THE UNIVERSITY
of EDINBURGH

Individual differences and their implications in educational and occupational outcomes

A thesis submitted to the University of Edinburgh for the degree of Doctor of Philosophy

Alicia Aranda Díaz

September 2021

Contents

List of Tables

List of Figures

Preface

Dedication and acknowledgments

Para ti, mamá.

(For you, mom)

Por cada sacrificio, por cada lágrima derramada en soledad, por todas las veces en los que te has caído – o te han hecho caer – y te has levantado más fuerte que nunca. Por todos los buenos momentos que nos has dado y todos los malos ratos de los que nos has protegido. Por todo lo que has hecho y todo lo que has dejado de hacer. Gracias por haberme ayudado a llegar hasta las cimas más altas. Una y otra vez, las hemos alcanzado todas. Nada de esto habría sido posible sin ti. A riesgo de sonar cursi, admito que eres mi mayor apoyo, mi mayor confidente, la persona a la que más admiro, la mejor madre del mundo. Supongo que estudiar es lo mío, y siempre he querido exprimirlo al máximo para demostrarte cuán agradecida estoy. Hoy hemos llegado al destino, mamá voy a ser Doctora.

I would also like to thank the experts that have accompanied me along the way: Dr. René Mottus, Dr. Tom Booth, Dr. Wendy Johnson, and Dr. M. Ángeles Quiroga. I also wish to thank the creators of the *bookdown* (Xie, 2021), R Markdown, and *knitr* (Xie, 2015) packages. I am immensely grateful to K.M., my supervisor in the shadows, for all the pieces of advice and all the times you've put up with me crying over the phone, for listening, and reassuring me when I needed it the most. My survival in this process would have not been possible with all my friends' support, especially the soon-to-be Dr. Brittany Blankinship, my brother, Manuel, and my very BFF, Nuria – I wish someday I become as inspirational as you guys are to me. I also owe a great deal to all the friends who have helped to organize my thoughts and to polish my chapters making this piece much more cohesive and articulate. Lastly, but most importantly, this thesis would have never been completed without the support of my partner, Fran, for not only giving me the courage to face my worse moments but for embracing and nurturing every single splinter on my bare-bones giving me the strength to grow bigger and greener.

Declaration of Authorship

1. I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

2. I confirm that this thesis presented for the degree of Doctor of Philosophy has

   i) been composed entirely by myself

   ii) been solely the result of my own work

   iii) not been submitted for any other degree or professional qualification

3. I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Alicia Aranda Díaz

Lay Summary

Concepts like "personality traits" or "cognitive ability" are often misused in pop-psychology articles. However, these constructs have been thoroughly studied for over a hundred years. Personality and Intelligence research are broadly referred to as "Individual Differences Psychology" because they represent the main aspects differentiating people. Both Personality and Intelligence have been proven to be powerful predictors of life outcomes like educational attainment, occupational status, mental and physical health, etc. This thesis summarizes some of this evidence and explores some new research avenues.

This thesis starts with an introductory chapter on Individuals Differences history and Intelligence research, followed by two chapters on intelligence and educational outcomes. Then, there is a chapter introducing Personality research, which aid contextualizing the next two chapters. Finally, the last chapter summarizes the research findings of these four studies and discusses four key themes present across them.

Chapter 1 introduces the concept and history of Individual Differences research, as well as the main milestones in Intelligence research history. Chapter 1 frames the next two chapters, exploring how verbal ability (i.e. an aspect of intelligence related to people's proficiency in spoken or written language) influences other life outcomes. Chapter 2 tackles the role of verbal ability in changes in student motivation when British students move from primary school into secondary school. Results did not show substantial change in student motivation but they reflected how intelligence might influence it. Chapter 3 tests whether changes in school grades affect occupational aspirations beyond the influence of verbal ability. Results showed that there was not mediation effects but that verbal ability predicted changes in occupational aspirations.

Chapter 4 summarizes the highlights of personality history and explains the approach to personality measures used in the following two chapters. Chapter 5 examines personality sex differences between four groups: heterosexual men, heterosexual women, homosexual men,

and homosexual women. It was found that, sometimes, homosexual women scored closer to heterosexual men than to heterosexual women. Conversely, homosexual men sometimes scored closer to heterosexual women than to heterosexual men. Chapter 6, studies if the personality sex differences influence the division of men and women in the labor market, especially as it relates to jobs related to Science, Technology, Engineering, and Mathematics (STEM). Results showed mixed evidence, meaning that personality sex differences might mirror personality differences depending on how occupation is defined. Lastly, Chapter 7 summarizes the main findings of these studies as well as their implications.

Technical Abstract

This thesis starts with an introductory chapter on Individuals Differences history and Intelligence research, followed by two chapters on intelligence and educational outcomes. Then, there is a chapter introducing Personality research, which aid contextualizing the next two chapters. Finally, the last chapter summarizes the research findings of these four studies and discusses four key themes present across them.

Chapter 1 introduces the concept and history of Individual Differences research. Chapter 1 also includes a summary of the milestones in Intelligence research history. Chapter 2 uses two waves from a British longitudinal study to explore the changes in student Self-Competence and Task-Value during the primary to secondary school transition according to students' verbal abilities (used as a proxy for cognitive ability) . In this study, change was conceptualized as Difference Scores and as Residuals. Although none offered a highly reliable measure of change, results show that verbal ability has a role in student motivation. The same sample is used in Chapter 3 to study the degree to which change in grades between ages 7 and 11 mediated the relation between verbal ability and changes in occupational aspirations between ages 11 and 14. The mediation path was not significant but verbal ability was relevant in predicting occupational aspirations change.

Chapter 4 summarizes the highlights of Personality research history and frames the methodological approach to personality measures used in the following two chapters. Chapter 5, uses Cohen's d and Mahalanobis D to study personality differences at facet and domain level between four groups: heterosexual men, heterosexual women, homosexual men, and homosexual women. Results show that homosexual participants did not always score as their heterosexual counterparts. Often, personality scores of homosexual participants shifted toward those of their opposite sex -e.g. homosexual men scoring more similar to heterosexual women than heterosexual men. In Chapter 6, I studied to which degree personality sex differences feed into the gendered nature of the labor market, this is, men tend to work in jobs related to Science, Technology, Engineering, and Mathematics (STEM), while women work

in Non-STEM related occupations. There was little evidence supporting this hypothesis. However, there was evidence suggesting that personality differences between occupational orientations (i.e. Prediger's dimension of people vs. things) do resemble personality sex differences. Lastly, Chapter 7 summarizes the key findings of these studies and their implications.

Chapter 1

A Brief Introduction to Individual Differences and Intelligence Research

1.1 Individual Differences

Psychology, as a research discipline, could be roughly divided into two methodological approaches: Experimental and Differential Psychology (Revelle et al., 2013):

> *"In contrast to experimental psychology, which seeks to discover general laws about human behavior, differential psychology seeks to understand how interindividual, intra-individual, and group differences in psychological characteristics interact with environmental affordances and demands, so as to produce differences in a variety of personal, occupational, educational, and social outcomes"*
>
> — Revelle and Bonaccio (2013, p. 187)

Differential Psychology has also been known as "psychology of Individual Differences" because it focuses on understanding how and why people differ. Two of the most prominent aspects distinguishing people are personality traits and cognitive abilities hence, Differential Psychology branches into Intelligence and Personality research. This thesis mirrors this division by including two introductions reviewing their individual historical trajectory followed by two studies per field.

Many would argue that Individual Differences research took off thanks to Sir Francis Galton's work around the mid-19 century. Galton, inspired by the work of his cousin Darwin, developed the idea that *talent* runs in families (Rigby, 2015; Spearman, 1904). He tried to test his hypothesis by setting, what he called, the *Anthropomorphic Laboratory*. Galton hoped to capture people's cognitive ability by developing a series of measures related to sensory perception, reaction time, "imagery ability", strength, etc. (Rigby, 2015). Altogether, Galton managed to record data from around 10 000 participants (who had even paid for the experience!). Although Galton's efforts did not allow him to formulate a theory on cognitive ability, he popularized the idea that individual differences were relevant and measurable. Moreover, he developed statistical tools that would later become the basis of current

Individual Differences research like the statistical regression and the bi-variate correlation (Revelle et al., 2013; Rigby, 2015).

Following Galton's contributions, the field of Individual Differences was largely shaped by three events: World War I and II, as well as the improvement of statistical tools. The World Wars imposed the need for a fast and reliable way of matching soldiers to positions of varying degrees of complexity. It is estimated that, only in World War II, around 1500 psychologists were involved in the US Army Air forces selection process. As it happened in other fields, both World Wars led to an explosive development in the methods for personality and cognitive assessment (Boake, 2002; Revelle et al., 2013; Vernon, 1947).

Another factor contributing to this rapid improvement was the development of dimensionality reduction tools like Cluster Analysis[1] and Factor Analysis[2] (FA) which allowed summarizing variables into fewer elements without losing excessive information, promoting the systematic study of specific cognitive abilities and personality traits. FA became especially popular towards the second half of the 20th century when computers became widely available. The use of FA guided researchers to conceptualize intelligence and personality as hierarchical structures. These models were debated but eventually became the most popular ones. Lately, the debate has reignited as new types of models are been proposed in both fields:

> *"Conceptually, networks are an attractive alternative for the dominant latent variable view on individual differences. With networks, it is not necessary to*

[1]Cluster analysis is a statistical tool that groups data according to several linear combinations of the measured variables. In each linear combination, variables are weighted according to their contribution.

[2]Factor Analysis (FA) developed by Spearman (1904), quickly became one of the most widely used in Individual Differences research (Revelle and Bonaccio, 2013). FA allowed grouping the common aspects that the observed variables shared (i.e. common variance), leaving behind their specificities (i.e. specific or unique variance) together with the error of measurement (Spearman, 1904). By doing this, FA unveiled the latent variables causing the observed variables (for reference see Figure 1.1). These latent causes could be grouped again into higher-order latent variables which, again, would represent the common variance that the variables in the previous level shared. Consequently, FA yield hierarchy-like models where observed measurement lay at the bottom of the hierarchy and latent constructs summarizing them occupied the upper levels. This makes FA computationally costly, so it was not widely used until the 1960s, coinciding with the increase in accessibility and power of computing hardware and software (Booth and Murray, 2018)

presume a common – but unidentified – cause of covariances between manifest variables"

— Van Der Maas et al. (2017, p. 14)

In response to these new models, experts in both fields have recently voiced the need to shift the focus from "big flashy theories" to enhancing and promoting research on more basic aspects of personality traits and intelligence (Deary et al., 2016; Deary and Sternberg, 2021; Eronen and Bringmann, 2021). In the end, this basic research is needed to shape better theories (Eronen and Bringmann, 2021). The implications of this debate around theories vs. individual constructs in both fields will be further discussed in Section 7.3.2.

Due to the strong connection between the development of Individual Differences research and the use of statistical tools, many have argued that constructs like personality traits and intelligence are "pure artifact" (Revelle and Bonaccio, 2013). However, it is important to remember that much like gravity, personality traits and intelligence are abstract concepts aimed to represent reality. All these constructs have been devised with the idea of simplifying nature so we can make predictions and exert stringent control over our surroundings (Revelle and Bonaccio, 2013). As it will be explored throughout this thesis, personality traits and intelligence have been proven very useful in predicting numerous life outcomes like educational attainment, occupational status, mental and physical wellbeing, etc. The remainder of this chapter is devoted to a brief review of Intelligence research history, which will be followed by an study examining the relationship between verbal ability and motivation changes through the primary to secondary school transition (Chapter 2). Following, Chapter 3 will explore the mediating role of grades between verbal ability and occupational aspirations in a cohort of young students. The next chapter, will summarize the history of Personality research (Chapter 4) so we can dive into the personality differences across sex and sexual orientation groups (Chapter 5), as well as the personality differences across occupations (Chapter 6). Lastly, Chapter 7, will review the results of the four studies together with some core themes present across them.

1.2 Intelligence

1.2.1 Intelligence structure and theories

I will now briefly summarize the main milestones across the history of Intelligence research. For those interested in dwelling on it, I would recommend Spearman (1904) for a detailed account of the first contributors and Revelle et al. (2013) for an overall review. Intelligence is defined as[3]:

> *"Cognitive abilities (or "mental abilities") are thus reflected as sources of variance in performance on a task which requires one to mentally process, comprehend, and manipulate information [...] Finally, a cognitive capacity can be said to be an ability (rather than a skill or knowledge), to the extent that it is relatively stable over time and predicts the acquisition of new skills or knowledge, but is itself relatively resistant to training or to explicit education"*
>
> — Revelle and Bonaccio (2013, p. 189)

Hence, there is a distinction between knowledge/skills and intelligence (or cognitive abilities). The former involves "familiarity gained from experience and association" providing information or proficiency at a task that can be used in certain scenarios (Revelle and Bonaccio, 2013, p. 189). The latter facilitates problem-solving regardless of the specificities of the situation (Johnson and Bouchard, 2005; Revelle and Bonaccio, 2013).

As previously mentioned, Galton was among the first authors to introduce the concept of "talent" which later derived onto mental or cognitive ability. Not long after, Ch. Spearman observed that tests measuring cognitive abilities were positively correlated, a phenomenon that he called *"positive manifold"*. According to Ch. Spearman, the positive manifold suggested that there was a common latent cause – a general factor of intelligence or g factor

[3]Although this definition might suggest that intelligence is highly genetically determined, evidence suggests that environmental factors can also influence an individual's intelligence.

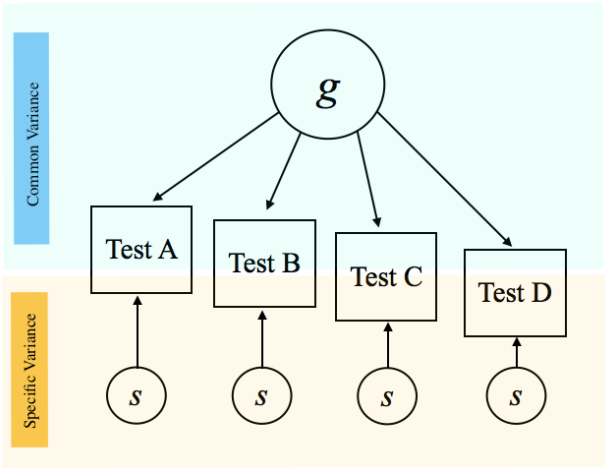

Figure 1.1
Spearman's model of intelligence.
Notes: Spearman's model of intelligence showing that the positive correlation among ability tests (i.e. the positive manifold) was determined by the degree of common information (i.e. common variance) that tests shared. According to Spearman, the only reasonable explanation was that all tests tapped onto the same latent construct – the general factor of intelligence or g Factor – along some abilities which were specific to each tests – the s Factor. The arrows reflect the directionality of causality. The strength of the positive manifold partly depended on the ratio on each test's common variance to specific variance (i.e. s). In this sense, correlation between Test A and B should higher than with Test D, for they are more loaded with common variance.

– tapped by all tests (Figure 1.1). According to Ch. Spearman, each test measured the *g* factor along with other specific abilities, the *s* factors. These specific abilities were precisely what differentiated tests:

> *"Whenever branches of intellectual activity are at all dissimilar, then their correlations with one another appear wholly due to their being all variously saturated with some common fundamental Function (or group of Functions)"*
>
> — Spearman (1904, p. 250)

Ch. Spearman argued that the strength of the correlation between tests partly depended on the degree to which they tapped onto the general factor compared to any specific ability. Spearman defined the *g* Factor as the "mental energy" or "mental power" responsible for (1) abstract ideas, (2) the ability to establish relationships between elements, and (3) the capacity to infer the consequences of those relationships (Spearman, 1904). With those three rules, Spearman devised a tool to measure *g* which was later developed by his student, J. Raven, into the famously known *Progressive Matrices* – i.e. a non-verbal test in which participants have to deduce the axioms ruling each matrix's design (Jensen, 1998). Spearman's idea of a general mental ability quickly popularized, and it still dominates intelligence research (Van Der Maas et al., 2017).

Other authors, like G. H. Thomson, advocated for alternative explanations for the positive manifold. As Thompson argued, the sets *a, b, c* and *c, d, e* would correlate for they share one of their components but, would not have a common latent cause. Similarly, a group of tests could correlate if they tapped onto the same independent abilities regardless of the existence of a general factor (Figure 1.2). Later in his career, Thomson nuanced his argument proposing that tests correlated to the degree they tapped on similar biological underpinnings or "bonds":

> *"There is not the slightest mathematical evidence so far forthcoming which will*

enable us to distinguish between overlapping Group Factors and a General Factor"

— Thomson (1916, p. 281)

Also, L. L. Thurstone refused the idea that there was such a general factor. He supported his arguments by factor analyzing data from 240 participants who had completed 56 different tests and found that tests were not grouped under a single factor but into several, depending on the specific ability they tapped onto (Thurstone, 1938). Thurstone called these specific abilities the *Primary Mental Abilities* (PMAs).

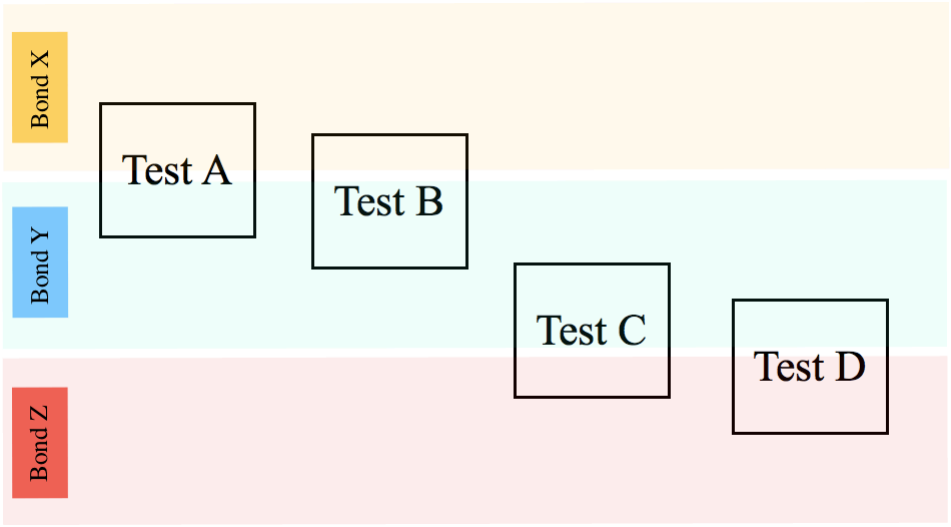

Figure 1.2
Thomson's model of intelligence.

Notes: Simplified diagram of Thomson's model of intelligence exemplifying that the positive correlation between ability tests is not necessarily due to a unique general factor. According to this model, the strength of tests' correlation would depend on the extent to which they tapped on the same "bonds". According to Thomson, all four tests on this diagram should be correlated because they all tap on Bond Y. However the pairs of tests A/B and C/D should be more highly correlated than any other pairs, for they tap on other similar bonds. Thomson was not certain on what "bonds" could be: mental abilities, neurological stratum, general speed of processing, etc.

By the mid 20th century, R. B. Cattell presented his G_f–G_c theory of intelligence. According to Cattell, cognitive abilities could be summarized by a general ability factor – the g factor – which could be roughly divided into *crystalized intelligence* (G_c) and *fluid intelligence* (G_f). The former was related to the declarative knowledge (e.g.: verbal ability, history knowledge, etc.) that people acquire through personal experiences, formal education, acculturation, etc. (Johnson and Bouchard, 2005; Horn and Cattell, 1966). Consequently, it was argued that G_c was largely influenced by socioeconomic status. It is important to highlight that G_c was not conceived to measure the degree to which people can study words or events but rather, their ability to infer them from spoken and written situations (Johnson and Bouchard, 2005). On the other hand, G_f allowed inductive reasoning and was thought to represent the genetically determined aspect of intelligence (Horn and Cattell, 1966; Johnson and Bouchard, 2005). Further studies showed that G_f and g correlated so strongly that they could be interchangeable. This has raised the question of whether the $G_f - G_c$ model makes a "useful theoretical contribution" (Johnson and Bouchard, 2005).

Cattell's ideas were further refined – and quite popularized – by his student, J. L. Horn (Johnson and Bouchard, 2005). Cattell-Horn model contains seven specific factors in addition to $G_f - G_c$. The main difference between Cattell's and Horn's models is that the latter did not acknowledge the existence of g. In this sense, Horn's model is reminiscent of Thurstone's PMAs model (Thurstone, 1938). Around the same time that Cattell was developing his $G_f - G_c$ model, P. E. Vernon proposed another prominent model. Vernon accepted the existence of a general factor explaining other abilities but emphasized that once the variance explained by this general factor was removed, tests could still be grouped in two categories which were further subdivided into many other "specialized factors" (Vernon, 1947). Thus, Vernon built a four-stratum model, meaning, a hierarchical model with four layers where the factors at each layer summarized some of the elements in the previous one. The observed variables (i.e. the tests) were at the base of Vernon's model. These were grouped into several specialized factors which, conversely, were subsumed into two latent factors: (1) *v:ed* (i.e. which was related to verbal, arithmetical, and educational abilities) and (2) *k:m*

(i.e. spatial, practical, and mechanical abilities). The g factor would sit at the top of the hierarchy, capturing both *v:ed* and *k:m* (Vernon, 1947).

At the end of the 20th century, J. B. Carroll took the grand endeavor of analyzing over 400 datasets containing information of various intelligence tests. His results converged in a three-stratum model headed by the g factor subsuming eight factors, each composed of other narrow abilities (Revelle and Bonaccio, 2013). Carroll's contribution has been compared to seminal works like Newton's *The Mathematical Principles of Natural Philosophy* (McGrew, 2009). It is important to highlight that Carroll used a tree-like diagram to represent his model for convenience but neither was he certain that intelligence actually followed a hierarchical organization, nor was he interested in unveiling which structure represented it best (Revelle and Bonaccio, 2013).

In 2005, W. Johnson and T. J. Bouchard used a sample that had completed 42 ability tests to compare the adequacy of three models: Vernon's model, the $G_f - G_c$ model, and Carroll's model. Their results showed that Vernon's model offered the best fit to the data yet, it could be improved. Johnson and Bouchard (2005) developed a model with three levels: g was at the top then, three intermediate factors (i.e. verbal, perceptual, and image rotation) summarized the eight factors sitting at the lowest level (Johnson and Bouchard, 2005). Johnson and Bouchard (2005) called it the verbal-perceptual-rotation (VPR) model, and it fitted the data better than the other three.

Soon afterward, McGrew (2009) observed that, although Cattell-Horn's model was an expanssion of Carroll's $G_f - G_c$ model, both remained notably similar. The main differences between these models laid in the existence of a general ability factor and the way some specific abilities were defined. Hence, McGrew (2009) merged the models proposed by Cattell, Horn, and Carroll creating the *CHC* model, which has become the most widely accepted model of intelligence (Figure 1.3). Like Carroll's model, the CHC was a three-stratum model. At the top of the hierarchy sat the g factor subsuming several specific abilities (including $G_f - G_c$), which were conversely summarizing several specific tests.

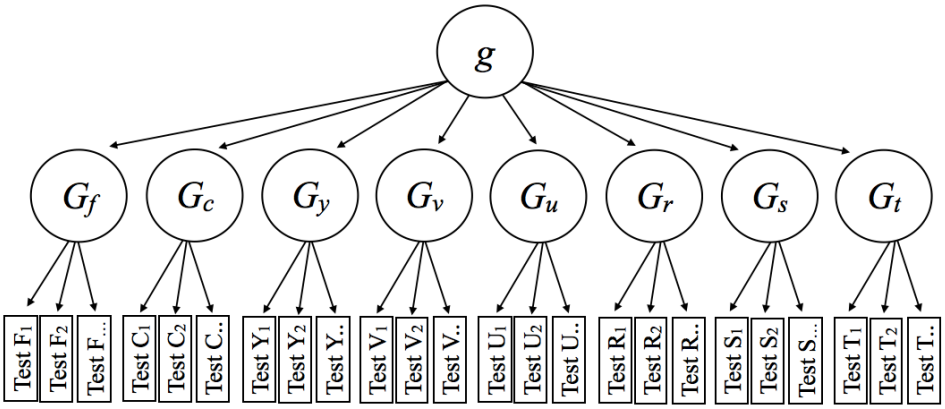

Figure 1.3
McGrew's Cattell-Horn-Carroll (CHC) model of intelligence.

Notes: McGrew's Cattell-Horn-Carroll (CHC) three-stratum model of intelligence. G_f= Fluid intelligence; G_c= crystalized intelligence; G_y= general memory and learning; G_v= broad visual perception; G_u= broad auditory perception; G_r= broad retrieval ability; G_s= broad cognitive speediness ; G_t= processing speed (reaction time).

Two new models have been recently proposed. Inspired by Thomson's bonds theory, Kovacs and Conway (2016) developed their *Process Overlap Theory*. According to Kovacs and Conway (2016), cognitive tests correlate because most require Executive Functions, which allow a top-down control promoting inhibition and goal-oriented behavior. In this sense, Kovacs and Conway (2016) argue that the strong correlation between g and G_f is a function of their overreliance on Executive Functions. In other words:

> *"The domain-general processes that are central to performance on cognitive tests are primarily the ones that are identified as executive processes in cognitive psychology in general […] Such processes are recruited by a large number of test items, alongside domain-specific processes, which are tapped by items appearing in specific types of tests only"*
>
> — Kovacs and Conway (2016, p. 161)

Although the Kovacs and Conway (2016) model has provided interesting insights, it has been heavily criticized for not being innovative enough and leaving several unresolved questions – e.g.: unknown but presumed sourced of general brain variance (for an extensive critique see Deary et al., 2016). Another recent model is the *unified model of general intelligence* of Van Der Maas et al. (2017), which combines a hierarchical organization headed by a general ability factor with a network model[4]. According to Van Der Maas et al. (2017) nodes are narrow abilities which, thanks to their connections, co-evolve. Additionally, each node in the network receives influence from the environment and genes.

