". Exercise frequency was "About how often do you do at least 30 minutes of exercise?". Both were also rated on 100-point sliding scales (1 = "Almost never"; 100 = "Every day"). The final variable was binary scored relationship status: "Are you currently in a romantic relationship?" (1 = "Yes"). All lifestyle markers were positively associated with SLS. Thus, there was evidence that they proxied for SWB. An added benefit was that they were anchored to tangible, real-world, circumstances.

4.4.1.1.6. Response Bias

Response bias was the tendency for participants to preferentially respond using either the minimum or maximum Likert scale endpoints, regardless of question wording. It was the average of participants' responses to the 10/30 NEO-IPIP facets that had an equal balance of two positively and two negatively worded items.[8] I preferred using exclusively balanced factors

[8] Prior to calculating response bias, I imputed the 0.11% of item responses that were missing using the multiple imputation procedure listed in the 'Data Preparation' section below. For expedience, I performed the procedure with participants from all countries rather than participants from each country separately. Due to computational constraints, I performed multiple imputation using six equal sized blocks of 15 items. Each block comprised 3

because—unlike imbalanced factors—scores mitigated artefacts that could have been caused by idiosyncratic responses to single items. Low and high scores suggested participants favoured the minimum and maximum scale endpoints respectively. There was a positive association between response bias and SLS. This was unsurprising because people high in SWB may respond to survey scales more agreeably (DeNeve & Cooper, 1998). Thus, I evaluated BMPN factor associations after *controlling for* response bias.

4.4.1.2. Data Preparation

Within each country, I again log transformed those continuous variables where skew/SE(skew) > |1|. Also, within countries, I converted all continuous variables to z-scores (M = 0; SD = 1). This meant that all scores were relative to participants' countrymen, rather than absolute. After z-scoring, all ICCs were approximately zero. Then for each country, I simulated all missing variables scores—except for the BMPN—using multiple imputation by chained equations (MICE). In the first instance, MICE assigns random values to all missing scores. Then, it sequentially generates prediction models for each variable using predictive means matching (PMM). The first stage of PMM is linear regression for continuous variables and logistic regression for binary variables. It yields predicted values for *all* non-missing and missing scores. Then, each case that originally had a missing value is randomly assigned the true value from one of the five non-missing cases with the nearest predicted values. Finally, MICE iteratively repeats this procedure (e.g.) five times—the fidelity of predicted scores for each missing value increasing with every iteration—and then substitutes the final randomly assigned values into the original data. There is evidence that MICE better preserves both original variable skewness and realistic random response variability than either simple means substitution or the raw predicted values from just one iteration of PMM (Buuren & Groothuis-Oudshoorn, 2011). Its effectiveness tends to plateau at/beyond five iterations of PMM.

Then, I evaluated the extent that MICE produced appropriate imputed variable scores for the present data. For continuous variables, I evaluated the percentage overlap in histograms of non-imputed and imputed scores. For binary variables, I evaluated the absolute difference in the percentage of non-imputed vs imputed cases that had the variable, subtracted from one. Results are in Table 4.5. Overall, there were between 93% and > 99% non-missing cases for each variable, and convergence ranged from 83% to 100%. Convergence may have been imperfect

items—from different facets—in each factor. There were negligible missing values, and thus minimal opportunity for imputed scores to influence overall response bias.

because (a) of fluctuations in PMM value assignment, (b) outliers in the imputed data suppressed true convergence rates, or (c) there were real differences in the profiles of cases that had missing vs non-missing data. Considering the generally high response rates, convergence scores and potential mitigating circumstances, I decided that multiply imputed variable scores were sufficiently accurate to be included in the final analysis.

Table 4.5

Multiple imputation diagnostics by variable

Variable	% Non-Missing	% Convergence
Age	93%	97%
Cheerfulness	> 99%	92%
Depression	> 99%	95%
Exercise Frequency	98%	85%
Fruit and Veggies	98%	85%
Happiness	98%	92%
Heterosexual	97%	97%
Male	97%	99%
Minority	98%	96%
Negative Affect	98%	94%
Physical Health	98%	83%
Positive Affect	98%	92%
Relationship Status	97%	100%
Religiosity	98%	96%
Satisfaction with Life	99%	91%
Social Class	98%	85%

Notes. % Non-Missing = The percentage of cases that originally responded to each variable. % Convergence = Overlap in kernel density plots of non-imputed and imputed scores for continuous variables, and one minus the absolute difference in the percentage prevalences of non-imputed vs imputed scores for binary cases.

4.4.1.3. Procedure

The aim of Study 2 was to evaluate the extent that BMPN needs satisfaction and thwarting captured real, construct valid, aspects of latent SWB. Construct validity is itself unobservable, and thus it must be inferred by triangulating indirect evidence. In the present study, I did this via convergent, discriminant, external, incremental and pragmatic validity.

Convergent validity was when the BMPN factors were associated with the other partial measures of SWB, both before and after controlling for response bias. For example, in needs

satisfaction convergent validity was when there were positive associations with the positively valanced scales, and negative associations with negatively valanced scales.

Discriminant validity was when the absolute magnitude of the convergent validity association for one BMPN factor was *larger* than the concomitant association for the other factor. For example, there was evidence for discriminant validity when the associations between needs satisfaction and other positively valanced SWB scales were *larger*—in absolute magnitude—than the associations between needs thwarting and these same scales. I expressed these as absolute ratios of point estimate associations and as conservative ratios—which used the 99% CI bounds of each association that were most likely to *reverse* the direction of effects. For example, values greater than one suggested associations were stronger for needs satisfaction than needs thwarting, and values less than one suggested the inverse. In another application, ratios greater than one suggested associations were larger for BMPN than SLS, and ratios less than one suggested in the inverse.

External validity was when there was convergent and discriminant validity for separate sub-samples comprising each of the 28 countries, and then also subsamples comprising exclusively women, men, the youngest-aged third, the middle-aged third, the oldest-aged third, non-minorities, minorities, non-heterosexuals and heterosexuals. Thus, results helped evaluate whether the heterogenous overall sample obfuscated incongruent sub-sample effects.

Pragmatic validity was when the BMPN measured more transitive aspects of SWB than SLS. Transitive variables were positive affect and negative affect, which both captured emotion experiences "… in the past few weeks". Challengingly, the raw magnitudes of associations for BMPN factor/s vs SLS could have been confounded by differences in overall scale measurement accuracy. Thus, I evaluated whether associations with transitive variables were larger than associations with the comparatively stable variables: happiness (experiences "In general…"), cheerfulness and depression (stable personality traits), and social class and religiosity (structural). For needs satisfaction, I focussed only the positively valanced measures—positive affect, happiness and cheerfulness—and social class and religiosity. For needs thwarting, I focussed only on the negatively valanced measures—negative affect and depression—and social class and religiosity. I evaluated all effects for SLS because it was originally designed to measure the whole of SWB.

Incremental validity was when BMPN factors explained real world phenomena over-and-above existing constructs. Thus, I evaluated the associations between BMPN and the lifestyle markers

of SWB after controlling for sex, age, response bias and SLS. I fit both needs satisfaction and needs thwarting in the same model, thus ruling out the possibility that results were driven by a single aggregate BMPN factor that was captured in their shared variation. This helped give further evidence for the parsimoniousness of the two-factor solution.

4.4.2. Results

I used five psychometric approaches to evaluate the construct validity of needs satisfaction and thwarting. My sample size was so large that—after adjusting for country-level variable skew—any remaining outliers likely had negligible leverage. Thus, I retained all available cases. Four study features also protected against the inflated Type 1 error that is potentially caused by performing multiple statistical tests. Specifically, I (1) evaluated aggregate effect *patterns*; (2) selected an error threshold ($p < .01$) that was conservative enough to conclude that most of each pattern was unlikely to be caused by sampling error; (3) used enough participants to negate the effects of random sample fluctuations; and, (4) likely used heterogenous enough participants to negate diverging sub-population effects.

4.4.2.1. Convergent Validity

First, I evaluated whether needs satisfaction and thwarting were associated with the other partial SWB scales, in the intuitive directions. To this end, I evaluated zero-order bivariate associations, as well as associations after controlling for response bias. Results are in Table 4.6. Needs satisfaction was positively associated with happiness, positive affect and cheerfulness, and it was negatively associated with negative affect and depression. Needs thwarting was negatively associated with happiness, positive affect and cheerfulness, and positively associated with negative affect and depression. Then, effect patterns and magnitudes were fully consistent after controlling for response bias. Thus, I concluded that both factors were robustly associated with convergent SWB scales.

4.4.2.2. Discriminant Validity

Next, I evaluated whether needs satisfaction and thwarting were differentially associated with positively and negatively valanced SWB. To this end, I computed absolute effect ratios by dividing each needs satisfaction association by its corresponding needs thwarting association. I repeated this process for the conservative ratio, using the CI bounds that were most likely to reverse the pattern of observed magnitudes. Ratios greater than one thus suggested that the needs satisfaction association was larger, and ratios less than one suggested that the needs

thwarting association was larger. Results are also in Table 4.6. Needs satisfaction was more strongly associated with happiness, positive affect and cheerfulness than needs thwarting. Needs thwarting was more strongly associated with negative affect and depression than needs satisfaction. All results held using both ratios and conservative ratios, both before and after controlling for response bias. Therefore, the needs satisfaction and thwarting factors were more associated with SWB experiences of flourishing and suffering respectively. Moreover, findings for needs thwarting suggested that the factor measured real underlying SWB; it did not emerge simply because its items were all originally designed to be reversed scored.

Table 4.6

Convergent and discriminant validity of psychological needs satisfaction and thwarting factors

BMPN Factor	Variable	Zero-order correlations		Controlling for response bias	
		R (99% CI)	Ratio (Cons)	r	Ratio (Cons)
Satisfaction	Happiness	0.51 (0.49, 0.53)	1.41 (1.28)	0.5 (0.49, 0.52)	1.16 (1.07)
	Positive Affect	0.54 (0.52, 0.55)	4.37 (3.61)	0.51 (0.49, 0.52)	2.28 (2.02)
	Cheerfulness	0.56 (0.55, 0.57)	1.99 (1.86)	0.53 (0.51, 0.54)	1.25 (1.19)
	Negative Affect	-0.24 (-0.26, -0.22)	0.39 (0.43)	-0.33 (-0.35, -0.31)	0.57 (0.62)
	Depression	-0.4 (-0.42, -0.39)	0.70 (0.74)	-0.47 (-0.48, -0.46)	0.85 (0.88)
Thwarting	Happiness	-0.36 (-0.38, -0.34)	-	-0.43 (-0.45, -0.41)	-
	Positive Affect	-0.12 (-0.14, -0.1)	-	-0.22 (-0.24, -0.2)	-
	Cheerfulness	-0.28 (-0.29, -0.27)	-	-0.42 (-0.43, -0.41)	-
	Negative Affect	0.61 (0.59, 0.62)	-	0.57 (0.55, 0.58)	-
	Depression	0.57 (0.56, 0.58)	-	0.55 (0.54, 0.56)	-

Notes. BMPN = Balanced measure of psychological needs. Variable: split by whether the convergent measure captured aspects of flourishing (positive) or suffering (negative). Ratio = The absolute ratio of each needs satisfaction vs thwarting association. Ratios > 1 suggested effects were larger for needs satisfaction, and vice versa. Cons = The absolute ratio of the CIs for needs satisfaction and thwarting effects that were most likely to cross the threshold of one, thus giving evidence against the discriminant validity of the BMPN factors.

4.4.2.3. External Validity

Next, I evaluated whether convergent and discriminant validity held when using deliberately unrepresentative participants. To this end, I replicated sample-wide findings separately for participants in each country, and separately again for women and men, the youngest, middle and oldest participants, non-minorities and minorities, and non-heterosexuals and heterosexuals. For simplicity, I focussed on zero-order associations. Results are in Table 4.7. For countries, I first found that internal consistency (Cronbach's alpha) for needs satisfaction (M = .84; CI = .83, .85) and needs thwarting factors (M = .80; CI = .77, .81) were both consistently strong. Then, I replicated the full pattern of convergent and discriminant associations for both factors. There was matching support for BMPN external validity using biased demographic samples. Internal consistency for needs satisfaction (M = .84; CI = .83,

.85) and needs thwarting factors (M = .82; CI = .81, .83) were both consistently strong. Then, I again replicated the full pattern of convergent and discriminant associations. Overall, results suggested that sample-wide effects did not obfuscate diverging sub-population effects.

Table 4.7

Convergent and discriminant validity of psychological needs satisfaction and thwarting factors for different, unrepresentative, sub-populations

Samples	BMPN Factor	Variable	R (99% CI)	Ratio (Cons. ratio)
Countries	Satisfaction	Happiness	0.51 (0.44, 0.57)	4.34 (1.33)
		Positive Affect	0.53 (0.46, 0.59)	12.02 (2.48)
		Cheerfulness	0.55 (0.49, 0.61)	4.86 (1.62)
		Negative Affect	-0.22 (-0.3, -0.14)	0.37 (0.57)
		Depression	-0.39 (-0.46, -0.32)	0.69 (0.92)
	Thwarting	Happiness	-0.36 (-0.43, -0.28)	-
		Positive Affect	-0.09 (-0.17, 0)	-
		Cheerfulness	-0.27 (-0.35, -0.19)	-
		Negative Affect	0.6 (0.54, 0.65)	-
		Depression	0.57 (0.51, 0.62)	-
Demographics	Satisfaction	Happiness	0.51 (0.48, 0.54)	1.55 (1.29)
		Positive Affect	0.54 (0.51, 0.56)	5.81 (3.62)
		Cheerfulness	0.56 (0.54, 0.58)	2.26 (1.93)
		Negative Affect	-0.22 (-0.25, -0.18)	0.36 (0.44)
		Depression	-0.39 (-0.41, -0.36)	0.68 (0.75)
	Thwarting	Happiness	-0.34 (-0.38, -0.31)	-
		Positive Affect	-0.11 (-0.15, -0.07)	-
		Cheerfulness	-0.26 (-0.29, -0.24)	-
		Negative Affect	0.6 (0.58, 0.63)	-
		Depression	0.57 (0.55, 0.58)	-

Notes. Countries = Each of the 28 countries sampled separately; Demographics = samples using just men, women, young, middle aged, old, minority, non-minority, heterosexual and non-heterosexual cases. Valence = whether the convergent measure captured negative/positive SWB. BMPN factors were only validated against scales with the corresponding valence. The 99% CI was the mean ± 2.58*SE because the sample size of effects was too small to provide bootstrapped CIs. Ratio = the absolute ratio of needs satisfaction to thwarting associations. Cons. ratio = the absolute ratio of the CIs for needs satisfaction effects that yielded values most likely to cross the threshold of one, and thus provide evidence against the discriminant validity of the BMPN scales.

4.4.2.4. Pragmatic Validity

Next, I evaluated whether needs satisfaction and thwarting were more associated with transitive than stable SWB, compared to SLS. As a preliminary step, I evaluated bivariate associations between the BMPN factors and SLS. There was a robust positive association for needs satisfaction (r = .49; CI = .58, .50; t(28,370) = 93.92, p < .001) and a robust negative association for needs thwarting (r = -.30; CI = -.32, -.29; t(28,370) = -53.57, p < .001). The magnitude of the needs satisfaction association may have been larger because SLS measured more positively than negatively valanced SWB.

Both BMPN factors were more associated with transitive SWB than stable SWB, compared to SLS. For needs satisfaction, I computed the ratio of associations between transitive positive affect and each of the other more stable positively valanced and structural measures of SWB. I did likewise for needs thwarting with transitive negative affect and each of the other more stable negatively valanced and structural measures of SWB. Then, I computed corresponding ratios for SLS. Finally, I divided each BMPN ratio by the corresponding SLS ratio. Thus, the ratios-of-ratios totally controlled for differences in scale accuracy, as well as different overall magnitudes of bivariate associations between BMPN and SLS. Scores greater than one suggested the BMPN had stronger relative associations with transitive SWB than life satisfaction, and scores less than one suggested the inverse. Results are in Table 4.8. Overall, both the ratio-of-ratio point estimate and conservative estimate—which again used the CI bounds that were most likely to reverse effects—were consistently above one for both needs satisfaction and needs thwarting. Thus, there was also consistent evidence that the BMPN factors captured more transitive aspects of SWB than SLS.

Table 4.8

Relative magnitude of needs satisfaction and thwarting associations for transitive vs stable convergent measures of SWB, compared to SLS associations

DV	Transitiveness	Convergent Measure	R (CI)	Num.	Rat. (Cons)	**R-of-R (Cons)**
NS	Transitive	Positive Affect	0.54 (0.52, 0.55)	-	-	-
	Stable	Cheer	0.56 (0.55, 0.57)	PA	0.96 (1)	**1.34 (1.32)**
		Happiness	0.51 (0.49, 0.53)	PA	1.06 (0.98)	**1.79 (1.56)**
		Social Class	0.16 (0.15, 0.18)	PA	3.38 (2.89)	**2.53 (2.46)**
		Religious	0.08 (0.06, 0.09)	PA	6.75 (5.78)	**2.06 (2.21)**
NT	Transitive	Negative Affect	0.61 (0.59, 0.62)	-	-	-
	Stable	Depression	0.57 (0.56, 0.58)	NA	1.07 (1.02)	**2.19 (1.81)**
		Social Class	-0.08 (-0.1, -0.07)	NA	7.62 (5.9)	**9.8 (6.67)**
		Religious	-0.02 (-0.03, 0)	NA	30.5 (19.67)	**15.98 (13.46)**
SLS	Transitive	Positive Affect	0.36 (0.34, 0.37)	-	-	-
		Negative Affect	-0.21 (-0.23, -0.19)	-	-	-
	Stable	Cheer	0.5 (0.49, 0.51)	PA	0.72 (0.76)	-
		Happiness	0.61 (0.59, 0.62)	PA	0.59 (0.63)	-
		Depression	-0.43 (-0.44, -0.41)	NA	0.49 (0.56)	-
		Social Class	0.27 (0.26, 0.29)	PA	1.33 (1.17)	-
				NA	0.78 (0.88)	-
		Religious	0.11 (0.1, 0.13)	PA	3.27 (2.62)	-
				NA	1.91 (1.46)	-

Notes. DV: NS = Needs satisfaction, NT = Needs thwarting, SLS = Satisfaction with life scale. Num. = Numerator used in calculating the ratio and ratio-of-ratios. It was always one of the two transitive convergent measures of SWB: positive affect or negative affect. The denominator was always the comparatively stable convergent measure. I computed an exhaustive set of pairwise ratios with transitive and stable effects, but only for those variables that had matching valence. Thus, the effect for positive affect was calculated relative to cheerfulness and happiness, and the effect for negative affect was calculated relative to depression. Social class and religiosity were relative to both positive and negative affect. Ratios were absolute. Cons = The absolute ratio of the CIs for needs satisfaction and thwarting, and SLS, effects that yielded values most likely to cross one. R-of-R = Needs satisfaction and deprivation thwarting ratios, divided by their corresponding satisfaction with life ratios. Scores greater than one meant the BMPN captured more transitive SWB than SLS. Conservative R-of-R's could sometimes be larger than the point estimates, when the widths of CIs for BMPN and SLS associations diverged.

4.4.2.5. Incremental Validity

Finally, I evaluated whether needs satisfaction and thwarting were associated with concrete lifestyle outcomes over and above other existing measures. To this end, I fit single multiple regression models for each outcome—physical health, fruit and vegetable consumption, exercise frequency and relationship status—controlling for sex, age, response bias, satisfaction with life and the shared variation in BMPN factors. Results are in Table 4.9. As a preliminary step, I evaluated the effects for just controls on physical health, which was the most heterogeneous of the target outcomes. Overall, being male and increasing SLS were both positively associated with health. Age and response bias were negatively associated with physical health. Then, I evaluated needs satisfaction and thwarting effects for each lifestyle outcome, with controls. There were consistent positive associations for needs satisfaction and consistent negative associations for needs thwarting. Therefore, there was also evidence that

both factors explained unique aspects of SWB that were not captured by criterion SLS, the controls or response bias.

Table 4.9

Incremental validity of needs satisfaction and thwarting associations

Outcome	Variable	B (99% CI)	T-value	P-Value
Physical Health	Intercept	-0.08 (-0.1, -0.06)	-9.55	< .001
	Age	-0.07 (-0.08, -0.05)	-11.71	< .001
	Response bias	-0.04 (-0.06, -0.03)	-7.17	< .001
	Sex	0.15 (0.12, 0.18)	13.17	< .001
	Satisfaction with life	0.32 (0.31, 0.34)	56.58	< .001
Physical Health	Satisfaction	0.17 (0.14, 0.2)	17.00	< .001
	Thwarting	-0.21 (-0.24, -0.19)	-20.59	< .001
Fruit and Vegetable Consumption	Satisfaction	0.21 (0.18, 0.23)	19.79	< .001
	Thwarting	-0.11 (-0.14, -0.09)	-10.51	< .001
Exercise Frequency	Satisfaction	0.19 (0.16, 0.21)	17.76	< .001
	Thwarting	-0.07 (-0.1, -0.04)	-6.38	< .001
Relationship Status	Satisfaction	0.1 (0.04, 0.16)	4.35	< .001
	Thwarting	-0.1 (-0.16, -0.04)	-4.21	< .001

Notes. SLS = Satisfaction with life. Control variables: age, response bias, sex and SLS. For expedience, the intercept and control effects were only reported for physical health, without including the BMPN. All BMPN effects were reported after fitting controls.

4.5. Discussion

In this chapter, I evaluated whether the BMPN could be rescored to measure overall SWB. I relaxed the assumption that it comprised three separate factors—autonomy, competence and relatedness—to instead find its endemic structure, using EFA. There was consistent evidence that two factors explained (a) item covariation better than a single factor that simply added all the items together, and (b) much *more* item covariation than subsequent factors. Two factors were also (c) least likely to produce additional unaccounted-for factors. Then, in the two-factor solution, I found that items clustered together to measure separate psychological needs satisfaction and psychological needs thwarting. All items loaded *more strongly* onto their own factor than the other factor. Moreover, 15/16 items were negatively associated with the opposing factor. This suggested that they measured largely distinct facets of SWB.

Then, I evaluated whether needs satisfaction and thwarting captured construct valid SWB. First, I evaluated the extent that they were associated with other, partial, SWB scales. Needs satisfaction was positively associated with positive affect, cheerfulness and happiness, and negatively associated with negative affect and depression. Needs thwarting was negatively associated with positive affect, cheerfulness and happiness, and positively associated with

negative affect and depression. Thus, all effects were exactly opposite and in the expected directions. Then, I found that needs satisfaction was *more strongly* associated with the positively valanced SWB scales than needs thwarting, and vice versa. Results held comparing both point estimate associations, and the 99% CI bounds of each association that were most likely to reverse the observed magnitudes. Thus, needs satisfaction and thwarting may disproportionately capture aspects of SWB associated with flourishing and suffering, respectively. All these convergent and discriminant associations emerged in both the sample at large, and in participants from each separate country and a variety of unrepresentative demographic strata (e.g. just young adults, just ethnic minorities). Therefore, the large heterogeneous sample also did not obfuscate diverging effects for noteworthy sub-populations.

Finally, I evaluated the utility of the BMPN factors compared to existing measures. I evaluated whether it captured more transitive SWB than life satisfaction. First, I examined the magnitude of associations between needs satisfaction and positive affect—and needs thwarting and negative affect—relative to associations from the more stable SWB variables and structural markers. Then, I evaluated each ratio relative to the corresponding ratio involving SLS associations. In every instance, both the point estimate and conservative ratios suggested that the BMPN factors were more strongly associated with transitive SWB than SLS. Finally, I evaluated whether the BMPN was associated with lifestyle outcomes even after controlling for possible confounds, as well as SLS. Specifically, I controlled for age, sex, response bias, SLS and the variance shared by both BMPN factors. Overall, there was fully consistent evidence that needs satisfaction was positively associated—and needs thwarting was negatively associated—with physical health, fruit and vegetable consumption, exercise frequency and whether individuals were in a relationship. Thus, there was consistent evidence needs satisfaction and thwarting both explained more transitive SWB than life satisfaction, as well as unique variation in concrete lifestyle markers of SWB.

4.5.1. Implications

This chapter may have measurement, theoretical and research design implications. From a measurement perspective, there was evidence BMPN can be rescored to form two superordinate factors that capture overall psychological needs satisfaction and thwarting. When used in this way, the BMPN may be one of the few short scales that captures total SWB—at least according to SDT—rather than just one or a few biased sub-component. It also has demonstrated construct validity in heterogenous countries and languages (Linton, Dieppe & Medina-Lara, 2016; Chen et al., 2015). As such, researchers might expand the uses of the

BMPN, as well as potentially other psychological needs scales, to measure omnibus SWB. Theoretically, BMPN scales captured two largely *dissociable*, construct valid components of SWB. In previous studies, any emergent factor with just negatively phrased items was considered an artefact. However, I found that both factor scores—each with exclusively positively or negatively worded items—converged with other SWB scales even after controlling for response bias. Moreover, both had real and superior explanatory power when accounting for SWB phenomena that matched their valence.

Practically, researchers may use needs satisfaction and/or thwarting as their primary measure of SWB. Although SLS may be particularly good at capturing the well-being implications of long-term structural SWB factors—like ethnicity and social class—results in this chapter suggest that it is less sensitive to *transitive* SWB phenomena than BMPN. In some circumstances, using BMPN might thus help establish the correct temporal sequence between the predictor and outcome, which mitigates the risks of confounding and reverse causation. Results also suggested that BMPN explained unique variation in lifestyle phenomena after accounting for SLS. As such, researchers may prefer the BMPN in some contexts.

4.5.2. Limitations and Future Directions

Nevertheless, there are at least four limitations. First, I established convergent and discriminant validity using exclusively self-report. Effects could have been conflated by common method variance or participants' general lack of self-awareness. Second, I focussed on relative rather than absolute effect sizes. Thus, I could conclude BMPN factors were *more associated* with one aspect of SWB over another, but not that they objectively measured that aspect. Third, in pragmatic validity there was only one transitive variable—positive/negative affect—for both needs satisfaction and thwarting. This increased the risk of confounding. Finally, I only compared the utility of the BMPN to SLS. Although SLS is a prevailing measure of SWB, there may now be more comprehensive criterion measures, such as the scales of general well-being (Longo et al., 2017). Thus, I could only conclude that the BMPN may be preferable to SLS, and not necessarily another more comprehensive scale, in some research contexts.

Future research can help increase the certainty of findings. For example, the construct validity of BMPN needs satisfaction and thwarting factors would be improved by (a) using both self- and peer-reported convergent variables, (b) evaluating results using ratio scales—where zero-values have meaningful interpretations (e.g. dopamine level)—which allow researchers to evaluate absolute and not relative effect patterns, and (c) measuring outcomes longitudinally

to better establish the correct temporal sequence (e.g. of incremental validity associations). Researchers would also benefit from using a fuller suite of *representative* (rather than *ad hoc*) convergent SWB scales. This would help confirm whether BMPN *unbiasedly* captures the entire construct. Finally, it is unclear whether two superordinate two factor BMPN structure generalises to other psychological needs scales.

4.5.3. Conclusion

The rescored BMPN addresses the need for a comprehensive measure of SWB. A product of the mature literature on human motivation, it was originally designed to capture the separate feelings of autonomy, competence and relatedness. When aggregated, however, I found extremely consistent evidence that it also captures the superordinate SWB experiences of flourishing and suffering. Then, I also found that the BMPN captures more transitive SWB than life satisfaction, and uniquely predicts various lifestyle outcomes over and above plausible covariates. Finally, both the superordinate BMPN structure—and support for its construct validity—may be extremely consistent across a variety of populations and iterative computational approaches. As such, there is evidence it is an especially robust, and novel, operationalization of SWB.

Chapter 5

Propensity Score Matching Increases the Internal Validity of Big Five Facets Effects on SWB

5.1. Abstract

A large body of research has investigated how personality differences predict SWB. However, it is difficult to investigate the role of individual big five personality *facets* because they have complex intercorrelations both within *and* between factors. Thus, controlling for all potentially covarying facets might increase multicollinearity—when correlated facets cancel each other out despite having real associations with SWB—while relaxing controls risks confounding. I propose that propensity score matching (PSM) mitigates this tradeoff. PSM is a sampling strategy that selects participants who differ on the facet of interest but are similar across the remaining facets. Thus, it may hold potentially confounding facets relatively constant without risking multicollinearity. Using the large multinational AXA sample (N = 36,498), I found that PSM held covariates 74% to 80% more constant than zero-order correlations, preserved non-negligible effect sizes and replicated established neuroticism and extraversion *factor* associations better than multiple regression. PSM also better isolated individual facet effects than the prevailing machine learning alternative: elastic net regression. Therefore, I used PSM to isolate the full range of facet associations with both needs thwarting and needs satisfaction SWB, as well as convergent SLS and health. There were consistent and noteworthy ($r > .10$) negative associations for depression and vulnerability, and consistent and noteworthy positive associations for cheerfulness, friendliness, gregariousness, self-efficacy and self-discipline. There was also evidence for different effect patterns in both agreeableness (morality, cooperation, and altruism) and conscientiousness (cautiousness and achievement-striving) across needs thwarting and needs satisfaction. Overall, PSM might lessen the tradeoff between confounding and multicollinearity, and thus offer a more internally valid approach to describing bivariate facet-SWB effects than conventional zero-order correlations and multiple regressions, as well as machine learning.

5.2. Introduction

In the previous two chapters, I evaluated univariate measurement issues concerning the big five facet predictors and the SWB outcomes. Now, I switch to their bivariate associations. These are an essential first-step in most empirical psychology research. They are an efficient way of assessing, preliminarily, whether theoretically plausible associations manifest in real world populations. That is, they can help inform whether an effect is large enough to warrant further investigation (Grissom & Kim, 2005). Moreover, they might give some indication of relative effect magnitudes, which can help direct research infrastructure to the most promising phenomena. When studying the big five and SWB, high-specificity facet-level analyses may be especially interesting because their effects are feasibly guided by discrete mechanisms, which might thus yield both actionable theory and precise applied insights.

However, to date methodological artefacts have obfuscated true facet-SWB associations. Although the big five are nominally orthogonal, facets in different factors may form overlapping or superordinate structures (Saucier & Ostendorf, 1999; Musek, 2007). That is, they may be correlated both within *and* between factors. This increases the risk of confounding because associations may be attributable to a wide plurality of unaccounted-for facets. Existing research has attempted to limit confounding by using stepwise and multiple regression. Problematically, these cannot tolerate the full range of facet covariates, which often causes multicollinearity (Thompson, 1995; Cohen et al., 2013). Multicollinearity is when the variation shared between two or more facets is totally removed from the statistical model, such that the target facet (a) is no longer representative of its underlying psychological construct, and/or (b) disproportionately comprises random response error (Kraha, Turner, Nimon, Zientek & Henson, 2012). Multicollinearity may cause effects to artificially change, and often to shrink. The consequence is that existing research inconsistently links between two facets—one of which is depression—and at least eight facets to SWB. It also contradicts the wider range of effects found in adjacent literatures.

Thus, I adapt propensity score matching (PSM) to lessen this confounding-multicollinearity tradeoff. PSM is underpinned by a simple premise: a covariate cannot confound when it does not vary (Rosenbaum & Rubin, 1985). It involves selecting participants for analysis who differ on the primary variable of interest but are similar on the remaining covariates. To date, PSM has been mostly used to increase the internal validity of experimental studies, especially when there is small sample size (Lu & Lemeshow, 2018). For example, Gupta, Han, Mortal, Silveri and Turban (2018) recently used it to find that women CEOs are subject to more shareholder

dissent than matched male CEOs. The present chapter aims to evaluate whether PSM improves the internal validity of *continuous* personality facet associations with SWB without increasing multicollinearity, compared to feasible alternatives. Then, I use PSM to evaluate the full pattern of more internally valid big five facet associations with SWB.

5.2.1. Big Five Personality and SWB

The big five may capture a relatively universal structure of personality. It comprises putatively orthogonal neuroticism, extraversion, openness, agreeableness and conscientiousness factors. It was a consolidation of the three prevailing theoretical traits at the time, and the two additional traits that consistently emerged in natural language (McCrae & Costa, 2017). Then, the big five structure was demonstrated in the major countervailing models of personality (see John & Srivastava, 1999). Since then, it has emerged in most cultures and languages tested (e.g. McCrae & Terracciano, 2005). The big five is also hierarchical. Costa & MacCrae (1992) found each of the big five has six subordinate facets. For example, neuroticism is the constellation of anxiety, anger, depression, self-consciousness, immoderation and vulnerability. This facet structure has also been widely replicated (McCrae, Costa, Del Pilar, Rolland & Parker, 1998; McCrae & Allik, 2002). Thus, the big five factors and their facets are sufficiently universal to be considered a gold-standard conceptualization of personality.

There has been extensive research on the big five *factor* associations with SWB. DeNeve and Cooper (1998) conducted the first large-scale review. Using prevailing personality constructs at the time, they found a negative association for neuroticism and a positive association for positive affect, which was captured in extraversion and agreeableness. In their updated meta-analysis, Steel et al. (2008) focussed exclusively on big five questionnaires. They found that associations were larger than previously observed, and especially strong for neuroticism (-) and extraversion (+) across a variety of different well-being measures. There were also positive associations for agreeableness and conscientiousness. More recently, Soto (2015) confirmed these findings longitudinally, using a large (N > 16,000) nationally representative Australian sample. Lamers, Westerhof, Kovács & Bohlmeijer (2012)—also using nationally representative panel data, from the Netherlands—highlighted the differential big five effects for suffering and flourishing. They found that low neuroticism disproportionally protected against mental illness, and high extraversion and then agreeableness disproportionately promoted positive mental health. Overall, findings implicate 4/5 of personality in SWB.

However, factor-level findings are likely too high-bandwidth to either be actionable or isolate single mechanisms. High-specificity facet-level associations are the most comprehensive remedy. As I outlined in the General Introduction, existing stepwise approaches implicate trait depression, and then either cheerfulness or achievement striving (Schimmack et al., 2004; Quevedo & Abella, 2011). Contrastingly multiple regression implicates up to 8/30 facets from neuroticism, extraversion and conscientiousness (Albuquerque et al., 2010; Anglim & Grant, 2016). Adjacent literatures suggest the additional SWB benefits of prosocial cooperation and altruism—perhaps through building social capitol (Helliwell, 2006). Among others, they also highlight the benefits of agency, through assertiveness and self-efficacy—perhaps because these traits promote environmental mastery (Sheldon & Elliot, 1999). Overall, there are a range of existing documented effects.