Overall, the CHC model of intelligence is the most widely used but the development of new models shows that this debate is not settled (e.g.: Deary and Sternberg, 2021). We will discuss this debate's implications in Section 7.3.2.

1.2.2 Intelligence measurement

So far, we have focused on the various theoretical approaches to intelligence. In this section, I will explore how robust is intelligence as a construct and why it is relevant in applied contexts like education.

Overall, intelligence researchers have developed robust g measures. In fact, g measures obtained through different test batteries[5] are roughly identical (they have been found to correlate .998!), as long as there is no range restriction in the sample[6] (Jensen, 1998; Johnson et al., 2004). This is still the case even when batteries do not contain exactly the same tests, although this might lower the correlation between g measurements (Thorndike, 1987). Furthermore, the g factor is considered a robust construct for being ubiquitous across sex,

[4]Network models work under the assumption that phenomena are organized into individual components, "nodes", which are connected to the rest according to the degree to which they depend on them. In this sense, any change in a single node would ripple and induce changes in the rest of them.

[5]A test battery is a collection of tests tapping on different abilities which are often administered together to have a more comprehensive measure of g.

[6]In this context, range restriction refers to samples in which participant selection is biased in terms of age, sex, socio-economic status, or any other variable that might reduce the heterogeneity among the participants involved.

racial, and ethnic groups (Jensen, 1998). This means that Intelligence researchers have a clear idea of what intelligence measures overall:

> *"Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings- 'catching on', 'making sense' of things, or 'figuring out' what to do"*
>
> — Gottfredson (1997, p. 14)

Although there is a clear idea of what intelligence measures, its structural organization is still debated. Further neurological research is needed to determine the biological underpinnings of *g* and its structure (Johnson and Bouchard, 2005; Kovacs and Conway, 2016; Revelle and Bonaccio, 2013; Van Der Maas et al., 2017):

> *"[W]e do not know how* g *results in intelligence test scores, how* g *operates in the brain, or how* g *develops"*
>
> —Van Der Maas et al. (2017, p. 6)
>
> *"[T]he effort to understand the psychobiology of intelligence has a resemblance with digging the tunnel between England and France: We hope, with workers on both sides having a good sense of direction, that we can meet and marry brain biology and cognitive differences"*
>
> — Deary et al. (2016, p. 195)

As we will see in the following chapters, Intelligence has been widely used because it is more robust and has a superior predictive power than other psychological constructs when

predicting mental health, aging, life expectancy (Deary et al., 2010; Jensen, 1998), educational attainment, occupational achievement (Deary et al., 2007; Hunt, 2010; Jensen, 1998; McCabe et al., 2020; Vernon, 1947), as well as prosocial and criminal behavior (Hunt, 2010; Jensen, 1998). Nevertheless, for reasons we will now explain, the construct of intelligence became unpopular in Educational research and both fields have become estranged.

1.3 Intelligence and education

The role of psychologists and, more specifically, of intelligence testing in the educational sphere was not contemplated until the end of the French revolution. In 1882, the French government passed a law requiring all children between ages 6 and 13 to go to school except if they were blind or deaf. The psychiatrist D. M. Bourneville had already been working on a method to detect otherwise healthy children who were still likely to encounter difficulties in school. According to D. M. Bourneville, these children could be hospitalized and educated in the same facilities (Nicolas et al., 2013). Psychologist A. Binet proposed a competing method requiring psychologists and educators to collaborate within each school so children would not need to be institutionalized. With this idea, Binet aimed to revindicate the role of psychologists in the educational system.

The French government invited several experts including Binet and Bourneville, to research what would be the best resolution to this issue. The final report did not involve any form of psychological testing but Binet did not give up (Nicolas et al., 2013). Soon later, Binet and his student T. Simon proposed another method to categorize these children through a medical, pedagogical, and psychological examination. Their proposal included an intelligence test, the *Binet-Simon Scale*, which classified students according to their cognitive ability (Binet and Simon, 1916). However, they had to provide much evidence on the benefits of their approach before their proposal was seriously considered by the administration. The *Binet-Simon Scale* became the foundation of current intelligence testing and even some of the original items are still used (Boake, 2002; Nicolas et al., 2013).

By the beginning of the 20th century, the Binet-Simon Scale was brought to the US to help accommodate the increasing number of immigrant children into the educational system. Intelligence testing quickly became an important tool in the educational system (Fass, 1980). L. Terman and other US researchers used the Binet-Simon Scale to develop questionnaires suited for large-scale testing (Boake, 2002; Fass, 1980; Terman, 1916). These large data sets prompted researchers to study intelligence differences across sex and races (Fass, 1980). Soon, there was a sound body of evidence using different intelligence measures and populations. In 1916, L. M. Terman published a book chapter exposing that US educational policies aimed at ensuring that all children performed equally well by buffering socio-economic inequalities failed because their working premise that all children were equally endowed was wrong. Terman (1916) argued that there was no scientific evidence showing that such policies could narrow the achievement gap between certain groups and, what was more, some of these ability differences seemed to unavoidably limit students' potential for life.

> *"Instead of wasting energy in the vain attempt to hold mentally slow and defective children up to a level of progress which is normal to the average child, it will be wiser to take account of the inequalities of children in original endowment and to differentiate the course of study in such a way that each child will be allowed to progress at the rate which is normal to him, whether that rate be rapid or slow [...] Failure crushes self-confidence and destroys the spirit of work. It is a sad fact that a large proportion of children in the schools are acquiring the habit of failure. The remedy, of course, is to measure out the work for each child in proportion to his mental ability."*
>
> — Terman (1916, p. 3)

Many other studies followed, consistently showing that (1) intelligence was a strong predictor of academic achievement and occupational success –in terms of high performance and suitability to perform complex jobs– and that (2) intelligence scores seemed to differ across sexes and races. This studies had a profound impact on the scientific community and the

society a as a whole but to understand why, it is important to frame this discussion in its socio-cultural context.

First, it is relevant to highlight how deeply rooted in American culture is individual and market freedom. Around the 1950s the American economy largely relied in the *laissez faire* policies[7] embodied by figures like Milton Freedman. These economic policies created a very competitive market with a broad job offer. These conditions led to the concept of the "American dream", which inspired individuals to "own their future and succeed". In terms of the organization of the American society, in 1896 "separate but equal" policies divided the population by race[8] across all age groups and settings (schools, busses, bathrooms, etc.). These and other policies remained in place until the second half of the 20th century, following the pressures of the Civil Rights movement led, among others, by M. Luther King who was assassinated in 1968. Race segregation was not unique of the US as in South Africa and Namibia, there was a similar system of segregation, the *Apartheid*, from 1948 to 1992. Also, in Germany and other parts of Europe, A. Hitler led a campaign to separate and control the Jews population from 1933 to 1945. Altogether, these events evidence the strong hierarchical division of the population throughout the 20th century, especially as it regarded race.

Considering the wider socio-cultural context in which Terman had made such claims, it seems evident now how these statements opposed American's core values and how that could spark a social crisis. On the one hand, this research seemed to have the power to remove all hope people had on improving their social status. On the other hand, and given the worldwide social scene, this situation led to a legitimate fear that this research would be used to further segregate and oppress certain groups. These tensions cooled down through the second half of the 20th century, but reminiscences of these controversies still linger. These

[7]The term *laissez faire* was coined by the French economist Vincent de Gournay in the 18th century to refer to economical policies preventing the government from interfering in the market.

[8]According to UNESCO (1950), "the term race designates a group or population characterized by some concentrations, relative as to frequency and distribution of hereditary particles (genes) or physical characters, which appear, fluctuate, and often disappear in the course of time by reasons of geographic and or cultural isolation." (p. 35). Note that the biological underpinings of race are still questioned (e.g.: Yudell et al., 2016).

debates permeated many areas in which measures of intelligence were being used. In 1950, UNESCO published a document discouraging the use of intelligence and personality research to categorize individuals, as well as the use of such research to inform educational policies. Among other reasons, the UNESCO report emphasized the impossibility of determining the degree to which genetic and environmental factors could influence individual differences. This report was one of many events that lead to the division of research on intelligence and education.

The debate over these research controversies has cooled down but it has never extinguished. Intelligence research still carries the burden of the social trauma that certain individuals feed. Until these social tensions are dissolved either through an overt confrontation or an indirect by-pass, it is unlikely that Intelligence research is used systematically in certain applied fields like education. However, intelligence research has proven to be essential to understanding educational achievement hence, efforts to foster this partnership should be made. The following two chapters propose a humble contribution to the reunification of these fields by using classic educational theories and complimenting them with intelligence measures.

1.4 Conclusion

Experts have theorized about intelligence for more than a century and, although researchers agree on its overall measure, its structure is still debated. Further advances in neurological research will help to define intelligence beyond its measures and its predictions.

Nevertheless, Intelligence is a robust psychological construct with a notable predictive power especially as it regards to educational and occupational outcomes. As verbal abilities have been a core construct across intelligence models we will use this measures as a proxy for Intelligence in the next two chapters. Nonetheless, future research should expand these studies exploring whether other specific cognitive competencies (e.g.: problem-solving competencies)

could have a more relevant role when predicting motivation changes in certain subjects or career aspirations.

Efforts to merge Intelligence and Educational research are crucial to develop comprehensive models of student learning. However due to various political controversies intelligence has often been ignored by Educational research. The next two chapters aim to bridge the gap between Intelligence and Educational research by showing how classical theories used in the educational sphere can be complimented with intelligence measures.

Chapter 2

The Influence Of Verbal Ability On Student Motivation Change After The Primary To Secondary School Transition

2.1 Introduction

Motivation is often conceptualized as a driving force pushing individuals to keep engaged in an activity. However, motivation cannot act in a vacuum, individuals must have certain abilities. Hence, academic achievement is likely to depend on aptitude as much as attitude.

Eccles's expectancy-value model has been so widely used to understand students' motivation that it has been part of the Program for International Student Assessment (PISA) since 2006 (OECD, 2007). Eccles's model considers the influence of sex and socio-economic status on student's motivation but it does not include students' actual cognitive abilities. The present study aims to examine changes in students' motivation during the primary to secondary school transition using an extended version of Eccles's expectancy-value model considering students' verbal ability as a proxy for cognitive ability.

Note that the the direction of causality between the different elements of Eccles's model is debatable and, although in this study we will assume that the effects ripple from socio-economic status and intelligence to motivation measures, this will be further discussed in Section 7.3.2.

2.1.1 Motivation, Cognitive Ability, and educational achievement

In Eccles's model, students' motivation is understood as a combination of perceptions of their abilities (i.e. self-competence) and the value they place on their education (i.e. task-value). According to Eccles (1983) Task-value depends on: (1) attainment value (i.e. caring of doing well or fulfillment), (2) intrinsic interest (i.e. enjoyment of the activity), and (3) utility value (i.e. the usefulness of a specific subject in the long run). It has been suggested that self-competence/task-value are subject-specific and scores should not be aggregated, however, there is not a consensus. The reader might refer to Table 2.1 for some examples on how these constructs are commonly measured.

Self-competence has been repeatedly found to depend on students' feedback (e.g.: grades)

and classroom experiences (Bachman and O'Malley, 1986; Marsh, 1984, 1987; Marsh et al., 1995; Marsh and Hau, 2003). Additionally, there is a strong correlation between academic achievement and cognitive ability (Deary et al., 2007; Roth et al., 2015). Chamorro-Premuzic et al. (2010) found that the relationship between academic achievement and student self-competence was reduced but still significant after adjusting for cognitive ability, and that self-competence and cognitive ability correlated between .21 and .35. Chamorro-Premuzic et al. (2010) results suggest that more cognitively able students probably experience success more often, which may prompt them to have higher levels of self-competence; nonetheless, self-competence predicts academic achievement beyond cognitive ability.

Task-value and self-competence are related (rs_{range} =.56 to .74) (Spinath et al., 2006). This means that students believing in their ability to succeed in a subject are more likely to value it. Additionally, Johnson et al. (2007) showed that school motivation (mainly measured as task-value) and cognitive ability were correlated (rs_{range} =.13 to .21); and they both were independent predictors of academic achievement. Similarly, Spinath et al. (2006) found correlations between cognitive ability and students' Mathematics, English, and Sciences self-competence (rs_{range} =.18 to .22) and task-value (rs_{range} =.08 to.13). These results were further replicated by Andersen and Cross (2014), who found that gifted children tend to show higher self-competence and task-value in both Mathematics and Science.

In summary, in Eccles's model academic achievement is a byproduct of a student's motivation (i.e. self-competence and task-value). However, Eccles's model does not consider how students' actual cognitive ability might interact with self-competence and task-value. Research has shown that these three elements are related, so it seems reasonable to assume they should be all part of the same framework.

Table 2.1
Sample Likert-Scale items used in literature to measure Self-competence and Task-value.

Authors	Self-competence	Task-value
Else-Quest, Linn and Hyde (2010)	"I learn Mathematics quickly"	"I am interested in the things I learn in Mathematics"
	Total number of items: 5	Total number of items: 4
Felson, R. B., and Reed (1986)	"How do you rate your-self in school ability compared to those in your grade at school?"	
	Total number of items: 1	
Fischer (2013)	"I always manage to solve difficult problems if I try hard enough"	
	Total number of items: 10	
Gaspard at al. (2017)		"I like doing . . ." (Students were instructed to insert the respective subject into the blank)
Guo, Marsh, Morin, Parker, and Kaur (2015)	"How do you rate in school ability com- pared to others?"	"How interesting are most of your courses to you?" "Are you studying hard to get good grades in school?'
	Total number of items: 3	
Jacobs et al. (2002)	Self percception of their abilities in each subject (five items per subject)	How fun/important/ useful each subject was
	Total number of items: 5	Total number of items: 3

Authors	Self-competence	Task-value
Johnson, McGue, and Iacono (2007)		Items included studying without being reminded, interest in school work, turning in homework, wanting good grades, and enjoying attending school, rated on a 4-point scale.
Spinath et al. (2006)	How good they were in specific activities in Mathematics/English/ Science	How much they liked each of the three different activities for English, Mathematics, and Science
	Total number of items: 9	Total number of items: 9
Wach et al (2015)	how well they thought they are in three different activities in Math (several items per subject with specific examples)	

2.1.2 Motivation, Cognitive Ability, and Sex

Sex has been found to be a relevant factor predicting cognitive ability and motivation. Sex differences at a cognitive ability latent factor are negligible (Spinath et al., 2008; Wach et al., 2015). However, small differences can be found in spatial ability in which boys override girls (Hyde, 2005) or verbal ability, in which the opposite is true (Steinmayr and Spinath, 2008; Spinath et al., 2014). These differences have been related to PISA sex gaps in reading and mathematics across the world (Machin and Pekkarinen, 2008), especially in more gender-equal countries (Stoet and Geary, 2013, 2018, 2020).

Some research shows that girls have higher self-competence in Languages over Mathematics,

whilst the opposite is true for boys (Eccles, 1983; Steinmayr and Spinath, 2008; Spinath et al., 2008; Wach et al., 2015; Watt et al., 2012), even in cases where there are no achievement differences (Else-Quest et al., 2010; Wach et al., 2015). Steinmayr and Spinath (2008) results suggest that sex differences in task-value seem to mimic those in self-competence. Nonetheless, not all studies support these sex differences.

In Spinath et al. (2010) there were no sex differences in either, self-competence or task-value. On the other hand, Fischer et al. (2013) suggested that boys define themselves as being more competent in all subjects, but girls have higher intrinsic motivation – i.e. task-value. Preckel et al. (2008) found that in groups without sex differences in Mathematics grades, there was a significant difference in self-competence and intrinsic interest (i.e., task-value), between high and average ability boys but not between high and average ability girls. These results suggest that other factors might prompt students to under- or over-estimate their abilities.

2.1.3 Motivation, Cognitive Ability, Sex, and Age

As students move forward in the educational system, they face more difficult challenges, they meet a wider variety of students, and they accumulate more feedback experiences. This dynamic has been associated with self-competence and task-value decline through primary and secondary school (Eccles, 1983; Jacobs et al., 2002; Stipek and Iver, 1989). The pattern of motivational changes across the school system becomes increasingly complex as models incorporate factors like sex, school year, or school subjects.

Some studies point towards a sex and subject-specific decline (e.g.: Gaspard et al., 2017; Jacobs et al., 2002; Watt et al., 2012) but not all (e.g.: Guo et al., 2015). Self-competence and task-value seem to decline slower for languages than mathematics (Jacobs et al., 2002). Some research suggests that task-values decline faster for boys in languages (Gaspard et al., 2017), but not in other subjects (Gaspard et al., 2017; Jacobs et al., 2002). Sex differences in change have also been found to vary across countries depending on the weight schools give to each subject (Watt et al., 2012). Other studies show self-competence being more stable

than task-value in late secondary school (e.g.: Marsh, 1984).

Most research converges showing a decline in self-competence and task-value across time. Changes in motivation would be expected to be more pronounced after school transitions when there are larger changes in school requirements, and less so in the latest years of formal education, when students have accumulated experiences informing them about their school performance (Hegna, 2014). Nonetheless, recent evidence has challenged these predictions: Chamorro-Premuzic et al. (2010) found self-competence to be quite stable during the primary to secondary school transition.

2.1.4 Motivation and Socio-Economic Status

Not many studies test the relationship between SES and self-competence/task-value. However, it is important to consider SES as a control variable, for research shows how it can lead to a cascade of effects influencing cognitive ability and academic achievement.

On the one hand, SES has been directly related to academic achievement (rs_{range} =.22 to .34) (Sirin, 2005; White, 1982). SES can influence academic achievement and motivation through its effects on the students' environment. High SES families are able to provide children with broader cultural knowledge, making school more accessible and attractive (Bourdieu and Passeron, 1977; Ditton et al., 2019). Indeed, teenagers of higher SES families have been found to have more school motivation (Bachman and O'Malley, 1986; Felson and Reed, 1986; Johnson et al., 2007; Marsh, 1987, 1984).

SES can also influence academic achievement through its effects on cognitive ability. Some research suggests that SES confounds the heredability or expression of cognitive ability so that the effects of the school environment (e.g.: teachers, school materials, etc.) have more impact on the cognitive development of low than high-SES children (Giangrande and Turkheimer, 2021; Turkheimer and Harden, 2013). Still, von Stumm (2017) showed that high SES predicted a higher academic achievement across primary and secondary school

even when controlling for cognitive ability.

2.1.5 Summary and Predictions

Being able to succeed in school does not imply doing it. There is much research predicting academic achievement from students' cognitive ability or motivational factors, but few have studied the relationship between these two. Sex and SES are also relevant factors contributing to motivation.

This study aims to study self-competence and task-value change after the primary to secondary school transition. For this purpose, self-competence and task-value change will be predicted separately from the interactions between (1) sex and either self-competence or task-value scores in primary school, and between (2) verbal ability (used as a proxy of cognitive ability) and either self-competence or task-value scores in primary school, controlling for sex, SES, and verbal ability (all three measured in primary school). We predict that:

- *Prediction 1 (P1):* Self-competence and Task-value will decline in the primary to secondary school transition.

- *Prediction 2 (P2):* Given previous literature, there might be a significant sex and Self-competence interaction when predicting change. Similarly, there will be a significant sex and Task-value interaction when predicting change.

- *Prediction 3 (P3):* SES will be positively related to Self-competence and Task-value and inversely related to Self-competence and Task-value decline so that high-SES students experience less decline in Self-competence and Task-value after the school transition.

- *Prediction 4 (P4):* Verbal Ability scores (used as a proxy of cognitive ability) will be positively related to Self-competence/Task-value in primary school.

- *Prediction 5 (P5):* Students with higher verbal ability will keep more motivated than less able students after the primary to secondary school transition. In other words, there will be an interaction between Verbal Ability and Self-competence/Task-value in primary school when predicting Self-competence/Task-value change.

2.2 Method

2.2.1 Sample

The Millennium Cohort Study (MCS) is a longitudinal study of children born in the United Kingdom between 2000 and 2001. The sample was stratified by within-nation electoral district clusters to ensure sample representativeness (Plewis et al., 2007). The MCS is not only superior to other longitudinal data bases for its size and representiveness, it is also one of the most recent cohort studies following children from their birth in the year 2001 to their teenage years. Hence, the MCS was preferred to other less comprehensive (e.g.: Education Longitudinal Study of 2002), more outdated (e.g. 1958 National Child Development Study or the 1970 British Cohort Study) or smaller (e.g. 2004 Growing-up in Scotland) longitudinal studies.

Participants in the MCS were recruited so that each electoral district cluster provided the same proportion of its families with children of the targeted ages. Disadvantaged and ethnic minorities, as well as families from Scotland and Northern Ireland, were oversampled to ensure population representativeness.

Families were found through the Child Benefit records (universally provided in the UK) provided by the Department of Social Security (later called Department for Work and Pensions; DSS/DWP). Sensitive cases were withdrawn by the DSS/DWP; these included families (a) whose children recently died, taken into care or under another person's ward, (b) being investigated for benefit fraud, or (c) who had completed the Survey of Low Income Families (Plewis et al., 2007). Health visitors that operate at a local level collaborated in contacting

Table 2.2
Sample in wave in each of the waves in the MCS.

Primary School (Age 11)	Secondary School (Age 14)
N= 13 287 (69% of the original sample) Mage=11.3 SDage = .50 Female=49.5%	N= 11 726 (61% of the original sample) Mage = 14.3 SDage = .50 Female=49.8%

families that had recently moved to different electoral district clusters and hence, were not included in the Child Benefit records (N=56). Families were contacted through letters and had the opportunity to opt out. The total sample size across countries was between 87-113% of the targeted sample.

Families were interviewed for the first time when cohort members were 9 months old and have been assessed every 2-3 years since then. This study used data from the fifth and sixth waves when cohort members were 11 and 14 years old respectively (Table 2.2) to study the motivation changes during the primary to secondary school transition. Note that in Scotland, children are allowed to start primary school at the age of 5 hence, children aged 11 might have already transitioned to secondary school. Additionally, students in the UK can skip a school year for other reasons (e.g. considerably high ability). In the MCS, 3.7% of the respondents aged 11 were already in secondary school. The following analyses were run with and without this subset of students but, results remained largely unchanged. This Chapter presents the results with the full sample, as students still faced more difficult challenges as they progressed in the educational system and their motivation was expected to change in the same direction.

There was 8% attrition between ages 11 and 14. Families that dropped out between these two waves had significantly lower initial equivalized income (d=.37); deprivation index χ^2(9, N=12783)= 99.89, $p < .00$, ϕ= 0.03; education χ^2 (8, N=11587)= 180.90, $p < .00$, ϕ= 0.04 ; and occupation χ^2 (4, N=10242)= 116.68, $p < .00$, ϕ= 0.05. Additionally, children who

Table 2.3
Comparison between percentage of workers per occupation group in UK and MCS sample.

Occupation Group	Percentage of workers in the UK census	Percentage of MCS total respondents	Percentage of MCS total respondents with reported occupation
White Collar (e.g.: senior officers, scientists)	29.6	28.8	42.0
Assotiative professionals, technical and administrative occupations (e.g.: IT technicians, secreatries, etc.)	25.0	10.6	15.4
Small employers and skilled technical occupations (e.g.: farmers, electricians, etc.)	10.8	9.5	13.9
Caring, leisure, and cotumer service occupations (e.g.: nursery staff, travel agents, salespersons, etc.)	17.3	4.3	6.3
Semi-routine and routine occupations (e.g.: plant machine operatives, road transport drivers, etc.)	17,4	15.3	22.4

dropped out had significantly higher Self-competence (d=.08) and Task-Value (d=.11) at age 11. All differences were very small to small according to Funder and Ozer (2019). Table 2.3 shows the percentage of MCS respondents per occupation group compared to the data from the UK census.

2.2.2 Materials

In this study, we used data from cohort members at ages 11 and 14 about their sex, verbal abilities, self-competence, task-value, and parental SES. Tables 2.4 and 2.5 contain these variables' descriptives.

Verbal abilities. Participants completed the Verbal Similarities subtest from the British

Ability Scales, Second Edition (BAS 2) at age 11 (Elliott, 1986). In this task, respondents are given three nouns and asked to say what they have in common (e.g.: For the list: apple, banana, and pear, the response considered correct is "fruits").

Student motivation. To measure self-competence, cohort members rated how well they agreed with the statements at ages 11 and 14: *"I am good in English", "I am good in Maths",* and *"I am good in Sciences"* on a four-point Likert scale. To measure task-value, cohort members rated how often they *"try their best in school" , "find school interesting",* and *"feel like school is a waste of time"* on a four-point Likert scale at ages 11 and 14. We reversed the last item so that a high rating meant high motivation. We aggregated Self-competence and Task-Value item scores by averaging their scores within each wave. Note that Items used to measure self-competence and task-value are not standardized but, as shown in Table 2.1, tend to be similar across studies. There is no standard number of items, but it is usually low, between 1 and 10.