5.2.2. Limitations in Existing Research

However, existing facet effects contradict each other. The only consensus is that trait depression reduces SWB. However, adjacent lines of research suggest more pluralistic associations. These contradictions may be caused by complex facet intercorrelations—both within and between the big five—that make it difficult to control for the full range of other, potentially confounding, personality facets.

Evidence for complex facet intercorrelations is that the factors are not actually orthogonal. During its conception, the big five hierarchical structure was challenged by circumplex approaches (Saucier & Ostendorf, 1999). Circumplex approaches suggest that facets may be differentially associated with their respective factors, and that each facet might uniquely covary with facets in different factors. For example, the influential agency-communion circumplex may capture predominantly aspects of conscientiousness and agreeableness respectively (Smith, Gallo, Goble, Ngu & Stark, 1998). Although the exact nature of the circumplex may depend on culture (Costa & McCrae, 1995), its prevalence still questions the strict hierarchical structure of the big five.

This criticism has since been corroborated by the emergence of a general factor. In their re-evaluation of two previous meta-analyses on the structure of personality, Rushton and Irwing (2008) found that a single meta-factor accounted for around 45% of total variation in the entire big five. Just's (2011) review found that each factor score had undesirable and desirable endpoints, and that responding was relatively constant across factors. In van der Linden, Dunkel and Petrides's (2016) updated review—which comprised self-reports, peer reports, and

observed behaviour—they found that a *single* personality factor consistently emerged across methodologies. It was even prevalent genetically (Riemann & Kandler, 2010). Thus, facets may cluster together both within *and between* factors. This may occur to such a degree that, in some circumstances, a general factor of personality more parsimoniously accounts for inter-facet variation than the separate big five factors.

Existing methods may thus be ill-equipped to account for the *full range* of facet confounds. As I mentioned in the General Introduction, the exact range of confounds may differ by the target facet, outcome and population of interest. When evaluating associations across entire personality, it is thus safest to control for all 29 facet covariates. However, such comprehensive controls remove almost all meaningful variation from the target facet. This may increase the preponderance of findings that are spuriously based on random errors (Cohen et al., 2013). To remedy, researchers often fit *a priori* but *partial* controls. In stepwise regression, they may arbitrarily assign theoretical precedence to (e.g.) the affect facets over commensurably-plausible agency facets (Thompson, 1995). Results are then self-fulfilling. Alternatively, they make potentially-false assumptions about the likeliest confounds in multiple regression. For example, they might neglect inter-facet confounds by only exerting intra-facet control. Conversely, they might control for all the superordinate big five factors. However, this completely removes the facet variance that *contributes* to the overarching factor, which is thus central to its construct validity. It is unfeasible to fully revert to findings from other, more distal, literatures because of fragmented operationalizations and controls, and because they require extra assumptions about how personality translates into manifest behaviour. Overall, most big five facet associations with SWB are still unclear.

5.2.3. The PSM Solution

In short, the problem with existing personality facet-SWB associations is sub-optimal internal validity. Associations may either involve facets that are confounded, or so degraded that they have lost all construct validity. In both cases, the consequence is untrustworthy effect estimates. These are especially compromised when the precise composition of appropriate controls differs according to both the target facet and outcome of interest, which produces unreliable *patterns* of effect magnitudes. This may be especially problematic during exploratory research, when the largest magnitude effects are considered the best candidates for more resource-intensive follow-up studies (Rozin, 2001). In such cases, any method—such as PSM—that incrementally boosts internal validity may help optimize resource allocation.

PSM mitigates the tradeoff between multicollinearity and confounding by matching participants across their covariates. At the outset, it generates propensity scores. They are predicted values from a logistic regression where the covariates—(e.g.) the 29 big five facets that are *not* the target—are the predictors and a binary version of the intended facet IV is the outcome. Thus, here propensity scores are *single numeric summaries* of the relationship between participants' target facet and remaining entire personality. Then, PSM matches pairs of participants who have similar propensity scores but different target facet scores.

For the big five, the numeric version of the target facet—with PSM weights—is then used to predict SWB. Whilst matching is likely imperfect, PSM may still be superior to unweighted alternatives when matched pairs' covariates partially cancel each other out without the need to explicitly fit other facets as controls. To minimise idiosyncratic matches, high scorers can be matched to multiple low scorers (or vice versa). Thus, low scorers with especially common covariates might be represented in multiple matches, while those with especially uncommon covariates might be discarded completely. Therefore, PSM is a sampling strategy that attempts to hold potential confounds constant, much as (e.g.) an experimenter might intend when using the *same* testing environment for all participants.

To date, PSM has been mostly used in fields outside social psychology. More specifically, PSM is a popular way of evaluating categorical effects when there is (a) small sample size and/or (b) quasi-experimental assignment to conditions. There, it is deployed as a means of raising internal validity to an acceptable minimum, to offset the potential confounding effects of non-random covariates. For example, Caliendo and Kopeinig's (2008) literature review found that PSM was commonly used to evaluate policy interventions. To this end, Hitt and Frei (2002) used PSM to find the effects of implementing online banking on company profitability, across matched country regions. PSM is also commonly used to evaluate medical interventions. For example, to find the merits of different surgery procedures (Appéré et al., 2017), and novel cancer drugs (Elshafei et al., 2018). However, existing research is largely confined to categorical predictors (there are some exceptions—although not in psychology—which I review below). Further, PSM may be equally or more effective in *large samples*, especially when there are too many plausible covariates to fit as explicit controls. In such conditions, sample size may increase affordances for the kind of extremely close matches that reduce the total number and/or leverage of problem covariates.

Indeed, PSM can feasibly be applied to the numeric personality facets. Although median splits—used when generating the propensity scores—have many pitfalls (MacCallum, Zhang, Preacher & Rucker, 2002), they might be appropriate in PSM because they only generate sample weights, and *not* final associations. Further, propensity scores can be generated using *multiple binary splits*, rather than just a single split, so that covariates are held constant across more ecologically valid levels of the target predictor. For example, very high scorers might only be matched with very low scorers, and moderately high scorers with moderately low scorers (i.e. quartiles). Even using these quartiles, splits almost approximate the ordinal five- or seven-point Likert scales typically used to measure the big five. To illustrate, very high scorers' median response to facet items may be "strongly agree", and very low scorers' median response may be "strongly disagree". Put another way, quartiles might already have endemic meanings that mitigate the artificiality of using median splits. Finally, there is also a second stage where PSM weights are applied to the original *numeric* versions of each facet. *Any* degree of matching that is preserved in the transition back to the numeric predictor would still hold covariates *more constant* than zero-order associations.

PSM has occasionally also been applied to survey research in the social sciences. For example, Dehejia and Wahba (2002) evaluated the effect of a work experience training program—with participants either randomly allocated to the program or a control condition—on future earnings. Then, they used PSM to fully replicate effects using convergent survey data. Foster (2003) used PSM to evaluate the extent that outpatient children suffering from mental health issues benefited from exposure to ongoing services. Outpatients were matched based on their symptomatology and previous exposure to therapy. Like personality, these services were conceptualised on an ordinal scale. Results suggested diminishing returns after 12-18 exposures. Further, Scherman, Arriagada and Valenzuela (2014) used PSM to match participants with different levels of social media usage on their socio-economic status and political engagement. Then, they found that social media usage was positively associated with subsequent protest behaviour. Therefore, PSM has already been used in the social sciences to boost the internal validity of correlational research.

5.2.4. Combined PSM and Elastic Net

Emergent machine learning techniques may also address the multicollinearity-confounding tradeoff. Most prominently, elastic net regression is designed specifically to *predict* outcomes from variables that have complex patterns of intercorrelations (Zou & Hastie, 2005). It is a combination of ridge and lasso regressions, which both have varying parameters designed to

mitigate ungeneralizable (i.e. overfitted) models (Cohen et al., 2013). Ridge regression down-weights large coefficients so that no single variable has excessive impact on predictions. Lasso regression down-weights small coefficients to converge on zero, thus reducing the likelihood that spurious variables impact predictions. Then, elastic net selects the systematically varying combinations of ridge and lasso parameters that yield the most explanatory model.

A combination of PSM and elastic net (PSM-ENET) *may* optimize the internal validity of personality facet associations with SWB. There are arguments against and for this proposition. Against elastic net is that it tends to retain clusters of correlated variables when any one constituent has a strong association with the outcome (Ryali, Chen, Supekar & Menon, 2012). Thus, it might yield a *de facto* factor solution that compromises the PSM matching on individual facets. In addition, to date the primary function of elastic net is to yield high fidelity *predicted scores*, often when the number of predictors converges with or exceeds the number of cases. Thus, ridge and lasso parameters may change realistic coefficients to optimize overall model fit. The argument in favour of PSM-ENET is that it uses complimentary approaches to mitigate multicollinearity. From the outset, PSM uses sample weights to reduce the overall likelihood of confounding. Then, the elastic net may account for whatever covariance remains by allowing controls to also be fit during modelling. That is, PSM may initially reduce the latent factors in the data, which thus allows elastic net to better isolate specific facet associations with SWB. Overall, it remains unclear whether elastic net can be used to help optimize the internal validity of PSM associations.

5.2.5. The Present Studies

The present studies represent a first-of-their kind application of PSM, and combined PSM-ENET, to numeric survey predictors in the psychological literature. First, I compared PSM to zero-order correlations and multiple regression. Specifically, I evaluated the extent (a) PSM weights held covariates more constant, (b) increasing PSM controls caused multicollinearity, and (c) PSM replicated canonical big five factor neuroticism and extraversion associations with SWB. The rationale for (c) was that established *factor*-level associations were likely to be robust because there were insufficient covariates to cause multicollinearity, and then that factor-level effects would manifest in their facet substrates. Then, I evaluated whether combined PSM-ENET better replicated neuroticism and extraversion effects, compared to PSM in isolation. Finally, I evaluated big five facet associations with SWB using the best-performing method, to find the *full* pattern of effects.

5.3. Study 1

Study 1 evaluated the efficacy of PSM. The control methods were initially zero-order correlations and multiple regression. I selected zero-order associations because they are the equivalent of fitting a single facet as the superordinate variable in a stepwise regression. Beyond this, I did not focus on stepwise regression because of its tendency to produce cascading statistical errors based on the variable/s arbitrarily chosen in the first step. In multiple regression, I focussed on models that accounted for incrementally increasing and, ultimately, all 29 facet controls. Finally, I evaluated whether PSM better replicated neuroticism and extraversion *factor* associations than either elastic net or PSM-ENET.

5.3.1. Method

The data analysed below are again from the self-report component of the AXA Research Project outlined in Chapter 1. The project was originally intended to generate prediction algorithms for multinational participants, which linked their self-report sociodemographic and psychological construct scores to their logged Twitter behaviour. Algorithms were then intended to be exported to massive databases of Twitter users who did not complete the concomitant self-reports. However, the survey data alone had sufficient power to permit PSM analyses without conflating Type 1 error (see 'Present Studies' in the General Introduction). Thus, I focussed exclusively on it because analyses using twitter-derived constructs would have meant compounded measurement and prediction errors, and to avoid any ethics complications concerning the use of data that were obtained without explicit prior consent.

5.3.1.1. Participants

Participants were 36,498 multinational internet panellists from 33 different countries—speaking 14 different languages—who completed a 15-minute battery of questionnaires that mostly assessed personality, SWB and demographic characteristics. Full participant descriptives are in Table 5.1. The mean sample size by country was 1,106 (SD = 373.49). As per Chapter 4, the participants were retained (90%) because they answered at least 70% of the questions and used multiple different response options when answering the 120 personality items. There were 50% men (ICC = .06) and the mean age was 34.55 (SD = 11.84; ICC = .08). I thus sampled a heterogenous and multinational adult population.

5.3.1.2. Materials

All survey items were administered using the prevailing back-translation approach in social psychology, which I described in Chapter 4 (Brislin, 1970). Some scales were only given to a subsample of countries. Although this meant that sample sizes differed across outcomes, all effects were still derived from > 29,000 participants. When appropriate, I also included other auxiliary scales from various country waves of the AXA project, to (a) help establish the construct validity of constituent of personality variables, and (b) enhance the internal validity of observed personality-SWB effects. These comprise a broad spectrum of attributes—included at the request of various collaborators—that range from attitudes towards climate change to sexual orientation. All sample sizes are reported alongside participant descriptives.

Table 5.1.

Participant descriptive statistics

Region	Language	ISO	N	RET	Sex	Age
Anglosphere	English	AUS	1,172	90%	54%	40.21 (12.77)
	English	CAN	1,367	91%	62%	36.04 (13.88)
	English	GBR	1,511	92%	64%	35.04 (12.87)
	English	USA	2,088	92%	81%	36.81 (13.25)
	English	ZAF	1,195	90%	53%	35.33 (11.44)
	SUBTOTAL		7,333	91%	65%	36.6 (12.92)
Asia	Mandarin	CHN	960	86%	43%	32.93 (8.78)
	Indonesian	IDN	2,141	89%	48%	30.68 (9.03)
	English	IND	996	91%	22%	30.54 (9.38)
	Japanese	JPN	458	83%	52%	42.11 (12.17)
	Korean	KOR	493	91%	52%	36.92 (11.11)
	Thai	THA	1,079	89%	51%	33.95 (9.14)
	Turkish	TUR	1,106	81%	28%	30.56 (9.48)
	Mandarin	TWN	1,025	84%	50%	34.22 (10.62)
	SUBTOTAL		8,258	87%	43%	32.78 (9.61)
Europe	German	AUT	1,240	93%	55%	39.89 (12.62)
	French	BEL	1,017	87%	51%	40.49 (13.4)
	German	DEU	1,128	94%	45%	37.53 (13.42)
	Spanish	ESP	1,020	94%	37%	33.82 (9.7)
	English	FIN	1,043	91%	52%	38.37 (12.43)
	French	FRA	1,103	93%	55%	37.2 (13.2)
	Italian	ITA	1,108	92%	51%	34.68 (10.78)
	Polish	POL	970	89%	34%	30.9 (11.23)
	Russian	RUS	1,174	94%	54%	36.87 (11.84)
	SUBTOTAL		9,803	92%	48%	36.74 (12.1)
South America	Spanish	ARG	1,144	88%	50%	36.24 (12.27)
	Spanish	BOL	169	79%	45%	33.19 (12.06)
	Portuguese	BRA	520	94%	39%	30.22 (9.03)
	Spanish	CHL	1,159	89%	50%	33.57 (11.15)
	Spanish	COL	1,100	94%	34%	30.55 (9.43)
	Spanish	ECU	1,198	84%	53%	34.17 (11.8)
	Spanish	MEX	1,210	96%	41%	30.09 (9.13)
	Spanish	PER	1,126	89%	39%	29.72 (9.64)
	Spanish	PRY	1,052	80%	53%	29.86 (9.04)
	Spanish	URY	1,301	85%	55%	35.89 (12.27)
	Spanish	VEN	1,125	94%	39%	32.41 (10.95)
	SUBTOTAL		11,104	89%	46%	32.48 (10.62)
GRANDTOTAL			36,498	90%	50%	34.52 (11.25)

Notes. N = Final number of participants. RET = Percentage retained.

5.3.1.2.1. Personality

I measured the big five for *all participants*, using the publicly available 120-item version of the NEO-AC from the International Personality Item Pool (Johnson, 2014). There were four items for each of the 30 facets. Items were rated on a scale from 1 = "strongly disagree" to 7 = "strongly agree". Descriptive statistics and bivariate associations with SWB are in Table 5.2. First, I evaluated internal consistency to confirm that items in each facet measured components of the same underlying construct. For 29/30 facets, Cronbach's alpha ranged from $\alpha = .40$ to $\alpha = .80$. I deemed this sufficient because scales were designed to capture the same construct heterogeneity as the original 300-item NEO-AC. Using the Spearman-Brown correction for survey length, this would have corresponded to $\alpha = .62$ to $\alpha = .91$. Thus, scores either approached or surpassed the conventional threshold ($\alpha = .70$) for internal consistency, even when the NEO-AC was administered to a plurality of different cultures and in multiple languages. The only exception was liberalism ($\alpha = .17$; projected $\alpha = .34$), which was the extent participants believed in social equality. As such, I evaluated its construct validity.

Table 5.2

Big five facet descriptive statistics and bivariate associations with SWB

FAC	No.	Facet/Subscale	N	M (SD)	α (300)	ICC	THWT	SAT
NUR	1	Anxiety	36,497	4.12 (1.31)	0.7 (0.85)	0.03	0.52*	-0.2*
	2	Anger	36,498	3.54 (1.4)	0.78 (0.9)	0.03	0.41*	-0.25*
	3	Depression	36,498	3.17 (1.36)	0.76 (0.89)	0.08	0.56*	-0.45*
	4	Self-consciousness	36,498	3.93 (1.15)	0.51 (0.72)	0.04	0.39*	-0.29*
	5	Immoderation	36,498	3.71 (1.05)	0.45 (0.67)	0.03	0.17*	-0.22*
	6	Vulnerability	36,497	3.39 (1.23)	0.65 (0.82)	0.08	0.49*	-0.36*
	Total		36,498	3.64 (0.91)	0.88 (0.95)	0.06	0.6*	-0.41*
EXT	1	Friendliness	36,498	4.64 (1.26)	0.72 (0.87)	0.06	-0.36*	0.43*
	2	Gregariousness	36,498	3.95 (1.33)	0.66 (0.83)	0.05	-0.24*	0.33*
	3	Assertiveness	36,497	4.7 (1.14)	0.66 (0.83)	0.13	-0.18*	0.5*
	4	Activity Level	36,498	4.09 (1)	0.4 (0.62)	0.07	-0.01	0.28*
	5	Excitement-Seeking	36,498	4.06 (1.16)	0.6 (0.79)	0.05	0.16*	0.27*
	6	Cheerfulness	36,498	5.22 (1.2)	0.79 (0.9)	0.14	-0.27*	0.6*
	Total		36,498	4.44 (0.81)	0.85 (0.93)	0.1	-0.23*	0.59*
OPN	1	Imagination	36,498	4.89 (1.28)	0.73 (0.87)	0.1	0.2*	0.21*
	2	Artistic Interests	36,497	4.96 (1.22)	0.63 (0.81)	0.06	-0.12*	0.31*
	3	Emotionality	36,498	4.84 (1.02)	0.47 (0.69)	0.04	-0.14*	0.23*
	4	Adventurousness	36,497	4.14 (1.06)	0.51 (0.72)	0.07	-0.25*	0.21*
	5	Intellect	36,498	4.58 (1.16)	0.57 (0.77)	0.03	-0.2*	0.2*
	6	Liberalism	36,498	3.85 (0.93)	0.17 (0.34)	0.05	0.06*	-0.03*
	Total		36,498	4.54 (0.63)	0.72 (0.87)	0.06	-0.12*	0.34*
AGR	1	Trust	36,497	4.28 (1.17)	0.71 (0.86)	0.04	-0.13*	0.26*
	2	Morality	36,498	5.68 (1.24)	0.8 (0.91)	0.1	-0.39*	0.17*
	3	Altruism	36,498	5.33 (1.07)	0.65 (0.82)	0.06	-0.24*	0.36*
	4	Cooperation	36,498	5.3 (1.25)	0.67 (0.84)	0.09	-0.4*	0.13*
	5	Modesty	36,498	4.2 (1.26)	0.67 (0.84)	0.1	-0.02*	-0.31*
	6	Sympathy	36,498	4.84 (1.08)	0.49 (0.71)	0.07	-0.12*	0.23*
	Total		36,498	4.94 (0.74)	0.83 (0.92)	0.07	-0.35*	0.21*
CON	1	Self-Efficacy	36,498	5.23 (1.08)	0.76 (0.89)	0.11	-0.23*	0.61*
	2	Orderliness	36,497	4.68 (1.43)	0.74 (0.88)	0.03	-0.33*	0.2*
	3	Dutifulness	36,498	5.41 (1.02)	0.6 (0.79)	0.07	-0.34*	0.3*
	4	Achievement-Striving	36,497	5 (1.12)	0.6 (0.79)	0.1	-0.3*	0.41*
	5	Self-Discipline	36,498	4.81 (1.13)	0.64 (0.82)	0.07	-0.42*	0.49*
	6	Cautiousness	36,498	4.65 (1.37)	0.79 (0.9)	0.04	-0.4*	0.12*
	Total		36,498	4.96 (0.84)	0.88 (0.95)	0.08	-0.48*	0.48*

Notes. No. = Original facet number (Johnson, 2014). α (300) = Observed Cronbach's alpha (projected alpha for original 300 item version of scale using the Spearman-Brown correction). ICC = Intraclass correlation coefficient. THWT = Bivariate association with psychological needs thwarting. SAT = Bivariate association with psychological needs satisfaction. * = $p < .001$

Despite its low internal consistency, liberalism showed convergent and discriminant validity. In this instance, I defined convergent validity as robust correlations, in the intuitive directions, with theoretically related variables. I defined discriminant validity as when the correlations were larger, in absolute magnitude, than those with theoretically unrelated variables. Convergent and discriminant variables were single scored items that I selected because they were (a) all binary (1 = "yes") to hold variable type constant, (b) had either strong or weak theoretical relationships with liberalism, and (c) were not used elsewhere in the present chapter.

The convergent variables were support for policies to mitigate climate change and increase immigration, and the discriminant variables were whether participants smoked and regularly donated to charity. Table 5.3 contains exact questions, variable wording, descriptive statistics and bivariate associations. Overall, convergent associations were larger than discriminant associations and their 99.9% CIs did not overlap. As such, I decided that liberalism was sufficiently construct valid to retain in the subsequent analyses.

Table 5.3

Convergent and discriminant variable associations with liberalism

Type	Variable	N	M	ICC	R (99.9% CI)
Convergent	Support for climate change	19,716	73%	0.05	0.08 (0.06, 0.1)
	Support for immigration	18,918	45%	0.12	0.18 (0.16, 0.2)
Discriminant	Current smoker	10,203	32%	0.01	0 (-0.03, 0.04)
	Donated to charity	35,275	42%	0.06	0.02 (0, 0.03)

Notes. Variable type = Whether there was a strong (convergent) or weak (discriminant) theoretical association with liberalism. M = % of participants who answered "yes". ICC = Intraclass correlation coefficient. R = Pearson's R correlation with liberalism.

Thus, I evaluated descriptive statistics for the all thirty facets in the big five. At the outset, I did this using original (non- group mean z-scored) scores. Overall, inspection of the ICC suggested 86% to 97% of variation in facet scores was attributable to individuals and not their countries of residence. In addition, there was no evidence for floor or ceiling effects. Thus, I next evaluated bivariate associations for facets *within* each factor. In support of the big five factor structure, 74/75 of the bivariate associations were positive (r_{mean} = .33; SD = .17). The only exception was between trust and modesty in agreeableness (r = -.12; CI_{95} = -.13, -.11). However, both these facets were positively associated with the remainder of agreeableness and thus they may have still belonged to the same latent factor. Then, I computed an exhaustive set of absolute bivariate associations between facets in *different* factors. The average absolute magnitude that was approximately 2/3 that of the intra-factor associations (r_{abs} = .21; SD = .13). Thus, I also confirmed that the facets were correlated between factors.

5.3.1.2.2. Subjective Well-Being

The primary measure was the 18-item BMPN (Sheldon & Hilpert, 2012). The BMPN was administered to 29,629 participants in 28/33 countries. In Chapter 4, I found evidence for the construct validity of separate needs thwarting and satisfaction factors—each comprising nine

items rated on a five-point Likert scale (1 = “Strongly disagree”; 5 = “Strongly agree”)—that captured feelings of suffering and flourishing respectively. Example items, descriptive statistics and bivariate associations between it and the other SWB variables are reported again in Table 5.4. Compared to the widely-used SLS, I found evidence that both factors measured transitive SWB—thus helping establish the correct temporal sequence with more stable facet predictors—and explained additional unique variation in real-world outcomes such as exercise frequency, fruit and vegetable consumption and relationship status. An added benefit was that BMPN items asked about discrete experiences (e.g. “I am currently experiencing some kind of failure…”). Thus, they were unlikely to be conflated with personality items (e.g. “often feel blue”), which measured more general tendencies. I highlight again that there was only a weak-moderate negative association between needs satisfaction and thwarting. This suggested that the BMPN factors formed largely independent components of SWB.

Table 5.4

SWB descriptive statistics and intercorrelations

Outcome	Variable	N	Mean (SD)	α	ICC	R-THWT	R-SLS	R-Health
Primary	Needs Satisfaction	28,940	3.73 (0.69)	0.84	0.11	0.22*	0.52*	0.28*
	Needs Thwarting	28,943	3.27 (0.8)	0.81	0.06	-	0.3*	0.21*
Secondary	SLS	35,737	4.38 (1.43)	0.87	0.07	-	-	0.35*
	Health	35,358	62.95 (23.55)	-	0.07	-	-	-

Notes. ICC = Intraclass correlation coefficient. R-TWT = Bivariate correlation with needs thwarting. R-SLS = Bivariate correlation with the SLS. R-Health = Bivariate correlation with health.

The secondary measures—administered to all participants—were SLS and physical health. Example items, descriptive statistics and bivariate SWB associations are also in Table 5.4. SLS captures mostly cognitive appraisals of SWB, and has demonstrated construct validity across cultures and languages (Pavot & Diener, 2008). It comprises five items that are each rated on a seven-point Likert scale (1 = “strongly disagree”; 7 = “strongly agree”). Physical health was “Please rate your health over the past 12 months” and scored on a 100-point sliding scale (1 = “Extremely poor”; 100 = “Extremely good”). Overall, effects that converged across primary and secondary outcomes were especially robust.

5.3.1.2.3. Controls

The comprehensiveness of personality facet covariates meant that they likely also accounted for a range of other individual-level variables. Thus, I only fit a limited number of additional controls—sex, age, social class, religiosity and response bias—which were each measured with single items. Although I introduced them in Chapter 4, I introduce them again here because the sample was larger and more heterogenous when using SLS and health secondary outcomes. Social class was "Where do you place yourself on the spectrum of social class compared to your countrymen?" and rated on a 100-point sliding scale (1 = "Very bottom"; 100 = "Very top") (M = 41.31; SD = 22.57; ICC = .10). Religiosity was "Are you currently practising a religion?" (1 = "yes"; 42%; ICC = .19). Response bias was the tendency for participants to prefer the minimum or maximum ends of Likert-style response scales. I centred response bias so that preferences for the minimum end were negative and preferences for the maximum end were positive (M = 0.18; SD = 0.48; ICC = .06). Then, I used multiple regression to evaluate the effects for all controls on both needs thwarting and needs satisfaction. Results are in Table 5.5. Response bias was by far the strongest predictor of both outcomes. Controlling for it, in particular, may have ensured parsimonious facet associations with SWB. I evaluated all effects relative only to participants' countrymen, and thus there was no need for multilevel models or country-level controls.

Table 5.5

Multiple regression effects for control variables on needs thwarting and satisfaction

Outcome	Variable	Beta (99.9% CI)	T (df)	p
Needs Thwarting	Sex	0 (-0.03, 0.03)	0.29(26,625)	.773
	Age	-0.01 (-0.01, -0.01)	-32.92(26,625)	< .001
	Social Class	0 (-0.01, 0)	-22.43(26,625)	< .001
	Religious	-0.04 (-0.07, -0.01)	-4.31(26,625)	< .001
	Response Bias	0.62 (0.58, 0.65)	63.32(26,625)	< .001
Needs Satisfaction	Sex	-0.01 (-0.04, 0.01)	-1.77(26,625)	.077
	Age	0 (0, 0.01)	11.91(26,625)	< .001
	Social Class	0 (0, 0.01)	26.76(26,625)	< .001
	Religious	0.15 (0.12, 0.18)	18.15(26,625)	< .001
	Response Bias	0.36 (0.33, 0.39)	41.5(26,625)	< .001

Notes. The controls were all fit together in a single multiple regression model for each of the primary outcomes.

5.3.1.3. Procedure

5.3.1.3.1. Multiple Imputation

I imputed all missing variables separately for each country using MICE. As per the procedure in Chapter 3, MICE assigns values to missing scores with plausible sampling error. To minimize the need for multiple imputation at the outset, multi-item scale scores were the *average* of all non-missing item responses. Thus, only scales that were completely unanswered, as well as single items, were missing (2% total). As an added precaution, I also performed multiple imputation using three independent sets of variables—demographic controls, big five facets and SWB—so that each set of imputed scores was uncontaminated by the other sets. To maximize the fidelity of imputations, I used all additional control and SWB variables available in each country. They are reported in Table 5.6. Averaging across all variables, there was 82% (SD = 10%) overlap in kernel density plots for non-imputed and imputed scores after multiple imputation. I decided this was sufficient—especially considering the small percentage of missing values—to accept the imputations.

Table 5.6

Additional demographic and SWB variables used during multiple imputation

Group	Item	M	M (SD)	ICC
Demographic	University degree	31,281	55%	.13
	Ethnic minority	23,456	17%	.11
	Romantic relationship	34,345	62%	.02
	Household income	24,239	37.37 (24.65)	.2
	Exercise regularly	35,301	46.46 (30.88)	.08
	Eat greens regularly	35,306	65.9 (24.74)	.05
	Recently donated to charity	35,275	42%	.06
	Heterosexual	32,303	86%	.02
	Current smoker	10,203	32%	.01
SWB	PANAS – Positive Affect	13,759	3.41 (0.78)	.12
	PANAS – Negative Affect	13,758	2.33 (0.87)	.06
	Type 2 diabetes	10,214	7%	.01
	High cholesterol	10,218	19%	.01
	High blood pressures?”	10,199	20%	.03

Notes. Group = Multiple imputation category. M (SD) = % of participants who answered “yes” for binary variables.

5.3.1.3.2. Measurement Equivalence

I evaluated measurement equivalence across countries using multigroup (MG) confirmatory factor analysis (CFA). Underlying CFA evaluates whether variables conform to a pre-defined factor structure. In practise, good CFA fit means unexplained variance is attributable exclusively to the *individual* variables, and not unaccounted-for variable clusters. MG-CFA then evaluates the extent this pre-defined structure continues to fit the data after introducing increasingly stringent assumptions about measurement equivalence across groups. These assumptions are configural invariance (the factor structure is appropriate in each separate group), metric invariance (factor loadings are also equal), scalar invariance (cases in different groups who score the same on each factor respond similarly to each facet) and residual invariance (CFA models have equal explanatory power in separate groups; Pendergast, von der Embse, Kilgus & Eklund, 2017). As the MG-CFA equivalence assumptions increase—meaning fewer CFA parameters are free to accommodate the actual data—CFA model fit meaningfully decreases when there is measurement *non*-equivalence. There is measurement equivalence when decreases are only negligible.

There are multiple ways of evaluating meaningful MG-CFA decreases. Often, researchers use the chi-squared goodness of fit statistic to determine whether the CFA model has more explanatory power than control-model alternatives. However, this metric is sensitive to sample size and thus inappropriate in conditions of high power. Instead, I focussed on the comparative fit index (CFI), which captures the percentage of shared facet variation explained by the CFA model. The threshold for adequate CFA model fit is often CFI > .90 (Pendergast et al., 2017). Corresponding delta statistics are the approximate SEs of model fits across countries.

I performed MG-CFA for the NEO-IPIP, the original BMPN facets and the superordinate BMPN SWB factors from Chapter 4. The original BMPN facets comprised experiences of thwarting and satisfaction in each of the three basic psychological needs domains (autonomy, competence, relatedness). All facets were z-scored within each country—which again accorded with Aguinis et al.'s (2013) best-practise recommendations for analyses using multi-level data. All results are in Table 5.7. I relaxed the threshold for good model fit in the NEO-IPIP because—as suggested in the General Introduction—facet intercorrelations may transcend factor boundaries in ways that do *not* have measurement equivalence. As such, there was poor CFA fit when the NEO facets were organised exclusively into the big five factors (CFI = 0.72). Nevertheless, there were only marginal further decreases in fit as MG-CFA added increasing

measurement equivalence assumptions. For the BMPN, the model specifying the originally-intended three factors was a poor fit (CFI = 0.62). Contrastingly, there was very good fit for my revised model (see Chapter 4), where negatively valanced facets collapsed into aggregate needs thwarting and the positively valanced facets collapsed into aggregate needs satisfaction (CFI = 0.97). Then, CFA models with increasingly stringent equivalence assumptions *all* remained above the threshold for adequate fit. There were only marginal incremental decreases. Therefore, I concluded that both the NEO-IPIP and my two-factor BMPN measure of SWB had sufficient measurement equivalence to proceed.

Table 5.7

Measurement equivalence of NEO-IPIP and BMPN needs thwarting and needs satisfaction factors

Equivalence	**NEO-IPIP**	**BMPN**
CFA	.72	.97
Configural	.70	.95
Loadings	.68 (.02)	.95 (< .01)
Intercepts	.68 (< .01)	.95 (< .01)
Residuals	.67 (.01)	.94 (.01)
Reduction	**7%**	**3%**

Notes. Equivalence = MG-CFA parameters. CFA = Original confirmatory factor analysis averaging across countries. Values are CFI (delta).

5.3.1.3.3. Accounting for Controls

Next, I apportioned all individual-level and country control variance from the target variables. To this end, I first created separate multiple regression models where the controls iteratively predicted each target personality facet, and each SWB outcome. Then, the target variable was assigned its concomitant residuals. This was the *exact equivalent* of fitting all control variables and facets in the same multiple regression model predicting SWB. The benefit of doing this procedure ahead of the main analysis was that I could exclusively focus results on the primary big five facet associations. In aggregate, controls explained small but fluctuating proportions of facet (M = 9%; SD = 6%) and SWB (M = 12%; SD = 4%) variance. Afterwards, none of these controls could impact effect estimates, at least as main effects. Then, I also removed all additional variation attributable to participants' country. Specifically, I group mean z-scored (M = 0; SD = 1) all facets and SWB so that the mean and SD was the same in every country.

5.3.1.3.4. PSM Procedure

Using PSM, I iteratively generated a separate set of participant weights for each personality facet. Weights were designed to retain the vast proportion of cases who scored high on the target personality facet and then oversample those cases who scored low on the target facet but had converging scores across the remaining 29 facets. Thus, weights were designed to hold the remaining facets more constant than unweighted alternatives. For clarity, I report my procedure as a series of numbered steps for each target facet:

1. I made a categorical version of the target facet by splitting scores into equal-sized "very low", "moderately low", "moderately high" and "very high" quartiles. Quartiles meant I could find separate matches for moderate (2^{nd} and 3^{rd} quartiles) and more extreme (1^{st} and 4^{th} quartiles) cases. This increased the likelihood that matches had roughly equal scores on the target variable, thus improving matching fidelity compared to a single median split. Quartiles already approximated the five- and seven-point Likert scales typically used to assess personality. In support, I found that quartiles explained 86% (SD = 1%) of the variation in original numeric scores. Thus, they were both high fidelity and contained sufficient cases (N = 9,124) for extremely close matches.

2. PSM is predicated on logistic regression and thus the predictor must be binary. As such, I generated *two* PSM models for each target facet. The first was where moderately high vs. moderately low scores was the DV, and the second was where very high vs. very low scores was the DV. In both cases, the IVs were the 29 other personality facets. Propensity scores were simply the predicted values—the likelihood participants would score "high" on the target facet based on their covariates—from these models. Thus, models yielded a unique propensity score for every participant. While the propensity scores themselves were subject to multicollinearity, shared covariance that was parsed from specific predictors was still captured in the concomitant model intercept and thus conserved (Cohen et al., 2013). Thus, despite potentially unreliable single coefficients, overall predictions were still an appropriate numeric summary of the entire relationship between *all* covariates and the target facet.

3. Then, I matched each moderately high scorer to the five moderately low scorers with the nearest propensity scores. I repeated this procedure for very high vs very low scorers. I decided on five matches to mitigate idiosyncratic pairings, and to maximize the large

sample size. I decided against a greater number to limit matching imprecision. I only formed matched groups when all propensity scores were < 0.5 pooled SDs.

4. Matches were then converted to weights. Every moderately high and very high scorer that was successfully matched had a weight of 1. Every moderately low and very low scorer was assigned a weighting of .20 every time they were matched. Thus, the weights of the five low scoring cases who matched with each high scoring case summed to one.

5. Weights were used in linear regression where the *numeric version* of the intended facet IV was used to predict SWB. Thus, quartiles were only used as a preliminary step to generate weights. Although matching was imperfect—because I matched cases across multiple covariates and switched from categorical to numeric variables—PSM was still potentially effective when it held covariates *more constant* than unweighted solutions.

5.3.1.3.5. PSM Diagnostics

Overall, PSM generated unique subsamples of participants for every facet who had similar scores across the remaining 29 facets. When each facet was the target, M = 89% (SD = 5%) of the cases had non-zero weight and were thus retained. As a preliminary step, I compared propensity score differences before and after the actual PSM weights were applied. Prior to applying weights, the average difference in raw propensity scores between high vs low responses was M = .29 (SD = .10). After PSM, it was M = .05 (SD = .02). This was an 83% reduction. Then, I also evaluated whether improvements were driven by middle or extreme cases. For the middle two quartiles, PSM weights reduced propensity score differences from M = .08 (SD = .04) to M = .01 (SD < .01), which was a 92% reduction. For the extreme two quartiles, they were reduced from M = .50 (SD = .17) to M = .05 (SD = .02), which was a 90% reduction. Thus, PSM equalised the distribution of propensity scores across all four quartiles.

5.3.1.3.6. PSM-ENET

Finally, I evaluated whether PSM-ENET better isolated single personality facet effects than PSM in isolation. Thus, I fit all 30 personality facets—in their numeric form and with PSM weights—together again in the same elastic net model. I evaluated which of 100 combinations of LASSO and ridge parameters (the L1 and L2 norms)—covering the full range of possible regularizations—maximized model accuracy, which I computed using 10-fold cross validation (10-FCV). 10-FCV randomly partitions the sample into deciles and then generates an exhaustive set of models using 9/10 of the deciles. I selected 10-folds because it was both

sufficient to mitigate the chances of spurious results, compared to fewer-folds, and often recommended as best practise in instructive machine learning texts (James et al., 2013; Kuhn, 2008). Then, model accuracy is the association between the true and predicted scores for the remaining decile. The final model is the average model accuracy—using optimized LASSO and ridge parameters—and coefficient loadings from all ten iterations of 10-FCV. Strengths of the approach are that model accuracy is not artificially inflated by using the same participants during training and testing, and idiosyncratic (i.e. overfitted) model coefficients are cancelled out across the folds (James et al., 2013). Thus, 10-FCV PSM-ENET was a rigorous test of whether PSM was compatible with the prevailing machine learning alternative.

There were separate elastic net models for every facet. Models used the PSM sample weights specific to each facet, when it was the target. Then, I saved only the beta coefficient for the target facet from each model. I also fit separate models for each SWB outcome. Thus, there were a total of 120 separate PSM-ENETs. On average, R^2 for facet models predicting the primary outcomes—needs thwarting (36%) and satisfaction (41%)—were consistent across models. R^2 for the models predicting the other two convergent outcomes—SLS (28%) and health (11%)—were lower but still also consistent across models. Thus, collectively the 30 personality facets had robust predictive power. However, it was still unclear whether this robustness was preserved in individual coefficient estimates.

5.3.2. Results

I compared PSM associations to zero-order correlations and multiple regression controlling for all non-target (i.e. all 29 other) facets. Full output—including beta coefficients, confidence intervals and overall variance explained—for these different models is in Table A5.1 (Appendix 5.1). These models are also referred to, and partially reported in tables, throughout the results. The focus was on evaluating whether PSM improved covariate score constancy and multicollinearity, and then replicated canonical big five factor associations. Then, I repeated this third approach comparing PSM to both elastic net and combined PSM-ENET. I mitigated Type 1 error by treating correlations as the unit of analysis. Thus, they comprised a *sample* of observed effects that were only subjected to inferential statistics at a second stage: as associations of associations. The significance threshold was $p < .001$.

5.3.2.1. Covariate Constancy

My first approach was to evaluate whether PSM improved the constancy of covariate facet scores across the different quartile levels of each target facet. Here, I focussed exclusively on

the facets and did not use SWB. Thus, there were 694 (6 target variables, split into quartiles and each with 29 covariates) separate means for the facets in each NEO-factor. As a preliminary step, I evaluated the extent mean covariate scores from across the quartiles summed to below or above zero after PSM.[9] Results are in Table 5.8. For neuroticism facets, participants were disproportionately matched when they scored low on the extraversion, openness, agreeableness and conscientiousness facets. This pattern was inverted for the facets in the remaining factors.

Table 5.8

Reduction in the 29 facet covariate scores across different levels of the target facet, after propensity score matching (PSM)

Factor	Covariate	M (SD)	Binary % (SD)	Middle % (SD)	Extreme % (SD)
Neuroticism	Within	0.25 (0.16)	85% (5%)	92% (3%)	82% (6%)
	Between	-0.14 (0.09)	77% (41%)	76% (64%)	71% (77%)
Extraversion	Within	0.21 (0.16)	82% (7%)	90% (4%)	80% (8%)
	Between	0.04 (0.13)	76% (69%)	88% (22%)	72% (58%)
Openness	Within	0.13 (0.14)	83% (10%)	89% (5%)	81% (12%)
	Between	0.06 (0.09)	80% (25%)	82% (31%)	69% (110%)
Agreeableness	Within	0.19 (0.14)	86% (8%)	91% (9%)	84% (10%)
	Between	0.06 (0.1)	77% (22%)	88% (10%)	74% (26%)
Conscientiousness	Within	0.28 (0.21)	85% (5%)	91% (3%)	83% (6%)
	Between	0.05 (0.12)	74% (52%)	82% (37%)	73% (30%)

Notes. Covariate: I separated results for facets in the same (within) and different (between) factors. M (SD) = Mean covariate score. Binary % = Mean percentage reduction (SD) in covariate means between high and low median splits after PSM. Middle % = Mean percentage reduction (SD) in covariate means between second and third quartiles after PSM. Extreme % = Mean percentage reduction (SD) in covariate means between first and fourth quartiles after PSM.

Then, I evaluated the percentage reduction in covariate differences after matching. Specifically, I evaluated the differences in mean covariate scores between low and high median splits, and between middling and extreme quartiles, both without and with PSM weights. Then, I calculated the percentage improvement using PSM weights. Results are also in Table 5.8. Averaging across the facets in each factor, results suggested that PSM reduced covariate score differences between low and high median splits by 74% to 80%. The range of percentage reductions was similar using just middling quartiles (76% =to 88%) and just extreme quartiles (69% to 84%). Results held across all five factors. Facet scores from the same factor (M = 84%;

[9] Prior to PSM, all covariate means summed to zero because they were z-scored (M = 0; SD = 1).

SD = 2%) were held more constant than facet scores from other factors (M = 77%; SD = 2%). Overall, PSM increased the constancy of covarying facets both within and between the big five.

5.3.2.2. Multicollinear Effects

Second, I evaluated whether PSM reduced effect multicollinearity compared to multiple regression. Specifically, I used the big five facets to predict each of the four SWB outcomes using randomly selected facet controls that increased from 1 to 29, in increments of 1. I used the same controls for PSM and multiple regression to avoid confounded results. Due to computational demands, I randomly sampled only 1,000 participants—on whom I performed PSM—at each increment. To stabilize coefficient estimates across SWB outcomes and randomly selected controls, I also repeated the procedure 10 times at each increment. Thus, there were 8,700 (30 target variables * 29 control levels * 10 iterations at each level) separate PSM and multiple regression models. As a preliminary step, I found that—averaging across all personality facets and SWB—the mean effect magnitude for PSM was 0.12 (SD = 0.10) and the mean effect magnitude for multiple regression it was 0.08 (SD = 0.08). This suggested that observed effects were larger for PSM. Then, I evaluated how effect magnitude changed as the number of controls increased.

Effect magnitude for PSM diminished by *less* as controls increased, compared to multiple regression. I evaluated effects with the number of controls as the predictor, and absolute mean effect magnitude as the outcome. After controlling for personality factor and outcome: the logarithmic effect for increasing controls (i.e. the forgetting curve) on effect magnitude was less negative for PSM than multiple regression ($b = .02$; $CI_{99.9} = .01, .02$; $t(571) = 8.59$; $p < .001$). Results are in Figure 5.1. Put more simply, as controls increased multiple regression effect magnitudes decreased more steeply than PSM. Both began to plateau, but only after a greater number of controls for PSM. PSM effects became significantly larger when there were only 3 controls ($b = .02$; $CI_{99.9} = < .01, .03$; $t(2{,}938) = 4.05$; $p < .001$). While this threshold was partly a function of sample size, it still highlights that magnitudes diverged even when attempting to exert modest control. Finally, I also evaluated the inflection point—the first instance where predicted scores were less than 50% of the largest observed magnitude—for both sets of effects. The inflection point was 8 controls for multiple regression, and 25 controls for PSM. Thus, multiple regression effects also decreased *more rapidly* compared to PSM. Finally, PSM effects decreased by < 1% from the inflection point to all 29 controls. This

suggested that effect size decay had almost stopped in PSM, and that models could tolerate the additional four controls beyond the inflection point with minimal extra multicollinearity.

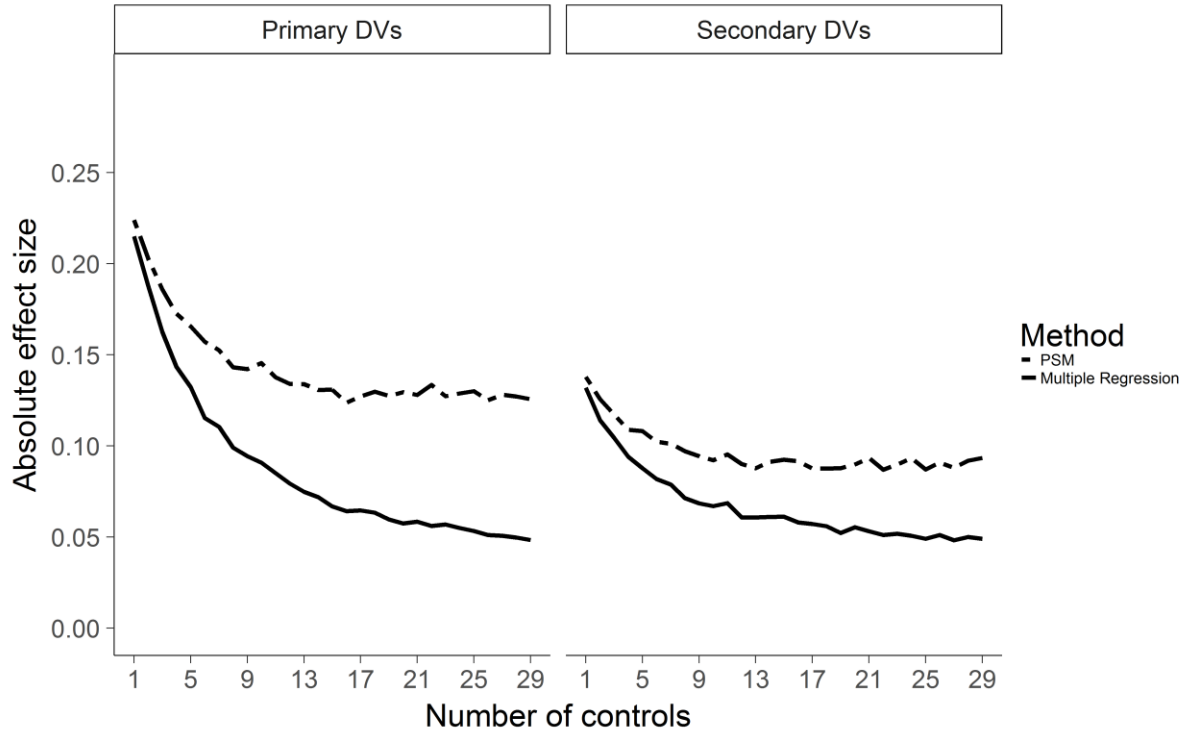


Figure 5.1. Effect size of personality facet-SWB associations with increasing control variables. Every personality facet was iteratively used as the target variable, for both primary outcomes (needs thwarting, needs satisfaction) and both secondary outcomes (SLS, health). Controls were randomly selected non-target personality facet covariates, which increased from 1-29 in increments of 1. I repeated each analyses at each number of controls 10 times—with randomly selected variables—to increase the stability of estimates. Absolute effect size was the average absolute beta coefficient of regression model effects where controls were either accounted for as PSM weights, or as explicitly defined variables in multiple regression models.

5.3.2.3. Replicating Established Big Factor Effects

Third, I evaluated which method best replicated *factor-level* personality effects on SWB. I reasoned that the more internally valid method would better show the established relative strength of neuroticism and extraversion effects (the target factors) on SWB, compared to the other factor effects (non-target factors). I reversed needs thwarting so that its associations were in the same direction as the remaining outcomes. I compared PSM results to zero-order correlations and partial correlations from multiple regression with all 29 covariates. As a preliminary step, I evaluated all facet associations with SWB outcomes in the target and non-target factors. All results are in Table 5.9. For all methods, target neuroticism facet associations

were negative and target extraversion facet associations were positive. Although zero-order effects appeared to be larger, they may still have failed to differentiate between target and non-target effects. Indeed, aggregate target vs non-target PSM effect proportions (3.51) were larger than zero-order (2.96) and multiple regression (2.40) proportions. Thus, PSM associations were better in aggregate at differentiating between effects in different facets.

PSM also better replicated established neuroticism and extraversion associations at the facet-level, compared to zero-order and multiple regression. To this end, I computed an exhaustive set of pairwise absolute proportions between facets in target versus non-target factors. Thus, there were 432 (6 target facets * 18 non-target facets * 4 SWB outcomes) unique pairwise ratios for both neuroticism and extraversion. The higher the value, the more unambiguously the target factor effects were isolated. The predictor was the method and the proportions were the outcome. There was a floor effect of absolute r = .005 to prevent unrealistically large ratios. I evaluated effects using separate Poisson regressions for neuroticism and extraversion, controlling for outcome. Compared to zero-order correlations, multiple regression proportions were roughly equal for neuroticism (b = -0.05; $CI_{99.9}$ = -0.16, 0.07; t(1288) = -1.40; p = .162) and lower for extraversion (b = .-0.28; $CI_{99.9}$ = -0.42, -0.15; t(1288) = -6.90; p < .001). Thus, I compared PSM to the better-performing zero-order correlations. PSM yielded higher proportions for both neuroticism (b = .44; $CI_{99.9}$ = .33, .54; t(1288) = 13.88; p < .001) and extraversion (b = .33; $CI_{99.9}$ = .22, .45; t(1288) = 9.51; p < .001). Thus, PSM most unambiguously replicated established personality factor effects for SWB.

Finally, I evaluated how ratios differed between non-target factors for PSM associations. Compared to conscientiousness, combined neuroticism and extraversion PSM proportions were especially large for openness (b_{ratio} = 1.73; $CI_{99.9}$ = 1.57, 1.90; t(861) = 35.36; p < .001) and then agreeableness (b_{ratio} = 0.96; $CI_{99.9}$ = 0.78, 1.13; t(861) = 18.01; p < .001). Therefore, PSM also best differentiated target facet effects when non-target facets came from the factors—openness and agreeableness—with the weakest documented SWB associations.

Table 5.9

Mean bivariate associations between SWB and neuroticism, extraversion and other facets using PSM and alternate methods

Method	Factor	R (SD)	R-Prop	B (SD)	B-Prop
PSM	Neuroticism	-0.13 (0.09)	1.94	-0.14 (0.09)	1.9
	Extraversion	0.1 (0.08)	1.57	0.12 (0.1)	1.65
	Other	0.07 (0.06)	-	0.07 (0.07)	-
Zero-Order	Neuroticism	-0.29 (0.11)	1.63	-	-
	Extraversion	0.24 (0.13)	1.34	-	-
	Other	0.18 (0.11)	-	-	-
Multiple Regression	Neuroticism	-0.04 (0.06)	1.54	-	-
	Extraversion	0.02 (0.05)	0.86	-	-
	Other	0.03 (0.03)	-	-	-
Elastic Net	Neuroticism	-	-	-0.05 (0.07)	1.82
	Extraversion	-	-	0.03 (0.06)	1.04
	Other	-	-	0.03 (0.03)	-
PSM-ENET	Neuroticism	-	-	-0.05 (0.07)	1.63
	Extraversion	-	-	0.03 (0.06)	1.26
	Other	-	-	0.03 (0.03)	-

Notes. All results were averaged across the four SWB outcomes after reversing needs thwarting correlations. Method: I compared PSM associations to zero-order correlations, multiple regression partial correlations with all 29 facets as covariates, elastic net and combined PSM and elastic net (PSM-ENET). Factor: Neuroticism and Extraversion were the target factors that had established associations with SWB; Other = Absolute effects for openness, agreeableness and conscientiousness factors. R (SD) = The mean correlation from constituent facet effects. R-Prop = Target correlation means proportional to non-target correlation means. B (SD) = The mean beta coefficient from constituent facet effects; it was computed in lieu of R because my version of elastic net did not allow partial correlations. B-Prop = Target beta means proportional to non-target beta means.

5.3.2.4. PSM vs Combined PSM and Elastic Net

Finally, I replicated Approach 3 comparing PSM to PSM-ENET. For comprehensiveness, I also compared results to elastic net in isolation. I switched from associations to beta-coefficient effect estimates because my elastic net procedure did not yield partial correlations. For methodological control, I generated each set of coefficients using the same 10-fold cross-validation protocol, via R's 'glmnet' package with 'caret' package interface. After reversing effects for needs thwarting, I again evaluated raw effects. Results are in also in Table 5.9. While they were all in the expected direction, they were larger for PSM than either elastic net or PSM-ENET. Aggregate target vs absolute non-target effect proportions for all three methods were approximately equal in neuroticism, and larger for PSM in extraversion. Overall, there was thus preliminary evidence that PSM outperformed elastic net and PSM-ENET.

PSM in isolation best replicated established factor associations with SWB. I again computed an exhaustive set of absolute pairwise proportions between target and non-target facet effects.

I created a floor effect of absolute b = .002 to prevent unrealistically large ratios. As a preliminary step, I compared PSM-ENET to elastic net in isolation (PSM-ENET = 1). They were equally good at differentiating neuroticism effects (b_{ratio} = -0.04; $CI_{99.9}$ = -0.12, 0.05; t(1,290) = 1.35; p = .178), and PSM-ENET better differentiated extraversion effects (b_{ratio} = 0.19; $CI_{99.9}$ = 0.08, 0.29; t(1,290) = 5.75; p < .001). Thus, I compared the better-performing PSM-ENET to PSM. PSM better differentiated effects than PSM-ENET (PSM = 1) for both neuroticism (b_{ratio} = 0.24; $CI_{99.9}$ = 0.16, 0.32; t(1,290) = 9.44; p < .001) and extraversion (b_{ratio} = 0.27; $CI_{99.9}$ = 0.18, 0.37; t(1,290) = 9.42; p < .001). Moreover, I replicated these effects using a series of different plausible models that accounted for potential methodological artefacts. These are in Table 5.10. Specifically, I fully replicated superior PSM effects when I (a) made the ceiling ratio 20 to mitigate the impact of outlier ratios; (b) used linear regression with log transformed ratios to mitigate the limitations of using Poisson regression with decimals; (c) omitted ratios involving conscientiousness because it has the next-most plausible factor-level associations with SWB; and (d) used only the largest ratio from each target facet to prevent artificially inflated results caused by using duplicate effects. The only exception to the overall consistent superiority of PSM was the effect for extraversion in (b), which trended in the same direction as the other effects but did not reach significance. In summary, PSM fully outperformed PSM-ENET when replicating established neuroticism and extraversion factor associations, across the primary and all but one plausible secondary analyses.

Table 5.10

Replicating established neuroticism and extraversion factor associations with SWB using PSM and PSM-ENET, across multiple methods

Method	Target Factor	B (99.9% CI)	T (df)	p
Truncated Outliers	Neuroticism	0.19 (0.08, 0.29)	5.77(869)	< .001
	Extraversion	0.12 (0.01, 0.24)	3.49(869)	< .001
Log Linear Regression	Neuroticism	0.77 (0.42, 1.13)	7.18(869)	< .001
	Extraversion	0.3 (-0.07, 0.68)	2.65(869)	.008
Conscientiousness Removed	Neuroticism	0.56 (0.47, 0.66)	19.36(571)	< .001
	Extraversion	0.61 (0.5, 0.72)	18.35(571)	< .001
No Duplicate Numerators	Neuroticism	0.43 (0.23, 0.63)	7.12(355)	< .001
	Extraversion	0.71 (0.47, 0.94)	9.9(355)	< .001

Notes. Method: Different permutations of my primary analysis that (a) created a ceiling value of 20 for observed effect ratios (Truncated Outliers), (b) used log-transformed ratios in linear regression (Log Linear Regression), (c) removed conscientiousness from the non-target effects because it had the next-most established relationship with SWB (Conscientiousness Removed), and (d) used only the largest ratio involving each target facet to mitigate undue influence from a single facet on the final model. Target factor: Whether pairwise ratios were computed with either neuroticism or extraversion facets as the numerator (and then non-target facets as the denominators). B (99.9% CI): Positive effects suggest that PSM ratios were larger than combined PSM-ENET.

5.4. Study 2

In Study 2, I evaluated every big five facet association with SWB. I used exclusively PSM because it yielded more internally valid individual facet-level associations with SWB than zero-order correlations, multiple regression, elastic net and PSM-ENET alternatives. I evaluated PSM-weighted big five effects using all 30 facets and across multiple measures of SWB. The primary outcomes were needs thwarting and satisfaction, which measured suffering and flourishing respectively. Secondarily, I evaluated convergent facet associations with SLS and health. To limit the number of additional hypothesis tests, I only evaluated effects for secondary outcomes when there were observed effects for the primary outcomes in the same direction. Converging effects further increased confidence that the primary effects did not emerge simply because of measurement or sampling error.

5.4.1. Method

5.4.1.1. Procedure

The participants, materials, data preparation and PSM were identical to Study 1. For each facet, I used PSM sample weights that specifically increased the internal validity of its association

with each SWB outcome. For simplicity, I computed each facet effect estimate using a single weighted correlation, rather than 10-fold cross validation. I examined facet effects on needs thwarting and satisfaction separately because they were dissociable SWB constructs. Thus, I evaluated 60 associations in total. Then, I also evaluated whether there were convergent effects for secondary outcomes. In all instances, the significance threshold was again $p < .001$.

5.4.2. Results

All associations were intended to be exploratory. Thus, I used my discretion to focus interpretations on those effects that (a) replicated across primary and secondary outcomes, (b) had CI bounds that were above the $r \approx .10$ threshold for noteworthiness, and/or (c) diverged from the majority of other effects in their factor. I grouped effects with overlapping CIs—and reported their extreme-most CI bounds—except in some borderline cases where effects had plausible theoretical discontinuities. Full results are in Figure 5.2.

5.4.2.1. Needs Thwarting

First, I evaluated facet associations with needs thwarting. All six neuroticism facets exacerbated needs thwarting, and all associations replicated across both SLS and physical health secondary outcomes. Effects for depression, and then anxiety and vulnerability (CI = .19, .23), were noteworthy. Four of six extraversion effects protected against needs thwarting, and they all replicated across both secondary outcomes. Effects for cheerfulness, and then friendliness and gregariousness (CI = -.18, -.12), were noteworthy. Four of six openness effects also impacted needs thwarting. Although none replicated across all three outcomes or were reliably above the threshold for noteworthiness, it was interesting to note that only imagination exacerbated needs thwarting. All six agreeableness facets impacted needs thwarting. Of these, morality and cooperation (CI = -.15, -.09) had noteworthy protective effects, although only the later replicated across all three outcomes. Although the effect for modesty was below $r = .10$, interestingly it alone exacerbated needs thwarting. Its effects also replicated across both secondary outcomes. Finally, all six conscientiousness facets protected against needs thwarting and replicated across both secondary outcomes. Effects for self-discipline, and then self-efficacy and cautiousness (CI = -.18, -.11), were all noteworthy. Thus, there were a range of different associations across the big five factors.

5.4.2.2. Needs Satisfaction

Then, I evaluated facet associations with needs satisfaction. All six neuroticism facets also inhibited needs satisfaction, and effects replicated across both SLS and physical health. Effects for depression and then vulnerability were both noteworthy. All six extraversion effects promoted needs satisfaction and replicated across both secondary outcomes. Cheerfulness, and then friendliness, gregariousness and assertiveness (CI = .13, .22), were all noteworthy. Five of six openness effects also impacted needs satisfaction. Although none replicated across all three outcomes or were reliably above the threshold for noteworthiness, it was interesting to note that imagination both exacerbated needs thwarting and promoted needs satisfaction. All six agreeableness facets also impacted needs satisfaction. Five of the six effects were positive, but only altruism was noteworthy. Its effect did not replicate across both outcomes. Interestingly, modesty *inhibited* needs satisfaction and its effect was both noteworthy and replicated across all four outcomes. Finally, all six conscientiousness facets promoted needs satisfaction and replicated across both secondary outcomes. There were cascading effect magnitudes: Self-efficacy, then self-discipline, then achievement-striving, and then dutifulness were all noteworthy. Overall, there were also wide ranging, and partially divergent, big five facet associations with needs satisfaction.

Figure 5.2. Heatmap of PSM weighted correlations between every big five facet and both the primary outcomes. Point estimate values are Pearson's R correlations with needs thwarting and satisfaction. Bracketed values are the 99.9% CIs. The darker the panel the more positive the association. The X axis contains the big five factors and the Y axis contains each of their six nested facets. The factors and facets (in order) are: Neuroticism (anxiety, anger, depression, self-consciousness, immoderation, vulnerability), Extraversion (friendliness, gregariousness, assertiveness, activity-level, excitement-seeking, cheerfulness), Openness (imagination, artistic interests, emotionality, adventurousness, intellect, liberalism), Agreeableness (trust, morality, altruism, cooperation, modesty, sympathy) and Conscientiousness (self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, cautiousness). * = $p < .001$ for the needs thwarting/satisfaction association. + = $p < .001$ for the secondary satisfaction with life association, when it also converged with the primary association. ^ = $p < .001$ for the secondary subjective health association, when it also converged with the primary association.

5.5. Discussion

Researchers typically use zero-order correlations, stepwise regression and multiple regression to evaluate personality facet effects on SWB. However, when doing so, they neglect intra- and inter-factor personality confounds or suppress effect estimates to the point they are meaningless. Until the present chapter, this trade-off was thought to be inevitable. However, I found that PSM—using pairs of participants who differ on the key variable of interest but have similar covariates—held all potentially confounding personality facet covariates more constant than existing alternatives, and with reduced multicollinearity. Then, I used PSM to better replicate established big five *factor* associations with SWB. Finally, I found that PSM in isolation outperformed both elastic net and combined PSM-ENET solutions. Thus, I used PSM to evaluate the full plurality of big five facet associations with SWB.

PSM outperformed conventional regression alternatives. First, I evaluated the difference in all 29 covariate facet means—where each facet was iteratively made the target—across quartiles before and after PSM. Even among cases from the first and fourth quartiles—the most difficult to match because they comprised the most dissimilar cases—there was still more than a 2/3 reduction in covariate means. In most instances, matching led to at least a 3/4 reduction in covariate means. Moreover, effects held across (a) the facets in all five factors, (b) covariates from the same and different factors, and (c) both original quartile and post-hoc median split target facets. Then, I evaluated the changes in absolute effect magnitudes—when each facet was used to predict SWB—as the number of covariates increased from 1 to 29. PSM effects decreased by less initially and then also stabilized at a higher absolute effect magnitude, compared to multiple regression. PSM effects became significantly larger than multiple regression when there were just three covariates. Even when fitting all 29 covariates, PSM effects were still conserved at approximately half the magnitude observed when fitting a single covariate. All effects held across the four SWB outcomes. Overall, I thus concluded that PSM held covariates more constant than zero-order correlations, and that it produced less multicollinear effects than multiple regression.

Then, I found that PSM better replicated established neuroticism and extraversion *factor* associations with SWB, compared to alternatives. Factor associations require fewer personality controls and thus do not suffer from the same tradeoff between confounding and multicollinearity as facet-level effects. That is, they are more trustworthy. Thus, these effects could serve as the criterion associations. First, I computed an exhaustive set of absolute pairwise ratios between the target replication and non-target facets. I initially found that PSM

ratios were larger than both zero-order and multiple regression ratios for both neuroticism and extraversion facets. Then, I replicated effects comparing PSM to the prevailing machine learning alternative. As a preliminary step, I found that elastic net in isolation and PSM-ENET performed equally well for neuroticism facet effects, and that PSM-ENET better differentiated extraversion facet effects. Thus, I compared the overall better-performing PSM-ENET to PSM. PSM in isolation outperformed PSM-ENET for both neuroticism and extraversion. Moreover, I replicated effects in four different ways that mitigated both the impact of outliers and methodological artefacts. Results again held across all four SWB outcomes. Therefore, I concluded that PSM yielded more internally valid big five facet-SWB associations than plausible regression and even machine learning alternatives.

5.5.1. Big Five Facet Associations with SWB

Thus, I used PSM to evaluate all big five facet effects on SWB. There were separate facet associations for needs thwarting and needs satisfaction. In addition to the remaining 29 facets, I also controlled for basic demographics and survey response bias. In support of Schimmack et al. (2004), depression from neuroticism and cheerfulness from extraversion had relatively large and robust effects on both suffering and flourishing SWB. However, there were also a range of other associations—of noteworthy magnitude (r = .10)—that emerged across both needs thwarting and needs satisfaction, as well as the secondary outcomes. They were vulnerability from neuroticism, friendliness and gregariousness from extraversion, and self-discipline and self-efficacy from conscientiousness. Results for vulnerability may suggest that sensitivity to frequent negative emotions both exacerbates suffering and impedes flourishing (Steptoe et al., 2007). Results for friendliness and gregariousness suggest the importance of both social support and social capital (Helliwell, 2006). Results for self-discipline and self-efficacy may highlight the effectiveness of gritty, goal-oriented, approaches (Sheldon & Elliot, 1999; Duckworth et al., 2007). There may be range of personality facet pathways that impact SWB.

Other associations may help illustrate subtle SWB effects for openness, agreeableness and conscientiousness. Although small, there were a variety of openness facet effects on both suffering a flourishing. They may support the mild SWB benefits of receptivity to new feelings, ideas and experiences (Keng, Smoski & Robins, 2011). The exception was imagination—which was associated with *both* exacerbated suffering and increased flourishing—perhaps because it promotes both rumination and creativity (Plante, Reysen, Groves, Roberts & Gerbasi, 2017). Agreeableness effects highlighted the SWB benefits of behaving prosocially (Keltner, Kogan, Piff & Saturn, 2014). However, *different* kinds of prosociality may protect

against suffering and promote flourishing. Indeed, effects for morality and cooperation suggest that civic behaviour protects against suffering, perhaps because it mitigates ostracism (Uskul & Over, 2017). Contrastingly, altruism may actively give meaning and purpose to life, and thus promote flourishing (Keltner et al., 2014). Importantly, modesty—combined shyness and tempered self-worth—showed the opposite pattern of effects to the other five agreeableness facets. Modesty may be detrimental to SWB because it stymies agentic behaviour in social contexts (Freidlin, Littman-Ovadia & Niemiec, 2017). Alternatively, modest people may simple report having lower SWB. Finally, all six conscientiousness facets had robust benefits that replicated across the four SWB outcomes. Interestingly however, the magnitude of cautiousness and achievement-striving was inverted when I switched from needs thwarting to needs satisfaction. This may support the relative protective benefits of risk aversion, and the relative enabling benefits of aspirational behaviour (Gross, 2015; Locke & Latham, 2002). Therefore, the emergent *pattern* of PSM associations may help reconcile the full range of big five facet effects on SWB.

5.5.2. Propensity Score Matching

PSM is particularly relevant to the study of personality because it helps account for the complex interrelationships between the big five facets. In theory, the 30 big five facets are nested in five orthogonal factors. In practise, both circumplex and general factor personality theories suggest facet relationships cross factor boundaries (e.g. Saucier & Ostendorf, 1999; van der Linden et al., 2016). Until now, this meant that researchers chose between multicollinear or confounded facet associations. Consequently, they found only a few noteworthy associations, or inconsistent associations that contradicted research in adjacent literatures. While previous research has focussed on using PSM to increase the internal validity of quasi-experiments, this chapter used PSM to overcome the trade-off between multicollinearity and confounding in numeric predictors. PSM achieves this by controlling for the complex web of potentially confounding personality facet covariates *during sampling* rather than statistical modelling. Specifically, PSM selects participants who differ on the predictor facet but have convergent scores across the remaining 29 facets. Thus, the remaining 29 facets are less likely to confound results because they have less actual covariance with the target predictor and outcome.