Family Socio-economic Status (SES). SES was computed as the sum score of several variables: the average of parents' education, highest parental occupation, OECD Equivalised Income, and official Deprivation Index of the family's residential area, all at age 11. More detailed information follows below but it is important to note that: (1) parents' education and occupation were computed using data from the main respondent (99.1% of the cases a biological parent) and his/her partner (87.8% of the cases was a biological parent and 10.2% the main respondent's current partner); and (2) the OECD Equivalised Income and the official Deprivation Index were computed for the whole household taking into account parents' data in addition to other sources of information.

Table 2.4
Frequencies and descriptive statistics of variables at Age 11.

Variable	N	Percent of the total sample	M	SD
Sex	13287	100.0		
Girls	6575	49.5	N/A	N/A
Boys	6712	50.5	N/A	N/A
Verbal ability	12994	97.8	58.7	10.1
OECD Equivalised weekly Income (GPB, £)	13287	100.0	404.94	177.05
Parent's Average Education	11982	90.2	3.1	1.1
NVQ Level 1	643	5.4		
Level 1,5	314	2.6		
NVQ Level 2	2588	21.6		
Level 2,5	895	7.5		
NVQ Level 3	2158	18.0		
Level 3,5	995	8.3		
NVQ Level 4	2948	24.6		
Level 4,5	826	6.9		
NVQ Level 5	615	5.1		
Parent's Highest Occupation	10567	79.5	3.8	1.5
NS-SEC re-coded Level 1	1847	17.5		
NS-SEC re-coded Level 2	560	5.3		
NS-SEC re-coded Level 3	1243	11.8		
NS-SEC re-coded Level 4	1580	15.0		
NS-SEC re-coded Level 5	5337	50.5		

Variable	N	Percent of the total sample	M	SD
Residential Area's Deprivation Index by Decline Decile	13280	90.2	5	3
<10%	1674	12.6		
10-20%	1621	12.2		
20-30%	1412	10.6		
30-40%	1255	9.5		
40-50%	1235	9.3		
50-60%	1226	9.2		
60-70%	1122	8.4		
70-80%	1106	8.3		
80-90%	1279	9.6		
>90%	1350	10.2		

Notes. Verbal ability was scored between 20 and 80, NS-SEC levels were re-coded so that *Level 1* that usually corresponds to White Collar occupations, represented Semi-Routine or Routine Occupations.

- *Average of parents' education.* Parents' educational level was measured following the UK National Vocation Qualification (NVQ) scheme has five levels. NVQ Level 1 corresponds to GCSE level and Level 5 requires a Postgraduate Qualification (Government Digital Service, 2020). The MCS dataset included two extra categories to account for "overseas qualifications" and "others". We treated these as missing. We used the mean of the two reporting parental figures when both were available and just the one when not.

- *Highest parent's occupation (measured at age 11).* Parents' occupations were coded

using the five categories of the UK *National Statistics Socioeconomic Classification*, NS-SEC (Office for National Statistics, 2021). The NS-SEC was re-coded so that Level 5 represented *higher managerial, administrative, and professional occupations* and Level 1 *semi-routine and routine occupations*. The highest occupation was retained for analysis (e.g.: Krieger et al., 1999).

- *OECD equivalised weekly average 'per-person' income (measured at age 11).* This was total family after-tax income divided by the number of "standard adults" in the household (Office for National Statistics, 2015). Adults are weighted 1 and children are given specific lower weights according to their ages.

- *Deprivation index of the family's residential area (measured at age 11).* These were the various official deprivation domains that each nation within the UK separately establishes and thus was not a homogeneously measured scale. All 4 nations in the UK included similar aspects of life quality such as income, employment, health deprivation, and disability, education barriers to housing and services, as well as crime and living environment (Noble et al., 2006). The overall deprivation index is a categorical variable divided into deciles of deprivation where decile 10% (Level 1) is the most deprived.

2.3 Analyses

2.3.1 Variables and scoring

Our independent variables were sex, SES, Verbal Ability, Self-competence, and Task-value scores (all at age 11). For both Self-competence and Task-value, sum scores were created across items measuring each construct. We standardized all variables except for sex. Our dependent variables were changed in self-competence and task-value between waves, this is, after the primary to secondary school transition.

Table 2.5
Descriptive statistics of longitudinal motivation variables at ages 11 and 14 years.

	Primary School (Age 11)			Secondary School (Age 14)		
	N	M	SD	N	M	SD
Self-competence						
English	12804	1.89	.69	11367	1.98	.71
Mathematics	12790	1.74	.75	11364	1.99	.77
Science	12659	1.96	.76	11356	2.0	.80
Task-value						
Try your best	12862	1.46	.59	11361	1.76	.62
Find school interesting	12838	2.14	.70	11363	2.51	.69
Consider school a time waste (reversed)	12811	1.57	.79	11355	1.77	.81

Notes. All items were self-reported on a four-point Likert scale.

2.3.2 Measuring change

It cannot be presumed that the same construct was measured between groups with different demographic or contextual characteristics – or in the same group over time as is the case in the current analysis – just because the same measurement tools and procedures were used (Horn and Mcardle, 1992). Group comparisons or changes across time would not yield meaningful results unless it is shown that the measurement of the construct of interest is comparable. *Measurement invariance* (MI) is a statistical concept that allows us to compare the construct been measured across groups or time. The degree of invariance can be assessed through Confirmatory Factor Analysis (CFA) for which an imposed model will constitute a robust description of a sample if their variance-covariance matrices are similar enough (Cheung and Rensvold, 2002). In the context of longitudinal analysis, invariance can be established by sequentially constraining model parameters to be equal across measurement occasions and assessing changes in model fit. Comparisons should be nested, meaning that more constraints should be added to subsequent models, building up levels of invariance to determine the extent of group's resemblance (Molenaar and Borsboom, 2013). Therefore, invariance at one level is a necessary condition for the next. The typical sequence of invari-

ance tests first places constraints on factor structure (i.e. *Configural Invariance*); (2) then factor loadings (i.e. *Metric/Weak Invariance*); (3) followed by residuals/unique variances (i.e. *Strict Invariance*); and/or (4) intercepts/means (i.e. *Strong Invariance*). In the context of longitudinal analyses, correlations between the same indicators over time are also included in the model specification.

The chi-square difference test ($\Delta\chi^2$), is often used to compare the different levels (i.e. models) of MI but it is sample-biased so even small differences are likely to yield significant results in large samples. The sample-adjusted Bayesian Information Criterion difference (ΔsaBIC) is better suited for large samples but it is unbounded. To determine if ΔsaBIC values reflect large differences between models they have to be compared to each model's original BIC.

Prior to fitting change models for Self-competence and Task-value, we first assessed measurement invariance. Tables 2.6 and 2.7 show it is not reasonable to assume that MI holds for both measures, Self-competence and Task-value. Further exploration of the modification indices led us to conclude that fixing the loadings left strong within-wave residual correlation which was not consistent across time for both, Self-competence and Task-value. This suggests that a latent variable of Self-competence (or Task-value) would unlikely reflect this construct consistently across waves. The results of the invariance analysis suggest the measurement of Self-competence and Task Value, may differ across time due to differences in measurement, not true change. For practical reasons, in the remainder of the analysis, we use a simple sum score for each construct.

Table 2.6

Self-competence Measurement Invariance. Chi-Squared Difference Test Between different Measurement Invariance Models.

	Df	CFI	RMSEA	AIC	saBIC	ΔsaBIC	Chisq	Chisq diff	Df	diff Pr(>Chisq)
Configural Model	5	.98	.05	154920.20	155015.80		202.62			
Metric Model	7	.97	.05	154962.52	155049.43	-33.63	248.93	46.32	2	>.001
Metric Model	7	.97	.05	154962.52	155049.43		248.93			
Strict Model	10	.95	.06	155145.79	155219.66	-179.24	438.21	189.27	3	>.001
Strcit Model	10	.95	.06	155145.79	155219.66		438.21			
Strong Model	12	.87	.08	155779.93	155845.11	-625.44	1076.34	638.13	2	>.001

Notes. p < .0001.

Table 2.7

Task-value Measurement Invariance. Chi-Squared Difference Test Between different Measurement Invariance Models.

	Df	CFI	RMSEA	AIC	saBIC	ΔsaBIC	Chisq	Chisq diff	Df	diff Pr(>Chisq)
Configural Model	5	1.00	.01	140432.07	140527.71		15.04			
Metric Model	7	.99	.04	140554.44	140641.38	-113.67	141.41	126.36	2	>.001
Metric Model	7	.99	.04	140554.44	140641.38		141.41			
Strict Model	10	.99	.04	140604.04	140677.94	-36.56	197.01	55.60	3	>.001
Strcit Model	10	.99	.04	140604.04	140677.94		197.01			
Strong Model	12	.96	.06	140922.74	140987.95	-310.01	519.71	322.71	2	>.001

Notes. p < .0001.

2.3.3 Models to measure change

Change was conceptualized in two different ways, as *Difference-Scores* and Regression *Residuals*. Each approach entails different assumptions, strengths, and weaknesses (Corder-Bolz, 1978). Hence, both were used to act as sensitivity analyses to one another, only considering results robust which were present across the methodological approaches. Measurement reliability is a common concern in national longitudinal studies, where the lack of focus often derives from few and uncomprehensive measures. In MCS, there were only six items about school motivation with unknown reliability which were only repeated twice across all waves.

In all analyses, we applied the specific weights developed by Fitzsimons (2017) to compensate for demographic stratification and clustering in the MCS sampling (i.e. the stratification variable "pttype2", the clustering variable "sptn00" and attrition weight "fovwt2"). Additionally, the maximum likelihood estimation with robust standard errors (MLR) was used and predictors were entered in a stepwise fashion to test which improved the model's fit.

2.3.4 *Difference Scores*

Difference Scores (D) estimate change as the difference between Time 2 (T_2) and Time 1 scores (T_1),

$$D = T_2 - T_1 \tag{2.1}$$

$$T_2 = \tau_2 + \varepsilon_2 \tag{2.2}$$

$$T_1 = \tau_1 + \varepsilon_1 \tag{2.3}$$

where each time point comprises a true score (τ) plus some measurement error (ε). By regressing *Difference Scores* on the independent variables (e.g.: sex), we can estimate those variables' influences on Self-competence and Task-value change during the school transition (Figure 2.1). *Difference Scores* have the advantage that they measure change directly as

the difference between the two time points. Additionally, they control unmeasured variables also contributing to the measured construct (e.g.: school anxiety), as long as its associations with these variables were stable (Gollwitzer et al., 2014).

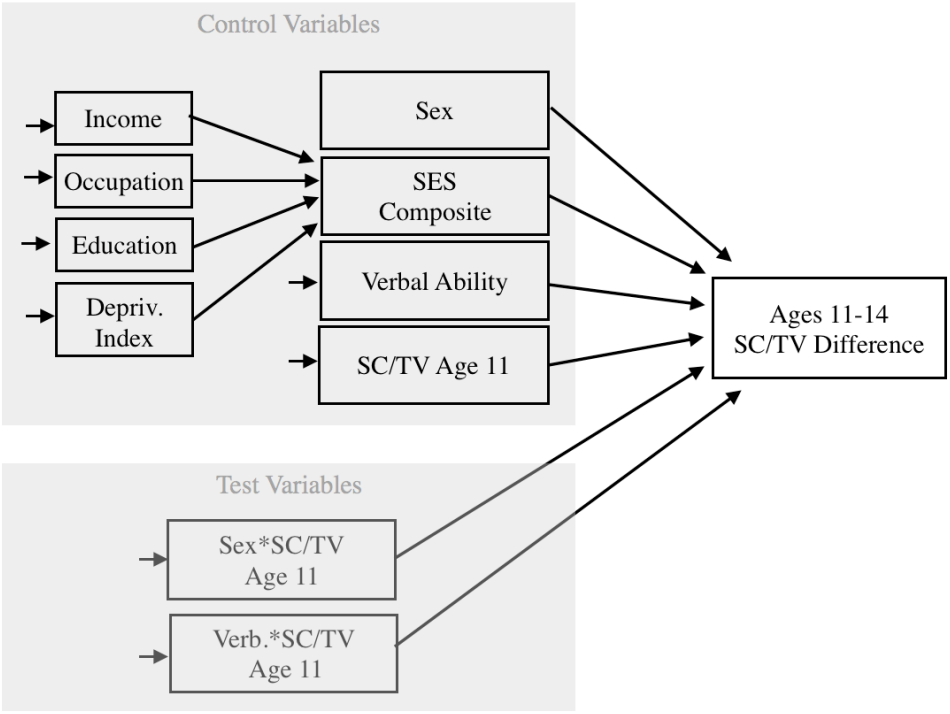

Figure 2.1
Regression Difference Score model.

Notes: In this model, the difference of Self-competence/Task-value scores between primary (age 11) and secondary school (age 14) is predicted by the interaction between sex and Self-competence/Task-value at age 11 and the interaction between Verbal Ability and Self-competence/Task-value at age 11; controlling sex, Socio-Economic Status, Verbal Ability, and Self-competence/Task-value at age 11. SES Composite=Sum Scores of Socio-Economic Status Variables, Verb*SC/TV Age 11=interaction between Verbal Ability and Self-competence/Task-value at age 11, SC/TV=Self-competence/Task-value.

However, *Difference Scores* are inherently less reliable (i.e. contain more measurement error) than the scores contributing to them because they contain the measurement error from both time points ($\varepsilon_2 + \varepsilon_1$). The reliability of *Difference Scores* can be further undermined in

specific situations. For instance, reliability of change (D) is especially low when true scores remain stable, because the T2 minus T1 difference is largely measurement error (Castro-Schilo and Grimm, 2018; Gollwitzer et al., 2014; Rogosa et al., 1982; Thomas and Zumbo, 2012).

The Reliable Change Index (RCI) can be used to assess *Difference Scores*' unreliability. The RCI estimates the percentage of change that can be considered reliable based on the measure's reliability (Christensen and Mendoza, 1986; Jacobson et al., 1984). In other words, the RCI would determine whether there was any actual change and, if so, in what direction. The RCI is computed as

$$RCI = \frac{x_2 - x_1}{\sqrt{2SD_T 1^2(1-r)}} \tag{2.4}$$

where x_2 is the score at T2, x_1 the score at T1, $SD_{T1}$is the Standard Error at T1 and r is the reliability of the measure. The RCI indicates if the magnitude of change is sufficient to discard other causes like regression to the mean and measurement error (Jacobson et al., 1984). The RCI considers two data points at different times significantly different if the intervals representing their standard errors of measurement do not overlap. Given that the RCI will be normally distributed if the used variables are, it produces a t-statistic so the number of cases having RCI values under -1.96 or over +1.96 can be used to estimate the percentage of reliable change with a 95% certainty level.

In the current analysis, the RCI was calculated to aid in the interpretation of the difference scores and is not part of the substantive models. In addition, as there was no reliability estimates for the Self-competence and Task-value items, so we computed RCI assuming reliabilities of 4., .6, and .8. Using different reliabilities levels also allowed to test how conservative was the RCI. Note that a tool's precision to detect small changes increases with its reliability thus, if the percentage of change was still small at high levels of reliability (i.e.: .8) it is likely that there was not much change.

Note that there are methods to estimate change in lieu of *Difference Scores* such as *Polynomial Regression Analysis* and *Latent Difference Scores*. *Polynomial Regression Analysis* are frequently used to estimate between-informant discrepancies (e.g.: de Haan et al., 2018; Edwards, 1994; Laird and De Los Reyes, 2013). *Polynomial Regression Analysis* represent an advantage when studying differences between informants because they allow to simultaneously test the influence of each informant report in addition to their interaction when predicting other outcome variables (e.g.: de Haan et al., 2018; Laird and De Los Reyes, 2013). Considering each measure individually rather than as a single measure of the difference is especially relevant when the effect of either variable is suspected to be (a) more prominent or (b) in a different direction (Edwards, 1994). *Difference Scores* will impose that both variables contributed equally and that their effects on an outcome variable should be opposite (e.g. in a regression predicting job satisfaction, Y, from the difference between actual - X_1 - vs. desired - X_2 - salary: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, you would constrain the β_1 and β_2 effects to be equal in magnitude but of opposing sing). However, the aim of this study was not so much focused on how both time points interacted and/or predicted other outcome variables but on obtaining a change measure and exploring how other variables influenced it. In other words, the change measure in this study was used as an outcome variable rather than a predictor. Moreover, motivation measures in the first wave were included in the analysis so baseline contribution to the change measures were already contemplated in the analyses. Therefore, *Difference Scores* were preferred over *Polynomial Regression Analysis*.

Latent Difference Scores allow researchers to obtained a cleaner measure of the *Difference Scores* by harvesting the common variance among the observed measures and discarding their measurement error. Although *Latent Difference Scores* offer a more reliable measure of change, their assumptions rely more heavily on the existence of measurement invariance (Castro-Schilo and Grimm, 2017). Given the lack of measurement invariance and the possibility of estimating the reliability of *Difference Scores* through the RCI, they were preferred over *Latent Difference Scores*.

2.3.5 *Regression Residuals*

The overall rate of change can be estimated by the slope coefficient from regressing T2 scores on T1 scores. Individual data points' residuals – the differences between actual T2 values and those resulting from the regression equation – are often considered estimates of change between the two time points. By regressing these *Residuals* on our independent variables, we can estimate their influences on Self-competence and Task-value change relative to the overall average change during the school transition (Figure 2.2).

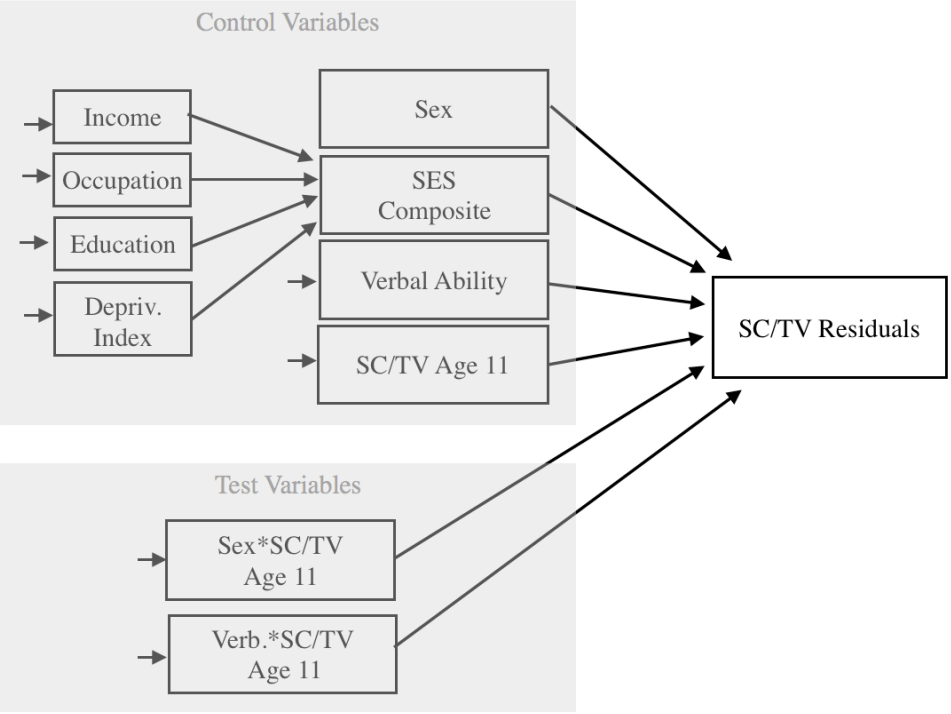

Figure 2.2
Regression of regression's Residuals model.

Notes: In this model, the residuals of Self-competence/Task-value between primary (age 11) and secondary school (age 14) are predicted by the interaction between sex and Self-competence/Task-value at age 11 and the interaction between Verbal Ability and Self-competence/Task-value at age 11; controlling sex, Socio-Economic Status, Verbal Ability, and Self-competence/Task-value at age 11. SES Composite=Sum Scores of Socio-Economic Status Variables, Verb*SC/TV Age 11=interaction between Verbal Ability and Self-competence/Task-value at age 11, SC/TV=Self-competence/Task-value.

Residuals have often been preferred over *Difference Scores*. They are considered more reliable for two reasons. First, the error in the T1 measure is regressed out, as well as whatever error at T2 that correlated with it. Second, the variance from stable intra-individual traits (e.g.: school anxiety) contributing to both time points to the same degrees is also partialled out. However, when regressing T2 to T1, any other independent variable affecting T2 measure, whether included in the regression or not (e.g.: school anxiety), it is assumed to lie at their sample means for each data point (Willett, 1988). Thus, this method measures change only relatively. Moreover, given that it is almost certain that no data point fits this profile, it is likely inaccurate for most of them, especially for those participants experiencing very small or large changes and/or with unusual values of other contributing variables given their values on those assessed.

2.3.6 Models' summary

Both our approaches have important strengths and limitations. *Residuals* are in some ways more reliable, but at best they measure only relative change and rely on unrealistic underlying assumptions. Therefore, the degree to which they reflect change experienced by individuals is questionable. On the other hand, *Difference Scores* capture individual change directly but contain measurement error from both data points. The RCI can be used to estimate the percentage of *Difference Score* change that can be considered reliable. We thus used both approaches intending to interpret only results that converged, but also to investigate why they differed as to increase understanding of the implications of the models' limitations.

2.4 Results

We used M*plus* version 8.3 (Muthén and Muthén, 1998) for our analysis. Table 2.9 shows the matrix of correlations between all the variables. Within each wave, correlations among Self-competence items (rs_{range}= .18 to .42) and Task-value items (rs_{range}= .34 to .48) were small to moderate. There was a similar range of effect sizes for the SES measures (rs_{range}=

Table 2.8
Reliable Change Index (RCI) for Self-competence and Task-value changes between ages 11 and 14.

	Reliability = .4		Reliability = .6		Reliability = .8	
	% Declines	% Increases	% Declines	% Increases	% Declines	% Increases
Self-competence	2.2	2.6	3.1	5.5	6.2	11.2
Task-value	2.6	3.0	4.1	4.9	9.0	9.1

.30 to .50).

Consistent with previous literature (Turkheimer and Harden, 2013), Verbal Ability was positively related to all SES measures (rs_{range}= .19 to .27). However, contrary to other studies (Marsh, 1987, 1984), there was a negative relation between students' SES measures, Self-competence, and Task-value at age 11 (rs_{range}= .00 to -.06) and age 14 (rs_{range}= -.01 to -.12).

The correlation between sex and Verbal Ability was significant but very small (r= .04). The correlation between sex and Self-competence items was small (rs_{range}= -.13 to .16), reflecting that girls had lower Self-competence in English than boys, but higher in Mathematics and Science in both, primary and secondary school. Sex also correlated with Task-value scores. Primary school girls had lower Task-value across school subjects (rs_{range}= -.13 to -.19), but these differences disappeared by secondary school in all but one item – i.e. "try my best".

At the between-persons level, there were no differences in Self-competence (t(10609)=-1.25, p=.62) or Task-value (t(10547)=-.49, p=.21) between ages 11 and 14. The absence of between-person change could be masking differences at a within-person level if children experiencing Self-competence/Task-value increases were being offset by those suffering declines. We further explored within-person level changes with the *Difference Scores* and *Residuals*.

2.4.1 Self-competence

The RCI showed very small percentages of reliable Self-competence change (Table 2.8), both for increases and decreases, so there was not a clear tendency on change direction. Even assuming rather high reliability (r=.8), the percentage of reliable change was small, between 6-11%. As previously mentioned, low reliability could be due to large measurement error or to true-score stability. The latter explanation is not supported due to the small correlations between the Self-competence items at age 11 and at age 14 (rs_{range}= .08 to .36; Table 2.9). In any case, the reliability of *Difference Scores* was not high and this should be taken into account when interpreting the results.

In both regressions, *Difference Scores* and *Residuals*, the control variables (sex, SES, and Verbal Ability) absorbed most of the outcomes' variance and the inclusion of the interactions did not improve the models' fits significantly (Tables 2.10 and 2.11, respectively). As explained previously, due to the nature of *Residuals*, they should not be correlated to T1 measures. However, *Difference Scores* and T1 scores could be correlated if both measures are loaded with error.

Model 2 in the *Difference Scores* regression (including sex, SES, Verbal Ability, and Self-competence at age 11 as predictors), was preferred for having a better fit than Model 1 and for being more parsimonious than Models 3 and 4, which included the interactions but did not have a better fit. Following a similar logic and taking into account the meaning of the correlation between T1 and the *Residuals*, Model 1 was chosen as the best in the *Residuals*' regression.