PSM may outperform elastic net, which is the prevailing machine learning alternative. Although these methods target distinct sampling and modelling components of the analysis, they may not be compatible because elastic net compromises individual effect coefficients to optimize the overall accuracy of predicted scores. While PSM weights did improve the extent

that elastic net isolated specific effect estimates, the combined procedure may have still down-weighted particularly large coefficients and rounded small coefficients to zero in order to optimize the extent clusters of correlated variables—and not individual variables—yielded robust predictions (Zou & Hastie, 2005). Of course, PSM in isolation is imperfect. Nevertheless, it may still yield *more* internally valid, consistently replicable *and* parsimonious big five facet-SWB effect estimates than both conventional and elastic net alternatives.

5.5.3. Limitations and Future Directions

There were also several limitations in the present chapter. Principally, I used cross-sectional surveys. Thus, I could not infer causation. In addition, some reported associations may have been artefacts of the self-report format. It was also unlikely that participants sampled were representative of their respective countries, or their countries representative of their world region. While I could have improved representativeness by fitting additional sample weights, this would have diluted PSM weights. Thus, despite being equipped with a large sample, I was still confronted with the tradeoff between using the more internally valid PSM solution, or an alternate, more externally valid, sample weighting strategy. Another alternative was to select a sub-sample of more representative participants. However, this may have compromised matching fidelity because fewer participants meant fewer dyads with closely covarying facet controls. PSM associations also have constraints. The greater the number of covariates, the less perfectly PSM can control for any single variable. Although covariates were usually held at least 3/4 more constant with 29 controls, changing intercorrelations and/or additional controls may reduce internal validity. Finally, internal validity would also be diluted when evaluating *interactions* between facets because models must then combine *both* sets of PSM weights. Thus, PSM may be most useful for *bivariate* descriptive research.

Nevertheless, PSM associations may still offer more parsimonious descriptive insights than alternatives in a range of contexts. For example, it could be used to evaluate the unique effects of different associated cognitions—like anxiety and rumination—on clinical depression. Further, high-fidelity PSM associations might help researchers identify differential patterns of effect sizes. This may allow researchers to better direct their attention to the most promising associations. From the chapter, researchers might be particularly interested in confirming patterns for neuroticism and extraversion effects, the different prosocial behaviours that protect against suffering and promote flourishing, the deleterious SWB effects of modesty and the differential SWB benefits of caution and achievement striving. Overall, PSM may help isolate the magnitude of different multicollinear effects on SWB, as well as other outcomes.

5.5.4. Conclusion

The big five is a comprehensive account of personality. However, until the present chapter there were contradicting big five *facet* associations with SWB, which mostly emphasised the importance of trait affectivity. This limited range of facet associations may have been an artefact of stepwise and multiple regression approaches. While both might sometimes mitigate confounding, they also make the burden of proof artificially high for some or all the other facets. Consequently, results are often incompatible with adjacent literatures—which also emphasise the SWB effects of (e.g.) friendliness and altruism facets, among others. To remedy, I deployed PSM to mitigate the multicollinearity-confounding tradeoff. PSM associations held covariates more constant than zero-order associations, and caused more realistic effect size estimates than multiple regression. PSM also better replicated established personality factor associations with SWB than both conventional regression and machine learning approaches. Then, I used PSM to reveal patterns of associations throughout the entire big five. Results suggested multiple pathways for SWB beyond mere negative and positive affect. Thus, I concluded that individuals with a wide range of different personality profiles might be protected against suffering and predisposed to flourishing.

Chapter 6

Twenty-Nine Way Interactions? Random Forest Constellations Isolate the Personality Facets that are Prevalent in Extreme SWB

6.1. Abstract

Social psychology approaches must simplify the individual for scientific rigor. This often involves the assumption that at most two or three other personality traits *change* (i.e. moderate) the relationship between any given target trait and the outcome. It thus fails to account for the full range of intrapersonal contingencies. To remedy, I used random forest to capture effects for each facet across differing levels of *all* remaining 29 facets. Of course, it is impossible to interpret these complicated dependencies. Instead, I used them to find the most robust facet main effects. Study 1 found that a single random forest model generalised to participants in different world regions and sociodemographic strata. I also found that it most accurately predicted SWB for participants with extreme 1% self-reported scores. In Study 2, I used *four billion* simulated cases to evaluate the facets that were most prevalent in cases with extreme 1% SWB. Results largely confirmed findings in Chapter 5: cases scoring high on most neuroticism facets, and low on most of the remaining facets had extreme needs thwarting. I observed the inverse pattern for needs satisfaction. However, Study 2 used *naturally occurring* trait constellations, which meant that some intrapersonal contingencies emerged more frequently than others. Thus, Study 3 used another four billion simulated cases where there was equal likelihood of *every* possible trait combination. Results suggested a *smaller subset* of 9 facet effects—such as the benefits of cooperation on needs thwarting and altruism on needs satisfaction—that were robust to *fully changing* intrapersonal contexts. They may be most robustly associated with extreme SWB because they are the most intransient. Overall, random forest (S1) found that single *universal* facet constellations were associated with suffering and flourishing, even when relaxing the assumption of non-complex effects; (S2) replicated results from Chapter 5 concerning the facet main effects that were most prevalent in the population sampled; and, (S3) isolated a subset of the most internally valid facets.

6.2. Introduction

Human experiences are multifaceted, rich and, importantly, unique. Even monozygotic twins show quantifiable and sometimes even quite extreme differentiation (Chen et al., 2018). Theories about socialization and, more recently, epigenetics have been recruited to explain this diversity (Witherington & Lickliter, 2017). Personality theories also acknowledge that individuals are unique. However, even when assuming that each of the big five facets has just three levels, there are still 2.06×10^{14} possible score profiles. Any of these may change (i.e. moderate) the effects of the target facet on outcomes like SWB. To date, the prevailing approach assumes that only small *a priori* defined subsets of these contingencies—usually comprising just one or two traits—act as moderators. It is appealing because it identifies a discrete set of effect contingencies that can be triangulated to isolate possible mechanisms.

However, this approach fails to account for how changes in *whole* constellations of personality traits impact SWB. Chapter 5 highlighted the difficulty accounting for the full range of plausible interrelated facet effects—which often transcend factor boundaries—even when only evaluating main effects, due to increased multicollinearity. This problem is compounded when moving to two- and three-way interactions, and then beyond. It is impossible for existing models to account for all the possible ways that other traits can *change* the magnitude and/or pattern of observed effects. Random forest machine learning—usually thought of as a black-box method to predict outcome variables—may remedy by capturing the complex patterns of dependencies between all 30 big five facets and SWB. I propose that it can be used to predict SWB for huge simulated databases of participants (e.g. billions) who have *comprehensive* ranges of plausible facet score constellations, and then isolate the cases with different stratum (e.g. bottom and top percentile) predicted SWB. These interactions—with up to 29 layers when using the 30-facet operationalization of the big five—are too numerous and too complex to interpret separately. However, effects that emerge across a large percentage of different contingencies may be immutable and thus driven by robust mechanisms. As such, random forest may help differentiate the most internally valid personality effects. Thus, I aim to (1) evaluate the universality of random forest predictions, (2) describe the facets that are most prevalent in real-world cases with extreme SWB, and (3) identify the subset of facets that have robust prevalences in fully random constellations, which are unlikely to occur in the real world.

6.2.1. Prevailing Interaction Approaches

Circumplex theories of personality are designed to capture two-way interactions. During Chapter 5, I noted that they demonstrate how personality facets cluster across factor boundaries, thus challenging the strictly hierarchical structure of the big five. Circumplexes—conceptualised as individuals' intersecting location on two trait scales and by virtue expressed as a *single* set of Cartesian coordinates—are also considered co-dependent and thus irreducible interactions (Woods & Anderson, 2016). They predict various SWB-relevant outcomes. For example, Smith et al. (1998) found that, among married couples, agency-communion circumplex coordinates were differentially associated with both cardiovascular reactivity to disagreements and aptitude tests. Similarly, Van Katwyk, Fox, Spector and Kelloway (2000) found that the trait valence-arousal affect circumplex predicted job satisfaction, stress and physical health. Schwartz & Boehnke (2004) located ten discrete human values on the trait openness-individualism circumplex. Then, using this circumplex, Joshanloo & Ghaedi (2009) found that individuals who clustered around achievement and traditionalism value orientation markers had an exceptionally high sense of purposefulness. In sum, circumplex approaches suggest that interacting trait pairs have predictive validity.

By contrast, continuous big five trait approaches can have both linear and non-linear effects. I discussed linear big five associations with SWB in Chapter 5 (e.g. for cooperation and self-discipline). Recently, non-linear combinations of big five variables have been linked to various forms of psychological impairment, especially in sub-clinical populations (Suzuki, Samuel, Pahlen & Krueger, 2015). For example, using a three-year prospective longitudinal design, Gershuny & Sher (1998) found that extraversion reduced the association between high neuroticism and global experiences of anxiety. Later, Naragon-Gainey & Simms (2017) found that *also* having high conscientiousness increased this protective effect. At the facet level, Kaplan, Levinson, Rodebaugh, Menatti and Weeks (2015) found that having low trust increased the negative association between openness and social anxiety. Allen et al. (2017) used the big five aspects—ten intermediary traits between the factors and facets—to find that withdrawal, industriousness and enthusiasm interacted with one another to predict depressive symptomology in both non-clinical and clinical populations. Overall, the big five factors and their sub-components may also have contingent associations with SWB-relevant traits.

6.2.2. Problems with the Literature

However, these approaches fail to account for the *whole* individual. Circumplexes presuppose exclusively two-way interactions when effects may have a different number of contingencies. For example, the valence-arousal model of affect may apply less to feelings of disgust—which also rely on gastrointestinal cues—than other emotion experiences (Eskine, Kacinik & Prinz, 2011). Further, big five approaches have difficulty selecting appropriate moderators due to complex-patterns of covarying facet scores. It is unclear whether facet contingencies are genuine, or whether they proxy for other facets. It is also unclear whether the same moderators generalise across socio-demographic and cultural strata. Consequently, issues in variable selection may result in fragmented and potentially biased moderation effects. These ultimately increase the risk of spuriousness. Finally, there are also interpretation difficulties. To illustrate, a four-way interaction is a contingency on the contingency of the contingency of the main effect. Results may be incomprehensible—and thus unactionable—long before they account for the full range of possible moderators.

Moreover, even robust and comprehensive patterns of contingencies do not necessarily help triangulate causal mechanisms. This is because trait prevalences are often yoked to one-another. For example, being high in friendliness might predispose someone to also being high in gregariousness. This causes range restriction, which is when the sample disproportionately comprises participants with predictor scores that are unrepresentative of the full range of possible values (Wiberg & Sundström, 2009). In complex interactions, range restriction occurs when levels of each predictor preferentially co-occur with certain levels of each moderator.[10] It means that each predictor exerts its impact on the outcome within a *bounded* and potentially unrepresentative subset of intrapersonal contexts. Results from cases with exceptional cooccurrence patterns—the kind needed to triangulate effects with genuinely robust mechanisms—are typically discounted as residual errors, which thus have little impact on the magnitude of observed effect sizes (although they may impact model precision). Put another way, range restriction is another source of confounding: emergent effects may cause the outcome, or they may be artefacts of the precise set of intrapersonal trait contexts prevalent in the sample. Emergent machine learning technologies help reconcile these limitations.

[10] It is important to note that range restriction concerns the *density* of score distributions, not their actual ranges.

6.2.3. Random Forest

Random forest is a machine learning technique that can capture extremely complex non-linear trait combinations in the population. Specifically, it generates decision trees that predict (e.g.) SWB on sub-samples of participants.[11] In every decision tree, a randomly selected predictor is binary split at the value that leads to the greatest reduction in model imprecision. Then, this procedure is repeated—within existing splits—using subsequent randomly selected predictors, until adding more randomly-selected predictors stops reducing model inaccuracy. All participants in the same nested set of splits are then given the same predicted score. Thus, the facets can form complex patterns of non-linear dependencies because they are nested under a range of superordinate facet splits. Then, this entire procedure is iteratively repeated—typically at least 500 times—with different random predictor orderings and different bootstrapped participants. Final values are the mean predictions from *all* trees. Given enough iterations, each predictor has an approximately equal chance of being modelled at each different level of the decision tree, and both with and without its most collinear variables. This helps random forest account for complex non-linear effects while mitigating both potential confounding, and the multicollinearity that would be caused by fitting all covariates together in a single iteration.

To date in psychology, random forest has been used as a proof-of-concept tool that shows groups of predictors are related to an outcome. For example, Mowery, Park, Bryan & Conway (2016) decomposed individual Tweets into their linguistic components and then used random forest to determine whether they indicated normal functioning, depressive symptomology or clinically diagnosed depression. Walsh, Ribeiro and Franklin (2017) used medical records from self-harming patients to random forest predict whether they attempted suicide, with 84% accuracy. Manesi, Van Lange, Van Doesum & Pollet (2018) used random forest to determine which variables associated with prosociality, socio-demographics and environmental primes were associated with charitable giving following a typhoon in the Philippines. They found that variables relating to empathy were most important to model accuracy.[12] Overall, random forest has so far been deployed because it may increase prediction accuracy over other linear and more simplified non-linear modelling approaches. Most recently in Manesi et al. (2018), it was also used to isolate the predictor variables that were most implicated in model accuracy. However, the resultant importance scores still do not offer empirical support that predictors are

[11] Subsamples are usually bootstrapped (i.e. sampled with replacement) to match the original sample size.
[12] Although effect direction may seem intuitive, this cannot be confirmed by random forest importance score alone because they simply indicate the extent that excluding a variable increases model imprecision.

implicated specifically in low or high outcome scores (e.g. they could have been more implicated in middling outcome scores). Therefore, existing random forest approaches can show that two phenomena are related but, to date, it does not isolate specific directional effects.

The present chapter is a first-of-its-kind attempt to isolate specific *directional* predictors from random forest models. Random forest has been viewed as a 'black box' approach until now because every variable can have both positive and negative effects on the outcome, depending on how it combines with the other variables. As such, it does not yield directional coefficient estimates. However, the algorithm can be *reverse-engineered* by examining variable *prevalences* in different strata of predicted SWB, such as the top and bottom percentiles. The mean prevalence is the variable score that combines with the most levels from other variables to yield scores in that stratum. Put another way, mean prevalence is the variable score that offers the most *affordances* for experiencing that stratum of SWB.

An added challenge of this approach, however, is that variable prevalences conflate predictors and co-occurring traits. That is, facets that artefactually co-occur with genuine predictors in the population will also be overrepresented in (e.g.) bottom and top strata predicted SWB. To date, difficulty isolating the predictors may be one cause of contradicting findings in theory-driven conventional approaches, and concomitant unempirical analyses (see Chapter 5). Random forest offers a comprehensive solution: genuine predictors will still have extreme low or high prevalences when SWB is predicted using participants where every constellation of facet scores is *equally likely*. This criterion has been made more salient during social psychology's replication crisis, which recognized that robust main effects must hold both within populations across testing environments, and in different populations (Shrout & Rogers, 2018). A classic example of meeting this robustness criterion is compliance in conditions of uncertainty. It emerges across demographics and cultures, and *almost always* occurs when requests are made by perceived legitimate authority (Packer, 2008). Applied to the big five and SWB, genuine predictors may have prevalences that are particularly *non-contingent* on the remaining facets.

6.2.4. Present Studies

The present chapter aimed to evaluate the facets most associated with SWB when accounting for individuals' whole constellation of big five personality traits. At the outset, it was unclear whether relaxing the assumption of linear controls meant a single statistical model still surmised the relationship between facet predictors and SWB throughout the entire population. Thus, Study 1 evaluated whether (1) models were more accurate for some sociodemographic

strata than others; (2) the same complex facet effects emerged across deliberately biased sub-samples; and, (3) predictions were more accurate for cases with extreme vs middling SWB. Then, Study 2 evaluated the (4) facets that were most prevalent in extreme (i.e. lowest and highest) population SWB. I used simulated cases to account for different realistic *real-world* patterns of facet co-occurrence, rather than relying on the single-observed and potentially idiosyncratic pattern in the sample. Finally, Study 3 (5) evaluated the extent prevalences replicated when I relaxed the assumption of facet co-occurrence, so that there was equal likelihood of every possible set of facet score combinations. This helped evaluate the facets that most likely caused extreme SWB. I expected a subset of facets from (4) to emerge in (5) because the latter were predicated on having effects that were robust to a much wider range of intrapersonal contingencies.

6.3. Study 1

6.3.1. Method

6.3.1.1. Procedure

I used the same data and preliminary approaches as Chapters 4 and 5. Specifically, participants were the 36,498 multinational internet panellists from 33 different countries and 14 language groups. Full details of the sampling procedure and methods are in Chapter 1. The 30 big five facets were measured using the 120-item public version of the NEO personality inventory (NEO-IPIP-120; N = 36,498), the primary outcomes were needs thwarting and satisfaction (N = 29,629), and the secondary outcomes were SLS and self-reported health (N = 36,498). I imputed the 2% of variable scores that were missing, within each country. Then, I removed all covariation from the data between the controls—sex, age, social class, religiosity and response bias—and the (a) 30 personality facets, and (b) 4 SWB outcomes. Finally, the facets and outcomes were z-scored so that all values were relative to participants' countrymen. Having already accounted for the controls, the analyses thus focussed exclusively on personality facets and SWB. Effects were relative to participants' countrymen rather than absolute.

6.3.1.2. Random Forests

I generated separate random forests for each outcome, using 10-fold cross validation (10-FCV). As in Chapter 5, 10-FCV separates participants into ten equal-sized groups. Then, it develops an exhaustive set of models on 9/10 of the groups and evaluates associations between true and

model predicted scores for the 10th, excluded, group. Ten folds is sufficient to mitigate spurious results (Kuhn, 2008). Each 10-FCV used R's 'randomForest' package default values for decision trees (500) and predictors randomly sampled at each split (1/3) because they have demonstrated robustness (Liaw & Wiener, 2002). Due to computational constraints, I generated model predictions using only 10,000 randomly sampled participants. To mitigate idiosyncratic findings, I generated three models for each outcome using different random samples. Predicted scores for participants involved in model development were from the 10-FCV iteration where they were not involved in model development. Then, I used the overall 10-FCV model to generate predicted scores for remaining participants. Thus, every participant had three predicted scores for each outcome.

Models explained noteworthy variation in each SWB outcome. I computed accuracy—as both the correlation and mean absolute error (MAE)—between self-reported and predicted scores. Results are in Table 6.1. For every outcome, Cronbach's alpha suggested that the three predicted scores were almost perfectly identical. Thus, I took their average. Models explained 35% and 41% variation in self-reported scores for needs thwarting and satisfaction respectively, and 28% and 11% for SLS and health respectively. Although model accuracy was approximately equal to the elastic nets used in Chapter 5—which evaluated main effects—random forests may have still yielded more parsimonious patterns of *specific* facet effects because it was designed to capture more ecologically valid facet interrelationships. Observed R^2 model accuracies were the equivalent of MAE ≈ 3/5 SDs in self-reported scores for the primary outcomes, and MAE ≈ 7/10 SDs in self-reported scores for the secondary outcomes. Put another way, self-report and predicted SWB were on average different by greater than 0.5 SDs. Despite this overall imprecision, results from Chapter 3 suggested that MAE may have been more informative at extreme prediction magnitudes.

Table 6.1.

Prediction accuracy for the big five facets on SWB, using random forest

DV	A	M (SD)	R	MAE (SD)
Needs Satisfaction	.99	-0.02 (0.61)	.64 (.63, .65)	0.59 (0.50)
Needs Thwarting	.99	0.00 (0.57)	.60 (.58, .61)	0.63 (0.50)
Satisfaction with Life	.99	0.00 (0.49)	.53 (.52, .55)	0.67 (0.52)
Subjective Health	.97	-0.01 (0.32)	.33 (.31, .35)	0.74 (0.59)

Notes. α = Cronbach's alpha computed from three random forest predicted scores, for each outcome. R = Correlation between average predicted score and self-report score, with 99.9% confidence intervals computed from effect SEs. MAE = Mean absolute error.

Finally, I evaluated the extent that each facet impacted model accuracy. Complex non-linear associations meant that every variable could have both positive and negative effects on SWB. Thus, the final models did not yield directional coefficient weights, but variable importance scores. These scores captured the extent overall model accuracy decreased in the subset of trees where the variable was randomly excluded. Results are in Table 6.2. For needs thwarting, the most important facets were depression, vulnerability, anxiety, cheerfulness and then self-discipline. For needs satisfaction, the most important facets were self-efficacy, cheerfulness, self-discipline, depression and then achievement-striving. For both secondary outcomes, the most important facets were depression and then cheerfulness. Despite putative convergence with PSM associations in Chapter 5, importance scores did not necessarily suggest the facets were implicated in *extreme* SWB.

Table 6.2

Big five facet importance for random forest prediction accuracy

Factor	Facet	Thwarting	Satisfaction	SLS	Health
NUR	F1	37%	13%	25%	53%
	F2	18%	13%	18%	39%
	F3	**100%**	44%	**100%**	**100%**
	F4	14%	14%	18%	38%
	F5	12%	13%	19%	42%
	F6	48%	18%	21%	50%
EXT	F1	15%	24%	26%	35%
	F2	12%	13%	22%	36%
	F3	12%	28%	19%	36%
	F4	12%	14%	21%	38%
	F5	12%	14%	18%	35%
	F6	29%	84%	91%	70%
OPN	F1	12%	14%	20%	35%
	F2	12%	16%	19%	35%
	F3	12%	17%	18%	37%
	F4	12%	13%	19%	36%
	F5	11%	16%	18%	36%
	F6	12%	12%	18%	37%
AGR	F1	12%	16%	29%	38%
	F2	19%	14%	20%	35%
	F3	12%	23%	17%	35%
	F4	16%	12%	18%	36%
	F5	12%	15%	23%	38%
	F6	12%	14%	19%	36%
CON	F1	17%	**100%**	29%	41%
	F2	13%	12%	19%	36%
	F3	15%	17%	20%	37%
	F4	13%	31%	19%	37%
	F5	28%	50%	41%	42%
	F6	18%	12%	18%	35%

Notes. Importance scores do not have an intuitive interpretation; thus, for each outcome, I transformed them into percentages of the maximum observed importance (in bold).

6.3.2. Results

Random forest with 10-FCV were performed using R's 'randomForest' package with 'caret' package interface. It yielded complex combinations of facets that predicted SWB. There were too many potential facet combinations to interpret (2^{29} > 500 million), and thus I focussed on (a) overall model accuracy, (b) facet importance scores, and (c) model accuracy for increasingly extreme predicted SWB. Accuracy for (a) and (c) was MAE because it yielded separate errors for every participant, and thus could both be fit as an outcome in individual-level prediction models, and used to create mean accuracies for participant subgroups (e.g. each percentile). I retained all outlier predictions because there was sufficient statistical power to mitigate their leverage. The significance threshold was $p < .001$.

6.3.2.1. Model Accuracy Across Controls

First, I evaluated whether models were equally accurate across sociodemographic strata. Although I apportioned all control *main effects* from the random forests models, they were still free to *interact* with the facets to change overall prediction accuracy. Put more simply, models may have had greater predictive power in some demographic groups (e.g. women) than others. Thus, I used multiple regression to evaluate whether participants' individual-level control characteristics, as well as their world region (Anglosphere, Asia, Europe and Latin America), were associated with their MAE. Anglosphere was the reference category because it yielded the lowest MAE for all four outcomes.[13] This maximized coefficients for the other three world regions, thus increasing the likelihood for evidence against the universality of my models. Results are in Table 6.3. Total model R^2 was < 1% for the models including just control characteristics, across both primary outcomes. Concomitant total model R^2 was ≈ 1% across both secondary outcomes. For primary outcomes, the largest observed effects were |b| = 0.02 for binary predictors and |b| = 0.03 for continuous predictors. Thus, for binary predictors the maximum total mean difference in MAE between groups was around 2% of the SD. For continuous predictors, every SD increase was associated with a maximum 3% SD change in the outcome. For secondary outcomes, the largest observed effects were approximately double this figure. However, their CIs also converged on zero, and thus the larger point estimates could have been caused by sampling error. Overall, observed magnitudes were small enough to suggest model accuracy was constant across different population sub-samples.

Table 6.3

Differences in overall random forest model accuracy across socio-demographic controls

Predictor	Satisfaction (CI)	Thwarting (CI)	SLS (CI)	Health (CI)
Sex	-0.01 (-0.03, 0.01)	-0.03 (-0.05, -0.01)	-0.01 (-0.03, 0.01)	0.01 (-0.01, 0.03)
Age	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
Religious	0.02 (0.00, 0.04)	0.01 (-0.01, 0.03)	0.00 (-0.01, 0.02)	0.02 (-0.01, 0.04)
Social class	-0.02 (-0.03, -0.01)	-0.01 (-0.02, 0.00)	-0.04 (-0.05, -0.03)	-0.05 (-0.06, -0.04)
Response bias	0.00 (-0.01, 0.01)	0.03 (0.02, 0.04)	0.03 (0.02, 0.04)	0.03 (0.02, 0.04)
Asia	0.01 (-0.02, 0.04)	0.03 (0.00, 0.06)	0.06 (0.03, 0.09)	0.00 (-0.03, 0.04)
Europe	0.00 (-0.03, 0.03)	0.02 (-0.01, 0.05)	0.02 (-0.01, 0.05)	0.02 (-0.01, 0.05)
Latin America	0.02 (-0.01, 0.05)	0.02 (-0.01, 0.04)	0.03 (0.01, 0.06)	0.01 (-0.02, 0.04)

Notes. Anglosphere was the reference category. Values are beta coefficients and 99.9% CIs (in brackets).

[13] This may have been because all scales were initially developed using English-speaking participants.

6.3.2.2. Predicted Scores in Different Sub-Samples

Next, I evaluated whether facet constellations were consistently associated with SWB, across deliberately biased sub-samples. Even though overall model accuracy was relatively constant across the control variables, sub-groups in the population could still have had different predictor relationships. Thus, I compared predicted scores from models generated using specific, biased, control variable strata. There were separate models for women, men, those currently practicing and not practising a religion and each world region. There were also models for low, middle and high thirds for age, social class and response bias. I generated a single random forest model using each stratum, where I randomly sampled a maximum 10,000 participants for 10-FCV (when relevant), or all participants in strata where N < 10,000.[14] Then, I generated predicted scores for all participants, as per the Procedure. Model summary statistics for the primary outcomes are in Table 6.4. As a preliminary step, I found that all models were approximately as accurate as the initial models using fully random participants. Then, for both primary outcomes, I found $\alpha > .99$ for the predicted scores from different strata. This corresponded to mean correlations between the predicted scores of $r = .98$. The mean correlation of these biased predicted scores to predicted scores obtained using fully random participants was $r = .99$ ($SD < .01$). Results also converged for SLS and health ($\alpha >= .99$; $R_{SLS} >= .91$; $R_{Health} = .94$). Therefore, I concluded that facets all combined in approximately the same way to predict SWB across the various biased subsamples.

[14] I did not fit multiple random forests models for each stratum (a) to conserve computational power, and (b) because internal consistencies observed in the procedure suggested that they were unnecessary.

Table 6.4.

Total random forest model prediction accuracies using different, biased, subsamples

		Needs Satisfaction		Needs Thwarting	
Variable	Level	R	MAE (SD)	R	MAE (SD)
Sex	Male	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
	Female	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
Religious	No	.64 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.5)
	Yes	.63 (.62, .64)	0.60 (0.50)	.59 (.58, .60)	0.63 (0.51)
Region	Asia	.63 (.62, .64)	0.60 (0.50)	.59 (.58, .59)	0.63 (0.51)
	Latin	.63 (.62, .64)	0.60 (0.50)	.59 (.58, .60)	0.63 (0.5)
	Anglo	.63 (.62, .64)	0.59 (0.50)	.59 (.58, .59)	0.63 (0.51)
	EU	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
Age	Young	.63 (.63, .64)	0.60 (0.50)	.59 (.58, .60)	0.63 (0.5)
	Middle	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
	Old	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
Social Class	Bottom	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
	Middle	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
	Top	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.5)
Response Bias	Low	.63 (.62, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
	Medium	.63 (.63, .64)	0.59 (0.50)	.59 (.58, .60)	0.63 (0.51)
	High	.63 (.62, .64)	0.60 (0.50)	.59 (.58, .60)	0.63 (0.51)

Notes. There was almost perfect convergence in model accuracies for models generated using different biased participant sub-samples. I excluded secondary outcomes from Table 6.4 because their effects almost totally converged with primary outcomes, and they detracted from overall table readability.

6.3.2.3. Classification Accuracy by Extremeness

Next, I evaluated whether predicted SWB scores were most accurate for extreme cases. To calculate extremeness, I took percentile rank of the absolute values of each self-reported and predicted SWB outcome (1 = "Least extreme", 100 = "Most extreme"). In the first instance, I was interested in MAE for true vs predicted score percentile. Result are in Figure 6.1. For both primary outcomes and SLS, MAE decreased and then plateaued as extremeness increased from 1 to ≈ 95; then, it decreased sharply from extremeness ≈ 96 to 100. It was lowest for the most extreme percentile cases, where MAE = 19.24 (SD = 24.63) for needs thwarting, MAE = 12.73 (SD = 18.22) for needs satisfaction, and MAE = 19.07 (SD = 24.04) for SLS. Thus, cases with the highest percentile predicted scores on average had self-report scores that were at least in the top 1/5. The exception was for health—which showed a negative quadratic trend across all extremeness values—perhaps because of the relative inaccuracy of its predicted scores. Thus, I omitted it from subsequent analyses.

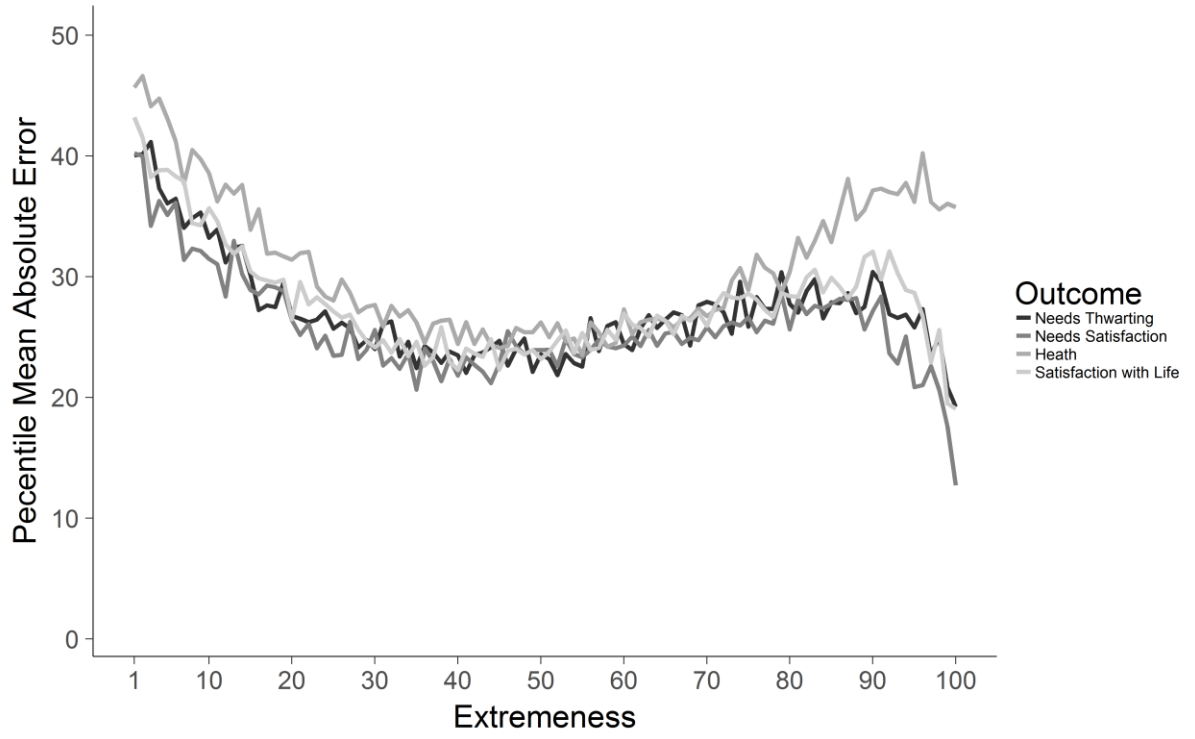


Figure 6.1. Mean absolute error by the percentile extremeness of predicted scores. The 1st percentile was the 1% absolute predicted scores closest to the mean for both primary outcomes (needs thwarting and needs satisfaction), as well as for both secondary outcomes (satisfaction with life (SLS) and health). The 100th percentile was the concomitant 1% absolute predicted scores furthest from the mean. Then, I also converted true scores to extremeness percentiles. Thus, the Y axis was the mean absolute difference between predicted and true score percentile.

Classification accuracy then increased for extreme cases when I categorized them into ecologically valid groups. Items for the primary outcomes and SLS were measured on Likert scales with seven-points (1 = "Strongly disagree"; 7 = "Strongly agree"). Thus, I bucketed true and predicted scores into these same seven categories. Results are in Figure 6.2. As extremeness increased, there was an exponential increase in the percentage of cases classified correctly. Thus, predicted scores were again most informative for the extreme cases. For the most extreme percentile, correct classification was 77% for needs thwarting, 86% for needs satisfaction, and 74% for SLS. Of the cases classified incorrectly, 9%, 7% and 16% respectively were incorrectly misclassified by *only one category*. This was the equivalent of wrongly predicting that cases "disagreed"/"agreed" with SWB items, on average. Thus, even a subset of the misclassifications may have also been interpretable. Overall, random forests may have thus yielded > 85% meaningful classifications—the majority of these correct rather than near misses—for the most extreme 1% of predicted scores.

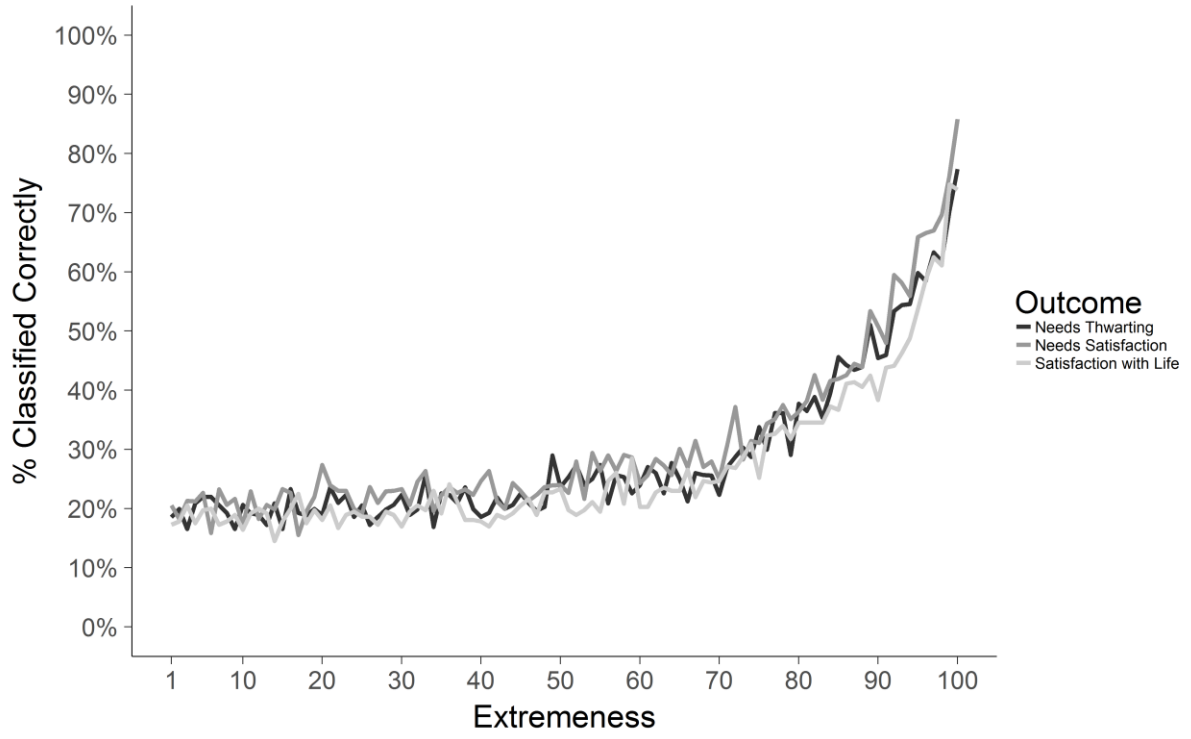


Figure 6.2. Classification accuracy by percentile extremeness of predicted scores. The 1st percentile was the 1% absolute predicted scores closest to the mean for both primary outcomes (needs thwarting and needs satisfaction) and satisfaction with life. The 100th percentile was the concomitant 1% absolute predicted scores furthest from the mean. Classification was into one of seven equal sized and ordered buckets. Classified correctly was the percentage of cases in each extremeness percentile with matching predicted and true score buckets.

6.3.2.4. Confusion Matrices for Top Percentile Extreme Classifications

Finally, I used confusion matrices to evaluate classification bias. As above, predicted scores were classified into seven equally sized buckets that ranged from 'lowest' to 'highest' scores. Table 6.5 contains the multigroup contingency matrices for *all* cases in the data, across both primary BMPN outcomes and SLS. Results confirmed that classifications were most accurate for the extreme categories (i.e. 'one' and 'seven') across all three outcomes. Classifications were least accurate for the middle three categories, likely because it was possible to make at least two classification errors in both directions. There were more misclassifications to adjacent buckets than non-adjacent buckets, and the buckets furthest away from each true score bucket had the fewest misclassifications. Finally, there was some evidence for asymmetric classification accuracy in the most extreme buckets, across all three outcomes. For example, in needs satisfaction there were more correct classifications into category seven than there were into category one, which I probed into further below.

Table 6.5.

Multigroup Confusion Matrix for Bucketing Predicted SWB

DV	Predicted Bucket	True Bucket						
		One	Two	Three	Four	Five	Six	Seven
TWT	One	**2040**	942	567	321	189	125	49
	Two	962	**1047**	848	584	423	239	130
	Three	521	877	**869**	745	571	423	227
	Four	337	579	714	**832**	788	611	371
	Five	166	384	605	726	**848**	887	617
	Six	119	240	385	625	829	**1015**	1020
	Seven	88	164	245	399	585	933	**1818**
SAT	One	**1892**	981	572	368	236	119	65
	Two	1062	**1119**	843	549	358	217	85
	Three	602	879	**951**	755	562	348	136
	Four	338	599	785	**870**	811	561	268
	Five	189	366	591	821	**931**	885	450
	Six	100	196	347	593	855	**1116**	1026
	Seven	50	93	144	276	480	987	**2202**
SWL	One	**2353**	1022	624	466	344	245	160
	Two	1172	**1161**	947	754	550	372	258
	Three	670	1004	**1002**	868	738	590	342
	Four	425	817	959	**972**	859	692	490
	Five	265	566	753	920	**1001**	952	757
	Six	202	408	575	741	982	**1175**	1131
	Seven	127	236	354	493	740	1188	**2076**

Notes. TWT = Basic psychological needs thwarting. SAT = Basic psychological needs satisfaction. SLS = Satisfaction with life scale. N = 29,629 for needs thwarting and satisfaction. N = 36,498 for SLS. There were an equal number of cases bucketed into each of the true and predicted score categories (ranging from "one" to "seven").

Then, I evaluated classification bias in extreme percentile predicted SWB. Cases in the 0.5^{th} predicted score percentile were necessarily bucketed in category 'one' and cases in the 99.5^{th} predicted score percentile were necessarily bucketed in category 'seven'. Thus, predicted categories within each percentile group did not vary, and I could only compare the number of correct vs incorrect classifications. Results are in Table 6.6. They confirmed the asymmetries observed in Table 6.5: classifications were more accurate for the bottom vs top 0.5^{th} percentile in needs thwarting and SLS, and for the top vs bottom 0.5^{th} percentile in needs satisfaction. However, differences were marginal for needs thwarting and SLS, and thus could have been due to sampling error. The larger difference for needs satisfaction could have simply been because participants were more accurate at self-reporting flourishing than the absence of flourishing. Nevertheless, I proceeded with the existing random forest classifications for needs

satisfaction, as well as the other SWB outcomes, because I planned to aggregated facet prevalences for 0.5th (reversed) and 99.5th percentile predicted SWB. As such, the less accurate pole likely widened facet prevalence CIs, which simply increased the stringency of existing Type 1 error-detection thresholds.

Table 6.6.

Correct classifications for extreme percentile predicted SWB

DV	0.5th Perc.	Correct		%
		Y	N	
Thwarting	Bottom	118	31	79%
	Top	112	36	76%
Satisfaction	Bottom	119	30	80%
	Top	135	13	91%
SWL	Bottom	142	41	78%
	Top	132	50	73%

Notes. SWL = Satisfaction with life. I only retained the bottom and top 0.5th percentile predicted scores. This was 297 total cases for both BMPN outcomes, and 365 cases for SWL.

6.4. Study 2

Results from Study 1 suggested that overall prediction accuracy, and the constellations of big five facets implicated in random forest SWB, were both universal. Then, it also evaluated differences in the accuracy of random forest models by the extremeness of participants' predicted scores. For needs thwarting, needs satisfaction and SLS, random forest prediction accuracy was highest for the most extreme 1% of cases. I discarded health because it did not show the corresponding accuracy improvements. Then, I evaluated the accuracy of remaining effects again, using seven ecologically valid categories. In the three retained outcomes, > 85% of classifications from the predicted scores meaningfully mapped onto participants' true scores. In Study 2, I thus evaluated the facet constellations implicated in extreme 1% predicted SWB.

When evaluating complex non-linear effects, parts of the facet intercorrelations in the sample may be attributable to sampling error. If so, they were unrepresentative of the population pattern of intercorrelations. This is especially problematic for assessing intra-facet contingencies because it may cause restriction of range. To remedy, I used Cholesky Decomposition (CD) to simulate facet scores with different plausible intercorrelations. In

psychology, CD is often used to transform different sets of variables—e.g. genetic and environmental—so that they are uncorrelated, thus allowing for orthogonal analyses (Archontaki, Lewis & Bates, 2013). However, the CD process can also be reversed to simulate variable scores that adhere to a pre-defined correlation structure (Davis, 1987). Thus, I iteratively simulated batches of cases using patterns of facet intercorrelations that were randomly sampled from the respective SE distributions of all the observed bivariate facet associations. Then, I used random forests to predict their SWB. I was interested in facet prevalences that were reliably below or above zero in the 1% most extreme cases, for 99.9% of the CD batches. This was an approximate bootstrapped CI, where there was $p < .001$ likelihood that population facet means were outside the observed prevalence ranges.

6.4.1. Method

6.4.1.1. Procedure

The participants and materials were the same as Study 1. To simulate cases, I first generated two 30*30 square matrices: (a) bivariate facet correlations, and (b) concomitant SEs. That meant each facet had its own column and corresponding row, correlations on the diagonal axis all equalled one and the lower and upper triangles of both matrices were symmetrical. For each cell in the lower triangle of the correlation matrix, I imputed one value from a randomly simulated sample of 1,000 normally distributed correlations where the mean was the observed point estimate and the SD was the observed SE. Simulated correlations > |3| SE from the point estimate were truncated to prevent unrepresentative matrices. Then, I transposed the lower triangle of the matrix onto the upper triangle. Thus, the full matrix comprised facet intercorrelations that were randomly sampled from the full range of plausible population effects. Then, I used CD to simulate facet scores for 200,000 cases from the new matrix. This meant that there were 1,000 cases for both the bottom and top 0.5^{th} percentiles, which was large enough to establish central tendency (Israel, 1992). Then, I also truncated facet scores > |3| SD from the mean, so that predictions applied to cases in the general population and not possible outliers. Overall, simulated scores were thus more representative of real-world populations.

I generated predicted scores for each outcome from a single random forest model, which comprised *all* participants.[15] Overall model accuracy converged with the models in Study 1 for needs thwarting ($R^2 = 35\%$), needs satisfaction ($R^2 = 40\%$) and SLS ($R^2 = 28\%$). For the bottom

[15]I did not use 10-FCV because (a) I did not need to make uncontaminated predictions for existing participants because models were exported to simulated cases; (b) 10-FCV and non-10-FCV models generally converge in large samples (James et al., 2013); and, (c) Study 1 suggested separate models would be almost identical.

and in the top 0.5%, I also removed cases with predicted SWB > |3| SDs from the mean for their stratum—again to mitigate outliers—and then took the facet means. I repeated this entire procedure 20,000 times for each outcome, so that there was both a representative range of facet intercorrelations and enough means to find stable 99.9% CIs in the extreme 1% of cases. As such, I simulated a total of four billion cases.

As a preliminary step, I used 1,000 simulations with the reported procedure to find the maximum possible facet mean for the most extreme 1% cases using *random normally distributed z-scores* with truncated outliers. It was M = |2.82| (SD = 0.01). When the mean converged with this ceiling, its high values combined with disproportionately *more* levels from the other facets that were prevalent in the population to promote extreme SWB. When the upper-bound CI failed to converge with this ceiling, other facets/facet combinations could override its SWB effects. Finally, prevalences in bottom and top 0.5% SWB were not necessarily exact opposites because random forest affects can be *asymmetric*. That said, for simplicity I averaged effects for the 99.5th percentile and reversed effects for the 0.5th percentile—by taking mean point and CI estimates—because the CIs overlapped in all but one instance. The exception was for depression, which was more prevalent in extremely high (M = 2.46; CI = 2.31, 2.60) than extremely low needs thwarting (M = 2.10; CI = 1.93, 2.27).

6.4.2. Results

All facet prevalences and 99.9% CIs for the most extreme 1% needs thwarting, needs satisfaction and SLS predicted scores are in Figure 6.3. Where prevalences converged with the results in Chapter 5, it increased confidence that they exerted genuine impacts on SWB in the presence of comprehensive real-world facet contingencies. Where they diverged, facets may have only exerted their impact on SWB in combination with constellations of other facets that happened to be overrepresented in the sample. I report facet prevalences in order of magnitude within each factor. In most cases, I grouped facets when there were overlapping CIs. The exception was when there was marginal overlap and possible theoretical discontinuity.

Figure 6.3. Heatmap of facet prevalence in cases with the most extreme 1% predicted needs thwarting and satisfaction. Point estimate values were mean z-score values (M = 0; SD = 1) of prevalence scores generated from 20,000 simulated populations with different plausible real-world patterns of facet covariance. Confidence intervals were from specific simulated populations with .0005th and .9995th mean facet prevalences. Thus, they formed a 99.9% bootstrapped CI. The darker the panel the more positive the association. The X axis contains the big five factors and the Y axis contains each of their six nested facets. The factors and facets (in order) are: Neuroticism (anxiety, anger, depression, self-consciousness, immoderation, vulnerability), Extraversion (friendliness, gregariousness, assertiveness, activity-level, excitement-seeking, cheerfulness), Openness (imagination, artistic interests, emotionality, adventurousness, intellect, liberalism), Agreeableness (trust, morality, altruism, cooperation, modesty, sympathy) and Conscientiousness (self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, cautiousness). * = The 99.9% facet prevalence CI for need thwarting/needs satisfaction did not cross zero. ^ = The 99.9% prevalence CI for satisfaction with life did not cross zero.

6.4.2.1. Needs Thwarting

Cases scoring *high* on all six neuroticism facets were overrepresented in extreme needs thwarting. The largest prevalences were for depression, anxiety and vulnerability. This was followed by anger and self-consciousness, and then finally immoderation. Cases scoring low on five of the six extraversion facets were also overrepresented. Low cheerfulness, friendliness, assertiveness and gregarious were most prevalent, followed by low activity-level. There was no effect for excitement seeking. Four openness facets were overrepresented. Low adventurousness, intellect and artistic interests, and high imagination, all had approximately equal prevalence. There were no effects for emotionality or liberalism. All six agreeableness facets were overrepresented. Low cooperation, morality, altruism, trust and sympathy, and high modesty, were overrepresented. Finally, cases scoring low on all six conscientious facets were overrepresented. Low self-discipline and self-efficacy were most prevalent, followed by low cautiousness, dutifulness, orderliness and achievement striving. All needs thwarting effects were also significant for SLS.

6.4.2.2. Needs Satisfaction

Cases scoring *low* on all six neuroticism facets were also overrepresented in extreme needs satisfaction. Low depression, vulnerability, self-consciousness, anxiety and anger were most prevalent, followed by immoderation. Cases scoring high on all six extraversion facets were overrepresented. High cheerfulness, friendliness, assertiveness and gregariousness were most prevalent, followed by high activity-level and excitement seeking. Four openness facets were overrepresented. High artistic interests, intellect, adventurousness and emotionality all had approximately equal prevalence. There were no effects for imagination or liberalism. All six agreeableness facets were overrepresented. High altruism had the largest prevalence. Then, high morality, trust, sympathy and cooperation, as well as low modesty, all had approximately equal prevalence. Finally, cases scoring high on all six conscientious facets were overrepresented. High self-efficacy and self-discipline were again the most prevalent, followed by high achievement-striving, dutifulness, orderliness and cautiousness. All needs satisfaction effects were also significant for SLS.

6.4.2.3. Differential SWB Prevalences

Finally, I evaluated whether there were any differences in the magnitude of needs thwarting and satisfaction facet prevalences. There were four facets that had non-overlapping CI magnitudes. In neuroticism, the anxiety effect was larger for needs thwarting (M = 2.14; CI =

1.94, 2.33) than needs satisfaction (M = -1.39; CI = -1.66, -1.12). In openness, the effect for emotionality was smaller for needs thwarting (M = -0.26; CI = -0.57, 0.06) than needs satisfaction (M = 0.92; CI = 0.63, 1.20). Finally, in conscientiousness the effects for both self-efficacy and achievement-striving were smaller for needs thwarting (C1: M = -1.63, CI = -1.88, -1.37; C4: M = -1.04, CI = -1.32, -0.75) than needs satisfaction (C1: M = 2.24, CI = 2.06, 2.41; C4: M = 1.58, CI = 1.33, 1.83). Overall, most facets had convergent prevalences for needs thwarting and satisfaction. However, there were still exceptions in three of the big five. This suggested partially distinct substrates for suffering and flourishing, in the population sampled.

6.5. Study 3

Study 2 suggested that personality constellations comprising low neuroticism facet scores, and high scores for most other facets, on average experienced top 1% SWB. It also observed the inverse pattern for bottom 1% SWB. However, these prevalences were not necessarily internally valid. Positive correlations between facets meant it was greater than chance that any participant randomly sampled from the population would score similarly on both facets. Across the 30 facets, this meant that certain constellations—corresponding to those typical of the population—were oversampled. Whilst Study 2 thus described the average profiles of cases experiencing extreme SWB in the *real world*, it also restricted the range of moderating facet levels. Facet prevalences that indicated robust, internally valid, SWB relationships may have retained their high prevalence when *every* personality constellation was equally likely.

6.5.1. Method

6.5.1.1. Procedure

Study 3 replicated the procedure from Study 2, except with simulated orthogonal rather than CD intercorrelated big five facet scores. For continuity, I again simulated 200,000 random normally distributed cases for each facet. I predicted the three SWB outcomes using the random forests models developed in Study 2, and retained the 1,000 cases with both the bottom and top 0.5% predicted scores. Facet prevalences were computed after removing cases with predicted scores > |3| SD from the mean for their stratum. Then, I again repeated this entire procedure 20,000 times, so there were enough means to compute 99.9% bootstrapped CIs. When facet means converged with the ceiling (M = |2.82|), they combined with the most levels from other facets—*regardless* of the likelihood that facets co-occurred in the population—to promote extreme SWB. When the upper-bound facet CIs failed to converge with the ceiling, it again

meant that other personality constellations could override SWB effects for the facet in question. I again aggregated mean prevalences for the 99.5th and reversed 0.5th percentiles—by taking mean point and CI estimates—because the CIs overlapped in *all instances*.

6.5.2. Results

All facet prevalences and 99.9% CIs for the most extreme 1% needs thwarting, needs satisfaction and SLS predicted scores are in Figure 6.4. There were far fewer noteworthy effects than in Study 2, and thus I report them by magnitude *across* the five factors. I again grouped prevalences when CIs overlapped, except in some marginal cases where there was also theoretical discontinuity. For extreme needs thwarting, high depression (N3), and then high anxiety (N1) and vulnerability (N6), and then low self-discipline (C5), cheerfulness (E6) and cooperation (A4) were most prevalent. For extreme needs satisfaction, high self-efficacy (C1) and cheerfulness, and then low depression and high self-discipline, assertiveness (E3) and altruism (A3) were most prevalent. I replicated all prevalences for SLS. Of these, the neuroticism facet effect magnitudes were larger for needs thwarting (N3: M = 1.82, CI = 1.61, 2.05; N1: M = 1.15, CI = 0.76, 1.51; N6: M = 0.94, CI = 0.59, 1.25) than for needs satisfaction (N3: M = -0.90, CI = -1.26, -0.49; N1: M = -0.09, CI =-0.41, 0.24; N6: M = -0.16, CI = -0.54, 0.21). Contrastingly, cheerfulness and self-efficacy effect magnitudes were larger for needs satisfaction (E6: M = -0.45, CI = -0.92, -0.04; C1: M = -0.30, CI = -0.65, 0.06) than for needs thwarting (E6: M = 1.34, CI = 1.04, 1.61; C1: M = 1.42, CI = 1.08, 1.73). As such, neuroticism facets may have been more prevalent in extreme needs thwarting than extreme needs satisfaction. There was also partial support for the converse: self-efficacy and cheerfulness were more prevalent in extreme needs satisfaction. Overall, results suggest smaller subsets of facets implicated in extreme needs thwarting and needs satisfaction. They emerged with more stringent (and unbiased) controls, for the *full range* of facet cooccurrence patterns.

Figure 6.4. Heatmap of facet prevalence in cases with the most extreme 1% predicted needs thwarting and satisfaction. Point estimate values were mean z-scores (M = 0; SD = 1) of prevalences generated from 20,000 simulated populations with different *random* patterns of facet covariance. Confidence intervals were from specific simulated populations with .0005th and .9995th mean facet prevalences. Thus, they formed a 99.9% bootstrapped CI. The darker the panel the more positive the association. The X axis contains the big five factors and the Y axis contains each of their six nested facets. The factors and facets (in order) are: Neuroticism (anxiety, anger, depression, self-consciousness, immoderation, vulnerability), Extraversion (friendliness, gregariousness, assertiveness, activity-level, excitement-seeking, cheerfulness), Openness (imagination, artistic interests, emotionality, adventurousness, intellect, liberalism), Agreeableness (trust, morality, altruism, cooperation, modesty, sympathy) and Conscientiousness (self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, cautiousness). * = The 99.9% facet prevalence CI for need thwarting/needs satisfaction did not cross zero. ^ = The 99.9% prevalence CI for satisfaction with life did not cross zero.

6.6. Discussion

The present chapter aimed to evaluate the constellations of personality facets that are associated with extremely low and high SWB. Until now, existing research was unable to capture complex facet interrelationships, possibly due to multicollinear predictors and interpretive constraints. I overcame these challenges using random forest, which captures the complex interdependencies between all thirty big five facets. In Study 1, I found that a single random forest captured a *universal* pattern of facet effects on SWB, and that results were especially accurate for cases with extreme 1% predicted cases. In Study 2, I found that cases with extremely high SWB had low neuroticism, and mostly high extraversion, openness, agreeableness and conscientiousness facet scores. I found the inverse pattern for extremely low SWB. In Study 3, I found the smaller subset of facets that had extremely robust internally valid relationships with SWB. Final results highlighted the importance of just 9/30 big five facets.

In Study 1, I found that a single universal constellation of facet effects predicted SWB, across both a range of sociodemographic groups, and cultures. At the outset, I used random forests models to find the associations between interdependent facet combinations and the SWB outcomes, which were needs thwarting and needs satisfaction (primary outcomes), and SLS and health (secondary outcomes). There were negligible associations between participants' sociodemographic characteristics and the accuracy of their random forest predicted scores. This suggested that models applied equally to a broad spectrum of different subgroups, which included men and women, the young and old, the working and upper classes, and participants from disparate world regions (e.g. Asia, Latin America). Then, I used each of these strata to generate separate random forests models. They yielded virtually identical predicted scores for the entire sample. This suggested there was also an *unchanging* pattern of associations between the personality facets and SWB. All results converged across both primary and secondary outcomes. Finally, I evaluated how random forest accuracy changed as a function of predicted score extremeness. As extremeness increased, prediction accuracy also increased for the primary outcomes and SLS. Participants with the most extreme 1% predicted SWB scores on average had the most extreme 1/5 self-report scores. They were also correctly classified into ecologically valid buckets (e.g. 'very strong') in upwards of 3/4 cases. Results did not converge for health—likely because of comparatively low random forest accuracy—and thus I excluded it from subsequent analyses. Whilst the retained predictions were imperfect, they could still identify a subset of facet constellations the were implicated in very high true SWB.

In Study 2, I found that people experiencing extremely low and high SWB differed from the average on a plurality of the big five facets. To this end, I used random forest models to predict SWB for *four billion* simulated cases. Using simulations—rather than just the original self-report participants—helped to account for a range of different feasible patterns of facet covariance in the population. After mitigating outliers, I evaluated each facet mean prevalence in the cases with the most extreme 1% predicted SWB. Cases scoring high on facets from neuroticism and low on most facets from the other four factors disproportionately experienced extreme needs thwarting. I also largely observed the inverse pattern for needs satisfaction. For example, facets associated with psychological impairment (e.g. depression, anxiety), sociability (e.g. friendliness, altruism) and self-control (self-efficacy and self-discipline) had relatively high prevalences in extreme cases for both primary outcomes. Nevertheless, anxiety was also more prevalent in needs thwarting, whilst emotionality, self-efficacy and achievement striving were all more prevalent in needs satisfaction. All results converged for SLS. Overall, Study 2 found the mean facet scores that combined with the most levels of other facets in cases to promote extreme SWB, in feasibly real-world populations.

In Study 3, I found the smaller subset of facets that had the most internally valid associations with extreme SWB. To eliminate possible confounding from covarying facet combinations, I evaluated when results from Study 2 replicated using another four billion simulated cases where *every* combination of facet scores was equally likely. Cases with extreme needs thwarting were characterised by their high depression, anxiety and vulnerability, and their low self-discipline, cheerfulness and cooperation. Cases with extreme needs satisfaction were characterised by their high self-efficacy, cheerfulness, self-discipline, assertiveness and altruism, and their low depression. Overall, results identified the subset of facet values that combined with the most levels of other facets to promote extreme SWB, regardless of their real-world co-occurrences. These may be the best candidates to have robust, internally valid, associations with SWB because they were *insensitive* to fully changing intrapersonal contexts. Put another way, they were the best candidates to have stable mechanisms.

6.6.1. Implications

Results are a first-of-their-kind demonstration that random forest can yield directional psychological insights for individual predictors. That is, this chapter ultimately found the specific facet effects that were most insensitive to the nested and thus interdependent patterns of other personality moderators (Asendorpf et al., 2013). Although random forest was originally developed as a ‘black-box’ method to generate high-fidelity predicted outcomes, I

found its predictions could also be reverse-engineered to investigate specific psychological phenomena. As a preliminary step, I found that the same constellations of personality facets were *universally* associated with SWB. This suggests that there are shared trait-based propensities to experience both suffering and flourishing, which apply at least across the heterogeneous adult population sampled. I also found that random forest was especially accurate for extreme predictions. This meant I progressed by accounting for the whole individual using extreme subpopulations, but still with normative statistics.

Individuals who experience extreme suffering and flourishing deviate from the average on most facets. To this end, Study 2 found the average profile of people experiencing top percentile SWB in the population. High convergence with the main effects in Chapter 5 suggests that most facets exert their impact on SWB independent of the other facets. However, convergent findings may still have only emerged because the facet exerted its effects on SWB in combination with other specific levels of covarying facets, which were overrepresented in the sample. This might explain some of the contradictory research in the literature to date: research might assess trends driven by these kind of sample-specific range restrictions, which cause artefactual moderation effects. Despite this ambiguity, results from Study 2 are still important for public policy because they help give high-fidelity descriptions of the unhappiest and happiest personality profiles in the population.

Ultimately, however, the key finding is in Study 3. Random forest also helped evaluate the facets that were robust predictors of SWB, across the full range of changing intrapersonal contexts. Findings from social psychology's replication crisis highlight the importance of this criterion when proposing direct, causal, main effect mechanisms (Shrout & Rodgers, 2018). In this chapter, findings suggested an exhaustive but not profligate account of nine big five facets that are robustly implicated in SWB. Depression, cheerfulness and self-discipline may drive changes in both suffering and flourishing SWB—if perhaps through different mechanisms because the two outcomes are largely uncorrelated. Then, low anxiety and vulnerability, and high cooperation may uniquely protect against suffering. Contrastingly, high assertiveness, altruism and self-efficacy may uniquely promote flourishing. As I discussed in the General Introduction, these traits may exert their influence on SWB because they (a) capture the direct sensitivity to experience certain SWB facets, (b) increase the likelihood of extracting SWB nutriments from their environment, and/or (c) increase the likelihood of positively reappraising existing circumstances (Schimmack et al., 2004; Ryan & Deci, 2001; Boyce & Wood, 2011). Although finding exact mechanisms is beyond the current scope, it is now a clearer next step.

6.6.2. Limitations and Future Directions

There are at least four limitations that especially impacted the present chapter. Participants were all adults who belonged to online survey panels. This may have reduced their heterogeneity, which limited the generalisability of the results. For example, co-dependent hunter-gatherers may benefit less from the propensity to cooperate because such behaviours are already enforced in their community structures (Hill et al., 2011). Methodologically, conclusions were limited to cases with extreme 1% predicted SWB. Results did not necessarily imply linear trends. Prevalences for other percentile scores may have also been logarithmic, curvilinear, more complex or even completely random. Next, random forest models may suffer from overfitting—at least compared to other tree-based alternatives, such as boosting—which could limit external validity (James et al., 2013). Finally, models still failed to explain more than half the variation in SWB. This suggests the personality facets may co-occur and/or interact with other stable socio-demographic variables, as well as learned values, abilities and the transient environment, to fully account for manifest SWB.

In addition, an interesting property of sorting cases into predicted SWB percentiles is that it forces rank order facet prevalences. To explain: the strongest predictor will have the smallest SD. Then, in random facet constellations, all other facets scores will tend to be *normally distributed* within this first, strong, predictor. Thus, there are simply fewer high scores to select in the second-strongest predictor, which dilutes its mean prevalence. This effect then cascades as predictor strength decreases. Prevalences are especially diluted when a stronger predictor carries a large proportion of the information contained in a weaker predictor. This property also variably manifests in non-random facet constellations (as in Study 2), depending on the extent of predictor intercorrelations. It is unclear whether this property is (a) an efficient way of parsimoniously selecting a discrete set of predictors, or (b) an additional source of bias.

Nevertheless, there are also promising directions for future research. Random forest complexity is only constrained by sample size: it may be feasible to account for non-linear relations between *both* stable and transitive phenomena, provided there are more subjects than input variables (e.g. > 5:1 predictor-to-outcome ratio is often recommended; Green, 1991). This may give an unprecedentedly comprehensive account of the psychological substrates that are robustly associated with SWB. Internally valid inferences can then be buttressed by using longitudinal designs with cross-lagged correlations, to confirm that personality and/or other predictors are indeed *antecedent* to SWB (Keller et al., 1987). Finally, potential mechanisms that drive effects may be tested experimentally, by inducing aspects of the causative traits in

participants with randomly distributed personalities. Of course, this entire research paradigm may also be applied to almost any other combination of IVs and DVs in social psychology.

6.6.3. Conclusion

Ultimately, this chapter reconciles discordant research on the personality predictors of SWB. Depressed, anxious, vulnerable, ill-disciplined, low cheer and uncooperative individuals may be uniquely predisposed to experience suffering. Efficacious, cheerful, self-disciplined, assertive, altruistic and non-depressed individuals may be uniquely predisposed to experience flourishing. Although findings may appear unremarkable, they isolate a subset of the full range of personality facets—which also include *putatively* robustly facets effects for friendliness, adventurousness and trust, among others—implicated in SWB. This may help to definitively isolate the subset of internally valid facet effects that have especially plausible mechanisms, and are thus worthy of future research.

Chapter 7

General Discussion

7.1. Abstract

Here, I integrate findings from across the general introduction and empirical chapters. Overall, the PhD addresses sequential aspects of the research process to ultimately identify the most robust big five facet correlates of suffering and flourishing SWB. I highlight the findings that: (a) online-predicted personality may be so imprecise that, counter-intuitively, it is viable for academic research; (b) systematic computational approaches can be used for expedient scale validation; (c) PSM-weighted associations between the big five facets and SWB, as well as random forest profiling of the most extreme 1% predicted SWB, isolate the average profiles of the unhappiest and happiest people throughout the population sampled; and, (d) random forest might also isolate the nine most internally valid facet effects. Implications focus on reconciling discordant facet-SWB effects in the existing literature, the benefits and constraints of my computational paradigm and how findings inform current privacy debates. I also address overarching limitations such as sample unrepresentativeness, cross-sectional surveys, establishing cross-cultural equivalence and sub-optimal construct operationalization. Finally, future directions focus on establishing causal associations using longitudinal designs and then experiments, isolating mechanisms, applying methods to other research areas and popularising an increasing range of computational methods in social psychology.

7.2. Chapter Summaries

Who is happiest? It is a seemingly mundane problem that psychologists have engaged with for decades. However, even provisional answers may depend on: (a) procuring sufficiently large data to remove sampling, methodological and statistical artefacts, thus enabling researchers to evaluate multiple permutations of their questions; (b) expediently finding construct valid variable operationalizations; (c) disentangling complexly interrelated personality trait predictors; and, (d) accounting for the full range of intrapersonal effect contingencies. Across the four empirical chapters, I showed that using computational psychology approaches—

centred on large online samples, simulations and iterative analyses—to intervene at various stages in the research process might yield more definitive personality-SWB associations.

The first problem was that computational approaches are power hungry. A promising solution is to predict variable scores at scale from online behaviour. However, recent ethics debates suggest this may be unviable when it intrusively profiles individuals—e.g. on their personality—without their explicit consent. Thus, I evaluated the extent these prediction algorithms applied to *specific* individuals. At the outset, I used pure simulations to generate perfectly unbiased 'predicted' scores with varying fidelity. I focused on three benchmarks—'best-case' (r = .90), 'demographic' (r = .60) and 'personality' (r = .30)—that reflected the future potential of the technology and current documented accuracies for different types of variables. Then, I replicated all results using real-world data and machine learning. Results suggested that best-case predictions could (a) consistently differentiate between participants with opposite extreme scores, as well as randomly drawn pairs; (b) be corrected to account for real-world thresholds (e.g. neutral vs extraverted); and, (c) be bucketed into thirds (i.e. low, medium, high). However, even they failed to correctly estimate the true magnitude of differences between people, profile individuals across multiple traits or differentiate edge cases. For comparison, individual predictions were only marginally better than chance at realistic accuracies. To illustrate, at the personality benchmark predictions failed to consistently differentiate between opposite personality extremes (e.g. highly introverted vs highly extraverted), were worse than simply assuming everyone was the average when corrected to capture real-world thresholds, and far more likely predict the entire big five 100% incorrectly than 100% correctly. Thus, I concluded that online-predicted personality does not apply to specific individuals.

Counter-intuitively, results may support the viability of predicted personality for academic research. Inaccurate predictions are still sufficient to evaluate normative trends, provided they are unbiased (prediction errors are fully random) and sample size is large enough to overcome increased measurement imprecision (Cohen, 1992). Then, any surplus power would enable additional, stratified, sub-population analyses. Further, new prediction algorithms can incrementally add construct scores to existing databases, thus increasing the richness of the data over time. All that said, I did switch to exclusively self-report data for the remaining chapters. This was (a) in keeping with the shifting focus of my PhD towards more fundamental associative research, which could be largely answered in surveys; (b) to avoid using data

obtained without explicit consent, regardless of its accuracy; and, (c) to avoid compounded prediction and original-scale measurement errors.