Table 2.9
Correlation matrix of all variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Sex	–																
2.Verbal Ability (age 11)	-.04	–															
3. Income		.27	–														
4. Parental Education		.22	.55	–													
5. Parental Occupation		.20	.50	.43	–												
6.Deprivation index		.19	.55	.30	.34	–											
7. SC English (Age 11)	-.13	-.10	-.04	-.06	-.04		–										
8. SC Math (Age 11)	.16	-.09		-.04	-.05	.03	.10	–									
9.SC Science (Age 11)	.10	-.13	-.06	-.06	-.04		.25	.23	–								
10. TV Try your best (Age 11)	-.17	-.04	-.03			.02	.21	.15	.15	–							
11. TV Find school interesting (Age 11)	- .13	-.04				.03	.28	.19	.23	.39	–						
12. TV Consider school a time waste (Age 11)	-.19	-.06	-.06	-.06	-.04	-.02	.22	.12	.14	.34	.48	–					
13. SC English (Age 14)	-.08	-.11	-.08	-.06	-.04		.25		.08	.07	.11	.10	–				
14. SC Math (Age 14)	.13	-.12	-.11	-.09	-.09	-.05		.36	.11	.06	.08	.08	.18	–			
15. SC Science (Age 14)	.10	-.18	-.12	-.12	-.07	-.05	.09	.17	.20	.05	.10	.09	.32	.42	–		
16. TV Try your best (Age 14)	-.09		-.05				.11	.07	.10	.26	.21	.19	.16	.18	.21	–	
17. TV Find school interesting (Age 14)		-.08	-.09	-.09	-.06	-.03	.13	.11	.12	.15	.26	.19	.17	.21	.24	.44	–
18. TV Consider school a time waste (Age 14)		-.06	-.09	-.08	-.06	-.05	.10	.07	.08	.12	.18	.27	.14	.17	.21	.34	.41

Notes. p < .01, two-tailed with no adjustments for multiple testing. SC= Self-competence, TV= Task-value.

Sex was a significant predictor of Self-competence change in both *Difference Scores* regression (β =-.03, $p < .001$) and *Residuals*' regression (β =-.05, $p < .001$) indicating that girls experienced slightly more change than boys. The interaction between sex and Self-competence at age 11 did not improve the model's fit but was significant in both regressions *Difference Scores* (β =-.04, $p < .001$) and *Residuals*' (β =-.06, $p < .001$), indicating that girls' tendency toward Self-competence change during the school transition was slightly tempered for those having higher Self-Competence at age 11.

SES was a significant predictor of Self-competence change. The correlation was small and positive in both *Difference Scores* regression (β =.06, $p < .001$) and *Residuals*' regression (β =.09, $p < .001$) reflecting that high-SES students tended to experience slightly more change.

Additionally, *Residuals*' regression indicated that more verbally able students tended to experience more change in Self-competence (β =.11, $p < .001$). This also converged with *Difference Scores* regression (β =.07, $p < .001$). Again, the interaction between Verbal Ability and Self-competence at age 11 did not improve the model's fit but was significant. The interpretation of this interaction would depend on the direction of the relation between change and Self-competence at age 11.

Self-competence at age 11 was negatively related in *Difference Scores* model (β =-.74, $p < .001$) but positively in the *Residuals* model (β =.07, $p < .001$). Hence, the interactions between Verbal Ability and Self-competence at age 11 (*Difference Scores* β =-.02, $p < .01$ and *Residuals* β =-.04, $p < .01$) would have different interpretations. These will be described in the Discussion section.

Overall, there was consistent convergence between *Difference Scores* and *Residuals*' regression results. There was a small percentage of reliable change, so there was no evidence supporting our Prediction 1 about an overall decline in Self-competence. Self-competence change was related to our control variables: sex, SES, and Verbal Ability. However, it is important to highlight that large data sets allow to find rare associations which, while being statistically

Table 2.10
Models Predicting Self-competence Difference Scores

	Model 1	Model 2	Model 3	Model 4
Intercept	.08 (.01)**	.03 (.01)**	.03 (.01)**	.03 (.01)**
Sex	-.07 (.01)**	-.03 (.01)**	-.03 (.01)**	-.03 (.01)**
SES	.06 (.01)**	.06 (.01)**	.06 (.01)**	.06 (.01)**
Verb. Abil.	.19 (.01)**	.07 (.01)**	.07 (.01)**	.07 (.01)**
SC at age 11		-.76 (.01)**	-.74 (.01)**	-.74 (.01)**
Sex*SC			-.03 (.01)**	-.04 (.01)**
SC*Verb. Abil.				-.02 (.01)**
CFI	1	1	1	1
RMSEA	<.05	<.05	<.05	<.05
saBIC	31696.40	22262.80	22253.12	22243.70
ΔsaBIC		-9433.6	-9.68	-9.42

p< .05 † p< .01 * p< .001 **. *Notes.* Standardized coefficients for simultaneous regression. SES= Socio-economic Status, Verb. Abl.= verbal ability at age 11, SC Age 11= Self-competence at age 11, Verb*SC= interaction between verbal ability at age 11 and Self-competence at age 11.

Table 2.11
Models Predicting Self-competence Residuals

	Model 1	Model 2	Model 3	Model 4
Intercept	.04 (.01)*	.04 (.01)*	.05 (.02)*	.04 (.02)**
Sex	-.05 (.01)**	-.05 (.01)**	-.05 (.01)**	-.05 (.01)**
SES	.09 (.01)**	.09 (.01)**	.09 (.01)**	.09 (.01)**
Verb. Abil.	.11 (.01)**	.11 (.01)**	.11 (.01)**	.11 (.02)**
SC at age 11		.03 (.01)†	.07 (.02)**	.07 (.02)**
Sex*SC			-.05 (.02)**	-.06 (.02)**
SC*Verb. Abil.				-.04 (.01)*
CFI	1	1	1	1
RMSEA	<.05	<.05	<.05	<.05
saBIC	22264.33	22262.80	22253.12	22243.70
ΔsaBIC		-1.53	-9.68	-9.42

p< .05 † p< .01 * p< .001 **.*Notes.* Standardized coefficients for simultaneous regression.SES= Socio-economic Status, Verb. Abl.= verbal ability at age 11, SC Age 11= Self-competence at age 11, Verb*SC= interaction between verbal ability at age 11 and Self-competence at age 11.

significant, might not have practical value. Hence, in these cases it is especially relevant to focus on the magnitude of the effect sizes. The strength of these predictors in this study was weak so it would be premature to argue that they had sizeable influences on Self-competence change during the primary to secondary school transition. Results do not support our Prediction 3, as high-SES students experience more change. Contrary to our Predictions 2 and 5, neither of the interaction terms, (1) between Self-competence at age 11 and sex, and (2) between Self-competence at age 11 and Verbal Ability, improved the models' fits. Potentially, more reliable measures of Self-competence could capture more change variance, allowing better estimation of the factors involved and their importance to Self-competence change over this important transition. Other factors, such as grades, might be more relevant to Self-competence change because they inform students about their performance and they serve as a base for peers' comparison.

2.4.2 Task-value

According to the RCI, there were small percentages of reliable Task-Value change (Table 2.8). Even assuming rather high reliability (r=.8), the percentage of reliable change was around 9%. Low reliability could be due to large measurement error or to true-score stability. The latter explanation is not supported due to the small correlations between the Task-value items at age 11 and at age 14 (rs_{range}= .12 to .27; Table 2.9). Hence, Task-Value *Difference Scores* were also unreliable.

In both regressions, *Difference Scores* and *Residuals*, the interactions did not improve the models' fits significantly (Tables 2.12 and 2.13, respectively). Following the same rationale as in Self-competence, Model 2 in the *Difference Scores* and Model 1 in the *Residuals*' regressions were preferred.

As in Self-competence, sex and SES were significant predictors of Task-value change in *Difference Scores* and *Residuals*' regressions. Low-SES students (*Difference Scores* β =-.07, $p < .001$ and *Residuals* β =-.09, $p < .01$) experienced slightly more Task-value change than

Table 2.12
Models Predicting Task-value Difference Scores

	Model 1	Model 2	Model 3	Model 4
Intercept	-.14 (.01)**	-.02 (.01)	-.02 (.01)	-.02 (.01)
Sex	.14 (.01)**	.03 (.01)*	.03 (.01)*	.03 (.01)*
SES	-.06 (.01)**	-.07 (.01)**	-.07 (.01)**	-.07 (.01)**
Verb. Abil.	.03 (.01)†	-.01 (.01)	-.01 (.01)	-.01 (.01)
TV at age 11		-.57 (.01)**	-.59 (.01)**	-.59 (.01)**
Sex*TV			.04 (.01)*	.04 (.01)*
TV*Verb. Abil.				.01 (.01)
CFI	1	1	1	1
RMSEA	<.05	<.05	<.05	<.05
saBIC	26910.64	22946.01	22939.35	22943.95
ΔsaBIC		-3964.63	-6.66	4.6

p< .05 † p< .01 * p< .001 **. *Notes.* Standardized coefficients for simultaneous regression.SES= Socio-economic Status, Verb. Abl.= verbal ability at age 11, TV Age 11= Task-value at age 11, Verb*TV= interaction between verbal ability at age 11 and Task-value at age 11.

Table 2.13
Models Predicting Task-value Residuals

	Model 1	Model 2	Model 3	Model 4
Intercept	-.03(.02)	-.03(.02)	-.02(.02)	-.02(.02)
Sex	.03 (.01)*	.03 (.01)*	.03 (.01)*	.03 (.01)*
SES	-.08 (.01)**	-.08 (.01)**	-.09 (.01)**	-.09 (.01)**
Verb. Abil.	-.02 (.01)	-.02 (.01)	-.02 (.01)	-.02 (.01)
TV at age 11		.00(.02)	.00 (.01)	-.03 (.02)†
Sex*TV			.05 (.01)*	.05 (.01)*
TV*Verb. Abil.				.01(.01)
CFI	1	1	1	1
RMSEA	<.05	<.05	<.05	<.05
saBIC	22939.95	22946.01	22939.35	22943.95
ΔsaBIC		6.06	-6.66	4.6

p< .05 † p< .01 * p< .001 **. *Notes.* Standardized coefficients for simultaneous regression.SES= Socio-economic Status, Verb. Abl.= verbal ability at age 11, TV Age 11= Task-value at age 11, Verb*TV= interaction between verbal ability at age 11 and Task-value at age 11.

their counterparts. Also, girls change less compared to boys (*Difference Scores* β =.03, $p < .01$ and *Residuals* β =.03, $p < .01$). The interaction between sex and Task-value at age 11 did not improve the model's fit but was significant in both regressions *Difference Scores* (β =.04, $p < .01$) and *Residuals* (β =.05 $p < .01$), indicating that girls' Task-value change during the school transition was even lower for those having low Task-value at age 11.

Unlike for Self-competence, Verbal Ability was not a significant predictor of Task-value change in either *Difference Scores* or *Residuals*' regressions. The interaction between Verbal ability and Task-value at age 11 was not significant either. This suggests that more able students might consider themselves competent in a subject but completely uninterested in it.

Overall, there was consistent convergence between *Difference Scores* and *Residuals*' regression results. There was a small percentage of reliable change, so there was no evidence supporting our Prediction 1 about an overall decline in Task-value. Task-value change was related to sex and SES, but not to Verbal Ability. Results support Prediction 3 because high-SES students experience less change. Agreeing with Prediction 2 the interaction between Task-value at age 11 and sex was significant but, contrary to Prediction 5, the interaction between Task-value at age 11 and verbal ability was not. Regardless, none improved the models' fit. As in Self-competence change, the strength of the predictors was weak so it would be premature to argue that they had sizeable influences on Task-value change during the primary to secondary school transition.

2.5 Discussion

In this study, we tried to expand Eccles' expectancy-value model by relating students' motivation (i.e. self-competence and task-value) to the interactions between its original level and either sex or verbal ability. For this purpose, we focused on changes in students' motivation during the primary to secondary school transition in a UK sample (i.e. the Millennium Co-

hort Study) controlling for relevant factors such as sex, SES, and verbal ability. Due to the lack of Measurement Invariance, we used sum scores and we capture change through two different statistical approaches – i.e. *Difference Scores* and *Residuals*. We then predicted these from sex, SES, Verbal Ability at age 11, Self-competence or Task-value at age 11, the interaction between sex and either Self-competence or Task-value at age 11, and the interaction between Self-competence or Task-value at age 11 and Verbal Ability. The results converged across models: Self-competence and Task-value change were predicted by sex and SES, but only Self-competence change was influenced by Verbal Ability. Additionally, neither of the interaction terms between (1) Self-competence/Task-value at age 11 and sex or (2) Self-competence/Task-value at age 11 and Verbal Ability, improved the models' fits.

The RCI showed small and similar amounts of reliable change in Self-competence and Task-value. Percentages of increases and decreases were even hence, results do not support Prediction 1 regarding a generalized decline in Self-competence/Task-value across time as previous research has found (e.g.: Gaspard et al., 2017; Guo et al., 2015; Jacobs et al., 2002; Watt et al., 2012).

Like in Spinath et al. (2010), but contrary to several other studies (Eccles, 1983; Else-Quest et al., 2010; Steinmayr and Spinath, 2008; Spinath et al., 2008; Wach et al., 2015; Watt et al., 2012), girls had lower Self-competence in English than boys, but higher in Mathematics and Science at ages 11 and 14. Girls also had lower Task-value than boys in all three subjects at age 11, although most sex differences in Task-value disappeared at age 14. Additionally, sex rose as a relevant factor predicting changes in Self-competence and Task-value: girls experienced more change in Self-competence and boys experience more change in Task-value. The interaction between sex and Self-competence or Task-value did not improve the model fit. However, it was it was significant for Self-competence, partially supporting Prediction 2. This interaction meant that girls tendency to change was not as prominent in those that already had higher Self-competence. If this interaction was stronger it might have suggested that high self-competence believes might be difficult to change for girls.

As expected, students' SES was related to their abilities. However, the data did not completely support Prediction 3 because (1) SES was negatively related to Self-competence and Task-value, meaning that high-SES students had lower interest and perceptions of their abilities, and (2) SES was positively related to Self-competence change but negatively to Task-value change. These effects were small but they do not have a clear explanation. Previous research has shown that high-SES families might provide students with broader cultural knowledge which would make school more accessible and attractive (Bourdieu and Passeron, 1977; Ditton et al., 2019). Only our results showing high-SES related to higher Task-value stability match these previous research.

Our Prediction 4 was not fully supported. Contrary to what we expected, Verbal Ability was negatively related to Self-competence and to Task-value at age 11 and 14, implying that actual ability negatively influences the perceptions of one's competence and their interest in school. This is, again, a surprising finding because it contradicts the idea that more able students are likely to think of themselves as more competent and to find school more rewarding. On the other hand, the relation between Verbal Ability and Self-competence was stronger than with Task-value at both waves, which does indicate that ability is more related to how good student think they are at schoolwork than how much they care about it.

We also found partial support for Prediction 5 as verbal ability predicted change in Self-competence but not in Task-value. This suggests that actual ability is more relevant to maintain students' perceptions of their abilities than their interest in school. Moreover, students having high Self-Competence at age 11 experienced less Self-competence change. This could be because they had a more realistic perception of their ability or because the feedback they received was similar throughout the school transition and they did not feel the need to changed their self-perceptions. Additionally, students having high scores in both, Self-competence and Verbal Ability at age 11, also experienced less change. This could suggest that verbally able students who already had high self-competence beliefs did not change as much as those verbally able students who started with low self-competence

beliefs. In this sense, feedback experiences might have lead this verbally able children to realized their potential hence, experiencing more change in their Self-competence scores. If this interaction had been stronger this would exemplify how actual ability can shape student experience and influence their motivation.

In conclusion, results from *Difference Scores* match *Residuals* models' results. Girls seem to change more in Self-competence but less in Task-Value as compared to boys. High-SES students tended to have more stable Task-value but less stable Self-competence across the primary to secondary school transition. Lastly, Verbal Ability was not related to Task-value change but it contributed to Self-competence stability. However, large data sets allow to detect statistically significant associations which might not be relevant in practice. For this reason, it is important to evaluate the magnitude of the effect sizes. In this study, the strength of the predictors was weak so it would be premature to argue that any had a crucial role in Self-competence or Task-value changes. Albeit the small effect sizes, this study shows that verbal ability – and by extension, cognitive ability – play a role in Eccles's expectancy-value model, especially as it relates to self-competence. Actual ability could reduce student changes in motivation by giving them a more realistic insight on their abilities. Moreover, as we will see in the next Chapter (Chapter 3), other prominent authors like Holland (1959) and Gottfredson (1981), have argued that properly gauging one's ability is crucial to develop a realistic occupational career path. In other words, the notion of self-competence accompanies individuals throughout their lifes and it might be especially relevant in their career development.

2.5.1 Strengths and limitations

It is important to bear in mind this study's limitations. First, the representativeness of the sample was compromised by a high drop-out rate (40% attrition between the first and the age-14 assessments), which was associated with SES, Self-competence, and Task-value. As a consequence of range restrictions in these variables, we have likely over- and under-estimated some predictors. Second, although the Self-competence and Task-value measures

were similar to those used in previous literature regarding total quantity and characteristics (for a comparison see Table 2.1), their retest reliability was unknown. Having better quality items could help to capture change more effectively. Third, change is difficult to estimate with only two waves (Rogosa et al., 1982) because change cannot be distinguished from the error of measurement, and overall trends of change cannot be detected. Additionally, the fact that the last measure of school achievement (i.e. grades) and the first measure of occupational aspirations were taken at the same time could have confounded the mediation role of change in grades between verbal ability and change in occupational aspirations.

Despite these limitations, this study had multiple strengths. First, the MCS has an extremely large sample (N>10 000) and extensive efforts have been made to ensure it was population-representative. Moreover, it is one of the newest longitudinal datasets including data from children born in the 2000s. Second, motivation measures were assessed using the same items before and after the primary to secondary school transition during which expectations of students, school routines, and students themselves (most experiencing puberty) change considerably. Motivation and school engagement often also change substantially, with some taking school more seriously and others finding it more alienating. Lastly, despite the data's limitations, results from both statistical methods converged depicting a consistent picture of factors involved in Self-competence and Task-value in primary and secondary school.

Chapter 3

Mediating Role of School Grades Change between Verbal Ability and Occupational Aspirations

3.1 Introduction

Educational and occupational aspirations refer to the educational attainment (e.g.: Compulsory Schooling, Bachelor, Master, etc.) and the occupational status (e.g.: managerial, manual, etc.) that students would want to eventually achieve. Research on educational and occupational aspirations has gained popularity due to its power to predict years of formal education and occupation status later in life (Hegna, 2014). In fact, educational aspirations at age 13-16 can predict differences in educational attainment, occupational status, and income 10-30 years later (Schoon and Parsons, 2002; Schoon and Polek, 2011), even when controlling for cognitive ability (Bergman et al., 2015).

Despite the general belief regarding individual's agency over their career plans, educational and occupational aspirations are not just a matter of free choice (Gottfredson and Becker, 1981; Lent et al., 2002). Many factors shape and limit students' aspirations, mainly: cognitive ability, self-competence, educational achievement, Socio-Economic Status (SES) background and opportunities, as well as cultural context, and sex stereotypes. These predictors intertwine, making educational and occupational aspirations the byproduct of a complex system of interactions. This study will focus on the mediating role that grades' change has between verbal ability and occupational aspiration change. The rest of the moderating factors will be further discussed in Section 7.3.2.

3.1.1 Theories on Education/Occupational Aspirations Change

Theories explaining the origin and evolution of educational and occupational aspirations can be roughly divided into social and comprehensive (for an extensive review see Hyde, 2014, or Steward-Williams, 2021). Social theories emphasize how social and cultural factors can shape individuals' educational and occupational aspirations through reinforcements (e.g.: Lent, Brown, and Hackett's, 2002, *Social Cognitive Career Theory*), power inequalities and the division of labor force (e.g.: Social Role Theory), as well as, sex-stereotypes and sex-roles (e.g.: Gender Socialization theory). Comprehensive theories like Eccles's (1983)

Expectancy-Value Theory (see Chapter 2), Holland's (1959) theory of vocational interests or Gottfredson's (1981) theory of Circumscription and Compromise, expose a more balanced view of both biological predispositions and social pressures. This wide range of theories shows the complex nature of occupational aspirations and reflects the idea that every model we use to describe them – even the most complex ones – will still be far from the reality. These issues will be discussed in more depth in Chapter 7 but, for now, I will introduce three theories around the origin and change of occupational aspirations.

Holland's (1959) theory of vocational interests, has been widely used to explore types of occupational aspirations. Holland's (1959) theory categorizes occupations into *environments*. Each environment is suited to individuals with a specific *interest*– i.e. personality traits, aptitudes, and interests. Holland developed six orientations (i.e. realistic, intellectual, social, conventional, enterprising, and artistic; see Chapter 6 for more details). Holland's orientations can be measured through questionnaires like the Vocational Preference Inventory. According to Holland, students' self-knowledge of their cognitive ability and skills would allow them to find their best occupation match.

Another approach is Gottfredson's (1981) theory of Circumscription and Compromise which has been widely used to explore changes in occupational aspirations. Her research showed that people perceive occupations similarly regardless of their sex, SES, educational level, and ethnicity, so there is a universal mind-map of occupations regarding their sex type and prestige. Additionally, people hold believes about what is appropriate and accessible for them according to their sex, cognitive ability, and SES.

According to Gottfredson (1981), occupational aspirations depend on two key processes and four developmental stages. These developmental stages change as individuals interiorize increasingly abstract notions of their identity: sex, social class, ability self-concept (i.e. how people would describe themselves), and personal interest or values. In early childhood (Stage 1), children name magical characters or adult roles as their occupational aspirations. Across primary school, students become aware of sex differences and, due to their dichotomous

thinking, they only ascribe to sex-typical behaviors and aspirations (Stage 2) without implying a devaluation of their sex. After multiple school feedback experiences and contact with a wider range of social groups, students gain insights on their social class and cognitive ability, which allows them to evaluate occupations according to social prestige and cognitive demand, in addition to sex stereotypes (Stage 3). Through secondary school students start valuing occupations compatible with their personalities, values, and interests (Stage 4). Throughout these four stages, there are two key processes that narrow down the alternatives: circumscription and compromise (Gottfredson and Lapan, 1997). Circumscription refers to the process by which individuals choose a set of acceptable occupations, this is, occupations within *tolerable-sex-boundaries, tolerable-effort-boundaries,* and *tolerable-prestige-boundaries*. Compromise refers to moments when individuals discard their preferred options when faced with limitations regarding occupations' accessibility and suitability (i.e. aspirations vs. expectations). Compromises tend to follow a specific hierarchy to preserve elements core to their self-concept: first people relinquish compatibility with their interests, then prestige, and finally, sex-stereotyped.

Another relevant model is Lent and colleagues' Social Cognitive Career Theory (SCCT). According to them, individuals start with certain predispositions determined by contextual factors like their SES, sex, and race. These predispositions lead students into specific learning experiences. Students start building interests on the activities in which they perceive themselves as more competent. Interest and self-efficacy (i.e. student's self-perceived ability to obtain a desired outcome through their actions) feed into each other through practice. This positive feedback loop further encourages or discourages students' interests on specific activities. Lent et al. (2002) hypothesize that this feedback loop continues developing throughout an individual's lifespan.

These three theories (and also Eccle's expectancy-value theory; see Chapter 2) share the hypothesis that feedback experiences shape student's interest. Therefore, feedback experiences should be considered a relevant factor influencing occupational aspirations. Another com-

mon denominator is the idea of a person-to-occupation match. According to Gottfredson, this match depends on some contextual factors as well as personality, interests, and actual cognitive abilities. Similarly, for Holland it has to do with an individual's personality, interests, and cognitive abilities. Lastly, Lent and colleagues considered environmental factors as well as the individual's perceived ability. The main difference between the theory of Circumscription and Compromise and the SCCT exemplifies what was already mentioned in Section 1.4 regarding the excision between cognitive ability and Educational research. Both theories acknowledge the role of abilities but address it differently: the SCCT theory uses self-competence while the theory of Circumscription and Compromise includes the actual cognitive ability.

3.1.2 Changes in educational and occupational aspirations

Low et al. (2005), meta-analyzed 66 studies and found that occupational preferences remained reasonably stable from primary school to adulthood. Specific degrees like art, architecture, and veterinary might become less popular across primary school, others keep stable (i.e. sciences and teaching), and others increase their popularity – i.e. social welfare and nursing (Gore et al., 2017; Gottfredson, 1979). The use of magical characters declines across primary and secondary school (Helwig, 2001). As primary school students grow, they no longer justify their aspirations based on their interest but on their abilities (Gore et al., 2015). Some studies report small declines in occupational aspirations through secondary school (Gottfredson, 1981; Helwig, 2001; Lee and Rojewski, 2012; Shapka et al., 2006). Additionally, research suggest that low-SES students have lower occupational aspirations, but they are as stable as for high-SES students (Gottfredson, 1981). Together, these studies suggest that primary school students already have occupational preferences, but they adjust them as they gain more insights on their abilities and on the working opportunities they might have (Gottfredson, 1979; Holland, 1959).

As previously mentioned, occupational aspirations have been used to predict educational attainment and occupational status in adulthood (e.g.: Croll, 2008; Schoon and Parsons,

2002). Thus, it is important to understand the factors influencing its changes like cognitive ability and educational achievement.

3.1.3 Cognitive ability, educational achievement, and aspirations

Following the theoretical frameworks previously discussed, occupational aspirations should be related to cognitive ability and educational achievement. On the one hand, cognitive ability has been consistently related to higher educational (Andersson et al., 2014; Dubow et al., 2009) and occupational aspirations (Cochran et al., 2011; Dubow et al., 2009; Schoon, 2001; Schoon and Polek, 2011). Additionally, Roth et al. (2015)'s meta-analysis showed a *.50* correlation between cognitive ability and school achievement, although some studies have found correlations as strong as *.74* (Andersson et al., 2014) or *.81* (Deary et al., 2007).