The second problem I addressed was finding appropriate construct operationalizations. There are multiple reasons constructs may be sub-optimally captured. In this PhD, there were a (surprising) lack of scales that comprehensively measured SWB. Thus, I used computational approaches to *expediently*, and *systematically*, repurpose the BMPN—which was originally intended to measure the three substrates of intrinsic motivation—as the primary outcome. EFA suggested two factors. Across bootstrapped factor loadings, there was consistent evidence that they were needs thwarting and needs satisfaction. Then, I evaluated their *exhaustive* associations with the other SWB and lifestyle variables in the data. Needs thwarting was more associated with negatively valanced SWB (i.e. suffering), and needs satisfaction was more associated with positively valanced SWB (i.e. flourishing). They also captured more *transient* SWB than life satisfaction. This was important for bivariate personality-SWB associations because BMPN thus better isolated facets effects that were sensitive to changes in the environment. The BMPN factors also explained variation in lifestyle outcomes—e.g. healthy eating, marriage—over and above criterion SLS. All results held across different country and sociodemographic strata. There was no evidence they were artefacts of response bias.

The third problem was finding internally valid bivariate facet-SWB associations. Facet-level analyses may be specific enough to isolate discrete mechanisms. However, IV fragmentation at this level of analysis increases the number of potentially confounding covariates beyond the threshold that can be tolerated using non-computational approaches. To remedy, I used PSM to extract pairs of participants who differed on the target facet but were similar across *all* 29 of the covarying facets. Thus, PSM mitigated the need to fit explicit controls because potentially confounding facets were held constant. Weighted PSM correlations better accounted for the facet covariates than zero-order correlations (i.e. the first tier in stepwise regression) and caused less artificial reduction in effect sizes than multiple regression. They also better reconstructed establish established *factor-level* effects—which are less fragmented and thus more definitive—than correlations, multiple regression, elastic net machine learning and even combined PSM-ENET. Then, using PSM, I found the full range of big five facet associations with SWB. There were neuroticism and extraversion facet effects beyond cheerfulness and depression, diverging effects within agreeableness and cascading SWB benefits for conscientiousness. These patterns may help reconcile the contradictory effects that currently exist in both the trait-SWB literature, and in multiple adjacent literatures.

The fourth problem was that even internally valid associations can change across intra-personal contexts. Even if PSM perfectly controlled for main effect confounds, results may still have been artefacts of ungeneralizable—sample-specific—facet interactions. To remedy, I thus replicated PSM findings using random forest. Random forest accounts for *nested* SWB effects involving all 30 facets. That is, it accounts for the *full constellations* of even complex personality contingencies. First, I found a single random forest model was equally accurate across various demographic strata, such as gender, age group (young, middle, old) and world region. Then, I generated separate random forest models using 17 different sample strata (e.g. just women, just Latin Americans). They yielded nearly identical predictions. This suggested that the facets combined in similar ways to predict SWB across diverging subsamples. Thus, a single, *universal*, combination of facets was still associated with SWB after relaxing the assumption of linear covariates. Then, I evaluated model accuracy for participants with different magnitude predicted scores. Accuracy was highest for the most extreme 1% predictions. On average, they had the most extreme 1/5 true scores, and were categorized either fully correctly or nearly correctly for > 85% of cases. Although predictions were imperfect, they were still sufficient at the extremes to identify very low and very high true scores.

Then, I evaluated the facets that were prevalent in extreme SWB across changing personality contingencies. Nested random forest interactions are far too complex to interpret. Instead, I evaluated facet prevalences in cases with extreme predicted SWB. However, results may still have been ungeneralizable due to idiosyncratic facet covariation in the sample. Thus, I predicted SWB for four billion simulated cases with facet scores that conformed to different, plausible, covariance patterns in the population. Many of the novel PSM findings held in the 1% most extreme of these predicted scores. For example, high depression consistently exacerbated needs thwarting more than it inhibited needs satisfaction. High self-efficacy promoted needs satisfaction more than it protected against needs thwarting. Modesty had consistent SWB consequences, whilst adventurousness and intellect had consistent SWB benefits. Cooperation was indeed more implicated than altruism in needs thwarting, and vice versa for needs satisfaction. Other PSM results may have been a consequence of range restriction. For example, adventurousness may have only exacerbated needs thwarting when combined with other facet levels that were overrepresented in the self-report data. Similarly, the sample covariance pattern may have suppressed the true protective benefits of activity-level on needs thwarting. It remained inconclusive whether morality, and cautiousness versus

achievement striving, had differential effects on needs thwarting compared to needs satisfaction. Overall, random forest clarified a range of PSM-derived SWB effects.

Triangulated PSM and random forest findings may be the most definitive, internally valid, account of who is on average happiest in the population. They better control for confounding facet effects, compared to conventional (e.g. multiple regression) and other machine learning alternatives (e.g. elastic net). Thus, they may allow researchers to construct the archetypal unhappiest and happiest personality profiles. These are in Figure 7.1. Unlike previous research, trait prevalences are from normative rather than case study data. Of course, effects may have still been caused by residual confounding (improvements were relative, not absolute). Nevertheless, they might offer better empirical grounds for future research than alternatives.

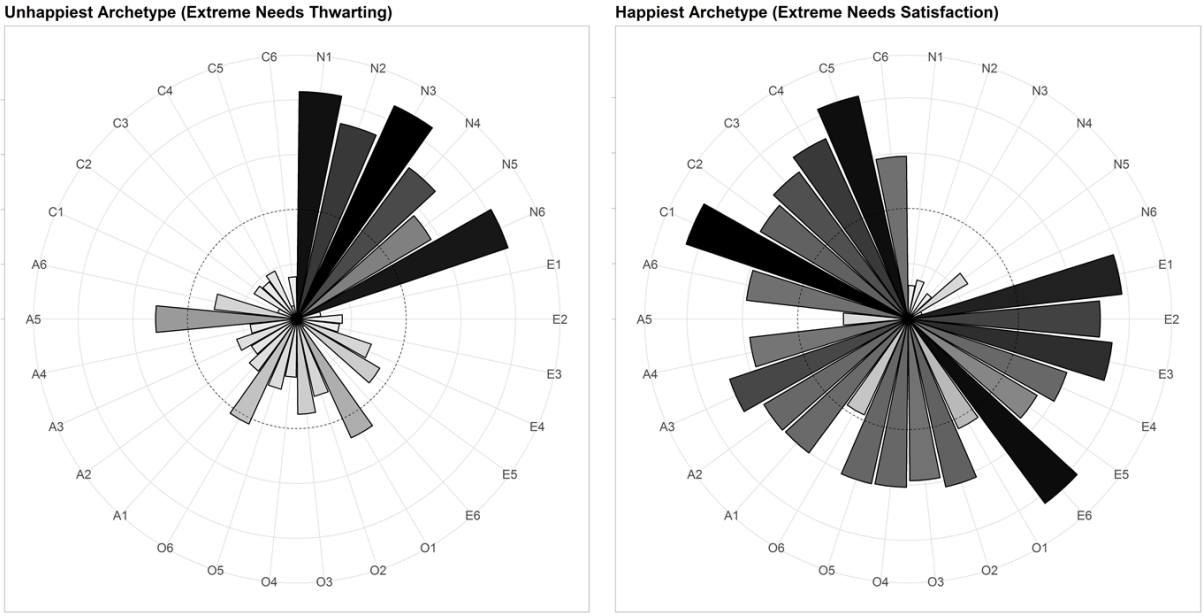


Figure 7.1. Mean facet prevalences in the most extreme 1% of *realistic* simulated cases from Chapter 6. N1-6 = Neuroticism facets (anxiety, anger, depression, self-consciousness, immoderation, vulnerability). E1-6 = Extraversion facets (friendliness, gregariousness, assertiveness, activity-level, excitement-seeking, cheerfulness). O1-6 = Openness facets (imagination, artistic interests, emotionality, adventurousness, intellect, liberalism), A1-6 = Agreeableness facets (trust, morality, altruism, cooperation, modesty, sympathy). C1-6 = Conscientiousness facets (self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, cautiousness). The dashed line represents the population mean (M = 0; SD = 1).

However, there is a ceiling to internal validity using feasibly real-world participants. In such cases, every facet still preferentially co-occurs with a different subset of other facets. Individual facet effects might change completely when there is *equal likelihood* of it combining with infrequently-occurring facet constellations. Observing facets in these conditions helps isolate

the main effects that are robust to an especially wide range of intrapersonal contingencies. Thus, I replicated the random-forest prediction method described above for another four billion simulated cases with *fully random* facet scores. A total of nine facets were implicated in extreme SWB. These are in Figure 7.2. Depression, cheerfulness and self-discipline were associated with both needs thwarting and satisfaction. High anxiety and vulnerability, and low cooperation, were exclusively implicated in needs thwarting. High assertiveness, altruism and self-efficacy were exclusively implicated in needs satisfaction. Collectively, results highlight the partly overlapping personality substrates of suffering and flourishing SWB. They transcend the affective facets, but still implicate fewer than 1/3 of the entire big five. Thus, results may parsimoniously describe the full range of facet-SWB effects.

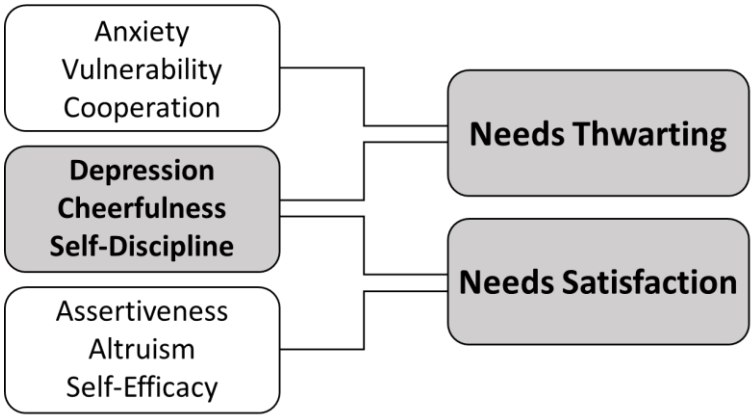


Figure 7.2. The nine big five facets that were robustly associated with extreme SWB. They emerged even after they were combined with random covarying facet constellations, which were both likely and unlikely to occur in real world populations. They had especially non-contingent associations with SWB, which may thus indicate that their effects are driven by stable mechanisms.

7.3. Implications

The most concrete PhD implications are for personality-SWB associations. First, I found evidence that emergent technologies yield high power samples without violating individual privacy. Then, I established a novel measure of omnibus SWB and isolated the full plurality of bivariate personality facet-SWB associations. In doing so, I mitigated many of the perceived trade-offs that constrain existing research in the field: (Chapter 3) samples from online behaviour are unacceptably intrusive; (Chapter 4) psychometric robustness necessitates using partial constructs, like tripartite SWB; (Chapter 5) establishing internal validity causes multicollinearity; (Chapter 6) normative statistical trends must oversimplify the individual by

ignoring intrapersonal contingencies. Ultimately, results helped find more definitive *exploratory* associations between all the granular personality facets and comprehensive SWB. Results are more parsimonious than both existing factor-level associations—which implicate up to four of the big five—and facet-level associations, which only reliably implicate depression. They reconcile adjacent research on the SWB consequences of certain coping-styles (anxiety, vulnerability), mastering the environment (self-discipline, self-efficacy, assertiveness) and prosociality (cooperation, altruism) (Gross, 2015; Ryan & Deci, 2000; Keltner et al., 2014). Therefore, results may inform *more* empirically based follow-up hypotheses, thus helping streamline cumulative science in the field.

More widely, findings show the value of large samples. The constructs measured using the particularly cheap, large and heterogeneous online AXA study sample were mostly robust. Specifically, they had adequate internal consistency, measurement invariance, and convergent and discriminant validity. This further supports the viability of online samples that transcend WEIRD populations (Henrich et al., 2010). In addition, power was so large that I could reuse the same sample throughout the entire PhD. By combining stringent thresholds for noteworthy effects (e.g. $r > .10$; $p < .001$) and non-parametric statistics (e.g. 99.9% CIs from resampled effects), I likely constrained the overall incidence of Type 1 error within acceptable bounds. Of course, any remaining spuriousness effects may have been conserved across chapters. Even so, findings may have been unusually robust for such ranging cross-sectional research.

I also evaluated whole populations of effects across iteratively changing samples, variables and models. These resampling approaches met recent methods recommendations to extract effects from multiple replications (Shrout & Rodgers, 2018). For example, bootstrapping and 10-fold cross-validation—which both replicate analyses across random subsamples—mitigated results caused by sampling error (James et al., 2013). Then, I also iteratively repeated the analyses across specific strata (e.g. just women) to increase external validity. This helped mitigate a limitation of increased sample heterogeneity: that aggregate findings can obfuscate disordinal subpopulation effects. For iteratively changing variables, comparing scales—e.g. across *every* convergent measure of SWB—helped quickly indicate how they captured the underlying construct, relative to convergent measures. For example, I found that the BMPN captured more transitive SWB than SLS. Iteratively changing model parameters—usually designed to maximize model explanatory power—increased ecological validity by extracting the highest fidelity associations from the data. That is, models were better attuned to real-world complexity

than relatively unadaptable non-machine learning alternatives (Yarkoni & Westfall, 2017). Overall, iterative approaches helped extract trends from *entire populations* of effects.

These approaches also yielded more precise effect estimates. The large sample meant that point estimates converged with even 99.9% CIs. Thus, often the only substantive sources of model error were the above-mentioned resampling techniques. This helped isolate differential *patterns* of effects. That is, I prototyped the post-Replication Crisis recommendation to move beyond falsely dichotomous p-value significance testing and evaluate effect magnitudes (Nelson et al., 2018). I also used participant weights to generate higher-fidelity effects. Although PSM weights were designed to boost internal validity by controlling for covarying facets, other weighting strategies could be used in a similar way to boost (e.g.) sample representativeness for various populations (Cohen et al., 2013). Finally, I used simulations to evaluate the extent that edge conditions—which are normally beyond the scope of survey studies—changed observed findings. Notably, I evaluated the extent unrealistically accurate personality predictions applied to specific individuals, and the extent different plausible predictor covariances changed observed facet-SWB associations. This helped isolate the most robust, plausible, effects.

Overall, results advance aspects of the computational paradigm in psychology. Once obtaining my large sample, I focussed largely on resampling and machine learning. These strategies may be particularly helpful at improving the certainty of findings in cross-sectional survey research, which is often the weakest form evidence used by empirical psychologists. Cross-sectional research may have especially fine margins between legitimate but still fickle effect variation across populations, and variation that is an artefact of arbitrary but (at least partly) unavoidable sampling, measurement and statistical errors (de Boeck & Jeon, 2018). And yet cross-sectional survey research is also essential. It is an expedient and often cheap way of finding the preliminary descriptive evidence needed to justify further time and research expenditure. It complements more in-depth, but also theoretically partial, qualitative research (Ponterotto, 2002). Computational psychology's potential to survey and then pick from whole populations of effects—combined with its potential internal and external validity gains—reduces the chances that follow-up research is guided by confirmation bias (Ioannidis, 2012). That is, researchers are less likely to find support for their preconceptions simply because they have neglected larger effects or confounds. While exploratory research will never be definitive, the computational paradigm can mitigate the large volume of false positives that limit progress.

However, findings also highlight that computational approaches are *not* a panacea. Even extremely accurate prediction algorithms may not yield psychological construct scores that apply to specific individuals. Instead, they compound with underlying scale measurement error to only ever capture traces of the true phenomenology. Thus, they may only be useful *in aggregate*, to evaluate probabilistic trends. Further, 'exhaustive' scale associations are still constrained by the variables measured, which are chosen by the researcher and thus partial. Even machine learning methods have ceiling effects. PSM—which used simple weighted correlations—was better at isolating specific facet-SWB associations than the more complex elastic net. This was even though elastic net was explicitly designed to account for highly intercorrelated predictors. Then elastic net—which only accounted for main effects—used cumulative personality to explain as much variation in SWB as more-complex-again random forest, which accounted for both main and non-linear effects. Moreover, Chapter 3 observed that machine learning prediction accuracy plateaued at around 80 predictors. It likely plateaued again when there were thousands or tens-of-thousands, not hundreds of thousands, of participants (see Kosinski et al., 2013). This suggests that current technologies may already be converging with best-case future hypotheticals. Thus, their projected future accuracy may be overstated. Finally, even random forest—until recently the gold-standard machine learning approach (e.g. Ahmad, Mourshed & Rezgui, 2017)—yielded predicted psychological construct scores that were perhaps only actionable for extreme cases. Although facet prevalences for bottom and top cases were often diametrically opposite, it was unclear whether they were linked by (e.g.) linear, cubic or logarithmic trends. Further, I was only able to generate normative claims for the extreme cases with the aid of extremely high-power simulations. These depend on potentially-wrong researcher assumptions about the rules that govern the phenomena, such as normally distributed variables. Overall, computational psychology approaches may thus incrementally, but not diametrically, improve current methods.

Finally, the potentials and constraints of the featured computational approaches can inform privacy debates. In the empirical chapters, higher fidelity computational methods still only yielded *normative* claims; I found no evidence they were intrusive. Further, findings from Chapter 3—about the non-applicability of predicted personality to specific individuals—also apply to the bivariate associations in subsequent chapters. To illustrate, there may only be marginally above-chance likelihood that any one individual with high trait depression has low SWB. Thus, survey research in the personality-SWB sub-field could continue to be regulated by ethics frameworks that mandate prior IRB approval, informed consent and minimal risk.

However, it is also sometimes unfeasible to obtain explicit informed consent for a specific study. Perhaps most topically, this occurs when archives of online behaviour are obtained from existing repositories (e.g. Twitter) and used in compliance with their existing terms and conditions (Golbeck et al., 2011). Despite recent landmark US Supreme Court cases—(e.g.) sanctioning the right to be forgotten and preventing law enforcement agencies from accessing GPS phone data without a search warrant (Grierson & Quinn, 2018; Liptak, 2018)—there is still no comprehensive regulatory framework for large digitally stored data (Athey, Catalini & Tucker, 2017). Thus, *independent* IRBs, researchers, private companies and other entities must exercise their own discretion. In practise, this means initiatives are often *reactive*. For example, Facebook only updated their privacy policy after widespread public allegations that their data helped influence the 2016 Presidential Election (Hsu & Kang, 2018). Contrastingly, in 2018 the EU introduced the GDPR. It mandates that every data repository provides accessible terms and conditions, obtains explicit consent, transparently discloses how data are used, and enables users to easily opt-out and permanently delete their data (eugdpr.org). It provides a pre-emptive and universal safeguard that helps align user privacy expectations with both current technological capacities and market incentives. Thus, it may complement existing protocols to better promote fully-consensual data usage.

7.4. Major Limitations

The PhD has structural, sampling and measurement limitations that go beyond the in-built limitations of the featured computational approaches. There are at least two overarching structural limitations. First, the empirical chapters only intervened at selected points in the research process. For example, I did not iteratively evaluate different ways of operationalizing personality or isolate *specific* plausible interaction effects (e.g. between neuroticism and conscientiousness facets; Naragon-Gainey & Simms, 2017). Ultimately, I prioritized topicality because total comprehensiveness was unfeasible in any single PhD. Second, there were inconsistencies across empirical chapters. For example, while Chapter 3 argued that online-predicted personality was sufficiently non-intrusive to use for psychological research, I still reverted to exclusively self-reported personality in subsequent chapters. In addition, PSM methods for more internally valid exploratory associations were immediately superseded by the potentially more definitive random forest methods used in Chapter 6. Ultimately, this was because chapters reflected the often non-linear and chaotic nature of academic research.

There were at least three sampling limitations. Although my sample was multinational and multilingual, it also comprised exclusively members of online survey panels. Although I controlled for some of their demographic characteristics (e.g. sex, age, social class) they still likely differed from non-panel members in the population on other qualities—for example, on higher-than-average introversion, and education attainment that transcended the binary measure of university degree attainment (MacCallum et al., 2002). This may have limited generalizability because of range restriction (Wiberg & Sundström, 2009). Participants were also unevenly distributed across world regions and language groups. Although I often confirmed there were consistent effects in separate countries, findings could have still been biased when (e.g.) the same scales measured somewhat idiosyncratic constructs in overrepresented strata. This could have been exacerbated by differing (non-random) survey structures and content across countries, which may have created artefactual order effects and increased sample unrepresentativeness respectively. Further, data were cross-sectional. I thus *assumed* that personality was antecedent to SWB. This may be problematic considering recent evidence supporting the transience of personality in adulthood (Soto et al., 2011). Moreover, in cross-sectional research there are simply more extraneous variables that covary with both the predictor and outcome (e.g. mood), compared to longitudinal data. This increases the absolute number of possible confounds.

Finally, there were at least three measurement limitations. Results were constrained by variable selection. Due to grant obligations, collaborator requests, time constraints and/or researcher error I often used sub-optimal scales and non-exhaustive construct operationalizations. For example, the most recent Big Five Inventory has an equal balance of positively and negatively phrased items in each facet (Soto & John, 2017). Thus, compared to the NEO-PI-R it may be both more resistant to positive response bias and capture more representative aspects of each construct. There were also non-exhaustive convergent measures of SWB. For example, I evaluated the transitiveness of BMPN using *only* convergent affect. This increased the chances that effect sizes were confounded by other forms of covariance that were not attributable to construct transience. In addition, group-mean z-scores assumed that the personality facets and SWB were equally prevalent in every country. This helped to control for *all* country-level main effects (e.g. GDP, education). It also equalized score variation, meaning countries with particularly wide score distributions were not overrepresented in the results (Aguinis et al., 2013). However, it also eliminated potentially real differences in national personality and SWB. For example, lower gregariousness may have been selected for in countries with greater

exposure to pandemic diseases because it reduces the incidences of widespread disease transmission (Schaller & Murray, 2008). That is, group-mean z-scores added another source of potential bias. Further, they did not eliminate cross-level interactions. National factors may have differentially *changed* individual-level associations. Finally, the fully self-report format meant that results were potentially driven by common method variance and introspection bias (Tellegen, 1985). Mixed peer, informant and behavioural measures would have yielded more comprehensive construct scores.

7.5. Future directions

Future research can evaluate causal facet-SWB associations. At the outset, this would involve switching to a prospective longitudinal design where both personality and SWB are measured at multiple time points. This would (a) help establish the correct temporal sequence where personality is antecedent to SWB, (b) control for personality change, and (c) reduce the number of possible confounds (which would have to covary with both the predictor and outcome at multiple time points). Longitudinal research also enables cross-lagged correlations, which establish whether the effect of personality measured at time one on SWB measured at time two is larger than the inverse (Kenny, 2005). This helps establish tentative causation. Then, researchers can evaluate whether specific facet nuances drive effects. For example, the effect of depression on needs satisfaction may be caused by anhedonia (Fava & Tomba, 2009). Nuances may correspond to such specific behaviours that they can be induced using experimental manipulations. This may be one way of establishing definitive causal effects. Results at such a stage may be sufficiently granular to evaluate mechanism—perhaps by iteratively testing effects for multiple candidate mechanisms—using preliminary cross-sectional and then longitudinal studies. Alternatively, each trait may also disproportionately exert its effects on specific sub-components of SWB. Thus, future research could evaluate patterns of facet effects on different SWB processes (e.g. affect, purposefulness), perhaps with the aid increasingly comprehensive SWB measures like the Scales of General Well-Being (Longo et al., 2017). This would help further isolate prospective mechanisms, because candidates would have to map onto whole patterns of facet effects.

The large sample and iterative resampling approaches can also be applied in other domains that have intercorrelated predictors. For example, they could help untangle the effects of the basic values (e.g. achievement, security, benevolence) on political orientation (Schwartz, Caprara & Vecchione, 2010). Studies are also not limited to psychology. For example, they can help

isolate the differential effects of intercorrelated variables like genetics, exercise, diet, socioeconomic status and access to healthcare on obesity (Rooney, Mathiason & Schauberger, 2011). Excitingly, findings may also be scalable to more than just the 30 predictors used in this PhD. Indeed, PSM weights and random forest tolerated all the facets with manageable effect suppression. All analyses could also be expediently performed (e.g. overnight) on a single personal computer. This makes it possible to *mix* a wider range of psychological and socio-demographic predictors in the same models, which allows researchers to control for reasonably comprehensive intra- and *inter*-personal contingencies. This would better fulfil the social psychological mandate to consider the individual in their wider social context.

Finally, I only demonstrate a small subset of emergent computational social science methods. The field is becoming increasingly accessible via plain-English instructive texts—for example, James et al.'s (2013) *Introduction to Statistical Learning*—and statistical software that packages often complex mathematical operations into easy-to-use functions, such as randomForest in 'R'. These tools often require only the kind of fundamental research skills (e.g. corroborating the instructive resources), statistics intuition and intermediate object-oriented coding (e.g. 'R', Python) that are already common/learnable in the social sciences. They provide less barriers to entry than differential and integrative calculus, matrix algebra and high-level coding (e.g. C++, Java), which were required by previous generations of researchers. Ultimately, increasing access means psychologists can increasingly marry computational methods with their more focussed training on distilling specific research questions from the inexorably complex real-world, theorizing only parsimonious complexity and evaluating the practicality of their effects. Rudimentary computational approaches—e.g. ridge and LASSO regression, and bootstrapping—are already commonplace. With ongoing interdisciplinary and thus translational work, simulations, iterative resampling and machine learning approaches may become equally prevalent. When melded with other emergent methods—e.g. consensually obtained *in vivo* event-sampling, live GPS tracking and natural language and image processing (Lazer et al., 2009)—they may better enable high-fidelity and comprehensive research throughout social psychology.

7.6. Conclusion

According to my findings, the happiest people have low trait depression, anxiety and vulnerability, and high trait cheerfulness, self-discipline, cooperation, self-efficacy, assertiveness and altruism. Other fragmentarily documented effects—e.g. for achievement striving, self-consciousness and friendliness (Quevedo & Abella, 2011; Anglim & Grant, 2016; Helliwell, 2006)—may be artefacts of intercorrelations with the most robustly internally valid facets. Alternatively, they might only be associated with SWB in the presence of certain restricted levels of other facets, or facet constellations, that are overrepresented in the population sampled. Despite extensive research into personality-SWB effects to-date, researchers can only begin to reconcile discordant findings in the field now, with the aid of emerging computational technologies. In my PhD, I focussed on very large samples and resampling approaches that utilize populations of different plausible effects. They helped to iteratively test multiple permutations of each research question, expediently find construct valid scales and isolate internally valid bivariate associations. They might conserve the fine margin between already-fickle psychological effects and other, artefactual, sources of error. Therefore, computational psychology may increase the internal validity of personality-SWB and perhaps other descriptive associations. In doing so, it could facilitate a new wave of fully evidence-based and thus rapidly-accumulating research.

References

Abranovic, W. A. (1997). *Statistical Thinking and Data Analysis for Managers*. Reading, MA (USA): Addison-Wesley.

Adams, J. U. (2015). Genetics: Big hopes for big data. *Nature, 527*, S108-S109.

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist* (online).

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modelling. *Journal of Management, 39*, 1490-1528.

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings, 147*, 77-89.

Alam, F., & Riccardi, G. (2014). Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. *ICASSP Conference Proceedings*, 955-959.

Albuquerque, I., de Lima, M. P., Matos, M., & Figueiredo, C. (2012). Personality and SWB: What hides behind global analyses? *Social Indicators Research, 105*, 447-460.

Aldao, A., Nolen-Hoeksema, S., & Schweizer, S. (2010). Emotion-regulation strategies across psychopathology: A meta-analytic review. *Clinical Psychology Review, 30*, 217-237.

Allen, T. A., Carey, B. E., McBride, C., Bagby, R. M., DeYoung, C. G., & Quilty, L. C. (2018). Big Five aspects of personality interact to predict depression. *Journal of personality*, *86*, 714-725.

Allik, I. (2002). *The Five-Factor Model of Personality Across Cultures*. New York (USA): Springer.

Allik, J., & Realo, A. (2017). Universal and specific in the five-factor model of personality. In T. A. Widger (Ed.) *The Oxford Handbook of the Five Factor Model*, 173-190. Oxford (UK):Oxford University Press.

Allport, G. W. (1937). *Personality: A Psychological Interpretation*. Oxford (UK): Holt.

Anglim, J., & Grant, S. (2016). Predicting psychological and SWB from personality: Incremental prediction from 30 facets over the Big 5. *Journal of Happiness Studies, 17*, 59-80.

Antonioni, D. (1998). Relationship between the big five personality factors and conflict management styles. *International Journal of Conflict Management, 9*, 336-355.

Appéré, F., Piardi, T., Memeo, R., Lardière-Deguelte, S., Chetboun, M., Sommacale, D., ... & Kianmanesh, R. (2017). Comparative study with propensity score matching analysis of two different methods of transection during hemi-right hepatectomy: Ultracision harmonic scalpel versus cavitron ultrasonic surgical aspirator. *Surgical Innovation, 24*, 499-508.

Archontaki, D., Lewis, G. J., & Bates, T. C. (2013). Genetic influences on psychological well-being: A nationally representative twin study. *Journal of Personality, 81*, 221-230.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*, 108-119.

Ashton, M. C., & Lee, K. (2018). How Well Do Big Five Measures Capture HEXACO Scale Variance? *Journal of Personality Assessment* (online).

Athey, S., Catalini, C., & Tucker, C. (2017). The digital privacy paradox: Small money, small costs, small talk. *National Bureau of Economic Research*, w23488.

Bardi, A., Guerra, V. M., & Ramdeny, G. S. (2009). Openness and ambiguity intolerance: Their differential relations to well-being in the context of an academic life transition. *Personality and Individual Differences, 47*, 219-223.

Barrett, L. F. (2017). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*, 1-23.

Baumeister, R. F. (2002). Yielding to temptation: Self-control failure, impulsive purchasing, and consumer behavior. *Journal of Consumer Research, 28*, 670-676.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407-425.

Bok, D. (2010). *The Politics of Happiness: What Government can Learn From the New Research on Well-Being*. New Jersey (US): Princeton University Press.

Bonneville-Roussy, A., Rentfrow, P. J., Xu, M. K., & Potter, J. (2013). Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *Journal of Personality and Social Psychology, 105*, 703-716.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*, 431-449.

Boyce, C. J., & Wood, A. M. (2011). Personality prior to disability determines adaptation: Agreeable individuals recover lost life satisfaction faster and more completely. *Psychological Science*, *22*, 1397-1402.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185-216.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5.

Burdea, G. C., & Coiffet, P. (2003). *Virtual Reality Technology*. New Jersey (USA): John Wiley & Sons.

Burgess, M. (2018, June 4). What is GDPR? The summary guide to GDOR compliance in the UK. *Wired*.

Busseri, M. A. (2015). Toward a resolution of the tripartite structure of subjective well-being. *Journal of Personality*, *83*, 413-428.

Busseri, M. A., & Sadava, S. W. (2011). A review of the tripartite structure of subjective well-being: Implications for conceptualization, operationalization, analysis, and synthesis. *Personality and Social Psychology Review*, *15*, 290-314.

Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1-67.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, *22*, 31-72.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, *36*, 1165-1188.

Chen, Y. C., Sudre, G., Sharp, W., Donovan, F., Chandrasekharappa, S. C., Hansen, N., ... & Shaw, P. (2018). Neuroanatomic, epigenetic and genetic differences in monozygotic twins discordant for attention deficit hyperactivity disorder. *Molecular Psychiatry*, *23*, 683-690.

Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E. L., Van der Kaap-Deeder, J., ... & Ryan, R. M. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and Emotion*, *39*, 216-236.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*, 155-160.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Abingdon (UK):Routledge.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*, 243-253.

Costa, P. T., & McCrae, R. R. (1976). Age differences in personality structure: A cluster analytic approach. *Journal of Gerontology*, *31*, 564–570.

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology, 38*, 668-678.

Costa, P. T., & MacCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional Manual*. Florida (USA): Psychological Assessment Resources (PAR)

Costa, P. T., & McCrae, R. R. (1995). Solid ground in the wetlands of personality: A reply to Block. *Psychological Bulletin, 117*, 216-220.

Costa, P. T., & McCrae, R. R. (2017). The NEO Inventories as instruments of psychological theory. In T. A. Widger (Ed.) *The Oxford handbook of the Five Factor Model*, 11-37. Oxford (UK):Oxford University Press

Costa, P. T., McCrae, R. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO personality inventory. *Personality and Individual Differences, 12*, 887–898.

Csikszentmihalyi, M. (1997). Happiness and creativity. *The Futurist, 31*, S8-S12.

Cumming, G. (2013). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Abingdon (UK):Routledge.

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the New Statistics: Estimation, Open science, and Beyond*. Abingdon (UK):Routledge.

Custers, B., van der Hof, S., & Schermer, B. (2014). Privacy expectations of social media users: The role of informed consent in privacy policies. *Policy & Internet, 6*, 268-295.

Daley, S. E., Burge, D., & Hammen, C. (2000). Borderline personality disorder symptoms as predictors of 4-year romantic relationship dysfunction in young women: Addressing issues of specificity. *Journal of Abnormal Psychology, 109*, 451-460.

Davies, H. (2015, December 11). Ted Cruz using firm that harvested data on millions of unwitting Facebook users. *The Guardian* (UK).

Davis, M. W. (1987). Production of conditional simulations via the LU triangular decomposition of the covariance matrix. *Mathematical Geology, 19*, 91-98.

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin, 144*, 757-777.

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings, 1644*, 97-104.

De Raad, B., & Mlačić, B. (2017). The lexical foundation of the Big Five factor model. In T. A. Widger (Ed.) *The Oxford Handbook of the Five Factor Model*, 191-216. Oxford (UK): Oxford University Press

Deci, E. L., & Ryan, R. M. (2011). Self-determination theory. In P. A. van Lange, A. W. Kruglanski & E. T. Higgins (Eds.) *Handbook of Theories of Social Psychology* (Vol. 1). London (UK): Sage

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*, 151-161.

DeNeve, K. M., & Cooper, H. (1998). The happy personality: A meta-analysis of 137 personality traits and SWB. *Psychological Bulletin, 124*, 197-229.

DeYoung, C. G., Carey, B. E., Krueger, R. F., & Ross, S. R. (2016). Ten aspects of the big five in the personality inventory for DSM–5. *Personality Disorders: Theory, Research, and Treatment, 7*, 113-123.

DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the big five. *Journal of Personality and Social Psychology, 93*, 880-896.

DeYoung, C. G., Weisberg, Y. J., Quilty, L. C., & Peterson, J. B. (2013). Unifying the aspects of the big five, the interpersonal circumplex, and trait affiliation. *Journal of Personality, 81*, 465-475.

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95*, 542-575.

Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist, 55,* 34-43.