On the other hand, research has shown that educational achievement is also related to higher occupational aspirations (Marjoribanks, 2002; Mau and Bikos, 2000; Patton and Creed, 2007a,b; Rojewski, 1996; Rojewski and Yang, 1997; Schoon and Parsons, 2002), even in primary school (Basler and Kriesi, 2019; Gore et al., 2015; Lee and Rojewski, 2012). Higher school achievement has even been related to specific occupations aspirations like engineering, architecture, or mathematics (Holmes et al., 2018). Patton and Creed (2007a) found that students with higher educational achievement not only have higher occupational aspirations but are also more informed on what that career path would entail and have clearer career-oriented goals. Similarly, Gore et al. (2015) showed that higher achievement in primary school was related to higher certainty in occupational aspirations.

3.1.4 The role of educational achievement in occupational aspirations change

Previous research has shown that cognitive ability, educational attainment, and occupational aspirations correlate. However, few studies have considered how these variables interact across time. From Gottfredson's, Holland's, and Lent and colleagues' theories, it could be

assumed that (1) cognitive ability influences school grades, (2) school grades provide feedback to students on their strengths and, in turn, (3) this shapes their occupational aspirations. This mediation role of grades change is largely supported by Marsh and colleagues' work. Through several studies, they showed that children in more intellectually selective groups had lower self-perceived ability or self-competence than children with the same ability in less selective classrooms (Bachman and O'Malley, 1986; Marsh, 1984, 1987; Marsh et al., 1995; Marsh and Hau, 2003). These findings led them to develop their "Big-Fish-Little-Pond" effect reflecting that the school context shapes students' believes in their own abilities. Following this idea, Davis (1966) emphasized that grades represent the most objective available standard against which students could compare each other.

Davis, as well as Marsh and colleagues research fed into the Circumscription and Compromise theory and into the SCCT, as both acknowledge the relevant role of feedback experiences on occupational aspirations. Although there is not a lot of research specifically correlating educational achievement *change* and occupational aspiration *change*, Shapka et al. (2006) found that low educational achievement around age 14 produced a more rapid decrease of occupational aspirations towards the end of secondary school. On the contrary, high educational achievement was related to a more gradual and linear decrease of occupational aspirations across high school. Additionally, DesJardins et al. (2019) found that grades improvement between ages 14 and 17 related to an increase in educational aspirations. Together, these studies show that grades *per se*, but also grades' change can influence students' occupational aspirations.

3.1.5 Summary and hypothesis

In summary, educational and occupational aspirations are not a matter of free choice, there are certain factors that shape them, like actual cognitive ability and feedback experiences. Several models have been proposed to show how these factors contribute to occupational aspirations. Gottfredson's (1981) theory of Circumscription and Compromise is among the most comprehensive ones, but Holland's vocational interests theory and Lent and colleagues

Social Cognitive Career Theory (SCCT), also consider the role of cognitive abilities and feedback experiences.

This study tests the mediating role of grades change between actual verbal ability and occupational aspiration change. Following previous research, I hypothesize that larger changes in grades should contribute to larger changes in occupational aspirations as students will have to re-evaluate whether they are a good match for those occupations.

3.2 Methods

3.2.1 Sample

For this study we used The Millennium Cohort Study (MCS), a longitudinal study of children born in the United Kingdom between 2000 and 2001 (Plewis et al., 2007). In the MCS, children have been tested every 2-3 years since they were nine months old. In this study we used data from the waves 4, 5, and 6, when participants were 7, 11, and 14 years old respectively. There was 40% attrition between the first and the sixth waves (3% between ages 7 and 11, 8% between ages 11 and 14, Mostafa and Ploubidis, 2017). Families that dropped out did not have significantly lower initial socioeconomic-status $\chi^2(9, N=8642)= 14.87$, $p > .05$. Additionally, there were almost no differences between children who participated in the three waves and those who dropped out on the last one in regards to grades average at age 7 (d=.04), grades at age 11 (d=-.07), or aspirations at age 11 (d=.06). All differences were very small according to Cohen (1992). For a detail description of the data collection process as well as this subsample demographics see Chapter 2.

3.2.2 Materials

In this study we used cohort members' data on their sex, verbal abilities, grades, and occupational aspirations. Table 3.1 contain the descriptive statistics of verbal abilities, grades, and occupational aspirations.

Verbal Ability. Participants completed the Word Reading subscale from the British Ability Scales (BAS) at age 7 (Elliott, 1986). In this task respondents are given a list of words that they have to read aloud. Words are organized in increasing difficulty levels.

Self-reported aspirations. Participants reported their occupational aspirations at age 11 and 14. This age range roughly corresponds to Stage 3 in Gottfredson's (1981) theory of Circumscription and Compromise in which students have (a) dismissed imaginary occupations, (b) interiorized sex-related stereotypes regarding occupational aspirations, and (c) are starting to use school feedback experiences to gauge what occupations could fit their abilities. Occupational aspirations were recorded through the question: *"When you grow up what would you like to be?"*. Responses were coded according to the Standard Occupational Classification (SOC2010) and a "not codable" category for answers like "worker", "dad", "mum", etc. In the UK, the Standard Occupational Classification was first published in 1990 to replace both the Classification of Occupations 1980 (CO 80) and the Classification of Occupations and Dictionary of Occupational Titles (CODOT). The SOC 90 was substituted for SOC2000 so that is matched the National Statistics Socio-economic Classification (NS-SEC) better. The increasing variety of jobs lead to some changes and the SOC2000 was substituted by the SOC2010. Both SOC2000 and SOC2010 have the same main groups, but subgroups in SOC2010 are slightly more nuanced (Great Britain et al., 2010): managers and senior officials; professional occupations; associate professionals and technical occupations; administrative and secretarial occupations; skilled trade occupations; sales and customer service occupations; process, plant and machine operatives; and elementary occupations.

Grades. Participant's grades were reported by their teachers through a 5 point Likert-scale ranging from "well above the average" to "well below the average". We used English, Mathematics, and Science grades at age 7 and 11.

Table 3.1
Descriptive statistics of the variables

	Mean	SD	Min.	Max.	Skewness	Kurtosis
Original Sample						
Grades English Writing (Age 7)	3.04	1.00	1.00	5.00	-.08	-.40
Grades Mathematics (Age 7)	3.24	.95	1.00	5.00	-.20	-.18
Grades Science (Age 7)	3.23	.81	1.00	5.00	-.10	.59
Grades English (Age 11)	3.34	1.02	1.00	5.00	-.24	-.39
Grades Mathematics (Age 11)	3.41	1.01	1.00	5.00	-.23	-.39
Grades Science (Age 11)	3.37	.87	1.00	5.00	-.17	.09
Occupational Aspirations (Age 11)	6.91	1.37	1.00	9.00	-1.64	2.63
Occupational Aspirations (Age 14)	7.17	1.31	1.00	9.00	-1.73	2.78
Composite Scores						
Grades sum scores (Age 7)	.00	.92	-2.39	2.00	-.17	.08
Grades sum scores (Age 11)	.00	.93	-2.47	1.70	-.24	-.17
Grades difference	-.01	.69	-4.48	2.65	-.05	.64
Occupational Aspirations difference	-.01	1.23	-5.52	5.73	.04	2.72

Notes. SD= Standard Desviation; Min.= Minimum possible value; Max.= Maximum possible value.

3.3 Analysis

Our independent variable was Verbal Ability at age 7 and our outcome variable was Occupational Aspirations change between ages 11 and 14. Both measures were standardized. Our mediator variable was change in Grades between ages 7 and 11. The overall grade at each wave was computed by standardizing (z-score) and averaging three Grades: English, Mathematics, and Science. At age 7 there were different grades for English writing, reading, and listening. We chose the English writing for having the strongest correlation with the other two subjects (r=.69-.73) compared with the listening and reading English (r=.66-.72).

We computed change in both Occupational Aspirations and Grades as the *Difference Scores* between two waves. On the one hand, *Difference Scores* are an intuitive way of computing change, they also have the advantage of controlling for non-accounted variables which might affect measured variables at both time points. On the other hand, *Difference Scores* are not

as reliable as other measures of change, for they absorb any measurement error embedded in each wave (see Chapter 2 for more details on *Difference Scores*).

As Figure 3.1 shows, the mediation model we tested involved a direct path (c) from Verbal Ability at age 7 to the change in Occupational Aspirations that happened between ages 11 and 14; as well as an indirect path from Verbal Ability at age 7 to Occupational Aspiration difference via Grades change between ages 7 and 11 (path $a * b$)[1].

In all our analyses we included the sample weights developed by Fitzsimons (2017) to compensate for demographic stratification and clustering in the MCS sampling.

3.4 Results

We used M*plus* version 8.3 (Muthén and Muthén, 1998) for our analyses. The bivariate relation from Verbal Ability to change in Occupational Aspirations was significant ($beta$= .03, p<.05). The strength of the relation was very weak so it might have reached significance due to the large sample size.

We then included the moderator, change in Grades, along with the indirect pathways – i.e. $a * b$. The indirect path was not significant ($beta$= .00, p=.32), meaning that change in Occupational Aspirations was not due to the moderating effect of Grades change. Additionally, the direct path (c) between Verbal Ability and Occupational Aspirations change was still significant. This implies that Grades change between ages 7 and 11, did not intervene in the relation between Verbal Ability and Occupational Aspiration change.

[1]Our original aim was to fit a latent difference score model to assess changes in both Grades and Occupational Aspirations. However, we encountered a number of issues in model estimation that will be further considered in the Discussion section.

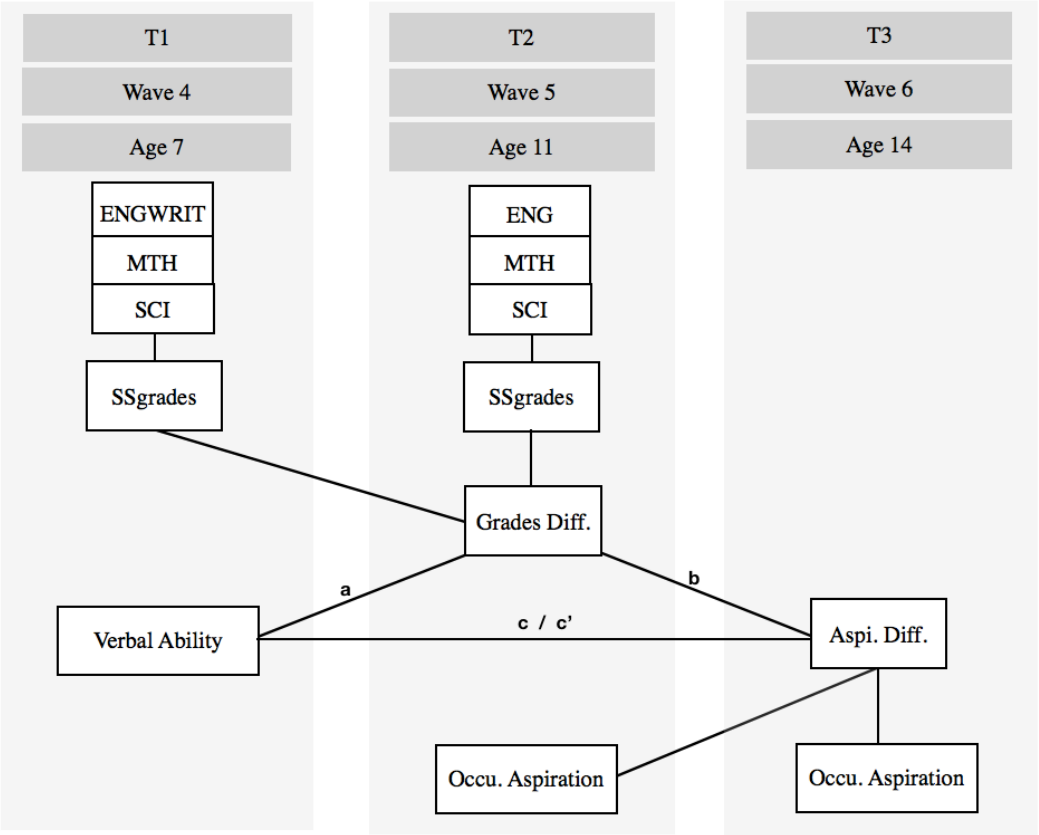

Figure 3.1
Mediation model of grades' change between verbal ability and occupational aspirations' change.

Notes: This diagram shows the mediating role (i.e. indirect path=Path a*Path b) of Grades change between the Verbal Ability and Occupational Aspirations change (Path c). ENGWRIGH= English writing grade; ENG= English grade; MTH= mathematics grades; SCI= science grade; SSgrades=grades sum scores; Grades Diff. = difference scores between both grades' sum scores; Occu. Aspirations= occupational aspirations.

3.5 Discussion

Several theories like Gottfredson's theory of Circumscription and Compromise, Holland's vocational interests theory and Lent and colleagues Social Cognitive Career Theory (SCCT), have explored the factors influencing students' occupational aspirations. The role of cognitive

abilities and grades is supported by the three theories. Furthermore, research evidence clearly support their hypotheses. This study aimed to test the mediating role of grades' change between verbal ability and occupational aspiration change using three waves from a longitudinal study in the UK, the MCS.

This study showed that Verbal Ability was related to changes in Occupational Aspirations between ages 11 and 14. However, contrary to what was hypothesized, this relationship was not moderated by changes in Grades between ages 7 and 11. The original aim was to operationalize change in both, Grades and Occupational Aspirations, through latent change scores which are more reliable than a raw difference score (Castro-Schilo and Grimm, 2018). M*plus* and R were used to fit these models, using a variety of different model identification constraints and starting values. However, all models encountered issues with the covariance matrices related to either of the change measures. Further exploration of the *Difference Scores* descriptive statistics suggest that there was comparatively little change between these waves, with the associated issues relating to low levels of variation in change.

Lack of change in Occupational Aspirations could be reflecting lack of change across time, which would be consistent with Low et al. (2005) metanalysis. However, it could also be due to poor reliability in both grades and aspiration measures. As thoroughly discussed in Chapter 2, *Difference Scores* tend to be more unreliable because they include the measurement error from both waves. A larger proportion of error in our Occupational Aspirations or Grades measures might have limited our power to detect a correlation between these two. Additionally, as some experts argue, changes in occupational aspirations might be better reflected through knowledge fields (e.g.: STEM vs. Non-Stem) rather than the codes that were used which represented occupational status (Gottfredson, 1979). On the other hand, it is also possible that the Occupational Aspirations categories were too broad, making it difficult to appreciate subtle changes. This issue could have also be applied to Grades.

Moreover, it is possible that there was little actual change in either aspirations or grades. For instance, primary school requirements are low so students might tend to have stable grades.

Similarly, students' occupational aspirations might stable because they are too young to be aware of how bast is the labor market.

3.5.1 Strengths and limitations

As it happened in Chapter 2, a strong limitation of this study was the number of waves. Having more data points might have allowed testing a more complex the model. Additionally, verbal ability might not have been a sufficiently accurate proxy for general ability. Another limitation was the lack of information about the reliability of the measures. Hence, it is difficult to estimate to what extent the lack of evidence for the mediation model was due to lack of lack of reliable change or lack of change all together.

On the other hand, the main strength of this study is not only the large sample size, but that it expanded over 7 years, from age 7 to 14. The MCS is not only a large and representative sample, it is also one of the most up-to-date longitudinal studies covering from childhood to teenage years. Therefore, the MCS is a relevant database for current educational research and it would even allow to extend this research beyond age 14 by using the recently released age 17 wave. Additionally, this study provided a new avenue to test Marsh and colleagues work regarding how feedback experiences can shape students' opinions of themselves. Lastly, this type of research is relevant because it explores how students conform their occupational aspirations which, in turn, can predict actual occupations latter in life (Schoon and Parsons, 2002; Schoon and Polek, 2011; Bergman et al., 2015). However, students might hold unrealistic aspirations (Helwig, 2001). Hence, it is crucial for vocational counselors to acknowledge the role of students' actual cognitive ability.

Chapter 4

A Brief Introduction to Personality Research

4.1 Origins of personality research

Research on intelligence and personality emerged simultaneously partially because they complement each other when predicting educational and occupational success. Hence, this thesis also incorporates two studies on personality differences across groups and occupations

Generally, personality is defined as "relatively enduring patterns of thoughts, feelings, and actions" (McCrae and Costa, 2008, p. 160). Ideally, we would open this introductory chapter with a more detailed definition. But, for reasons the reader will soon understand, tackling the specificities of this definition is precisely what ignites debate in the field of Personality research. After all:

> *"The task faced by the personality psychologist would seem more difficult than that faced by the early biologists. Biologists classify individual exemplars according to their attributes; in personality taxonomies, the exemplars are the attributes themselves"*
>
> — Jonh and Angleitner (1988, p. 172)

This chapter will start by reviewing one of the earliest attempts to understand personality, the *lexical approach*, by which researchers explored personality traits studying words used to describe people – e.g.: anxious, bubbly, etc. Then, we will move on to consider the currently dominant and related, *psychometric approach*, which developed more elaborate questions (i.e. items) to tackle personality traits. Both approaches benefited from the improvements in computer software and statistical methods – e.g.: Cluster and Factor Analysis (for more detail on these methods see Chapter 1). The lexical and psychometric approaches, developed in parallel but it was the effort to combine both that catalyzed the development of personality models and theories. Their strengths and weaknesses will be discussed along with some of the most relevant personality models that have been proposed. Towards the end of the chapter, we will dive into the latest personality models. Recent research has challenged

numerous core assumptions in popular personality models and suggested different analytic focuses based on the level of specificity in personality structure. This final discussion will help to frame the analysis methods used in this thesis.

It is important to highlight that this chapter does not aim to be an exhaustive record of personality research history, but rather, a *comprehensive enough* discussion on the most influential studies, as well as those relevant to the methodological framework used in this thesis. For those interested in dwelling on the history of personality research, I would recommend John et al. (1988) and Block (1995) reviews, for being comprehensive and balance in terms of the advantages and limitations of personality research.

4.2 First steps: the Lexical approach

It is often considered that F. Galton provided the first formal effort to conceptualize personality by developing the lexical approach (John et al., 1988; Revelle et al., 2013). The lexical approach is based on the idea that personality is a "salient phenomenon". Hence, relevant personality traits should sediment in natural language in the form of nouns, verbs, adverbs, and, especially, adjectives – e.g.: "anxious", "controlling", "bubbly", "curious", etc… (Allport and Odbert, 1936; Briggs, 1989; John et al., 1988; Matthews et al., 2009). In this way, single words, phrases, and idioms would become the "personality descriptors" at the base of the *personality taxonomy*.

The *personality taxonomy* was conceptualized as a system distinguishing, ordering, and grouping personality characteristics according to a set of rules describing how each aspect related to the rest (John et al., 1988; Norman, 1963). The first personality researchers prioritized a theory-driven personality taxonomy, for it facilitated the communication and accumulation of knowledge (John et al., 1988). As explained in Chapter 1, experts in the field aim to uncover such taxonomy through the lexical and the psychometric approach.

Following the ideas of the lexical approach, researchers started by reviewing the dictionary,

gathering, filtering, and subsetting the personality descriptors yielding the most parsimonious personality structure. Researchers thought that the personality descriptors found to be invariant across cultures would be at the core of human personality (Matthews et al., 2009). Allport and Odbert (1936) did one of the most relevant initial contributions to the lexical approach by selecting and categorizing around 18 000 words from the Webster's New International Dictionary:

> *"The criterion for inclusion consists in the capacity of any term to distinguish the behavior of one human being from that of another [...] In many cases the application of this criterion involved a considerable degree of arbitrariness"*
>
> — Allport and Odbert (1936, p. 24)

Other authors followed Allport and Odbert (1936) method and derived different sets of personality descriptors. This body of research provided interesting insights but was too exhaustive to inform personality theories. Moreover, adjectives were sometimes not sufficiently specific, other times they were too rare for participants to understand them. R. B. Cattell pioneered the idea of integrating the lexical and the psychometric approaches with the hope that each would buffer the other's disadvantages: the lexical approach would provide a relevant set of personality descriptors that could inspire the items created by the psychometric approach so they tapped on specific personality traits.

Cattell started with Allport and Odbert (1936)'s list of personality descriptors, he eliminated those terms he considered less relevant and he paired opposite concepts to create bipolar dimensions (e.g.: Cattell, 1947). Through several studies, Cattell run rudimentary Cluster Analyses and kept reducing his list of personality descriptors, first to 67-69 clusters, then to 35. Cattell factor analyzed these 35 clusters obtaining 12 factors, which he then used to create the 16 Personality Factors Questionnaire (16PF). At that point, Cattell had eliminated 99% of the original personality descriptors used by Allport and Odberg's without following an objective or systematic method (Block, 1995).

Cattell's research was praised for its novelty, but it was also very criticized for not being replicable (McCrae and Costa, 1985a; Grice, 2001). As it will explain in later sections, there are several reasons why all attempts to replicate Cattell's results failed (e.g.: type of factor extraction and rotation, sample types, factor indeterminacy, the threshold for factor loadings, etc.); of particular concern was the fact that he selected items idiosyncratically and was inconsistent when reporting the number of personality descriptors or clusters he subset (Block, 1995; Eysenck, 1991; John et al., 1988; Norman, 1963; Tupes and Christal, 1961).

Overall, the lexical approach was a steppingstone on personality research:

> *"By far the most general efforts to specify the domain of phenomena on which to base such a system have proceeded from an examination of the natural language"*
>
> — Norman (1963, p. 574)

The lexical approach has continued to be used, although now it is mostly focused on the personality structure of adjectives (e.g.: Saucier and Iurino, 2020). Nevertheless, the lexical approach has also received a lot of criticism. Firstly, as Allport and Odbert (1936) themselves recognized, there was a certain degree of subjectivity. Secondly, the lexical approach puts the focus on lay people's interpretation of personality, overemphasizing its expressive and evaluative elements which might not be of scientific use (John et al., 1988; Pervin, 1994). Conversely, there could be relevant aspects of personality that are not encoded into language: "[n]o one would imagine that an analysis of common English terms for parts of the body would provide an adequate basis for the science of anatomy; why should personality be different?" (McCrae and Costa, 1985b, p.711). Thirdly, there was often not a one-to-one translation of some personality descriptors into other languages. Additionally, some studies have shown differences in the degree to which adjectives code information about the individual – e.g.: Indian adjectives refer to personality traits based on actions and rely more on the social context where they occurred than on the person acting (Pervin, 1994).

4.3 Refined methods, refined models

As we already mentioned, the lexical approach provided evidence that personality was not only cognoscible, but it was deeply rooted in social interactions. However, its solutions were too complex to be systematically tested, let alone to be translated into a personality taxonomy. Certain statistical methods of dimensionality reduction like Cluster Analysis and FA were the perfect fit for this challenge, for they allowed researchers to *objectively* filter and categorize personality descriptors. Furthermore, FA allowed researchers to abstract a taxonomy-like structure from the data, as it yields hierarchical models where the observed variables laid at the bottom and latent constructs rose across the upper layers (see Chapter 1).

Following Cattell's attempts to reduce Allport and Odberg's original personality descriptors list, Tupes and Christal (1961) factor analyzed 8 samples which had been tested using different sets of personality descriptors, and found 5 analogous factors reoccurring in all but one of the samples: (a) Surgency, (b) Agreeableness, (c) Dependability, (d) Emotional Stability, and (e) Culture. According to Tupes and Christal (1961), their Surgency factor resembled what other authors had called "Extroversion", which was related to aspects like assertiveness, sociability, energy, and cheerfulness. Their Agreeableness factor was also similar to a trait previously identified relative to emotional maturity, mildness, cooperativeness, and trustfulness. Tupes and Christal (1961) thought that their Dependability factor was associated with characteristics like conscientiousness, perseverance, responsibility, and orderliness. Their Emotional Stability was close to what earlier researchers had called "Emotionality" and it included calmness, as well as not neurotic or hypochondriacal behaviors. Lastly, their Culture factor seemed equivalent to other already identified traits related to curious, cultured, imaginative, and independent-minded individuals. Slowly, these five factors gain popularity and started to be referred to as the *Big Five*.

Interestingly, two of Tupes and Christal's (1961) samples had been previously used by Cattel, who identified 12 rather than 5 factors (Norman, 1963). Tupes and Christal (1961) study

seemed to provide sound evidence that personality could be consistently mapped with just a few traits, yet their study was far from perfect. First, Tupes and Christal (1961) samples were mostly composed of young men, most of them working on the Air Force, so their results' generalizability was questionable. Second, Tupes and Christal (1961) guided their analysis to obtain as many factors as they did in the first sample, clearly introducing a strong bias in their results (Block, 1995). Regardless of the criticism, Tupes and Christal's (1961) study became a seminal piece, especially after Norman (1963) replicated their five-factor structure using the two samples that Cattel and Tupes and Christal (1961) had used. Norman (1963) was seen as a confirmation that personality could be accurately described through a handful of traits. However, as the next section will explain, these results are less outstanding when considering the FA's limitations.

Around the same time, H. J. Eynseck developed his "PEN" model – i.e. Psychoticism, Extroversion, and Neuroticism – which focused on the biological basis of personality traits (Eysenck, 1947, 2016; McCrae and Costa, 1985a). At that point, there was a considerable amount of evidence supporting the existence of traits related to Extroversion and Neuroticism (i.e. "Emotional Stability"), but no so much for Psychoticism (Eysenck, 2016; McCrae and Costa, 1985a). In the 70s, P. Costa and R. McCrae started a life-long collaboration to uncover the best personality model, which rose them as two of the most prominent psychology researchers of the 20th century. Costa and McCrae (1976) started by cluster analyzing the results of a large sample of men who had completed Cattell's 16PF and found that two analogous factors rose in the three age groups: "Adjustment-Anxiety" (which was reminiscent of Tupes and Christal (1961)'s "Emotional Stability", and which was later called "Neuroticism", N) and "Introversion-Extraversion" (later labeled "Extraversion", E). The third cluster differed significantly across groups: in the youngest group, this factor related to "openness to feelings and aesthetic sensibility", in the middle-aged group, it was closer to "openness to new ideas and values", and, lastly, in the older group this factor had more to do with openness to "affective and cognitive experiences". Costa and McCrae (1976) argued that across the three groups, this cluster reflected a version of openness to "cognitive and

affective" experiences, hence they labeled it "Openness", O. These three constructs formed the basis of P. Costa and R. McCrae's "NEO model".

Costa and McCrae (1976) noted that the Adjustment-Anxiety cluster resembled Eynksen's Neuroticism trait, and their Extraversion cluster was analogous to Eynksen's Extraversion trait. However, their Openness factor remained controversial and they made specific efforts to refine its conceptualization. Costa and McCrae (1978) factor analyzed several scales and inventories – e.g.: the Experience Inventory, the scales of Fantasy, Aesthetic experience, and abstract ideas – together with other three self-generated scales: "Feelings", "Actions", and "Non-dogmatic values". According to their results, Openness was a solid domain, equivalent to factors found in other models like Norman's, as well as Tupes and Christal's "Culture" factor, or Zukerman's New Experience subscale of Sensation Seeking (McCrae and Costa, 1985a,b).

In the mid-1980s, P. Costa and R. McCrae kept carrying out studies unifying the lexical and the psychometric approach by cluster or factor analyzing adjectives and items in questionnaires like the 16PF, EASI-III, and the Experience inventory, together with a list of other personality descriptors (McCrae and Costa, 1985a,b). Their findings showed that, across questionnaires, all solutions overlapped with their NEO measure (McCrae and Costa, 1985a,b). Shortly after, they expanded their model by adding two new domains "Agreeableness" and "Conscientiousness", giving rise to their NEO-PI model, which would later become the NEO-PI-R after some items were updated (Costa et al., 1991). Costa et al. (1991) devised the Agreeableness and Conscientiousness factors *ad-hoc* to match the results from previous research that had consistently identified them. Altogether, the NEO-PI-R measures 5 domains, each subsuming 6 facets, which are conversely measured through several items: Neuroticism (composed of the following facets: Angry Hostility, Impulsiveness, Depression, Self-Consciousness, Anxiety, and Vulnerability); Extraversion (Positive Emotions, Assertiveness, Gregariousness, Excitement Seeking, Activity, and Warmth); Openness to Experience (Values, Ideas, Feelings, Actions, Aesthetics, and Fantasy); Agreeableness

(Trust, Tender-Mindedness, Modesty, Compliance, Altruism, and Straightforwardness); and Conscientiousness (Deliberation, Discipline, Dutifulness, Achievement Striving, Order Self-, and Competence).

Eysenck was very critical with both Agreeableness and Conscientiousness as they had been developed atheoretically and without "solid empirical evidence" (e.g.: Eysenck, 1992). Additionally, given that both were strongly negatively correlated with Eynseck's Psychoticism, he argued that both should be subsumed into the latter (Eysenck, 1992; McCrae and Costa, 1985a). Conversely, McCrae and Costa (1985a) argued that Psychoticism should be considered part of Extroversion.

4.3.1 Reviewing the Big Five

By the end of the 20th century, personality researchers were frustrated with the ongoing debate on what factors should compose the definite personality model (Revelle et al., 2013). Eventually, P. Costa and R. McCrae's Five-Factor model became the most widely used personality model; partly due to the prolific and convincing evidence they provided, and partly because it was in the field's interest to use a single model across studies to guarantee systematicity and cross-research comparisons. Nevertheless, the Big Five has never been declared *The* personality model (Revelle et al., 2013).

There are, in fact, numerous advantages to the Big Five. Firstly, it is comprehensive: the NEO-PI-R has been shown to capture personality traits measured by many other personality models, scales, and questionnaires devised by experts using different approaches like Cattell, Eysenck, Wiggins, Murray, the Jungian Myers-Briggs Type Inventory, and the Occupational Personality Questionnaire (Matthews et al., 2009; McCrae and Costa, 1985b). Secondly, it is concise, as it summarizes the content of several other models in a manageable number of factors (McCrae and Costa, 1985a,b). Thirdly, the Big Five as constructs – rather than as specific scales in a questionnaire – are temporally stable, valid across observers, acknowledged for the layperson and personality theorists, as well as present across sexes, ages, and cultures

(Costa and McCrae, 1992a).

Despite all these assets, Big Five –as a questionnaire and as constructs –have received considerable criticism. Among them, it is often highlighted that, as the Big Five did not source from a personality theory, there were no *a priori* predictions on how these factors would relate to perceptions, cognitions, and behaviors. Hence, the Big Five have often been defined as "mere descriptions" rather than a "full model of personality" (Block, 1995). This debate over the relevance of theories will be further discussed in section 7.3.2. Additionally, the factors in the Big Five are said to be independent but, in reality, they are all interrelated (Block, 1995). Moreover, the convergence across measures of the Big Five has often been overstated, as Briggs (1989) argued, "whereas there is a general resemblance in the factors that replicate from sample to sample and investigator to investigator, the resemblance is more fraternal than identical" (p. 24). This means that, even if different tests aim for similar traits, their scores might not be interchangeable (Matthews et al., 2009).

In addition, there were substantial limitations inherent to the statistical tools used which were often dismissed, even though they had a pervasive impact on research results. For instance, FA yields hierarchical structures, but this might not be the best fitting personality structure (Booth and Murray, 2018). Also, FA relies on correlations and, as such, it also depends on: the error of measurement (i.e. reliabilities), how much do variables overlap (i.e. collinearity), and how varied is the sample – i.e. range restrictions (Block, 1995). These three factors were not always considered before running the analyses. As we have mentioned, on the one hand, researchers tended to pre-screen the variables they included. On the other hand, they often used samples mostly composed of young white men. Another notorious limitation of FA is "factor indeterminacy", which means that the FA solution is not definite for the unknown parameters to compute the solution surpasses the number of known parameters. Therefore, there is an infinite number of solutions compatible with the pattern of relations between the observable variables and the latent variables (for a full description see Grice, 2001). Lastly, the number of factors that should be extracted from a FA

is not unequivocal, but rather a subjective decision (Block, 1995; Booth and Murray, 2018). Therefore, is not just the number and kinds of variables that determine the factor solution but also the number of factors the researcher decides to extract. Different measures have been proposed to limit the impact of these limitations, however, most cannot be tackled, only acknowledged (Block, 1995; Booth and Murray, 2018).

Furthermore, there were also issues related to how researchers used FA. For instance, the FA solution inexorably depends on the variables it is fed, which is often a subjective decision (Block, 1995; Booth and Murray, 2018). Consequently, the chances of finding the same factors structure are higher when studies used the same variables. Thus, the convergence between Tupes and Christal (1961), Norman (1963), and Costa and McCrae (1976) results – which were seminal pieces on which large personality models and theories were built – might not be as impressive as it was originally thought (Booth and Murray, 2018). Additionally, it was a common practice at the time to select variables that overlapped and which were already known to cluster together before running the analysis (i.e. "pre-structuring"). This practice casts doubts on the reliability of the findings, as researchers could decide the structure of the factor solution before running an analysis (Block, 1995; Booth and Murray, 2018). If researchers were still not satisfied with the results, they could use different types of axis rotation to force the variables to be clustered closer or further away from each other, determining whether they remained correlated (i.e. oblique rotation) or independent – i.e. orthogonal rotation (Block, 1995; Booth and Murray, 2018). Block (1995) argued that axis rotation is precisely what guarantees factors' independence as they are, in fact, correlated: "N and C scales now correlate -.53 and the E and O scales now correlate .40, both of these figures being uncorrected for attenuation" (Block, 1995, p. 206).

4.3.2 Summary: Too many models, too much uncertainty

As we have reviewed in the previous section, during the 20th century there were several competing personality theories, some derived from the lexical approach (i.e. which focused on deriving personality descriptors for natural language) and from the psychometric approach

(i.e. based on the creation of personality items). Both research lines were aided by statistical tools like Cluster and Factor Analysis. Among the most influential models, there were: (1) Cattell's 16 factors which were derived using a highly modified version of Allport and Odberg's original personality descriptors list; (2) Eynseck's PEN (i.e. Extroversion, Neuroticism, and Psychoticism) model; and (3) P. Costa and R. McCrae's Five-Factor model (i.e. Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness) which were derived factor analyzing Cattell's 16 factors.

All these models had important weaknesses. On the one hand, most were derived from Cattell's list of personality descriptors and clusters, which he had chosen idiosyncratically. On the other hand, they carried the strong limitations of the statistical techniques with which they were drawn, even if these were downplayed or ignored– e.g.: factor indeterminacy, collinearity, subjective factor extraction, etc. (Block, 1995; McCrae and Costa, 1985b). These models also had strengths but there was no way of determining which offered the best trade-off (and, for sure, none of these prominent researchers would have wanted to give in after a lifetime of commitment to their models):

> *"How does one conclude that a description is 'reasonably sufficient' or 'comprehensive' or 'fully adequate' ?"*
>
> — Block (1995, p. 187)

This question has haunted personality research ever since (Eysenck, 1991; John et al., 1988; Matthews et al., 2009; Mõttus et al., 2020; Pervin, 1994), slowing down the progress of personality assessment (Goldberg, 1981; McCrae and Costa, 1985a). Hence, the latest definition of personality traits is not much more nuanced than the one that opened this chapter:

> *"[W]e define traits as dimensions of any kind of relatively stable psychological (affective, cognitive, motivational and behavioral) differences among people, independent of their content, breadth, or expected importance"*

— Condon et al. (2021, p. 926)

4.4 New approaches to personality research

As it happens, time and distance allowed researchers to reframe the number-of-factors conundrum. And, as it happens, new generations of researchers have started appealing for a change in paradigm. Some endeavours have focused on personality structure, others on improving personality measures.

4.4.1 New approaches in terms of structure: non-hierarchical and network structures

As previously discussed, personality theories are mostly based on hierarchical models (Condon et al., 2021). Hierarchical models have traditionally been preferred because of the extraction methods that were used and because experts have prioritized defining the structure of personality (Cattell, 1943; Eysenck, 1991, see Chapter 1).

Condon et al. (2021) argued that hierarchical models have three main weaknesses. First, existing personality hierarchies contain jingle/jangle fallacies as it is not infrequent for different aspects to have similar names, and for similar aspects to have different names. As Block (1995) once said when describing Agreeableness, it is a "[b]road, bland, impressively unincisive umbrella of a label" (Block, 1995, p. 194). Second, current hierarchical models do not guarantee that every level of aggregation is exhaustive or comprehensive enough. Third, hierarchical models discard items' unique variance and recent evidence shows that it is as relevant as items' common variance to capture personality traits.

Loehlin and Goldberg (2014) re-analyzed two samples, one that had been tested using an adjective list and another that had completed several personality inventories. Loehlin and Goldberg (2014) explained that hierarchical structures have a salient structure in which new layers are generated when the factors in previous layers split. On the other hand, in list-

like organizations, factors in each layer should be mostly unrelated to those in contiguous layers. This list-type organization is precisely what their results showed, suggesting that the hierarchical structure of the Big Five might not reflect personality as accurately as it has been claimed.

In spite of these criticisms, advocates of hierarchical models argue that they have the great advantage of reducing redundancy hence, being more parsimonious (Condon et al., 2021; Mõttus et al., 2020). Yet, parsimony is not the only factor that should be considered. In fact, broadening our evaluation criteria might "help resolve the apparent dilemma of parsimony versus comprehensiveness" (Saucier and Iurino, 2020, p. 1188). This change in paradigm is not capricious, but rather a consequence of the computer revolution. Thanks to the increase in computational power researchers can (1) easily collect larger amounts of data, (2) safely store big datasets, and (3) effortlessly carry out complex data analysis. These improvements have allowed to test more complex personality models and to rise the standards when assessing their suitability.

Cramer et al. (2012), for instance, proposed a personality network model (see Chapter 1) and built an R-package, *qgraph* (Epskamp et al., 2011), to test it. In Cramer et al. (2012) model, each personality trait is composed of effective, cognitive, and behavioral components. And it is the interaction between these components across an individual's lifespan that gives rise to personality traits. For instance, Neuroticism is not a latent cause but, rather, "groups of tightly inter-connected components". Consequently, traits that are conceptually synonyms (or antonyms) should co-evolve – e.g.: increasingly "enjoy being accompanied" might cause an increase in "liking going to parties", together with a decrease in "disliking people". Interactions between individuals and the environment should further shape personality traits. Cramer et al. (2012) results showed that items were more highly interconnected than what would be expected had they been caused by independent latent traits, suggesting that their network model reflected the structure of personality traits better than hierarchical models. Network models are becoming increasingly popular, recent efforts have even been made to

devise ways of comparing network and hierarchical models (e.g.: Kan et al., 2020).

4.4.2 New approaches in terms of focus:

Recently researchers have started questioning whether it is worth to continue investing resources in uncovering the *right* personality taxonomy over other aspects, like improving its predictive power[1] (e.g.: Condon et al., 2021; Mõttus et al., 2020). Research prioritizing predictive power should focus on more granular levels of analysis by incorporating the items' information that is lost in higher levels of aggregation –i.e. unique variance – thus, moving towards a "bottom-up" approach (Mõttus et al., 2017a, 2020). The next two chapters serve as an example of how this new approach could be applied in personality research, as well as the benefits of employing it.

In the Big Five model, the lowest level of aggregation are nuances, which have been formally defined as:

> *"[T]he lowest level at which patterns of responses to items continue to have reliable specific variance"*
>
> — Condon et al. (2021, p. 6)

Condon et al. (2021) proposed that to achieve a high-quality collection of nuances we should, first, identify a sufficiently comprehensive item pool like Goldberg (1999)'s International Personality Item Pool (IPIP) which includes thousands of freely available items that researchers have been testing, translating, and updating. According to Condon et al. (2021) researchers should aspire to collect large amounts of items while embracing that it might be impossible to procure an exhaustive pool. Second, Condon et al. (2021) recommended that the quality of the items should be assessed using a wide range of techniques including qualitative evaluations through focus groups and experts' discussions, as well as quantitative evaluations

[1] The predictive power of a single variable (or a group of variables) refers to their capability of making accurate predictions on other variables or future outcomes – e.g.: a person's behaviors, occupational outcomes, etc.

of item's clarity, brevity, and readiness. The deprecated items should be recorded together with their properties and the reasons why they were discarded. Third, items should be selected according to their inter-rater[2] and retest reliability[3] as they would inform on its validity[4] and temporal stability. Fourth, item pools and participant information should be aggregated so items' behaviors across samples can be compared and improved. Fifth, final item selection should be done according to their their predictive power.

Such a demanding quality control process on single items would reduce their measurement error, improving the item's reliability. Thus, following Condon et al. (2021) definition, such items should roughly correspond to nuances (McCrae, 2015; Mõttus et al., 2020).

So far, personality nuances have been operationalized in two ways: (1) a single item's residual variance (e.g.: Mõttus et al., 2014, 2017b) and (2) a single item's total variance (e.g.: Achaa-Amankwaa et al., 2020; Mõttus et al., 2017a) and. The former account implies that nuances would equate to the variance being left after removing whatever variance that was absorbed into facets and domains. Hence, the item's residual variance refers to an item's unique variance plus the measurement error. Although an item's residual variance would contain a higher proportion of measurement error (and, hence, it is likely to yield lower correlations with external criteria – e.g.: cross-rater agreement compared to items, facets, or domains) it might still carry relevant information about personality traits (Mõttus et al., 2014). However, the amount of relevant information left in an item's residual variance would depend on what variance is removed from the item. Following the latter definition, nuances would be equivalent to single items, which would include item's unique variance together with the common variance they share with other items (which would be absorbed into facets and domains), as well as the measurement error. Note that this is how nuances have been conceptualized in this thesis.

[2]Inter-rater reliability is the degree to which two or more people – i.e. raters – agree when labelling or ranking specific materials. In this case, we refer to the degree of agreement over what items should be subsumed under what facets/domains.

[3]Retest reliability measures whether an item (or a questionnaire, a scale, etc.) captures a construct consistently across time.

[4]An item's or test's validity refers to its capability to reflect the construct they aim to measure.

Research has shown that personality scores' predictive power increases considerably when using several items together without aggregating them[5]. It could be argued that this improvement in predictive power is merely due to the increased number of predictors. For instance, Achaa-Amankwaa et al. (2020) used personality scores at different levels to predict participants' nationality; they correctly identified 60% of participants when using domain scores, 75% if using facets, and 89% when using nuances. However, researchers supporting this approach argue that the predictive power increase is due to the relevance of the unique information provided by each facet or nuance which has always been discarted (Seeboth and Mõttus, 2018). Moreover, research using nuances, either operationalized as items or as item's residuals, is limited but there is evidence suggesting that they contain genuine information about personality traits (McCrae, 2015; Mõttus et al., 2014, 2020; Seeboth and Mõttus, 2018), and they share domains' properties of cross-cultural presence, temporal stability, observability across raters, and genetic basis (e.g. Achaa-Amankwaa et al., 2020; Mõttus et al., 2014, 2017b).

There is still a more intuitive advantage to using facets and nuances beyond the *techinical* rationales we have reviewed: domain scores might obscure patterns at lower levels of aggregation (Booth and Irwing, 2011; Del Giudice et al., 2012; Kaiser, 2019; Mac Giolla and Kajonius, 2019). This could become particularly relevant when comparing different populations (e.g.: men and women) as contradicting results at nuance level might translate as "no differences" at domain level. This example will be further explored in Chapter 5. From this dicussion we can conclude that, although this fine-tuned approach is not interested in developing a personality taxonomy, it does provide more solid foundations as it samples the "persome – the universe of variables capturing personality variability – more broadly than currently available measures do" (Mõttus et al., 2020, p. 3).

[5]Seeboth and Mõttus (2018) explained this concept by comparing nuances with single nucleotide polymorphisms (SNPs), because individually, nuances have small predictive power but combined into into polygenic scores (i.e. "polynuanced scores") have a strong predictive power.

4.5 Limitations to a bottom-up approach

Nuances research is promising but is also facing a considerable amount of resistance. In this section, we will focus on the main criticisms and myths surrounding this new approach.

4.5.1 Why would we trust single items if they are much more unreliable than scales or questionnaires?

As McCrae (2015) explains, classical approaches to scales construction ensured that questionnaires had high internal consistency[6]. This was often attained by selecting items with overlapping content. Therefore, classic personality measures had high internal consistency but did neither guaranteed that constructs were being measured comprehensively nor that they had a strong predictive power.

Experts are increasingly inclined to change this paradigm of developing and assessing personalty measures so that lower levels of aggregation are properly measured and prove useful to predict external criteria. McCrae (2015) proposed that individual items should no longer be chosen according to the percentage of information they share with others but to their ability to measure the same construct consistently across time (i.e. retest reliability). This approach does not only ensures that each nuance represents accurately a specific portion of a personality trait, but also that all personality traits are thoroughly mapped. Additionally, these personality measures would have higher predictive power because each item would capture a unique aspect of the external criteria we aim to predict (Chapman et al., 2016).

4.5.2 Using too many items would just overfit the data, models will not be generalizable...

Doubts on the model's generalizability are based on the idea that the more predictors are feed into a model (or the higher is the ratio of observations to predictors), the more likely

[6]Internal consistency measures the degree to which all items included in a scale, inventory, or questionnaire tackle the same construct.

it is that they will pick up on variance from the specific sample at hand. This phenomenon is known as overfitting (Chapman et al., 2016; Yarkoni and Westfall, 2017). In overfitted models, the values of parameters' estimates are closer to those of the sample at hand than to the true population values, jeopardizing the model's generalizability.

Overfitting seems unavoidable when using hundreds of nuances to predict a single outcome. Nonetheless, a model's overfitting can be tackled through *k-fold cross-validation*, where the sample is split into several samples of equal size and the parameter estimates of the model are computed in all the samples except for one, which reserved to test the model's generalizability (for more details see Yarkoni and Westfall, 2017). Several iterations of this process allow to fine-tune the parameters' estimates while testing for generalizability and assessing the model's error. Additional precautions can be included by dividing the original sample into a training sample with which to do the k-fold cross-validation, and a testing sample to examine the performance of the final model (Chapman et al., 2016). Note that cross-validation tackles generalizability within the sample being used. Thus, although it prevents extreme overfitting, there is no guarantee that the model might still not generalize to another population.

Although cross-validation informs about the degree to which the model is overfitting the data, it might not prevent it. There are different approaches to avoid overfitting, like: Supervised Principal Components Analysis (SPCA), Regularization, and Boosting (see illustrated examples in Bair et al., 2006; Chapman et al., 2016; Yarkoni and Westfall, 2017). Therefore, there are methods to tackle data overfitting[7].

4.5.3 There would be an excessive number of items to choose from and cross-study comparisons would not be possible...

Interestingly, the convergent validity across questionnaires supposedly measuring the Big Five is lower than expected, in fact, they only shared 36% of their variance (Pace and

[7]Given that in this thesis personality scores are not used as predictors, overfitting will not be an issue. However, this is a relevant limitation of this methodological approach that must be acknowledged and discussed.

Brannick, 2010). This means that results using different Big Five measures are not as comparable as most researchers would want. This issue has been ignored because tests have not often been compared (Block, 1995; Goldberg, 1999).

The objective of this new approach to personality measures is not that all studies test all nuances, but that studies focus on finding the subset of nuances that consistently predict the same outcome. Only these nuance subsets would be "theory-worthy". This nuance-hunt would be more efficient if researchers shared and updated their items on the same site (Condon et al., 2021). Arguably, having all items' information in the same place would help researchers find the items that suit their research question best.

4.5.4 How would a collection of independent and unrelated items be translated into a personality theory?

Theories are valued for providing a falsifiable framework which can further guide research (Johnson and Bouchard, 2005). However, theories' relative worth also depends on the extent to which they lead to accurate predictions (this issues will be further discussed in Section 7.3.2). For this reason, good theories (regardless of the field) need to be drawn from robust patterns emerging from various sources. Eronen and Bringmann (2021) recently asked to what degree psychologists have accumulated enough knowledge to produce a high-quality theory of personality. According to Eronen and Bringmann (2021) we not only need evidence on what are the best personality traits to predict certain outcomes, we also need to know which are not. Focusing on items' predictive power will not only provide more information about the usefulness of that item, it would also help inform personality theories.

Arguably, personality theories and models often underperform when predicting life outcomes because they were not built for that purpose (e.g.: Mõttus et al., 2017a). In fact, the properties of many psychological measures are not excursionized enough, hence many measures are not as good as they claim to be (Eronen and Bringmann, 2021; Goldberg, 1999). Evidently, using these measures would lead to poor theories even if psychologist were to collect large

sets of data. A suitable alternative could be based around the items used to measure personality. This approach would also facilitate keeping personality theories updated, as it is easier to recognize the need for a change and to adopt it at the base of the hierarchy than at the top (Condon et al., 2021; Eronen and Bringmann, 2021). This nuanced approach does not guarantee to (nor did it ever intend to) find the perfect personality taxonomy but allows to sample the personality universe that current measures have not been able to (Mõttus et al., 2017a).

Lastly, advances in computer software able to explore more complex personality theories should be used to avoid simplifying its nature[8]. It is important to highlight that this "bottom-up" approach it is not a matter of mathematizing psychology, but of increasing the quality of basic research.

4.6 Conclusion

As an overall conclusion, personality is a collection of stable individual characteristics that are likely to be partially shaped by the environment as well as a person's genetic makeup (Matthews et al., 2009). Previous research using the lexical approach has shown that "individuals in a culture hold preexisting ideas about what types of attributes go together, and that the five-factor solution is a reflection of these culturally defined conceptual similarities" (Briggs, 1989, p. 249). However, the Big Five are not the only way in which we can conceptualize personality traits. In fact, there are (there have always been) several competing models organizing traits and their relationships differently– e.g.: as a hierarchy, a network, etc. Previous research has been limited by material contingencies like the amount of data that could be collected or the time it took to collect it and to process it. However, current advances allow researchers to study larger sets of personality markers that can be evaluated to guarantee their accuracy and predictive power.

[8]This is not to say that theories should be exclusively built on the basis of statistical software outputs, but as new statistical tools allow us to explore the suitability of personality models beyond the hierarchical structures (e.g.: network models, Cramer et al., 2012), we should consider all possible options.

These recent advances have given rise to a "bottom-up" approach to personality research, meaning creating theories focused on its most basic building blocks – i.e. personality nuances. Such an approach does not only guarantee that personality models would be optimized according to their predictive power, but that they would be easily updated. Nevertheless, this approach has found some resistance. The chapters that follow will embrace this debate, as they explore and compare results at different aggregation levels when using personality measures on topics like sex differentiation[9] and occupational choice. Hence, the present chapter helps to frame the methods used in the next two chapters, as both use single items as nuances and average sum scores of nuances as facets.

[9]Sex differentiation refers to the process by which men and women develop physical and psychological characteristics that distinguish them.