Diener, E., & Chan, M. Y. (2011). Happy people live longer: Subjective well-being contributes to health and longevity. *Applied Psychology: Health and Well-Being, 3,* 1-43.

Diener, E., Diener, M., & Diener, C. (2009). Factors predicting the subjective well-being of nations. In E. Diener (Ed.) *Culture and Well-Being: The Collected Works of Ed Diener*, 43-70. Dordrecht (NL): Springer.

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment, 49*, 71-75.

Diener, E., Heintzelman, S. J., Kushlev, K., Tay, L., Wirtz, D., Lutes, L. D., & Oishi, S. (2017). Findings all psychologists should know from the new science on subjective well-being. *Canadian Psychology, 58,* 87-104.

Diener, E., Lucas, R. E., & Oishi, S. (2018). Advances and open questions in the science of subjective well-being. *Collabra: Psychology* (online).

Diener, E., & Ryan, K. (2009). Subjective well-being: A general overview. *South African Journal of Psychology*, *39*, 391-406.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*, 1246-1256.

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.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PloS One* (online).

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*, 1087-1101.

Ellison, C. G. (1991). Religious involvement and subjective well-being. *Journal of Health and Social Behavior, 32,* 80-99.

Elshafei, A., Hatem, A., Parsons, J. K., Polascik, T., Tay, K. J., Given, R., ... & Jones, J. S. (2018). MP30-20 propensity score matching (PSM) comparison of the salvage focal to salvage total cryoablation of the prostate. *The Journal of Urology, 199*, e383.

Epstein, R. (1984). The principle of parsimony and some applications in psychology. *The Journal of Mind and Behavior*, 119-130.

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science, 22,* 295-299.

Everett, J. A. C., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in Psychology*, *6*, 1152-1156.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual of the Eysenck Personality Questionnaire*. San Diego, CA (US): EdITS.

Fava, G. A., & Tomba, E. (2009). Increasing psychological well-being and resilience by psychotherapeutic methods. *Journal of personality, 77*, 1903-1934.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Foster, E. M. (2003). Propensity score matching: An illustrative analysis of dose response. *Medical Care, 41*, 1183-1192.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS One* (online).

Freidlin, P., Littman-Ovadia, H., & Niemiec, R. M. (2017). Positive psychopathology: Social anxiety via character strengths underuse and overuse. *Personality and Individual Differences, 108,* 50-54.

Gavin, J., Keough, M., Abravanel, M., Moudrakovski, T., & Mcbrearty, M. (2014). Motivations for participation in physical activity across the lifespan. *International Journal of Wellbeing, 4*, 46-61.

Gerber, A. S., Huber, G. A., Doherty, D., & Dowling, C. M. (2011). The big five personality traits in the political arena. *Annual Review of Political Science, 14*, 265-287.

Gershuny, B. S., & Sher, K. J. (1998). The relation between personality and anxiety: Findings from a 3-year prospective study. *Journal of Abnormal Psychology, 107*, 252-262.

Golbeck, J., Robles, C., Edmondson, M., & Turner, K. (2011). Predicting personality from twitter. *IEEE Conference Proceedings, 3¸149-156.*

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*, 1216-1229.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist, 48*, 26-34.

Gomez-Uribe, C. A., & Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems, 6*, 13-32.

Gosling, S. D., Augustine, A. A., Vazire, S., Holtzman, N., & Gaddis, S. (2011). Manifestations of personality in online social networks: Self-reported Facebook-related behaviors and observable profile information. *Cyberpsychology, Behavior, and Social Networking, 14*, 483-488.

Gosling, S. D., Gaddis, S., & Vazire, S. (2007). Personality impressions based on Facebook profiles. *ICWSM, 7*, 1-4.

Gosling, S. D., Ko, S. J., Mannarelli, T., & Morris, M. E. (2002). A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Personality and Social Psychology, 82*, 379-398.

Grassegger, H., & Krogerus, M. (2017, January 28). The data that turned the world upside down. *Motherboard*.

Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research, 26*, 499-510.

Grierson, J., & Quinn, B. (2018, April 13). Google loses landmark 'right to be forgotten' case. *The Guardian*.

Grissom, R. J., & Kim, J. J. (2005). *Effect Sizes for Research: A Broad Practical Approach*. Mahwa, NJ (US):Lawrence Erlbaum Associates

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry, 26*, 1-26.

Gunnell, K. E., Crocker, P. R., Wilson, P. M., Mack, D. E., & Zumbo, B. D. (2013). Psychological need satisfaction and thwarting: A test of basic psychological needs theory in physical activity contexts. *Psychology of Sport and Exercise, 14*, 599-607.

Gupta, V. K., Han, S., Mortal, S. C., Silveri, S. D., & Turban, D. B. (2018). Do women CEOs face greater threat of shareholder activism compared to male CEOs? A role congruity perspective. Journal of Applied Psychology, *103*, 228-236.

Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology, 104*, 354-370.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29-36.

Headey, B., Kelley, J., & Wearing, A. (1993). Dimensions of mental health: Life satisfaction, positive affect, anxiety and depression. *Social Indicators Research, 29*, 63-82.

Heintzelman, S. J. (2018). Eudaimonia in the contemporary science of subjective well-being: Psychological well-being, self-determination, and meaning in life. In E. Diener, S. Oishi, & L. Tay (Eds.) *Handbook of Well-Being.* Salt Lake City, UT (US): DEF.

Heiphetz, L., Spelke, E. S., & Banaji, M. R. (2013). Patterns of implicit and explicit attitudes in children and adults: Tests in the domain of religion. *Journal of Experimental Psychology: General, 142*, 864-879.

Heller, D., Judge, T. A., & Watson, D. (2002). The confounding role of personality and trait affectivity in the relationship between job and life satisfaction. *Journal of Organizational Behavior, 23*, 815-835.

Helliwell, J. F. (2006). Well-Being, social capital and public policy: What's new? *The Economic Journal, 116*, C34-C45.

Helliwell, J. F., R. Layard, & J. D. Sachs. (2018). *World Happiness Report*. New York (US): United Nations

Henning, P. J. (2017, July 10). Digital privacy to come under Supreme Court's scrutiny. *The New York Times*.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61-83.

Herzberg, F. (1964). The motivation-hygiene concept and problems of manpower. *Personnel Administration, 27*, 3-7.

Hill, K. R., Walker, R. S., Božičević, M., Eder, J., Headland, T., Hewlett, B., ... & Wood, B. (2011). Co-residence patterns in hunter-gatherer societies show unique human social structure. *Science, 331*, 1286-1289.

Hirsh, J. B., DeYoung, C. G., Xu, X., & Peterson, J. B. (2010). Compassionate liberals and polite conservatives: Associations of agreeableness with political ideology and moral values. *Personality and Social Psychology Bulletin, 36*, 655-664.

Hitt, L., and F. Frei (2002): Do better customers utilize electronic distribution channels? The case of PC Banking. *Management Science, 48*, 732–748.

Hoeyberghs, L., Verté, E., Verté, D., De Witte, N., & Schols, J. (2018). Hopelessness, life dissatisfaction and boredom among older people. *British Journal of Community Nursing*, *23*, 400-405.

Holbrook, M. B. (2001). The millennial consumer in the texts of my times: Exhibitionism. *Journal of Macromarketing*, *21*, 81-95.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications*. Abingdon (UK):Routledge.

Hsu, T., & Kang, C. (2018, March 26). Demand grows for Facebook to explain its privacy policies. *New York Times*.

Hu, W., Singh, R. R., & Scalettar, R. T. (2017). Discovering phases, phase transitions, and crossovers through unsupervised machine learning: A critical examination. *Physical Review, 95*, 1-14.

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*, 645-654.

Ispas, D., Iliescu, D., Ilie, A., & Johnson, R. E. (2014). Exploring the cross-cultural generalizability of the five-factor model of personality: The Romanian NEO PI-R. *Journal of Cross-Cultural Psychology*, *45*, 1074-1088.

Israel, G. D. (1992). *Determining Sample Size.* Agricultural Education and Communication Department, University of Florida, IFAS Extension.

James, W. (1890). *The Principles of Psychology*. Redditch (UK):Read Books Ltd.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York (US): Springer.

Jang, K. L., Livesley, W. J., Angleitner, A., Riemann, R., & Vernon, P. A. (2002). Genetic and environmental influences on the covariance of facets defining the domains of the five-factor model of personality. *Personality and Individual Differences*, *33*, 83-101.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.) *Handbook of Personality: Theory and Research (Vol. 2),*102-138. New York (USA): Guilford Publications

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, *51*, 78-89.

Joshanloo, M., & Ghaedi, G. (2009). Value priorities as predictors of hedonic and eudaimonic aspects of well-being. *Personality and Individual Differences, 47*, 294-298.

Jovanović, V. (2015). A bifactor model of subjective well-being: A re-examination of the structure of subjective well-being. *Personality and Individual Differences, 87*, 45-49.

Just, C. (2011). A review of literature on the general factor of personality. *Personality and Individual Differences, 50*, 765-771.

Kaplan, S. C., Levinson, C. A., Rodebaugh, T. L., Menatti, A., & Weeks, J. W. (2015). Social anxiety and the big five personality traits: The interactive relationship of trust and openness. *Cognitive Behaviour Therapy, 44*, 212-222.

Keller, M. B., Lavori, P. W., Friedman, B., Nielsen, E., Endicott, J., McDonald-Scott, P., & Andreasen, N. C. (1987). The longitudinal interval follow-up evaluation: A comprehensive method for assessing outcome in prospective longitudinal studies. *Archives of General Psychiatry, 44*, 540-548.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods, 39*, 979-984.

Keltner, D., Kogan, A., Piff, P. K., & Saturn, S. R. (2014). The sociocultural appraisals, values, and emotions (SAVE) framework of prosociality: Core processes from gene to meme. *Annual Review of Psychology, 65*, 425-460.

Keng, S. L., Smoski, M. J., & Robins, C. J. (2011). Effects of mindfulness on psychological health: A review of empirical studies. *Clinical Psychology Review, 31*, 1041-1056.

Kenny, D. A. (2005). Cross-lagged panel design. In B. Everitt & D. Howell (Eds.) *Encyclopedia of Statistics in Behavioral Science*. Hoboken, NJ (US):John Wiley & Sons

Kern, M. L., Waters, L. E., Adler, A., & White, M. A. (2015). A multidimensional approach to measuring well-being in students: Application of the PERMA framework. *The Journal of Positive Psychology, 10*, 262-271.

Kim-Cohen, J., Caspi, A., Taylor, A., Williams, B., Newcombe, R., Craig, I. W., & Moffitt, T. E. (2006). MAOA, maltreatment, and gene–environment interaction predicting children's mental health: New evidence and a meta-analysis. *Molecular Psychiatry, 11*, 903-913.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Cemalcilar, Z. (2014). Investigating variation in replicability. *Social psychology, 45*, 142-152.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS, 110*, 5802-5805.

Koutrika, G. (2018). *Recent Advances in Recommender Systems: Matrices, Bandits, and Blenders*. Presentation, Athens (GR).

Kraha, A., Turner, H., Nimon, K., Zientek, L., & Henson, R. (2012). Tools to support interpreting multiple regression in the face of multicollinearity. *Frontiers in Psychology* (online).

Kuhn, M. (2008). Caret package. *Journal of Statistical Software, 28*, 1-26.

Lamers, S. M., Westerhof, G. J., Kovács, V., & Bohlmeijer, E. T. (2012). Differential relationships in the association of the big five personality traits with positive mental health and psychopathology. *Journal of Research in Personality, 46*, 517-524.

Lapowsky, I. (2017, October 26). What did Cambridge Analytica really do for Trump's campaign? *Wired*.

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods, 42*, 871-876.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: The coming age of computational social science. *Science, 323*, 721-723.

Levelt, W. J., Drenth, P. J. D., & Noort, E. (2012). Flawed science: The fraudulent research practices of social psychologist Diederik Stapel. Presentation, Nigmegen (NL).

LeVine, R. A. (2018). *Culture, Behavior, and Personality: An Introduction to the Comparative Study of Psychosocial Adaptation*. Abingdon (UK):Routledge.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*, 18-22.

Linton, M. J., Dieppe, P., & Medina-Lara, A. (2016). Review of 99 self-report measures for assessing well-being in adults: Exploring dimensions of well-being and developments over time. *BMJ Open* (online).

Liptak, A. (2018, June 22). In ruling on cell phone data, Supreme Court makes statement on digital privacy. *New York Times*.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*, 705-717.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*, 584-585.

Lomas, N. (2018, April 24). Kogan: 'I don't think Facebook has a developer policy that is valid'. *Tech Crunch.*

Longo, Y., Coyne, I., & Joseph, S. (2017). The scales of general well-being (SGWB). *Personality and Individual Differences, 109*, 148-159.

Lu, B., & Lemeshow, S. (2018). Survey sampling and propensity score matching. In P. Irwing, T. Booth & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (95-111). New Jersey (US): John Wiley.

Lucas, R. E., & Diener, E. (2015). Personality and subjective well-being: Current issues and controversies. In M. Mikulincer, P. R. Shaver, M. L. Cooper, & R. J. Larsen (Eds.) *APA Handbook of Personality and Social Psychology (Vol. 4)*. Washington, DC (US): American Psychological Association.

Lykken, D., & Tellegen, A. (1996). Happiness is a stochastic phenomenon. *Psychological science, 7*, 186-189.

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research, 46*, 137-155.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40.

Manesi, Z., Van Lange, P. A., Van Doesum, N. J., & Pollet, T. V. (2018). What are the most powerful predictors of charitable giving to victims of typhoon Haiyan: Prosocial traits, socio-demographic variables, or eye cues? *Personality and Individual Differences* (online).

Manrique, P., Qi, H., Morgenstern, A., Velasquez, N., Lu, T. C., & Johnson, N. (2013). Context matters: Improving the uses of big data for forecasting civil unrest. *Proceedings of IEEE International Conference*, 169-172.

Mantelero, A. (2013). The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review, 29*, 229-235.

Martela, F., & Ryan, R. M. (2016). The Benefits of benevolence: Basic psychological needs, beneficence, and the enhancement of well-being. *Journal of Personality, 84*, 750-764.

Maslow, A. H. (1943). Hierarchy of needs: A theory of human motivation. *Psychological Review, 50*, 370-396.

Matherly, T. (2018). A Panel For Lemons? Positivity bias, reputation systems and data quality on MTurk. *European Journal of Marketing* (online).

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *PNAS, 114*, 12714-12719.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*, 181–190.

McAdams, D. P. (1997). A conceptual history of personality psychology. In R. Hogan, J. Johnson & S. Briggs (Eds) *Handbook of Personality Psychology*, 3-39. Cambridge, MA (US):Academic Press

McClelland, D. C. (1965). Toward a theory of motive acquisition. *American Psychologist, 20*, 321-333.

McCrae, R. R., & Allik, J. (2002). *The Five-Factor Model of Personality Across Cultures*. Berlin (DE): Springer

McCrae, R. R., & Costa, P. T., Jr. (1992). Discriminant validity of NEO-PI-R facets. *Educational and Psychological Measurement, 52,* 229–237.

McCrae, R. R., Costa Jr, P. T., Del Pilar, G. H., Rolland, J. P., & Parker, W. D. (1998). Cross-cultural assessment of the five-factor model: The revised NEO personality inventory. *Journal of Cross-Cultural Psychology, 29,* 171-188.

McCrae, R. R., & Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: data from 50 cultures. *Journal of Personality and Social Psychology, 88,* 547-561.

McLeod, S. (2007). Maslow's hierarchy of needs. *Simply Psychology* (online).

Morrison, M., Tay, L., & Diener, E. (2011). Subjective well-being and national satisfaction: Findings from a worldwide survey. *Psychological Science, 22,* 166-171.

Mõttus, R., McCrae, R. R., Allik, J., & Realo, A. (2014). Cross-rater agreement on common and specific variance of personality scales and items. *Journal of Research in Personality, 52,* 47-54.

Mowery, D. L., Park, A., Bryan, C., & Conway, M. (2016). Towards automatically classifying depressive symptoms from Twitter data for population health. *Proceedings of the Workshop on Computational Modeling of PEOPLES,* 182-191.

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18,* 107-118.

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality, 41,* 1213-1233.

Naragon-Gainey, K., & Simms, L. J. (2017). Clarifying the links of conscientiousness with internalizing and externalizing psychopathology. *Journal of Personality, 85,* 880-892.

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied, 7,* 171-181.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual review of psychology, 69,* 511-534.

Neubauer, A. B., & Voss, A. (2016a). The structure of need fulfilment: Separating need satisfaction and dissatisfaction on between- and within-person level. *European Journal of Psychological Assessment, 34,* 220-228.

Neubauer, A. B., & Voss, A. (2016b). Validation and revision of a German version of the balanced measure of psychological needs scale. *Journal of Individual Differences, 37,* 56-72.

Newman, D. B., Tay, L., & Diener, E. (2014). Leisure and subjective well-being: A model of psychological mechanisms as mediating factors. *Journal of Happiness Studies, 15,* 555-578.

Obar, J. A., & Oeldorf-Hirsch, A. (2016). The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Proceedings of the 44th Research Conference on Communication, Information and Internet Policy,* 1-20.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97,* 307-315.

Packer, D. J. (2008). Identifying systematic disobedience in Milgram's obedience experiments: A meta-analytic review. *Perspectives on Psychological Science, 3,* 301-304.

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81,* 524-539.

Pavot, W., & Diener, E. (2008). The satisfaction with life scale and the emerging construct of life satisfaction. *The Journal of Positive Psychology, 3,* 137-152.

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology*, *60*, 65-82.

Plante, C. N., Reysen, S., Groves, C. L., Roberts, S. E., & Gerbasi, K. (2017). The fantasy engagement scale: A flexible measure of positive and negative fantasy engagement. *Basic and Applied Social Psychology, 39*, 127-152.

Ponterotto, J. G. (2002). Qualitative research methods: The fifth force in psychology. *The Counseling Psychologist, 30*, 394-406.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717-731.

Prentice, M., Jayawickreme, E., & Fleeson, W. (2018). Integrating Whole Trait Theory and Self-Determination Theory. *Journal of Personality*, 1-14.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1-15.

Quevedo, R. J. M., & Abella, M. C. (2011). Well-being and personality: Facet-level analyses. *Personality and Individual Differences, 50*, 206-211.

Reis, H. T., Sheldon, K. M., Gable, S. L., Roscoe, J., & Ryan, R. M. (2000). Daily well-being: The role of autonomy, competence, and relatedness. *Personality and Social Psychology Bulletin, 26*, 419-435.

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology, 84*, 1236-1256.

Riemann, R., & Kandler, C. (2010). Construct validation using multitrait-multimethod-twin data: The case of a general factor of personality. *European Journal of Personality, 24*, 258-277.

Robbins, B. D. (2008). What is the good life? Positive psychology and the renaissance of humanistic psychology. *The Humanistic Psychologist, 36*, 96-112.

Robinson, (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351-357.

Rocchi, M., Pelletier, L., Cheung, S., Baxter, D., & Beaudry, S. (2017). Assessing need-supportive and need-thwarting interpersonal behaviours: The interpersonal behaviours questionnaire (IBQ). *Personality and Individual Differences*, *104*, 423-433.

Rooney, B. L., Mathiason, M. A., & Schauberger, C. W. (2011). Predictors of obesity in childhood, adolescence, and adulthood in a birth cohort. *Maternal and Child Health Journal, 15*, 1166-1175.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*, 33-38.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*, 2-14.

Rushton, J. P., & Irwing, P. (2008). A general factor of personality (GFP) from two meta-analyses of the big five: Digman (1997) and Mount, Barrick, Scullen, and Rounds (2005). *Personality and Individual Differences, 45*, 679-683.

Ryali, S., Chen, T., Supekar, K., & Menon, V. (2012). Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage, 59*, 3852-3861.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist, 55*, 68-78.

Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology, 52*, 141-166.

Ryan, R. M., & Deci, E. L. (2017). *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York (US):Guilford Publications.

Ryan, R. M., & La Guardia, J. G. (2000). What is being optimized? Self-determination theory and basic psychological needs. In S. H. Qualls & N. Abeles (Eds.) *Psychology and the Aging Revolution: How I Adapt to Longer Life*, 145-172. Washington, DC (US): American Psychological Association.

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.

Ryff, C. D., Love, G. D., Urry, H. L., Muller, D., Rosenkranz, M. A., Friedman, E. M., ... & Singer, B. (2006). Psychological well-being and ill-being: Do they have distinct or mirrored biological correlates? *Psychotherapy and Psychosomatics*, *75*, 85-95.

Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the big five personality factors: A cross-language replication. *Journal of Personality and Social Psychology, 76*, 613-627.

Schaller, M., & Murray, D. R. (2008). Pathogens, personality, and culture: Disease prevalence predicts worldwide variability in sociosexuality, extraversion, and openness to experience. *Journal of Personality and Social Psychology*, *95*, 212-221.

Scherman, A., Arriagada, A., & Valenzuela, S. (2015). Student and environmental protests in Chile: The role of social media. *Politics, 35*, 151-171.

Schimmack, U., Oishi, S., Furr, R. M., & Funder, D. C. (2004). Personality and life satisfaction: A facet-level analysis. *Personality and Social Psychology Bulletin, 30*, 1062-1075.

Schulenberg, J. E., Sameroff, A. J., & Cicchetti, D. (2004). The transition to adulthood as a critical juncture in the course of psychopathology and mental health. *Development and Psychopathology, 16*, 799-806.

Schwartz, S. H., & Boehnke, K. (2004). Evaluating the structure of human values with confirmatory factor analysis. *Journal of Research in Personality, 38*, 230-255.

Schwartz, S. H., Caprara, G. V., & Vecchione, M. (2010). Basic personal values, core political values, and voting: A longitudinal analysis. *Political Psychology, 31*, 421-452.

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*, 513–523.

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality* (online).

Seligman, M. E., & Csikszentmihalyi, M. (2014). *Flow and the Foundations of Positive Psychology*. Rotterdam (NL): Springer.

Sheldon, K. M., & Elliot, A. J. (1999). Goal striving, need satisfaction, and longitudinal well-being: The self-concordance model. *Journal of Personality and Social Psychology, 76*, 482-497.

Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology, 80*, 325-339.

Sheldon, K. M., & Hilpert, J. C. (2012). The balanced measure of psychological needs (BMPN) scale: An alternative domain general measure of need satisfaction. *Motivation and Emotion, 36*, 439-451.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual review of psychology, 69*, 487-510.

Siddique, J., de Chavez, P. J., Howe, G., Cruden, G., & Brown, C. H. (2018). Limitations in using multiple imputation to harmonize individual participant data for meta-analysis. *Prevention Science, 19*, 95-108.

Skowron, M., Tkalčič, M., Ferwerda, B., & Schedl, M. (2016). Fusing social media cues: Personality prediction from Twitter and Instagram. *Proceedings of the 25th International Conference on the World Wide Web*, 107-108.

Smith, T. W., Gallo, L. C., Goble, L., Ngu, L. Q., & Stark, K. A. (1998). Agency, communion, and cardiovascular reactivity during marital interaction. *Health Psychology, 17*, 537-545.

Soto, C. J. (2015). Is happiness good for your personality? Concurrent and prospective relations of the big five with subjective well-being. *Journal of Personality, 83*, 45-55.

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*, 117-143.

Soto, C. J., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology*, *100*, 330-348.

Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and SWB. *Psychological Bulletin, 134*, 138-161.

Steptoe, A., Hamer, M., & Chida, Y. (2007). The effects of acute psychological stress on circulating inflammatory factors in humans: A review and meta-analysis. *Brain, Behavior, and Immunity, 21*, 901-912.

Strickhouser, J. E., Zell, E., & Krizan, Z. (2017). Does personality predict health and well-being? A metasynthesis. *Health Psychology*, *36*, 797-810.

Sumner, C., Byers, A., Boochever, R., & Park, G. J. (2012). Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets. *Proceedings of Machine learning and Applications, 11*, 386-393.

Sun, M., Li, C., & Zha, H. (2017). Inferring private demographics of new users in recommender systems. *Proceedings of the ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems, 20*, 237-244.

Suzuki, T., Samuel, D. B., Pahlen, S., & Krueger, R. F. (2015). DSM-5 alternative personality disorder model traits as maladaptive extreme variants of the five-factor model: An item-response theory analysis. *Journal of Abnormal Psychology, 124*, 343-354.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*, e2000797.

Tam, N. T., Huy, N. T., Thoa, L. T. B., Long, N. P., Trang, N. T. H., Hirayama, K., & Karbwang, J. (2015). Participants' understanding of informed consent in clinical trials over three decades: Systematic review and meta-analysis. *Bulletin of the World Health Organization, 93*, 186-198H.

Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma & J. D. Maser (Eds.) *Anxiety and the Anxiety Disorders*, 681-706. Hillsdale, NJ (US): Lawrence Erlbaum Associates.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Education and Psychological Measurement, 55*, 525-534.

Unanue, W., Dittmar, H., Vignoles, V. L., & Vansteenkiste, M. (2014). Materialism and well-being in the UK and Chile: Basic need satisfaction and basic need frustration as underlying psychological processes. *European Journal of Personality, 28*, 569-585.

Uskul, A. K., & Over, H. (2017). Culture, social interdependence, and ostracism. *Current Directions in Psychological Science, 26*, 371-376.

Van den Broeck, A., Ferris, D. L., Chang, C. H., & Rosen, C. C. (2016). A review of self-determination theory's basic psychological needs at work. *Journal of Management, 42*, 1195-1229.

Van der Linden, D., Dunkel, C. S., & Petrides, K. V. (2016). The General Factor of Personality (GFP) as social effectiveness: Review of the literature. *Personality and Individual Differences, 101*, 98-105.

Van Katwyk, P. T., Fox, S., Spector, P. E., & Kelloway, E. K. (2000). Using the job-related affective well-being scale (JAWS) to investigate affective responses to work stressors. *Journal of Occupational Health Psychology, 5*, 219-230.

Vanhove-Meriaux, C., Martinent, G., & Ferrand, C. (2018). Profiles of needs satisfaction and thwarting in older people living at home: Relationships with well-being and ill-being indicators. *Geriatrics & Gerontology International*, *18*, 470-478.

Verdugo, R. R. (2002). Race-ethnicity, social class, and zero-tolerance policies: The cultural and structural wars. *Education and Urban Society*, *35*, 50-75.

Volkova, S., Bachrach, Y., Armstrong, M., & Sharma, V. (2015). Inferring latent user properties from texts published in social media. *Proceedings of AAAI*, 4296-4297.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (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, 3*, 426-432.

Walsh, R. (2011). Lifestyle and mental health. *American Psychologist, 66*, 579-592.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science, 5*, 457-469.

Walter, S. L., Seibert, S. E., Goering, D., & O'Boyle, E. H. (2018). A tale of two sample sources: Do results from online panel data and conventional data converge? *Journal of Business and Psychology* (online).

Wang, Y., & Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114*, 246-257.

Watson, C. (2018, April 11). The key moments from Mark Zuckerberg's testimony to Congress. *The Guardian*.

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.

Weiss, L. A., Westerhof, G. J., & Bohlmeijer, E. T. (2016). Can I increase psychological well-being? The effects of interventions on psychological well-being: a meta-analysis of randomized controlled trials. *PloS One*, *11*, e0158092.

White, N. P. (2008). *A Brief History of Happiness*. Hoboken, NJ (US): John Wiley & Sons.

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation, 14*, 2-10.

Wibowo, M. R. F., Yudiana, W., Reswara, I. P., & Jatmiko, B. W. (2017). Reliability and Validity of the Indonesian Version of Big Five Inventory. *UI Proceedings on Social Science and Humanities* (online).

Witherington, D. C., & Lickliter, R. (2017). Transcending the nature-nurture debate through epigenetics: Are I there yet? *Human Development, 60*, 65-68.

Woods, S. A., & Anderson, N. R. (2016). Toward a periodic table of personality: Mapping personality scales between the five-factor model and the circumplex model. *Journal of Applied Psychology, 101*, 582-604.

Wright, A. G. (2017). Factor analytic support for the five-factor model. In T. Widger (Ed.) *The Oxford Handbook of the Five Factor Model*, 217-242. Oxford (UK): Oxford University Press.

Yang, J., & Coughlin, J. F. (2014). In-vehicle technology for self-driving cars: Advantages and challenges for aging drivers. *International Journal of Automotive Technology, 15*, 333-340.

Yang, Y., Zhang, Y., & Sheldon, K. M. (2017). Self-determined motivation for studying abroad predicts lower culture shock and greater well-being among international students: The mediating role of basic psychological needs satisfaction. *International Journal of Intercultural Relations, 63*, 95-104.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*, 1100-1122.

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *PNAS, 112*, 1036-1040.

Zecca, G., Verardi, S., Antonietti, J. P., Dahourou, D., Adjahouisso, M., Ah-Kion, J., ... & Dougoumalé Cissé, D. (2013). African cultures and the five-factor model of personality: Evidence for a specific pan-African structure and profile? *Journal of Cross-Cultural Psychology, 44*, 684-700.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*, 301-320.

Appendices

Appendix 1.1: Final Approved AXA Ethics Application (PRE.2016.027.V8)

Question 1: Title of the study

Notes: The title should be a single sentence

Cross-national study of social relationships, prosociality, well-being, health and political preferences using big data.

Question 2: Primary applicant

Notes: The primary applicant is the name of the person who has overall responsibility for the study. Include their appointment or position held and their qualifications. Primary applicants cannot be research students or junior research assistants. For studies where students and/or research assistants will undertake the research, the primary applicant would normally be their supervisor.

Dr Aleksandr Kogan (now Dr Aleksandr Spectre)
University Lecturer, Department of Psychology, University of Cambridge (2012-Present)
Ph.D. in Psychology, University of Hong Kong (2011)
B.A. in Psychology, University of California Berkely (2008)

Question 3: Co-applicants

Notes: List the names of all researchers involved in the study. Include their appointment or position held and their qualifications

Question 4: Corresponding applicant

Notes: Give the name of the person to whom correspondence regarding this application is to be addressed. This person should be the primary applicant or one of the co-applicants. An email address for correspondence must be provided.

Primary: Dr Aleksandr Kogan (now Dr Spectre; ak823@cam.ac.uk)
Secondary: Mr Matthew Samson (mjs268@cam.ac.uk)

Address (for both): Department of Psychology, Downing Street, Cambridge, CB2 3EB

Question 5: In which Department(s) or Research Unit(s) will the study take place?

Notes: Indicate where the study procedures will take place as well as the location for the storage and analysis of data. If the study will use National Health Service facilities, give a contact name and address of the Trust R&D office.

Study Procedures: Department of Psychology, University of Cambridge
Data Storage: Raw data will be stored on secure servers at Philometrics Inc., 191 Goodwin Avenue-Suite 5, Wyckoff, NJ, USA, 07481-2052. A fully anonymised version of this data will then be made available to the applicant and Co-applicants listed above, and stored on a secure server in the Cambridge Prosociality and Well-Being lab (which is headed by Dr. Kogan; now Dr Spectre).
Data Analysis: Department of Psychology, University of Cambridge

Question 6: What are the start and end dates of the study?

Notes: If exact dates are unavailable, explain why and give approximate dates.

Start: Immediately upon receiving ethics approval.
Finish: December, 2019

Question 7: Briefly describe the purpose and rationale of the research

Our goals are as follows:

(a) Develop and validate methods for making forecasts about a wide variety of demographic and psychological variables using tweets and Twitter users' account information, which is in compliance with Twitter's User terms and conditions and existing data sharing agreements to which our collaborator – Philometrics Inc. - is subjected.. We anticipate that while accuracy of the forecasted scores at the individual level for the variables will range from weak (for complex psychological states like well-being) to strong (for demographic variables), the correlations between forecasted variables and their aggregates (e.g. state or country level averages) will be highly accurate—that is, very similar to findings using traditional surveys.

(b) Develop and validate methods for making these same forecasts about a wide variety of demographic and psychological variables using participants' self-reported personality item responses.

(c) Once we have developed and validated the machine learning approaches, we will investigate how a wide variety of psychological and demographic variables are related to well-being and health, and how these relationships vary across countries. This work will be guided by existing theory within social psychology on the social determinants of well-being and health (e.g. conservatives are happier than liberals; kinder people tend to be happier and healthier; having a large number of social contacts predicts better health), and also exploratory analyses at the cross-national level aimed at understanding cultural variability in these effects.

(d) Once we have constructed the broad dataset of forecasted scores (for millions of people) and actual survey responses (for thousands of people), we will make the dataset available to the broader scientific community for secondary analysis.

The results of the initial data collection phase of our project will be a survey dataset comprising approximately 40,000 respondents, up to 15,000 of whom had also disclosed a sufficient amount of their twitter behavior to generate forecasts. In addition, we also have a dataset of hundreds of millions of Twitter users who did not complete any surveys, to whom we will forecast survey responses. We will make fully anonymized versions of both the original survey dataset and the much larger forecasted dataset publicly available for academic research. This is to ensure we gain maximum positive impact from our allotted grant resources. The protocol is listed in Question 9 of this application.

Question 8: Who is funding the costs of the study?

Notes: Give the name and address of funding bodies or other sponsorship (other than the University of Cambridge) involved in providing resources for the study.

AXA Research Fund
UK/Ireland, Mediterranean region & Latin America
Life & Health Risks
GIE AXA 25, avenue de Matignon 75008
Paris, France
Tel.: +33 1 40 75 39 86/ E.Fax : +33 1 56 69 93 29

Please find successful grant application in Appendix A.

Question 9: Describe the methods and procedures of the study

The proposed program of research is a collection of numerous small studies that all follow the same basic procedure. Advertisements will be posted on Twitter that tells viewers they can receive feedback about their various psychological traits (e.g. "Find out your personality", "Find out your happiness"). Participants who click on the ads will be taken to a consent form that informs them that their data will be used for research purposes and we will model how their twitter activity predicts their responses. It will also outline various data usages (such as sharing the data with other researchers in an anonymous form). Participants will then (a) provide us their twitter account name and (b) will complete a set of questionnaires (e.g., personality questionnaire). We will then provide them immediate feedback on how they scored. For instance, if they complete a Big 5 personality inventory, we will give them feedback on how they scored along each of the 5 dimensions. Thus, our incentive to the participants will not be monetary in nature; rather it will be in the form of insight about them from the questionnaires. In some cases, we will only give them feedback on a single questionnaire, but ask them to complete several short other questionnaires. Below we detail which questionnaires are used, what platform will be used to collect the data, and how access to twitter data will be handled.