Chapter 5

Personality differences between sex and sexual orientation groups

5.1 Introduction

Personality differences between men and women have been thoroughly studied. Research suggests that sex differentiation happens on a continuum, hence it is an oversimplification to study it using just two categories – i.e. men and women. However, sexual orientation might map this spectrum better, yielding a more detailed picture of personality differences between men and women (Lippa, 2001). Research using sexual orientation is limited but provides evidence that personality scores of homosexual men tend to move away from their heterosexual counterparts toward heterosexual women. Similarly, personality scores of homosexual women tend to move away from heterosexual women toward heterosexual men. Additionally, it has been observed that personality shifts in homosexual participants tend to mirror personality differences between heterosexual men and women. This finding would support the idea that personality sex differences actually occur on a continuum.

Before discussing previous findings, it is important to recognize the distinction between sex and gender. According to the World Health Organization (WHO), gender is defined "the characteristics of women and men that are socially constructed" while sex "refers to those [characteristics] that are biologically determined (WorldHealthOrganization, 2022). Given how early in life children are socialized, the impact that sex and gender might have on other variables become inextricably entangled. It could be argued that certain personality differences are rooted in specific biological strata, but some might be heightened (or simply changed) according to our socializing experiences. These nature-nurture interactions might be particularly relevant when considering sexual orientation. Although theoretically, the terms "sex" and "gender" reflect very different ideas it is difficult to differentiate them in practice. To acknowledge the complexities embedded in this distinction between sex and gender, the term "sex/gender" will be used hereon. Furthermore, "men and women differences" should be also interpreted as "sex/gender differences".

5.1.1 Personality differences between men and women

As mentioned, sex/gender differences are present at an early age at different biological (e.g.: testosterone levels, prenatal exposure to androgen levels, contrast sensitivity, pain sensitivity) and psychological levels (e.g.: responsiveness to social cues and emotional expressions). Some of these differences might be the predecessors of wider personality sex/gender differences later in life (Alexander and Wilcox, 2012). For instance, girls being more sensitive to social cues may explain why women tend to score higher in Agreeableness.

Personality differences between men and women have been thoroughly studied. It has been argued that, on average, differences are small (Hyde, 2005). However, the percentage of overlap between both groups varies substantially across populations and it is especially large among developed countries (Costa et al., 2001; Kaiser, 2019; Mac Giolla and Kajonius, 2019; Schmitt et al., 2008). Charles and Bradley (2009) argue that personality sex/gender differences heightened in developed countries due to the rise in postmodern values promoting self-expression and the promotion of woman's university enrollment coupled with their increased participation in the labor market.

Research on the Big Five converges in showing that women score higher in Neuroticism and, to a lesser extent, in Extraversion and Agreeableness. Results on sex/gender differences on Conscientiousness and Openness to experience (Openness, for short) remain unclear. Specifically, women have been found to score higher in Extraversion in samples across the world, with effect sizes ranging (d_{range}) from .06 to .29 (Costa et al., 2001; Heineck, 2011; Kajonius and Mac Giolla, 2017; Lippa, 2005, 2008, 2010; Mac Giolla and Kajonius, 2019; Schmitt et al., 2008). Sex/gender differences in Agreeableness favoring women are slightly larger and have also been reproduced in cross-country samples – i.e. d_{range}= .15 - .60 (Costa et al., 2001; Heineck, 2011; Kajonius and Mac Giolla, 2017; Lippa, 2005, 2008; Mac Giolla and Kajonius, 2019; Schmitt et al., 2008). The largest sex/gender difference across studies tends to be in Neuroticism – i.e. d_{range}=.36 - .51 (Costa et al., 2001; Heineck, 2011; Kajonius and Mac Giolla, 2017; Lippa, 2005, 2008; Mac Giolla and Kajonius, 2019; Schmitt et al.,

2008). Sex/gender differences in Conscientiousness favor women in some studies – i.e. d_{range}= .04 - .17 (Heineck, 2011; Kajonius and Mac Giolla, 2017; Lippa, 2005; Schmitt et al., 2008; Mac Giolla and Kajonius, 2019), and men in others (Kajonius and Mac Giolla, 2017), but not all studies have found significant differences (e.g.: Costa et al., 2001). There are also contradicting results on sex/gender differences in Openness to experience: some show small differences favoring women (Costa et al., 2001; Kajonius and Mac Giolla, 2017; Lippa, 2005; Mac Giolla and Kajonius, 2019; Schmitt et al., 2008), others favoring men (Heineck, 2011; Kajonius and Mac Giolla, 2017; Schmitt et al., 2008).

As discussed in Chapter 4, describing sex/gender differences at the domain level might obscure differences at lower levels of aggregation like facets or nuances (Booth and Irwing, 2011; Del Giudice et al., 2012; Kaiser, 2019; Mac Giolla and Kajonius, 2019). Feingold's (1994) meta-analysis suggested that sex/gender differences in Neuroticism were mainly due to the *anxiety* facet. Additionally, sex/gender differences in Agreeableness were mostly found in the *tenderness* and *trust* facets. Lastly, Feingold (1994) found that differences in Extraversion were driven by sex/gender differences in the *assertiveness* facet. Costa et al. (2001) found that differences in Neuroticism and Agreeableness were similar across the domain and facet level. On the contrary, they found that not all facets in Extraversion and Openness showed significant differences between men and women. Altogether, this evidence suggests that sex/gender differences at the facet level could explain the contradicting results on certain domains like Extraversion and Openness (Booth and Irwing, 2011; Costa et al., 2001; Feingold, 1994).

Section summary

The research coincides in that women tend to score higher than men in Extraversion, Agreeableness, and Neuroticism. Often, women are found to score higher in Conscientiousness, but research results are less consistent. Results on Openness's sex/gender differences also remain unclear. Most research on personality sex/gender differences has focused on the Big Five at the domain level, which might conceal differences at the facet level. Moreover, the

effect size of sex/gender differences varies across samples and tends to be larger in developed countries.

5.1.2 Personality differences across sexual orientations

It is hypothesized that personality sex/gender differences occur on a spectrum because the biological factors contributing to them (e.g.: genes, hormones, etc.) are likely to operate on a continuum (Bao and Swaab, 2011). Operationalizing sex/gender as two categories does not reflect this spectrum, but sexual orientation could (Lippa, 2001). R. Lippa exemplified this through his *Shift Hypothesis*. According to the Shift Hypothesis, personality traits could be either represented as (1) two continuums – for men and women independently – with heterosexuality at one end and homosexuality at the other, or as (2) a bipolar dimension with heterosexual men and heterosexual women standing at each extreme. In either case, homosexual individuals would shift away from the heterosexual pole and bisexual individuals should lay between the two, being closer to homosexuals or heterosexuals depending on the degree to which they are attracted to the opposite sex[1]. In this sense, evidence for the Shift Hypothesis on personality differences could serve as evidence that personality sex/gender differences mirror the continuum on which certain biological precursors operate.

The Shift Hypothesis was likely inspired by the *Inversion Hypothesis*, which is based on Freud's (1953) idea of *gender inversion*. According to Freud (1953), homosexual men should show female traits and homosexual women should have male traits. Hence, in a way, the Inversion Hypothesis could be understood as a specific case of the Shift Hypothesis but it does not allow to study sex/gender differences as a continuum. There is some evidence supporting the Inversion Hypothesis, but it seems insufficient to explain personality differences across sexual orientations (Greaves et al., 2017).

[1]Actually, we have developed this assumption about the bisexual population based on the American Psychological Association (2012) definition of bisexuality as an attraction – in any degree – to either sex.

Table 5.1
Summary of studies testing the Shift and Inversion Hypotheses

Study	Shift vs. Inversion Hypotheses
Lippa (2005)	Evidence for the Shift Hypothesis and partial evidence for the Inversion Hypothesis in the women sample
Lippa (2008)	Partial evidence for the Shift and the Inversion Hypothesis
Zheng et al. (2008)	Inversion Hypotheses not tested
Greaves et al. (2017)	Support for the Inversion Hypothesis but not for the Shift Hypothesis
Allen and Robson (2020)	Inversion Hypotheses not tested

5.1.2.1 Personality differences between heterosexual and homosexual samples

Few studies have tackled personality differences across sexual orientations and, those which have, only used the Big Five at the domain level. Moreover, just a few studies have directly tested the Inversion Hypothesis (see Table 5.1). For instance Lippa (2005) and Lippa (2008) tested the Shift Hypothesis by comparing same sex/gender heterosexual and homosexual participants and tested the Inversion Hypothesis by correlating domain-level personality differences between heterosexual men and women with the personality differences between homosexual participants and their heterosexual counterparts. Lippa (2005) found a correlation of *.24* and *.89* in the male and women samples, respectively. Hence, Lippa (2005) argued that there was only evidence for the Inversion Hypothesis in the women sample. Later, Lippa (2008), found higher correlations (*.95* and and *.94* in the men and women samples), showing a stronger support for the Inversion Hypothesis. However, these correlations might be driven by differences in certain traits so, it is not clear to what degree Lippa's results reflected the Inversion Hypothesis. On the other hand, Greaves et al. (2017) directly compared all sex/gender and sexual orientation groups and found consistent evidence for the Inversion Hypothesis (see Table 5.2 for a summary of these studies).

Lippa (2005) found no differences in Extraversion between sexual orientation groups but three years later, Lippa (2008) he found evidence supporting the Shift Hypothesis in the

women sample. Zheng et al. (2008) and Greaves et al. (2017) did not find differences in Extraversion between any of the groups. On the contrary, Allen and Robson (2020) found evidence supporting the Shift Hypothesis for men and women in their Australian sample, but only for women in their meta-analyses.

Lippa (2005) showed evidence for the Shift Hypothesis in their male samples for Agreeableness. Greaves et al. (2017) found that the Agreeableness scores of homosexual men were no different from those of heterosexual women – i.e. homosexual man's scores were inversed. Lippa (2008) and Allen and Robson's (2020) Australian sample provided evidence for the Shift Hypothesis in both, the men and women sample. Allen and Robson (2020) meta-analysis of representative samples found evidence for the Shift Hypothesis only in the women sample. However, Zheng et al. (2008) and Allen and Robson (2020) non-population representative samples did not find any difference between groups.

Lippa (2005), Lippa (2008), and Zheng et al. (2008) showed evidence for the Shift Hypothesis in Neuroticism. Allen and Robson (2020) found evidence supporting the Shift Hypothesis for men in their meta-analyses but not in their Australian sample. Lastly, Greaves et al. (2017) only found inversion of the homosexual male's scores. These results are interesting because they contradict the popular *Social Stress Theory*, which argues that non-heterosexual individuals are more likely to experience traumatic events – e.g.: bulling – and thus, should score higher on Neuroticism (Meyer, 2010).

Most research has found women scoring higher than men in Conscientiousness but there are some inconsistent findings. Hence, evidence for the Shift and the Inversion Hypothesis should be considered sample-specific. If in a given sample women have higher Conscientiousness, the Shift Hypothesis would predict that homosexual men would score higher than heterosexual men, and homosexual women would score lower than heterosexual women. This is precisely what Lippa (2005) and Allen and Robson (2020) australian sample showed. Allen and Robson (2020) also found evidence supporting the Shift Hypothesis for men and women in their meta-analysis of population representative samples.However, Zheng et al. (2008) and

Allen and Robson (2020) meta-analysis of non-population representative samples found no difference between groups. Lastly, Greaves et al. (2017) found evidence for the Inversion Hypothesis only in the women sample.

Similarly to Conscientiousness, there is no consensus on Openness sex-differences. Zheng et al. (2008) and Greaves et al. (2017) found no difference between groups however, Lippa (2005) and all studies in Allen and Robson (2020) converged showing evidence for the Shift Hypothesis in both, male and female samples. Homosexual participants scoring higher than their heterosexual counterparts may be explained by the unconventional nature of those who defy the heteronormativity[2] by defining themselves as homosexuals.

Section summary

Overall, studies using the Big Five model have found some evidence supporting the Shift Hypothesis for Extraversion, Agreeableness, and Neuroticism. Evidence for the Shift Hypothesis on Conscientiousness is less consistent. Moreover, there has been no evidence of Shift Hypothesis on Openness to experience – although research results converged on showing that homosexuals score higher than heterosexuals. However, there is also reasonable evidence supporting the Inversion Hypothesis across traits. Together, these results show that there are personality differences across sexual orientation groups but it is not clear which hypotheses represent them better (see Table 5.2 for a summary of previous research findings).

[2]According to Merriam Webster dictionary, heteronormativity is "of, relating to, or based on the attitude that heterosexuality is the only normal and natural expression of sexuality" (Heternormativity, 2021).

Table 5.2
Summary of facets and domains driving differences between groups

Sample type	Extraversion	Agreeableness	Conscientioussness	Neurocisim	Openness
Expected given previous research	W > M	W > M	Unclear	W > M	Unclear
Lippa (2005). Meta-analysis US college student samples previously collected by the author	No differences between Ho and He	Shift for Ho-M	Shift for Ho-M	Shift for Ho-M and Ho-W	Higher in homosexual men and women
Lippa (2008). International sample	Shift for homosexual women.	Shift for Ho-M and Ho-W	NA	Shift for Ho-M and Ho-W	NA
Zheng et al. (2008). Chinese sample	No differences between homosexuals and heterosexuals	No differences between homosexuals and heterosexuals	No differences between homosexuals and heterosexuals	Shift for Ho-M and Ho-W	No differences between homosexuals and heterosexuals

Sample type	Extraversion	Agreeableness	Conscientioussness	Neurocisim	Openness
Greaves et al. (2017). New Zeland sample	No differences between homosexuals and heterosexuals	Inversion for Ho-M	Inversion for Ho-W	Inversion for Ho-M	No differences between homosexuals and heterosexuals
Allen and Robson (2020). Australian sample	Shift for Ho-M and Ho-W	Shift for Ho-M and Ho-W	Shift only for men. No differences between homosexual and heterosexual women	No differences between homosexuals and heterosexuals	Higher in homosexual men and women
Allen and Robson (2020). Meta-analysis of international non-population representative samples	Shift for Ho-W	No differences between homosexuals and heterosexuals	No differences between homosexuals and heterosexuals	Shift for Ho-M	Higher in homosexual men and women

Sample type	Extraversion	Agreeableness	Conscientioussness	Neurocisim	Openness
Allen and Robson (2020). Meta-analysis of international population representative samples	Shift for Ho-W	Shift for Ho-W	Higher for homosexual men than heterosexual men and lower for homosexual women than heterosexual women. This pattern suggests a shift	Shift for Ho-M	Higher in homosexual men and women

Notes. He= Heterosexual, Ho= Homosexual, M= Men, W= Women.

5.1.2.2 Personality differences between heterosexual, homosexual and bisexual samples

It is difficult to gauge Shift and Inversion Hypotheses' predictions for bisexual men and women. The disposition of bisexuals along the spectrum of sexual orientation could depend on the degree to which they are attracted to the opposite sex/gender – e.g bisexual women would be closer to homosexual women the more the latter were attracted to women and closer to heterosexual women the more homosexual women were attracted to men (Lippa, 2008, 2020). Therefore, the Shift Hypothesis will predict that bisexuals' trait levels will lay in-between heterosexuals and homosexuals of the same sex. However, the Inversion Hypothesis will predict that bisexuals' personality scores would equate those of the sex/gender they are less attracted to. Given the strict assumptions of the Inversion Hypothesis, it is easier to study the personality shifts of bisexual individuals using the Shift Hypothesis. Hence, this section will only briefly review the evidence for the Shift Hypothesis using bisexual samples.

In Allen and Robson (2020) Australian sample there were no differences in Extraversion between bisexuals and heterosexuals or homosexuals – except for homosexual women who were less extraverted than bisexuals. Similarly, in both Allen and Robson's (2020) meta-analyses, there were no differences between bisexuals and heterosexuals or homosexuals except for heterosexual women who were more extraverted than bisexuals. Lippa (2008) found that bisexual men scored in Extraversion less than homosexual and heterosexual men, but bisexual women scored in between heterosexual and homosexual women. Hence, Lippa (2008) found evidence for the Shift Hypothesis for bisexual women but not for men, contrary to Allen and Robson (2020) Australian sample in which there was evidence for the Shift Hypothesis for bisexual men but not for women.

In Allen and Robson (2020) Australian sample bisexuals scored in Agreeableness between heterosexuals and homosexuals, supporting the Shift Hypothesis. However, these results were not corroborated in either of their meta-analyses. Lippa (2008) also found support for the Shift Hypothesis in bisexual men but not in bisexual women.

In Allen and Robson (2020) Australian sample bisexual men scored in Conscientiousness similarly to heterosexual men, and both lower than homosexual men. Additionally, bisexual women scored lower in Conscientiousness than their heterosexual and homosexual counterparts. There was no support for the Shift Hypothesis in Conscientiousness in neither of Allen and Robson (2020) meta-analyses.

Bisexuals and homosexuals showed equal levels of Neuroticism in all Allen and Robson (2020) analyses, so there was no support for the Shift Hypothesis. Additionally, in the Australian sample bisexuals scored higher in Neuroticism than heterosexuals. In Allen and Robson (2020) meta-analyses, bisexual men scored higher than heterosexual men, but there were no differences between bisexual and heterosexual women. Lippa (2008) results did not support either of the hypotheses because bisexual men scored like homosexual men and bisexual women scored higher than their heterosexual and homosexual counterparts. Overall, neither Allen and Robson (2020) nor Lippa (2008) strongly support the Shift Hypothesis or the *Social Stress Theory* for Neuroticism.

Interestingly, Allen and Robson (2020) found bisexuals scoring higher in Openness to experience than heterosexuals across their three analyses. In their Australian sample, there were no differences between homosexuals and bisexuals in Openness to experience. In their meta-analyses, there were no differences between homosexuals' women and bisexual women; but bisexual men scored higher than homosexual men.

5.2 Summary and hypotheses

Given that sex/gender differentiation happens on a continuum, sex/gender orientation as an independent variable might provide a more nuanced approach to understand personality differences. Research on personality differences across sexual orientation is scarce but, so far, it shows that personality scores of homosexuals tend to move away from their heterosexual counterparts.

There are two competing hypotheses to trying to explain these results: the Shift and the Inversion Hypotheses. The Shift Hypothesis implies that the personality scores of homosexual participants move away from their heterosexual counterparts. On the other hand, the Inversion Hypothesis proposes that homosexual men should score as heterosexual women and homosexual women should score as heterosexual men.

So far, evidence for the Shift Hypothesis is especially consistent for Extraversion, Agreeableness, and Neuroticism. However, there is also evidence for the Inversion Hypothesis across traits. Our goal in this study is to test the Shift and the Inversion Hypotheses at the facet and domain level. Following previous research I hypothesize that:

- *Hypothesis 1:* Discrepancies between group differences at facet level may explain inconsistencies in sex/gender differences reported in previous studies at domain level.

- *Hypothesis 2:* Homosexual participants score higher than heterosexual participants in Neuroticism and in Openness.

- *Hypothesis 3:* The personality scores difference between homosexual and heterosexuals might depend on how large the difference is between heterosexual men and women.

- *Hypothesis 4:* Given that the Shift Hypothesis makes less rigid claims in terms of what the degree of personality change about personality shifts according to sexual orientations, it is likely that we will find more evidence supporting it.

5.3 Methods

5.3.1 Sample

Together with other colleagues, I created a life-outcomes and a personality questionnaire. The survey was uploaded to Formr (Arslan et al., 2020) and published the link in various social media sites (e.g.: Twitter, facebook, Instagram, etc.) as well as in Prolific. A total of

Table 5.3
Demographic variables of the sample

			Age		Education Level		Income Range	
Sex	Sexual Orientation	N	Mean	SD	Mean	SD	Mean	SD
Men	Heterosexuals	88	35.72	12.78	4.72	1.04	3.80	1.49
	Homosexuals	131	36.53	12.17	4.73	1.01	3.70	1.55
Women	Heterosexuals	221	36.19	12.61	4.70	0.99	3.42	1.43
	Homosexuals	102	33.07	11.28	4.71	1.01	3.41	1.33

Note:
Education Level includes: 1= Primary School, 2= Secondary School, 3= Vocational Education / Apprenticeship, 4= Undergraduate Education, 5= Postgraduate education (e.g.:Master, Ph.D., etc.); Income Range includes 1= Less than 10 000 pounds a year, 2=Between 10 000 and 18 000 pounds a year, 3= Between 18 001 and 26 000 pounds a year, 4= Between 26 001 and 44 000 pounds a year, 5= More than 44 001 pounds a year

630 people landed on the survey. Participants under 18 years old were filtered out together with those who did not identified as men, women, heterosexual or homosexual (N= 80). After inspecting the data's quality with R's *careless* package (Yentes and Wilhelm, 2018), I further eliminated 8 participants that gave the same answer 10 or more consecutive personality times. This left 542 participants (60% Women). Table 5.3 shows our sample demographics.

5.3.2 Materials

Sexual orientation. Measured through a single item asking participants to self-identify as heterosexual/straight, homosexual/gay/lesbian, bisexual, asexual or other. Due to the limited number of participants self-identifying as bisexual, asexual or other, only the data from heterosexual and homosexual participants was analyzed.

Personality. A total of 90 items were selected from the pool of items created by Henry, Mõttus, & Vainik (in preparation), who screened the items to maximize their re-test reliability. Items were mainly retrieved from the International Personality Item Pool (IPIP, Goldberg, 1999) and the Synthetic Aperture Personality Assessment (SAPA, Condon, 2018). Some

items were written by the authors. Items were selected to maximize their re-test reliability (M=.71, SD=.09, Values range=.60-.85), their clarity, readability, as well as to minimize their redundancy. Following Condon et al. (2021) advice, items were selected and categorized into facets according to: (1) their re-test reliability, (2) feedback from Dr McCrae, and (3) fit within the IPIP's facet definitions. Domain scores were computed as the average across six facet scores, each computed as the sum scores of three items.

5.4 Analyses

5.4.1 Cohen's *d* and Mahalanobis Distance

The main goal of this study was to show how personality group differences are overlooked when considering the domain level because these differences are reduced or canceled out when aggregated. For this reason, my analyses focus on the difference or distance between groups rather than on the consequences of specific personality differences at domain or facet level.

Personality sex/gender differences are usually measured individually for each of the Big Five using Cohen's *d*, which estimates the difference between the group's means in standard deviation units (Del Giudice, 2019). Del Giudice (2021) explained the existence of a direct relationship between the number of variables and the size of Cohen's *d*, suggesting that sex/gender differences might have been underestimated (e.g.: Hyde, 2005). Using more indicators – i.e. facets or nuances – would result in a more accurate estimate of personality sex/gender differences. This approach would raise the question of which variables to consider and how to combine them. For instance, the distance between two cities could be estimated using latitude, altitude, and longitude (Del Giudice et al., 2012). Choosing just one of these indicators will likely under or over-estimate the *actual* distance between these cities. Additionally, how these attributes are combined would also have a direct impact on the measurement's accuracy.

There are different methods to estimate group differences based on multiple indicators (for an extensive discussion see Del Giudice, 2021). First, the *city-block distance* represents the sum of the absolute distance on every dimension between two groups, but it tends to be highly influenced by the dimension having the largest distance. Second, the *shape distance* works under the assumption that the correlations between variables are equal in both groups and focuses on indicators' "elevation", "shape", and "scatter" which, for the sake of simplicity, can be understood as their mean and variance. And third, the *Mahalanobis Distance (D)* is the multivariate generalization of Cohen's d and represents the distance between the centroids (multivariate means) relative to the standard deviation along that distance (Del Giudice, 2019; Mahalanobis, 1936). Contrary to Cohen's d, Mahalanobis D, takes into account the pattern of correlation between the measured variables, which is especially relevant in personality research where traits tend to be correlated. D has become a popular method for estimating group differences but it has three main limitations: (1) lack of psychological interpretability of the axis in which the distance is computed; (2) overestimation of the distance when there is a small number of observations per indicator; (3) uncertainty on which variables have more weight on the final result; and (4) susceptibility to violations of normality assumptions.

Regarding the first limitation, researchers can ensure that D is interpretable by providing theoretical justifications for each indicator before examining group differences. Additionally, Del Giudice (2019) offered some complementary measures to aid D's interpretation like the Overlapping Coefficient (OVL), which represents the proportion of the distribution that is shared by both groups ranging from 0 to 1. Concerning the second limitation, Del Giudice (2021) proposed the following formula to correct for the bias induced when having few observations per indicator (i.e. less than a hundred observations per trait):

$$D_{Mu} = \left[max\left(0, \frac{N_1 + N_2 - k - 3}{N_1 + N_2 - 2} D_M^2 - k\frac{N_1 + N_2}{N_1 N_2}\right)\right]^{1/2} \tag{5.1}$$

where k is the number of variables, and the Ns represent the two sample sizes. Del Giudice (2018) also offered bias-corrected alternatives for OVL -i.e. OVL_u.

To address the third limitation involving the extent to which each indicator contributes to the distance's magnitude, Del Giudice (2019) proposed the Heterogeneity Coefficient (H_2) and the Equivalent Proportion of Variables coefficient (EPV_2). Coefficient H_2 ranges from 0 to 1, where 0 means that all variables contributed equally and 1 indicates that a single variable is driving the group differences (Del Giudice, 2017, 2018). The EPV_2 coefficient ranges from 0 to 1, and represents the proportion of variables that contributing equally, would produce the same degree of heterogeneity if the rest of the variables did not contribute (Del Giudice, 2017, 2018). For instance, EPV_2 =.40 would represent the heterogeneity level expected in a hypothetical case in which 40% of the indicators contributed equally to D and 60% did not contribute at all. Therefore, models having large EPV_2 and low H_2 values are more desirable.