In circumstances where Twitter advertisements and/or non-monetary incentives are unfeasible, we propose collecting data with existing survey panel providers. Under such circumstances we may not provide customised feedback because this is not necessarily expected or desired by professional survey takers.

Questionnaires: For each study, questionnaires will be selected from the full list of scales in Appendix B and Appendix B.1. They are convergent measures that relate to the following psychological phenomena: autism, depression, emotions experiences, personality, well-being, health, political orientation, demographic factors, psychological needs, prosocial behaviour, factors that promote resilience, feminist identity, dark triad personality, basic values and morality. All items have demonstrated construct validity when administered to random samples of adult participants without mental illness. During recruitment, we will target participants from 35 countries. For each country, we will present surveys in the native or fully fluent language of the participants.

When implemented, feedback will only be given for questionnaires that have nominal and *not* ordinal valenced endpoints. For example feedback will be given above participants' self-reported score on Introversion to Extraversion, and on Republican to Democrat (for US participants), because in both examples one endpoint is not more obviously positively/negatively valenced than the other. Also by virtue, feedback will *not* be given on scales that measure (e.g.) Depression because there are differences in the valence of endpoints (e.g. participants will likely perceive being depressed as more negatively valenced than not being depressed), which could potentially cause distress. When we intend to give survey feedback, this information is included immediately below its associated survey in Appendix B. Finally, participants will be invited to re-Tweet the survey link to our survey after receiving feedback. This will increase our chances of receiving self-report information from participants' followers, thereby maximising survey dissemination.

The proposed program of research is a collection of numerous small studies that all follow the same basic procedure. Advertisements will be posted on Twitter that tells viewers they can receive feedback about their various psychological traits (e.g. "Find out your personality", "Find out your happiness"). Participants who click on the ads will be taken to a consent form that informs them that their data will be used for research purposes and we will model how their twitter activity predicts their responses. It will also outline various data usages (such as sharing the data with other researchers in an anonymous form). Participants will then (a) provide us their twitter account name and (b) will complete a set of questionnaires (e.g., personality questionnaire). We will then provide them immediate feedback on how they scored. For instance, if they complete a Big 5 personality inventory, we will give them feedback on how they scored along each of the 5 dimensions. Thus, our incentive to the participants will not be monetary in nature; rather it will be in the form of insight about them from the questionnaires. In some cases, we will only give them feedback on a single questionnaire, but ask them to complete several short other questionnaires. Below we detail which questionnaires are used, what platform will be used to collect the data, and how access to twitter data will be handled.

In circumstances where Twitter advertisements and/or non-monetary incentives are unfeasible, we propose collecting data with existing survey panel providers. Under such circumstances we may not provide customised feedback because this is not necessarily expected or desired by professional survey takers.

Questionnaires: For each study, questionnaires will be selected from the full list of scales in Appendix B and Appendix B.1. They are convergent measures that relate to the following psychological phenomena: autism, depression, emotions experiences, personality, well-being, health, political orientation, demographic factors, psychological needs, prosocial behaviour, factors that promote resilience, feminist identity, dark triad personality, basic values and morality. All items have demonstrated construct validity when administered to random samples of adult participants without mental illness. During recruitment, we will target participants from 35 countries. For each country, we will present surveys in the native or fully fluent language of the participants.

When implemented, feedback will only be given for questionnaires that have nominal and *not* ordinal valenced endpoints. For example feedback will be given above participants' self-reported score on Introversion to Extraversion, and on Republican to Democrat (for US participants), because in both examples one endpoint is not more obviously positively/negatively valenced than the other. Also by virtue, feedback will *not* be given on scales that measure (e.g.) Depression because there are differences in the valence of endpoints (e.g. participants will likely perceive being depressed as more negatively valenced than not being depressed), which could potentially cause distress. When we intend to give survey feedback, this information is included immediately below its associated survey in Appendix B. Finally, participants will be invited to re-Tweet the survey link to our survey after receiving feedback. This will increase our chances of receiving self-report information from participants' followers, thereby maximising survey dissemination.

Security and Anonymity: Since a key part of the project is to model how tweets map onto survey responses, linkage between user survey responses and twitter accounts is needed. Thus, initially data will not be anonymous. In order to protect the identity of the participants, all data analysis on raw tweets (which will not be anonymous) will occur on the Philometrics internal servers that are encrypted and password protected. Once a modelling approach has been settled and raw tweets are no longer necessary, we will generate derived dimensional scores for each user and anonymize the data. These dimensions, while based on the original tweets, have the advantage that the same scores can be arrived from many different combinations of tweets (of which there are near-infinite combinations). Thus having only the dimensional scores is not enough to reverse the anonymization.

The derived dimensions data will be made available to the CPW lab (through a data sharing agreement with the Research Office) by Philometrics in connection with an anonymized version of the survey results (the link between surveys and the derived scores will be facilitated by a randomly generated ID). Thus, by the point at which any data reaches the lab, it will be in fully anonymous form. The derived scores will be used in generating forecasts expeditiously —see modelling steps below for more details.

Eventually, some of the data will be made available through a website to other researchers for secondary data analysis. To ensure security, we will make available only (a) the actual survey responses without the Twitter ID link and (b) forecasted scores for several million people for whom we have Twitter data, but who never participated in any survey. For the data in (b), only forecasted results will be made available. No original twitter data of any kind will be made available.

Group (a) has given consent for their data usage and thus their inclusion in the proposed study raises minimal ethical concerns. For group (b), the scores are derived for individuals who never actively participated in any way in our research and thus could not have consented. Thus, more care should be taken in evaluating the ethical implications of their data usage. In our view, since the data that is used to generate scores is publically posted for anyone to see and use, and users of Twitter can be reasonably expected to understand this, there is not a necessity to gain consent for our particular application. Furthermore, we minimize any potential harm to the users through (a) anonymization of the data and (b) providing only forecasted scores which, as we describe below, are relatively inaccurate at the individual level (but provide rather accurate aggregate scores and information about how variables are correlated). Access to the original raw tweets is further guarded and only occurs within Philometrics by a small number of researchers (named the co-investigators on this application). By the time any data reaches university servers, it has already been abstracted to a point where working backwards to de-anonymize the records becomes extremely difficult.

Modelling Approach: We plan on using four aspects of tweets in our modelling: (a) mentions of other popular users (typically brands or celebrities), (b) hashtags, (c) the language used, and (b) who users are following. For all four, our first step will be to reduce the data down to a small set of dimensions (e.g. 20-100). This is done by examining the co-occurrence of different types of mentions, hashtags, words and followers throughout Philometrics's entire database of Tweets, and then collapsing clusters of similar values into singular aggregate variables. Once done, we build models by taking the tweet-derived dimensions of the users who provided survey responses, and entering these dimensions as predictors in various types of linear and non-linear models (e.g., regression lasso/ridge, neural networks) with the outcome being the survey response. We build a separate model for each self-report variable (e.g., a model for well-being, a model for agreeableness, a model for whether the person smokes or not, etc.). Accuracy is tested through cross-validation—that is, we leave out a group with both tweets and survey responses from the modelling process, the use forecasting models developed with other participants to forecast the scores for this left-out group, and then finally comparing the left-out group's forecasted scores to their true responses. Given sufficient model validity (non-zero positive effect between forecasted and self-report scores; evidence for the normal distribution of errors), we then apply the same forecasting method to millions of cases for whom we only have Twitter data. We note that the users we forecast for will have never interacted with us; rather, they are from the database of 120 million or so users supplied to Philometrics by Twitter.

The Accuracy of Predictive Models: Initial pilot studies using data from Facebook data suggest that, at best, such models will predict only around 1/3 of the variance in self-report scores for any given psychological construct, with most effects much lower – at around 1/10. This suggests that at an individual level, the accuracy of the forecasts is rather imprecise. However, the strength of the method is not in individual analyses. Rather, our method aims to make prediction errors normally distributed around participants' true scores. If this precondition is fulfilled - as we consistently find that it is - then inferences about the population average can still be valid even when scores are very imprecise, provided there is enough power. That is, even inaccurate scores produce a highly accurate population average (e.g. nation average, state average, or city average) assuming that there are enough data points. Furthermore, we find that correlations between variables are very similar in the forecasted and actual survey data. Thus, even though individual scores are rather inaccurate, the types of data most useful for researchers - population averages and relationships between variables - are highly robust. We view this as an optimal circumstance as it strongly reduces any possibility of data abuse (which is an especially poignant risk at the individual level) while maintaining the scientific value.

To make the aforementioned data publically accessible, we propose creating a project website using Google Sites. In addition to featuring grant-related research, this will contain an application form that interested parties can complete to gain access to the data. We feel that such an application is prudent to ensure that data are used exclusively for academic purposes and comply with data protection protocols listed in this application. To this end, we have created an application template (Appendix G) that is modelled on the one used by the Out of Service project, which is an existing publically available personality dataset administered by Dr Jeff Potter and colleagues (www.outofservice.com).

The protocol for evaluating an application is as follows: Dr Spectre and at least one listed collaborator will determine whether (a) the request is for exclusively academic purposes; (b) granting access will not bring the reputations of the university, the grant provider, Dr Spectre or collaborators into disrepute; (c) the applicant can be reasonably expected to use the data in an exclusively ethical way. Dr Spectre and all listed collaborators who evaluate the application must agree that these criteria are met before the data are shared. They may also request application revisions or reject an application outright, at their discretion. Each applicant will then be given a unique login to a password-protected page, which contains an indexed version of the data that is available for download. The highly sensitive variable "participant zipcode" will not be made publically available except under exceptional circumstances, and only then after ethics committee approval – via an amendment - for each specific request. Specific Twitter, Philometrics and survey-supplier account ID information will not be made publically available under any circumstance, except as mandated by law.

Question 9a: Does the study involve any pharmaceutical or other compounds with physiological effects?

Notes: This includes all compounds licensed under the Medicines Act. However, some compounds may be considered as Investigational Medical Products and studies of them, therefore, as clinical trials (CTIMPs). If there is any ambiguity, investigators should contact the Medicines and Healthcare Products Regulatory Agency (MHRA) for guidance. Include any response from the MHRA in your application. CTIMPs must seek NRES approval.

No.

Question 10: What ethical issues does this study raise and what measures have been taken to address them?

Notes: Describe any discomfort or inconvenience that participants may experience. Include information about procedures that for some people could be physically stressful or might impinge on the safety of participants, e.g. noise levels, visual stimuli, equipment; or that for some people could be psychologically stressful, e.g. mood induction procedures, tasks with high failure rate. Indicate what procedures are in place if clinically relevant information arises from the study (e.g. from brain scans or questionnaire responses that might indicate that a participant is at risk).

The studies have minimal risk to participants. Those who give self-report information will be made fully aware of the aims and implications of the present study. Moreover, participation will be online and thus participation will occur in a comfortable and convenient environment. Nevertheless, there are additional concerns associated with our collaboration with Philometrics, as well as with the application of our machine learning method to the larger database of Twitter profiles.

Collaboration with Philometrics: Dr Kogan (now Dr Spectre) is co-founder and active member of Philometrics, and also primary applicant on this document. To mitigate any potential conflict of interest, Philometrics will provide the above-mentioned dataset to the applicants listed in this document free of charge. Philometrics cannot provide raw data to the lab because this breaches their agreement with Twitter. Whilst the applicants listed here will use the Philometrics survey platform to collect data, this is for entirely practical reasons: Philometrics offers the capacity to deliver detailed customized feedback to participants – an important incentive – that is unavailable by competing survey platforms such as Qualtrics, Google Forms and Survey Monkey either at all or in any sort of easy to use manner.

Participant Anonymity: Self-report participants will provide their Twitter username, data from their twitter account and self-report information. Thus, data in their raw format are not anonymous. To mitigate, versions containing Twitter account information will only be used in the preliminary stages - when we establish the best way to reduce the tweet information into dimensions. This stage will only be undertaken by those co-applicants who are interning at Philometrics. Furthermore, we will take several steps to ensure data protection at the various stages of its usage, both in the lab and beyond. For further reference please see the above section on study procedures.

Public dissemination of available data: Please see Question 9

Question 11: Who will the participants be?

Notes: Describe the groups of participants that will be recruited and the principal eligibility criteria and ineligibility criteria. Make clear how many participants you plan to recruit into the study in total.

Eligibility criteria: Aged 18 and older.
Participants: We aim to recruit at least 1,000 from each of the following 35 countries (resulting in 35,000-40,000 total participants): Argentina, Australia, Austria, Belgium, Brazil, Bolivia, Canada, Chili, China (mainland), China (Hong Kong), Colombia, Ecuador, Finland, France, Germany, India, Indonesia, Israel, Italy, Japan, Mexico, Paraguay, Peru, Poland, Portugal, Russia, South Africa, South Korea, Spain, Taiwan, Thailand, Turkey, UK, Uruguay, USA, and Venezuela.. Finally, we also aim to collect another 1,000 responses from Arabic-speaking participants throughout the Middle East. Surveys that are not relevant to each of these countries will be omitted (e.g. US political orientation for countries outside the US). Surveys will be translated into at least one official non-English language for each country, when the English survey is not appropriate.

Question 12: Describe the recruitment procedures for the study

Notes: Gives details of how potential participants will be identified or recruited. Include all advertising materials (posters, emails, letters etc.) as appendices and refer to them as appropriate. Describe any screening examinations. If it serves to explain the procedures better, include as an appendix a flow chart and refer to it.

Recruitment will take place through Twitter's advertising platform. Please see Appendix C for the recruitment flyer. Alternatively, recruitment will take place via a University sanctioned survey panel provider.

Question 13: Describe the procedures to obtain informed consent

*Notes: Describe when consent will be obtained. If consent is from **adult participants**, give details of who will take consent and how it will be done. If you plan to seek informed consent from **vulnerable groups** (e.g. people with learning difficulties, victims of crime), say how you will ensure that consent is voluntary and fully informed.*

*If you are recruiting **children or young adults** (aged under 18 years) specify the age-range of participants and describe the arrangements for seeking informed consent from a person with parental responsibility. If you intend to provide children under 16 with information about the study and seek agreement, outline how this process will vary according to their age and level of understanding.*

How long will you allow potential participants to decide whether or not to take part? What arrangements have been made for people who might not adequately understand verbal explanations or written information given in English, or who have special communication needs?

If you are not obtaining consent, explain why not.

Consent will be gathered at the beginning of the study. We will present the participants with the information sheet (Appendix D) and then ask them to consent to partake in the study (Appendix E).

Question 14: Will consent be written?

*Notes: If **yes**, include a consent form as an appendix. If **no**, describe and justify an alternative procedure (verbal, electronic etc.) in the space below.*

Guidance on how to draft Participant Information sheet and Consent form can be found on the Psychology Research Ethics Committee website.

Yes. Consent will be written in electronic format at the beginning of the study (see Appendices D and E).

Question 15: What will participants be told about the study? Will any information on procedures or the purpose of study be withheld?

Notes: Include an Information Sheet that sets out the purpose of the study and what will be required of the participant as appendices and refer to it as appropriate. If any information is to be withheld, justify this decision. More than one Information Sheet may be necessary.

The purpose of the proposed research will be provided by the Information Sheet at the very beginning of the study. Feedback will be given at the end of the study to maximise comprehension (Appendix F). No information will be withheld.

Question 16: Will personally identifiable information be made available beyond the research team?

Notes: If so, indicate to whom and describe how consent will be obtained.

We will collect twitter user ID as part of the procedure. This information will not be made available beyond the research team.

Question 17: What payments, expenses or other benefits and inducements will participants receive?

Notes: Give details. If it is monetary say how much, how it will be paid and on what basis is the amount determined.

No payment will be made directly to any participants. We will instead compensate them in the form of survey feedback. Alternatively, they will be compensated via payment to a university-sanctioned survey panel provider.

Question 18: At the end of the study, what will participants be told about the investigation?

Notes: Give details of debriefings, ways of alleviating any distress that might be caused by the study and ways of dealing with any clinical problem that may arise relating to the focus of the study.

The aims of the study will be made fully transparent from the outset. Participants will be reminded of these aims upon completing the study via the feedback form (Appendix F).

Question 19: Has the person carrying out the study had previous experience of the procedures? If not, who will supervise that person?

Notes: Say who will be undertaking the procedures involved and what training and/or experience they have. If supervision is necessary, indicate who will provide it.

Yes. Administration is done using conventional online survey methods, which are familiar to all listed applicants and co-applicants. Further, these people are all also fully aware of the appropriate procedure to conduct a study and how to strictly follow all corresponding rules and regulations. Dr. Kogan (now Dr Spectre) also has extensive experience with secure database storage and management, as well as all other procedures listed above.

Question 20: What arrangements are there for insurance and/or indemnity to meet the potential legal liability for harm to participants arising from the conduct of the study?

Notes: Insurance would normally be provided by the University's or Medical Research Council's insurance for persons employed by them or working in their institutions. Please contact the appropriate Insurance Office to arrange for insurance. If you do not have an appropriate institutional affiliation, say how you will provide public indemnity insurance, including insurance against non-negligent injury to participants. Evidence of insurance is required before a Letter of Approval can be issued.

Dr. Kogan (now Dr Spectre) has affiliation with the University of Cambridge and thus falls under the University's insurance.

Question 21: What arrangements are there for data security during and after the study?

Notes: Digital data stored on a computer requires compliance with the Data Protection Act; indicate if you have discussed this with your Departmental Data Protection Officer and describe any special circumstances that have been identified from that discussion. Say who will have access to participants' personal data during the study and for how long personal data will be stored or accessed after the study has ended.

We comply with the Data Protection Act. All data will be collected on a secure server. All data will be stored indefinitely on the server. The data are accessible by only research team members.

All prospective collaborators requesting access to our data will be asked to signed a disclaimer saying they will fully accord with the UK data protection act (Appendix G). To aid comprehension, this disclaimer also highlights many of its aspects that are most poignant to individual researchers and the present study, and provides a hyperlink to the full Act.

Appendix 3.1: 'R' Code for Study 1 and 2 Simulated Correlations

trueValues = Random normally distributed scores—of any size, mean and SD—generated using rnorm().

desiredR = Target correlation between trueValues and simulated variable.

noise.incr = Amount of noise—as proportion of true score SDs—to iteratively add to trueValues. Larger proportions decrease processing time. The default specified was sufficient to simulate accurate correlations to two decimal places (e.g. r = .31).

equal.sds = Should SDs be left as they are or corrected to reflect the shrinkage that tends to occur in predicted scores from real world machine learning models? If 'F', SDs are shrunk so that they are proportional to desiredR.

```
simCors <- function(trueValues, desiredR, noise.incr = .05, equal.sds = F){
 predictedValues <- trueValues   # duplicate true values
 # generate noise as a function of SD of trueValues
 noise <- rnorm(10000, 0, sd(trueValues)*noise.incr)
 # iteratively add noise to trueValues until desired correlation is reached
 for (z in 1:2000000){
  predictedValues <- predictedValues + sample(noise, length(predictedValues), replace = T)
  # peg min and max predicted values to min and max trueValues
  predictedValues[predictedValues > max(trueValues)] <- max(trueValues)
  predictedValues[predictedValues < min(trueValues)] <- min(trueValues)
  predictedValues <- round(predictedValues, 2)
  # find accuracy
  myR <- cor(trueValues, predictedValues)
  if (myR < desiredR) break
 }
  # adjust SDs
 if (equal.sds == T) {
  values <- predictedValues - mean(predictedValues)
  predictedValues <- values*SD(trueValues)/SD(predictedValues) + mean(trueValues)
 } else {
  predictedValues <- (predictedValues - mean(predictedValues))/sd(predictedValues)
  var <- sd(trueValues)*desiredR
  predictedValues <- predictedValues*var + mean(trueValues)
 }
 results <- as.data.frame(cbind(trueValues, predictedValues))
 names(results) <- c("true","predicted")
 return(results)
}
```

Appendix 3.2: Confusion Matrices to Evaluate Bias in Category Assignment

Here, I decompose correct classifications for predicted psychological characteristics at the three pre-defined accuracy benchmarks—personality (r = .30), demographic (r = .60) and best-case (r = .90)—when scores are bucketed into thirds (i.e. "low", "medium", "high"). This can be done with confusion matrices, which evaluate whether classifications are either correctly or incorrectly (true, false) bucketed into target or non-target categories (positive, negative). First, I generated a multigroup (i.e. > 2 categories) confusion matrix, where frequencies on the top-left to bottom-right diagonal were correct classifications. These are in Table A3.1. It confirmed that there was a higher rate of correct classifications as predictions moved towards Best-Case accuracy. Across all accuracies, cases were more likely to be misclassified into the adjacent category, rather than the opposite category. This trend became more pronounced as accuracy increased. Cases with true scores in the middle third had a roughly equal chance of being misclassified into top and bottom thirds, regardless of accuracy. Correct classifications were roughly equal for bottom vs top thirds at all three accuracies. Put together, results suggested that classifications were unbiased.

Table A3.1

Confusion matrices when bucketing true and predicted scores

Benchmark	Predicted Third	True Third		
		Low	Mid	High
Personality (r = .30)	Low	**1515**	1133	686
	Mid	1157	**1145**	1032
	High	663	1055	**1616**
Demographic (r = .60)	Low	**2032**	1016	287
	Mid	1018	**1363**	952
	High	285	954	**2094**
Best Case (r = .90)	Low	**2706**	614	13
	Mid	613	**2116**	604
	High	15	603	**2716**

Notes. Benchmarks were the correlation between true and predicted continuous variable scores. These scores were then both bucketed into thirds, with equal N. Frequency estimates are the average from 10 iterations of 10,000 simulated predicted scores at each benchmark. Bold frequencies reflect true classifications.

Aside from total accuracy rates (reported in-text), the key confusion matrix performance metrics are precision, sensitivity, specificity and false positive rate. Precision is the proportion of true positives to total positives. Recall is the proportion of true positives to combined true

positives and false negatives. Specificity is the proportion of true negatives to total negatives. Finally, false positive rate is the proportion of false positives to total negatives. They can all be calculated from two-by-two confusion matrices. Thus, I collapsed the multigroup confusion matrices by evaluating each category in a separate matrix, against the other two aggregated categories. An advantage of collapsing the matrices in this way was that I could compare the performance of specific categories to one another. Results are in Table A3.2.

Table A3.2

Confusion matrix accuracy metrics when bucketing true and predicted scores

Benchmark	True Third	Precision	Recall	Specificity	False Positives
Personality (r = .30)	Low	.45	.45	.73	.27
	Mid	.34	.34	.67	.33
	High	.48	.48	.74	.26
Demographic (r = .60)	Low	.61	.61	.80	.20
	Mid	.41	.41	.70	.30
	High	.63	.63	.81	.19
Best-case (r = 0.90)	Low	.81	.81	.91	.09
	Mid	.63	.63	.82	.18
	High	.81	.81	.91	.09

Notes. Benchmarks were the correlation between true and predicted continuous variable scores. The multiclass confusion matrices from Table A3.1 were collapsed into a series of two-by-two confusion matrices were each true third was iteratively the target (positive), and the other two thirds were aggregated together to form the non-target category (negative). Precision = true positives / total positives. Recall = true positives / (true positives + false negatives). Specificity = true negatives / total negatives. False positive = false positives / total negatives.

Results confirmed that classifications were unbiased. At each benchmark, precision was approximately equal for low and high thirds, and lower for the middle third. That suggested that the proportion of correct classifications was unrelated to whether true scores were in the low vs high third. In every case, recall also exactly matched precision. That suggested the extreme thirds were equally sensitive to catching every case belonging to an extreme category. Specificity and false positives were again almost exactly equal for both extreme categories, at each benchmark. This suggested that classifications were equally accurate when assigning non-target categories at low and high thirds. It was unsurprising that all accuracy metrics were lower for true scores in the middle third, across the benchmarks. This reflected results from Table A3.1, which suggested that classification errors were more likely to be in the adjacent rather than opposite category. That is, the high rate of middle third misclassifications can be explained

by it having two adjacent categories and no extreme categories. Finally, confusion matrix performance metrics improved exponentially as prediction accuracy progressed to the best-case benchmark. This could be explained by the concomitant non-linear increases in prediction R^2. Therefore, the confusion matrices suggested that predictions performed equally well for bottom and top third true scores, and the performed worse for middling scores. Finally, they also showed increased bucketing success rates as prediction accuracy increased.

Appendix 3.3: 'R' Code for Study 3 Simulated Correlations

true = Original scores for the target variable

predicted = Machine learning predicted scores for true, which are generally obtained with predict()

desiredR = Target correlation between true and predicted scores

noise.range = Amount of noise—as proportion of true score SDs—to iteratively subtract from predicted scores to reduce prediction error. Larger values speed processing time. The default specified was sufficient to simulate accurate correlations to two decimal places (e.g. r = .91).

equal.sds = Should predicted score SDs be left as they are or corrected to reflect the shrinkage that tends to occur in real world machine learning models? If 'F', SDs are *expanded* so that they are proportional to desiredR.

```
upCors <- function(true, predicted, desiredR, noise.range = seq(0,.01, .00001), equal.sds = F){
 cor <- cor(true, predicted)
 # find appropriate inflation factor for SDs
 sd.ratio <- sd(predicted)/sd(true) + (1-sd(predicted)/sd(true))*((desiredR - cor)/(1 - cor))
 # iteratively remove random portions of noise from predicted values
 for (z in 1:2000000){
  model <- lm(predicted ~ true)
  resid <- model$residuals
  noise <- sample(noise.range, length(predicted), replace = T)
  predicted <- predicted - (resid*noise) # correct predicted scores
  # peg min and max predicted values to min and max trueValues
  predicted[predicted > max(true)] <- max(true)
  predicted[predicted < min(true)] <- min(true)
  myR <- cor(true, predicted)
  if (myR > desiredR) break
 }
 # fix final predicted score values
 if (equal.sds == T) {
  # equalize means and SDs of true and predicted values
  values <- predicted - mean(predicted)
  predicted <- values*SD(true)/SD(predicted) + mean(true)
 } else {
  # equalize means and inflate SDs using sd.ratio
  predicted <- (predicted - mean(predicted))/sd(predicted)
  var <- sd(true)*sd.ratio
  predicted <- predicted*var + mean(true)
 }
 results <- data.frame(true, predicted)
 return(results)
}
```

Appendix 5.1: Comparison Table for Conventional and PSM Models

Table A5.1

Comparison table for zero-order, multiple regression and PSM model performances

		Zero-Order		Multiple Regression		PSM	
	Facet	Thwarting	Satisfaction	Thwarting	Satisfaction	Thwarting	Satisfaction
NUR	1	-0.1 (-0.19, -0.16)	0.24 (0.22, 0.26)	-0.01 (-0.03, 0.01)	0.04 (0.03, 0.06)	-0.07 (-0.09, -0.05)	0.11 (0.09, 0.13)
	2	-0.26 (-0.28, -0.25)	0.27 (0.25, 0.28)	-0.06 (-0.08, -0.04)	0.01 (-0.01, 0.04)	-0.16 (-0.18, -0.13)	0.09 (0.06, 0.11)
	3	-0.2 (-0.21, -0.18)	0.36 (0.34, 0.38)	0.03 (0.01, 0.05)	0.06 (0.04, 0.08)	-0.07 (-0.09, -0.04)	0.2 (0.17, 0.22)
	4	-0.27 (-0.29, -0.25)	0.21 (0.19, 0.23)	-0.05 (-0.07, -0.03)	-0.02 (-0.04, 0)	-0.12 (-0.15, -0.1)	0.03 (0.01, 0.05)
	5	0.11 (0.1, 0.13)	-0.2 (-0.22, -0.18)	-0.01 (-0.03, 0.01)	-0.05 (-0.07, -0.04)	0.05 (0.03, 0.07)	-0.11 (-0.13, -0.09)
	6	-0.12 (-0.13, -0.1)	0.24 (0.22, 0.26)	0.01 (-0.01, 0.03)	0.01 (-0.01, 0.03)	-0.02 (-0.05, 0)	0.09 (0.07, 0.12)
EXT	1	-0.36 (-0.38, -0.34)	0.52 (0.5, 0.53)	-0.04 (-0.06, -0.02)	0.17 (0.15, 0.19)	-0.19 (-0.21, -0.16)	0.32 (0.3, 0.35)
	2	-0.24 (-0.26, -0.22)	0.22 (0.2, 0.24)	-0.01 (-0.03, 0.01)	-0.03 (-0.05, -0.01)	-0.09 (-0.11, -0.07)	0.04 (0.02, 0.06)
	3	-0.28 (-0.29, -0.26)	0.32 (0.3, 0.34)	-0.03 (-0.05, -0.01)	0.04 (0.02, 0.06)	-0.11 (-0.13, -0.09)	0.12 (0.1, 0.14)
	4	-0.25 (-0.27, -0.23)	0.4 (0.38, 0.41)	0 (-0.02, 0.02)	0.06 (0.04, 0.08)	-0.08 (-0.1, -0.06)	0.18 (0.16, 0.2)
	5	-0.4 (-0.41, -0.38)	0.47 (0.45, 0.49)	-0.06 (-0.09, -0.04)	0.09 (0.07, 0.11)	-0.24 (-0.26, -0.22)	0.27 (0.25, 0.29)
	6	-0.29 (-0.31, -0.27)	0.21 (0.19, 0.23)	-0.04 (-0.06, -0.02)	-0.01 (-0.03, 0.01)	-0.14 (-0.16, -0.11)	0.06 (0.04, 0.08)
OPN	1	-0.33 (-0.35, -0.32)	0.39 (0.38, 0.41)	-0.02 (-0.04, 0.01)	0.03 (0, 0.05)	-0.2 (-0.22, -0.17)	0.23 (0.21, 0.25)
	2	-0.25 (-0.27, -0.23)	0.29 (0.27, 0.31)	-0.02 (-0.04, 0)	0 (-0.03, 0.02)	-0.16 (-0.19, -0.14)	0.17 (0.15, 0.19)
	3	-0.26 (-0.28, -0.25)	0.41 (0.39, 0.43)	0.01 (-0.01, 0.03)	0.07 (0.05, 0.09)	-0.09 (-0.11, -0.07)	0.19 (0.17, 0.21)
	4	-0.09 (-0.11, -0.07)	0.22 (0.2, 0.24)	0.02 (0, 0.03)	0.02 (0.01, 0.04)	0 (-0.02, 0.02)	0.06 (0.04, 0.08)
	5	-0.02 (-0.04, 0)	0.16 (0.14, 0.18)	0.04 (0.02, 0.06)	0.02 (0, 0.04)	0.01 (-0.02, 0.03)	0.07 (0.05, 0.09)
	6	-0.4 (-0.42, -0.39)	0.5 (0.48, 0.52)	-0.06 (-0.09, -0.04)	0.14 (0.11, 0.16)	-0.24 (-0.27, -0.22)	0.32 (0.3, 0.34)
AGR	1	0.42 (0.4, 0.43)	-0.27 (-0.29, -0.25)	0.12 (0.1, 0.14)	-0.01 (-0.03, 0.01)	0.23 (0.21, 0.25)	-0.11 (-0.13, -0.09)
	2	0.34 (0.32, 0.36)	-0.27 (-0.29, -0.26)	0.01 (-0.01, 0.03)	0 (-0.02, 0.02)	0.11 (0.09, 0.14)	-0.08 (-0.1, -0.06)
	3	0.52 (0.51, 0.54)	-0.45 (-0.46, -0.43)	0.26 (0.23, 0.28)	-0.1 (-0.12, -0.08)	0.37 (0.35, 0.39)	-0.26 (-0.28, -0.24)
	4	0.32 (0.3, 0.34)	-0.3 (-0.32, -0.28)	0.04 (0.02, 0.06)	0 (-0.02, 0.02)	0.12 (0.1, 0.14)	-0.1 (-0.12, -0.08)
	5	0.19 (0.17, 0.21)	-0.17 (-0.19, -0.15)	0.01 (-0.01, 0.02)	-0.02 (-0.04, 0)	0.04 (0.02, 0.07)	-0.04 (-0.06, -0.02)
	6	0.44 (0.43, 0.46)	-0.38 (-0.39, -0.36)	0.09 (0.06, 0.11)	-0.02 (-0.04, 0)	0.22 (0.2, 0.25)	-0.15 (-0.17, -0.13)
CON	1	0.07 (0.05, 0.09)	0.1 (0.08, 0.12)	0.03 (0.01, 0.04)	0.03 (0.01, 0.04)	0.05 (0.03, 0.07)	0.03 (0.01, 0.05)
	2	-0.11 (-0.13, -0.09)	0.28 (0.26, 0.3)	0.05 (0.03, 0.06)	0.03 (0.01, 0.05)	0 (-0.02, 0.02)	0.11 (0.09, 0.13)
	3	-0.1 (-0.12, -0.08)	0.26 (0.24, 0.28)	-0.01 (-0.03, 0.01)	0.05 (0.03, 0.07)	-0.04 (-0.07, -0.02)	0.12 (0.1, 0.14)
	4	-0.19 (-0.21, -0.17)	0.22 (0.2, 0.24)	0 (-0.02, 0.01)	-0.01 (-0.03, 0)	-0.06 (-0.08, -0.04)	0.07 (0.05, 0.09)
	5	-0.15 (-0.17, -0.13)	0.25 (0.24, 0.27)	0 (-0.02, 0.01)	0.03 (0.01, 0.05)	-0.04 (-0.06, -0.02)	0.09 (0.07, 0.11)
	6	0.02 (0, 0.04)	0 (-0.02, 0.02)	0 (-0.02, 0.02)	0.01 (-0.01, 0.02)	0 (-0.02, 0.02)	0.01 (-0.01, 0.03)
R^2		**7% (SD = 7%)**	**10% (SD = 7%)**	**1% (SD = < 1%)**	**< 1% (SD = < 1%)**	**2% (3%)**	**2% (2%)**

Notes. Values are beta coefficients and 99.9% CIs (in brackets) from linear regression models. For zero order and PSM, there were separate models for each facet, and for each outcome. The only difference was that PSM contained weights designed to equalize all covariate facet scores across every level of the target facet. For multiple regression, there was a single model including all facets, for each outcome. R^2 for zero order and PSM was the mean percentage of covariation between each facet and the outcome. R^2 for multiple regression was the percentage of covariation *unique* to each facet, and the outcome (i.e. the squared partial correlation); negligible values highlighted the extent of multicollinearity when using all 29 other facets as controls. Overall R^2 for multiple regression was 36% for needs thwarting and 41% for needs satisfaction.