Lastly, D sensitivity to violations of normality assumptions calls for a rigorous examination of the data to ensure homogeneity of means and variances across groups. If means and variance homogeneity cannot be proven, Cohen's d or its unbiased estimate (d_u) must be used. The OVL or OVL_u can be computed for Cohen's d and d_u. However, neither H_2 nor EPV_2 can be computed for Cohen's d or d_u, for they works under the assumption of multivariate normality and homogeneity of variances.

5.4.2 *Interpreting Cohen's d and Mahalanobis D*

The distance measured through Cohen's d and Mahalanobis D is often interpreted according to Cohen (1992) cut-off points by which .1 would suggest a small distance (or effect size), .25 a medium distance and .4 a large distance. Recently, Funder and Ozer (2019) reviewed the suitability of Cohen's standards and proposed new ones based on the most commonly found effect sizes in psychological research. According to Funder and Ozer (2019) labeling effects sizes could start from .05 (i.e. very small distance), and continue in .10 intervals from

.10 (i.e. small distance) to .40 (i.e. very large distance). It could be argued that Funder and Ozer (2019) standards follow circular logic: the findings on the field are ordered in size – and relevance – according to the labels of the effect sizes, and the labels of the effect sizes are determined by the most common effect sizes on the field. Nevertheless, their proposition must be appreciated as it reflects the need to adapt the measures of effect sizes to every field. In this sense, it is fair to argue that sex/gender differences in personality research tend to be, on average, larger than in other psychology disciplines (see Section 5.1.1). Hence, in this chapter we will use small, moderate, high, and very high for ranges (-.2, .2), [|.2, .5|), [|.5, .8|), and $\geq$ |.8|, respectively.

5.4.3 Multivariate Normality and homogeneity of variance-covariance matrices

Testing for multivariate normality (MVN) implies estimating whether a collection of variables follow an underlying normal distribution so that their linear combination is also normally distributed. It is usually recommended to use different MVN tests because there is not a single one that performs well under all circumstances (e.g.: Del Giudice, 2019; Mecklin and Mundfrom, 2003). In this study, we will use Mardia's, Doornik-Hansen's, Henze-Zirkier's, Royston's, and the Energy tests.

Mardia's and Doornik-Hansen's (DH) tests are based on measures of multivariate skewness and kurtosis, meaning they focus on the distribution's spread and density. Henze-Zirkier's (HZ) test is the multivariate generalization of Epps and Pulley's univariate normality test and Royston's test is the multivariate generalization of Shapiro and Wilk's test. Both work similarly, as they test how similar is the distribution at hand from an ideal distribution. The HZ test actually compared the distance between the function representing the actual distribution of the data and a function representing a hypothetical distribution (Alpu et al., 2016). On the other hand, Royston's test is based on an approach similar to the goodness-of-fit tests which allow researchers to determine the suitability of a model by comparing

the variance-covariance matrix of an ideal data set - in this case, a multivariate normally distributed data set - against that of the data at hand. Lastly, the Energy test is a nonparametric test estimating the equality between multivariate distributions based on Euclidean distance between each samples' units (Alpu et al., 2016).

MVN tests are evaluated in two ways. First, tests are judged according to their rate of false positives (i.e. Type I error) when using MVN distributions at specific α levels depending on the number of observations and variables. Second, tests are compared based on their power to detect MVN violations depending on (1) different distributions and (2) the ratio between observations and variables. The first requirement helps to determine if tests are working as they should and the second is to gauge how good these tests are when facing a gradient of deviations from normality.

The HZ, Royston's, DH, and Energy tests have consistent and adequate Type I error rates with samples over 50 observations, although HZ outperforms them when using small samples (n=20) (Alpu et al., 2016; Farrell et al., 2007; Mecklin and Mundfrom, 2003, 2005). Mardia's test usually has a very low Type I error rate for skewness but not for kurtosis regardless of the combinations of sample sizes and the number of variables (Mecklin and Mundfrom, 2003, 2005).

Each of the MVN tests' power can be improved by increasing the sample size and variables number (Farrell et al., 2007), except for the Energy test in specific distributions, for which large samples might decrease its power (Alpu et al., 2016). Simulations show that the power of each test varies depending on the type of distribution being used, which is the reason why it is important to compare several tests (Alpu et al., 2016; Farrell et al., 2007; Mecklin and Mundfrom, 2003).

Overall Mardia's test performance seems to be the most inconsistent (Mecklin and Mundfrom, 2003, 2005). The rest of the tests outperform the others in specific scenarios. However, several reviews have found HZ and Royston's tests as the most balanced test in terms of

strengths and weaknesses. Additionally, both provide consistent results across different violations of MVN (Alpu et al., 2016; Farrell et al., 2007; Mecklin and Mundfrom, 2003, 2005). The Energy test is not as frequently used but Alpu et al. (2016) showed that it had a high power in various scenarios.

Multivariate normality tests were run at the facet (N=30) and domain level (N=5) across all the pair combinations that could create out of the four groups – i.e. heterosexual men, heterosexual women, homosexual men, and homosexual women. Hence, the sample was first divided (N= 542) into four different matrices with the raw personality item scores. There were 88 heterosexual men, 221 heterosexual women, 131 homosexual men, and 102 homosexual women.

Lastly, Box's M was used to test the homogeneity of variance-covariance matrices. Homogeneity of variance-covariance matrices guarantees that the spread of the distribution is similar across groups, which is relevant when estimating the distance between them. Note that Box's M is sensitive to normality departures.

5.5 Results

5.5.1 Comparisons at domain level

5.5.1.1 Multivariate Normality and homogeneity of variance-covariance matrices at domain level

We run the MVN tests on the domain scores for each of the four groups. Aggregation tends to compensate for normality violations at lower levels. Table 5.4 shows that most MVN tests confirmed that domain scores did not violate multivariate normality assumptions.

Table 5.5 shows the results from the Box's M test for homogeneity of variance-covariance matrices. As the p-values were above the significance level of .05, there was evidence for homogeneity of variance-covariance matrices at domain level in all of the group comparisons

Table 5.4

Multivariate normality tests at domain level for each sex and sexual orientation group

Test	p-value
Heterosexual Men	
Henze-Zirkler	0.5
Royston	0.76
Doornik-Hansen	0.87
E-statistic	0.47
Mardia Skewness	0.2
Mardia Kurtosis	0.53
Heterosexual Women	
Henze-Zirkler	0.24
Royston	0.06
Doornik-Hansen	<.05
E-statistic	0.09
Mardia Skewness	0.17
Mardia Kurtosis	0.14
Homosexual Men	
Henze-Zirkler	0.2
Royston	0.85
Doornik-Hansen	0.73
E-statistic	0.14
Mardia Skewness	<.05
Mardia Kurtosis	0.64
Homosexual Women	
Henze-Zirkler	0.45
Royston	0.18
Doornik-Hansen	0.86
E-statistic	0.13
Mardia Skewness	0.08
Mardia Kurtosis	0.76

Table 5.5
Box's M, a test for homogeneity of variance at domain level for each group comparison

	Parameter	Df	p-value
Heterosexual men vs. Heterosexual women	12.12	15	0.67
Heterosexual men vs. Homosexual men	15.11	15	0.44
Heterosexual men vs. Homosexual women	17.13	15	0.31
Heterosexual women vs. Homosexual women	32.11	15	<.01
Heterosexual women vs. Homosexual men	24.21	15	0.06
Homosexual women vs. Homosexual men	16.04	15	0.38

except for Heterosexual women vs. Homosexual women (χ^2 (15, N_{He_W}=221, N_{Ho_W}=102)= 32.11, $p = 0.01$). These results might lead to an underestimation of the distance between these two groups, so results should be interpreted with caution.

5.5.1.2 Comparisons between heterosexual men and heterosexual women at domain level

Given that domain scores had MVN and homogeneity of variance we used D_{Mu} to measure the overall personality difference at domain level between the four groups, along with d_u to estimate the distance at each domain and other measures of overlap -i.e. $OVLD_u$ - and heterogeneity -i.e. H_2 and EPV_2. Table 5.7 gathers the D_{Mu}, $OVLD_u$ and H_2 values across group comparisons.

The overall multivariate distance, D_{Mu}, between heterosexual men and women personality scores was 0.88. As mentioned in the introduction, this shows the positive relationship between the distance size and the number of variables simultaneously considered. The overall overlap across domains was moderate: $OVLD_u = 0.66$.

The univariate overlap across domains $OVLd_u$ ranged from 0.77 to 0.99, indicating a high to very high overlap between both groups. Moreover, H_2 was 0.7, meaning that the difference between these groups was not being driven homogeneously by all the domains. This was confirmed by the EPV_2 which was 0.44, meaning that the same level of heterogeneity would

be achieved in an hypothetical scenario where 44% of the variables were contributing equally while 56% of the variables did not contribute at all. We explored each domain's d_u to assess which were responsible for the largest differences. As table 5.6 shows Agreeableness and Neuroticism lead the differences between Heterosexual men and Heterosexual women.

5.5.1.3 Comparisons between heterosexual men and homosexual men at domain level

D_{Mu} between heterosexual and homosexual men was 0.55, with a high overall overlap of 0.78. The univariate overlap across domains ranged from 0.85 to 0.97, also indicating a high to very high overlap between both groups. Note that the $OVLd_u$ range was smaller than in previews sections.

Moreover, H_2 was 0.64 meaning that the difference between these groups was mainly driven by a few domains, albeit not as unbalanced as in the previous case. The EPV_2 confirmed that the same level of heterogeneity would be achieved in an hypothetical scenario where 49% of the variables were contributing equally while 51% of the variables did not contribute at all (Table 5.7). Table 5.6 shows that Agreeableness and Neuroticism lead the differences between heterosexual and homosexual men.

5.5.1.4 Comparisons between heterosexual men and homosexual women at domain level

D_{Mu} between heterosexual men and homosexual women was 0.87, which was as high as the distance between between heterosexual men and women. The univariate $OVLd_u$ across domains ranged from 0.75 to 0.97, which indicated a high to very high overlap. H_2 for heterosexual men and homosexual women distance was 0.61, meaning that the difference between these groups was quite homogeneous. The EPV_2 was notably smaller than in the previous two sections (EPV_2= 0.51), showing that most variables did not contribute at all (Tables 5.7 and 5.6).

Table 5.6
d_u sizes for each domain across goup comparisons

Domain	Distance size
Heterosexual man vs. Heterosexual woman	
Agreeableness	-0.58
Neuroticism	-0.49
Conscientiousness	-0.13
Extraversion	0.08
Openness	0.04
Heterosexual man vs. Homosexual man	
Neuroticism	-0.37
Agreeableness	-0.36
Extraversion	0.22
Openness	-0.16
Conscientiousness	0.09
Heterosexual man vs. Homosexual woman	
Neuroticism	-0.64
Extraversion	0.39
Agreeableness	-0.31
Openness	-0.26
Conscientiousness	0.09
Heterosexual woman vs. Homosexual woman	
Extraversion	0.35
Openness	-0.30
Conscientiousness	0.22
Agreeableness	0.22
Neuroticism	-0.19
Heterosexual woman vs. Homosexual man	
Conscientiousness	0.22
Agreeableness	0.20
Openness	-0.19
Extraversion	0.15
Neuroticism	0.08
Homosexual woman vs. Homosexual man	
Neuroticism	0.25
Extraversion	-0.17
Openness	0.10
Agreeableness	-0.02
Conscientiousness	0.00

Table 5.7
Distance and additional measures for each sex and sexual orientation group

	Du	OVLu	H2	EPV2
Heterosexual Men vs. Heterosexual Women	0.88	0.66	0.70	0.44
Heterosexual Men vs. Homosexual Men	0.55	0.78	0.64	0.49
Heterosexual Men vs. Homosexual Women	0.87	0.66	0.61	0.51
Heterosexual Women vs. Homosexual Women	0.50	0.80	0.61	0.51
Heterosexual Women vs. Homosexual Men	0.38	0.85	0.16	0.87
Homosexual Women vs. Homosexual Men	0.18	0.93	0.75	0.40

Note:
Du= Corrected Mahalanobis D reflects the multiariate distance (i.e. considering the five domains simultaneously) between each group; OVL= Overlapping Coefficient represents the proportion of the distribution that is shared by both groups ranging from 0 (no overlap) to 1 (full overlap); H2= Heterogeneity Coefficient ranges from 0 (equal contribution across variables) to 1 (completelly unequal contribution); EPV2= Equivalent Proportion of Variables coefficient ranges from 0 (few variables drive differences) to 1 (all variables contributed to the differences)

5.5.1.5 Comparisons between heterosexual women and homosexual women at domain level

D_{Mu}, between heterosexual and homosexual women was 0.5, which was similar to the overall distance between heterosexual and homosexual men. The multivariate $OVLd_u$ was 0.8 and the univariate $OVLd_u$ ranged from 0.86 to 0.92, which was similar to the distance and overlap found in previous sections. The H_2 was 0.61, meaning that the difference between these groups was produced by several – but not all – domains. The EPV_2 value was 0.51, so roughly half of the variables contributed equally (Table 5.7). Table 5.6 shows that the domains driving this difference were Extraversion and Openness, which was also accentuating the differences between heterosexual men and women.

Note that that there was no evidence of homogeneity of variance-covariance matrices between these two groups. This might affect the estimation of the distance between them, so these results should be evaluated with caution.

5.5.1.6 Comparisons between heterosexual women and homosexual men at domain level

D_{Mu} between heterosexual women and homosexual men was 0.38, with $OVLD_u = 0.85$ and $OVLd_u$ ranging from 0.91 to 0.97. The H_2 was 0.16, meaning that the difference between these groups was produced homogeneously by all domains. The EPV_2 value was 0.87 suggesting that only half of the variables contributed to this difference (Table 5.7). Conscientiousness, Agreeableness, and Openness were the most relevant domains (Table 5.6).

Interestingly, the difference in Neuroticism between heterosexual women and homosexual women was half of that between heterosexual women and homosexual men. Neuroticism as a domain still rose as one of the main differences between men and women.

5.5.1.7 Comparisons between homosexual women and homosexual men at domain level

The overall distance between homosexual women and homosexual men was the smallest, D_{Mu}=0.18, and their overall overlap was larger than in previous sections, $OVLD_u = 0.93$. The univariate $OVLd_u$ ranged from 0.9 to 1. Their heterogeneity coefficient was the highest (H_2=0.75), showing that this difference was due to few domains. In fact, it was as if 60% of domains were not contributing to it. Looking further into the individual d_u values, Neuroticism stood as the domain with the highest difference. This offers further support to the fact that Neuroticism is a relevant personality variable referencing men and women.

5.5.1.8 The Shift Hypothesis vs. the Inversion Hypothesis at domain level

The largest overall distance at domain level was between heterosexual men and woman (D_{Mu}= 0.88) and the smallest was between homosexual men and women (D_{Mu}= 0.18). The distance between heterosexuals and homosexuals ranged between .4 -.9. Taken together, these results suggest a personality shift according to sexual orientation. Table 5.6 allows us to

explore the Shift and the Inversion Hypotheses at domain level by comparing the univariate distance in each personality domain across group comparisons.

Agreeableness offered strong support for the Shift Hypothesis as heterosexual women scored the higher, heterosexual men the lowest ($d_{uHeM-HeW}$= -0.58) and the scores of homosexual groups laid in between ($d_{uHoW-HoM}$= -0.02).

Neuroticism results at domain level partially supported the Shift and the Inversion Hypotheses. On the one hand, homosexual women scored higher than heterosexual women ($d_{uHeW-HoW}$= -0.19), which also partially supports the Social Stress Theory. And, on the other hand, homosexual men scored as heterosexual women ($d_{uHeW-HoM}$= 0.08). This could also be interpreted as a partial support for the *Social Stress Theory*, at least, as it regards homosexual men.

The differences at domain level in Conscientiousness were small. Heterosexual women scored higher than the rest of the groups ($d_{uHeM-HeW}$=-0.13, $d_{uHeW-HoW}$=0.22, and $d_{uHeW-HoM}$=0.22). Contrary to Neuroticism results, there was a unilateral inversion of homosexual women's scores, providing partial evidence to the Inversion Hypothesis.

Results in Extraversion and Openness do not support either of the hypotheses. However, they showed that personality scores of homosexual participants tended to deviate from their heterosexual counterparts, especially when they score similarly. Heterosexual men and women scored similarly in Extraversion. Homosexual women scored the lowest and homosexual men scored in between homosexual women and heterosexual participants. On the contrary, and following previous research, homosexual men and women scored similarly and higher than both heterosexual groups.

In summary, at domain level we observed some support for the Inversion Hypothesis but not so much for the Shift Hypothesis. There was a shift in homosexuals' Agreeableness scores for both, men and women. There was a unilateral inversion of homosexual men scores

in Neuroticism. Conversely, homosexual women inverted their scores in Conscientiousness domain scores. It is possible that the distance between heterosexual and homosexual women at domain level was underestimated. Nonetheless, it is unlikely that homosexual women experienced inversion more often than homosexual men because the (multivariate and univariate) distances at domain level between heterosexual men and homosexual women were larger than between heterosexual men and homosexual men. Lastly, results in Extraversion and Openness were similar: heterosexuals and homosexuals scored likewise but differently from their heterosexual counterparts. Perhaps, it should not be surprising to find lack of evidence for the Shift or the Inversion Hypothesis in domains were there was a small difference between heterosexual men and women. What it is noticeable is that, in such cases, homosexual's scores consistently displaced in the same direction.

5.5.2 Comparisons at facet level

5.5.2.1 Multivariate Normality and homogeneity of variance-covariance matrices at facet level and

First, we explored facets' skewness and kurtosis individually. There were no facets with skewness or kurtosis values higher than 1 or lower than -1.

Tables 5.8 shows MVN tests' results for each group. Given that few of the tests' values surpassed the significance level of .05, there was no evidence that the facets in any of the groups followed a multivariate normal distribution. Consequently, only Cohen's d could be used to estimate personality differences across groups.

Table 5.9 shows the results from the Box's M test for homogeneity of variance-covariance matrices. Given the lack of multivariate normality, it is unlikely that Box's M would represent an accurate estimate of variance-covariance matrices homogeneity. As Table 5.9 shows, few of the p-values were above the significance level of .05 hence, there was no evidence of homogeneity of variance-covariance matrices at facet level in any of the group comparisons.

Table 5.8
Multivariate normality tests at facet level for each sex and sexual orientation group

Test	p-value
Heterosexual Men	
Henze-Zirkler	0.32
Royston	<.05
Doornik-Hansen	0.54
E-statistic	<.05
Mardia Skewness	<.05
Mardia Kurtosis	0.06
Heterosexual Women	
Henze-Zirkler	<.05
Royston	<.05
Doornik-Hansen	<.05
E-statistic	<.05
Mardia Skewness	<.05
Mardia Kurtosis	<.05
Homosexual Men	
Henze-Zirkler	<.05
Royston	<.05
Doornik-Hansen	<.05
E-statistic	<.05
Mardia Skewness	<.05
Mardia Kurtosis	<.05
Homosexual Women	
Henze-Zirkler	<.05
Royston	<.05
Doornik-Hansen	<.05
E-statistic	<.05
Mardia Skewness	<.05
Mardia Kurtosis	0.11

Table 5.9
Box's M, a test for homogeneity of variance at facet level for each group comparison

	Parameter	Df	p-value
Heterosexual men vs. Heterosexual women	521.12	465	0.04
Heterosexual men vs. Homosexual men	532.39	465	0.02
Heterosexual men vs. Homosexual women	569.38	465	<.01
Heterosexual women vs. Homosexual women	590.26	465	<.01
Heterosexual women vs. Homosexual men	581.05	465	<.01
Homosexual women vs. Homosexual men	557.46	465	<.01

5.5.2.2 Comparisons between heterosexual men and heterosexual women at facet level

Each sex/gender and sexual orientation group had few participants – i.e. observations – per facet. Thus, d_u was used to estimate the distance between groups. The average d_u between heterosexual men and women across all facets was 0.31 with SD=0.21, indicating a moderate distance, which was much smaller than the overall distance at domain level. Table 5.10 shows that the facets with the largest d_u –and similarly to results at the domain level – were related to Agreeableness and Neuroticism (Table 7.1 in the Appendix shows the d_u values for all facets across group comparisons).

To further explore the relation between domains and d_u, facets were grouped according to the distance between men and women – i.e. very high (above .8 or below -.8), high (between absolute values of .5 and .8), moderate (between absolute values of .2 and .5), and small (between -.2 and .2). Figure 5.1 shows the relative contribution of each domain to these distance categories. According to Figure 5.1 Agreeableness facets tended to be responsible for larger differences. As Figure 5.1 shows, there was not a single domain driving large differences between heterosexual men and women.

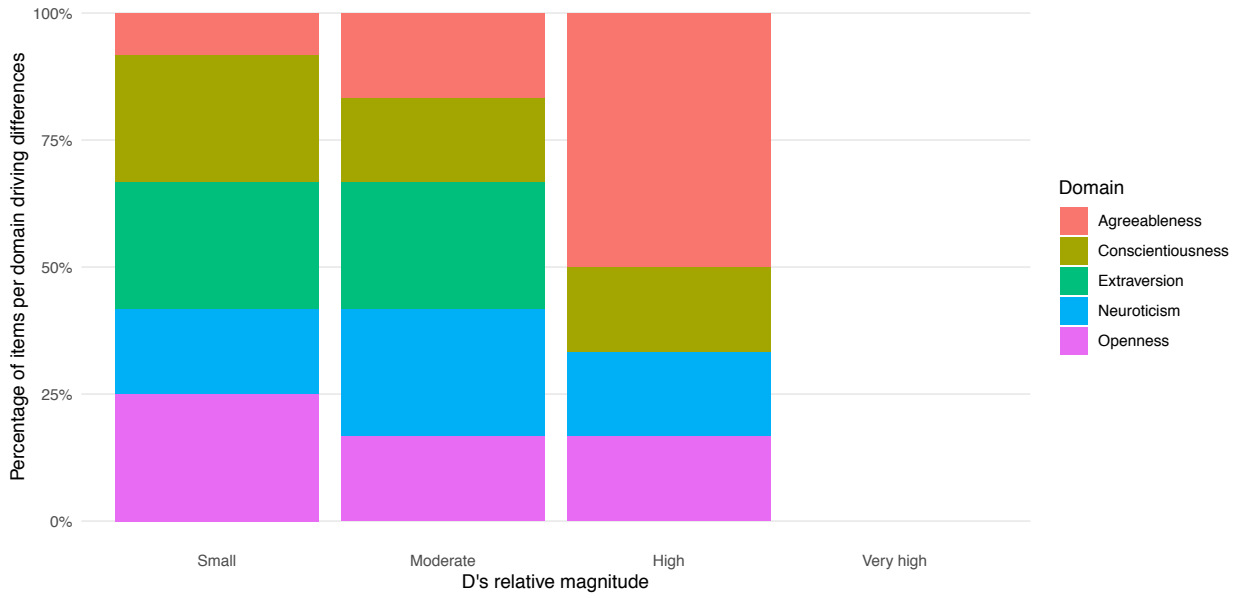

Figure 5.1
Percentage of facets per domain composing each of the distance levels (small, moderate, high, and very high) between heterosexual men and women. 'Small' distance represents du values = (-.2, .2); 'Moderate' distance represents du values = [|.2, .5|); 'High' distance represents du values = [|.5, .8|); 'Very high' distance represents du values equal or higher than |.8|

5.5.2.3 Comparisons between heterosexual men and homosexual men at facet level

The average d_u between heterosexual and homosexual men was half of that between heterosexual men and women (M=0.21; SD=0.11), indicating a small to moderate distance. Heterosexual and homosexual men appear to differ specially on Agreeableness and Neuroticism facets (Table 5.10). However, there were other facets form different domains yielding high d_u.

5.5.2.4 Comparisons between heterosexual men and homosexual women at facet level

The average d_u between heterosexual men and homosexual women was almost identical to that between heterosexual men and heterosexual women (M=0.31; SD=0.18). As found in

the previous section, Neuroticism facets were prominent in the difference between heterosexual men and homosexual women (Table 5.10).

5.5.2.5 Comparisons between heterosexual women and homosexual women at facet level

The average d_u between heterosexual women and homosexual women was 0.19, SD=0.15. This difference was slightly smaller than that found either between heterosexual and homosexual men, or between heterosexual men and homosexual women. Heterosexual women differed from homosexual women on various facets, specially those grouped under the Openness domain (Table 5.10).

5.5.2.6 Comparisons between heterosexual women and homosexual men at facet level

The average d_u between heterosexual women and homosexual men was 0.17, SD=0.12. This difference was similar to the difference between heterosexual women and homosexual women. This means that homosexual men were more similar to heterosexual women, than to heterosexual men. There did not seem to be a specific domain driving the differences between heterosexual women and homosexual men (Table 5.10).

5.5.2.7 Comparisons between homosexual men and homosexual women at facet level

The average and distribution of d_u between homosexual women and homosexual men was the smallest across the groups comparisons: M= 0.14, SD=0.1. As this was a smaller distance than that found between heterosexual men and heterosexual women, this already suggests some personality score shifts among homosexual participants.

Figure 5.2 is a summary of the percentage of facets – organized by domains – in each distance group (i.e very high, high, moderate, and small) and across group comparisons. As Figure 5.2 shows, differences between homosexual women and homosexual men were mainly driven by

in Openness facets. This was surprising because both have been found more open than their heterosexual counterparts and previous research has assumed high Openness as a common characteristic among homosexual individuals. For more details see Table 5.10.

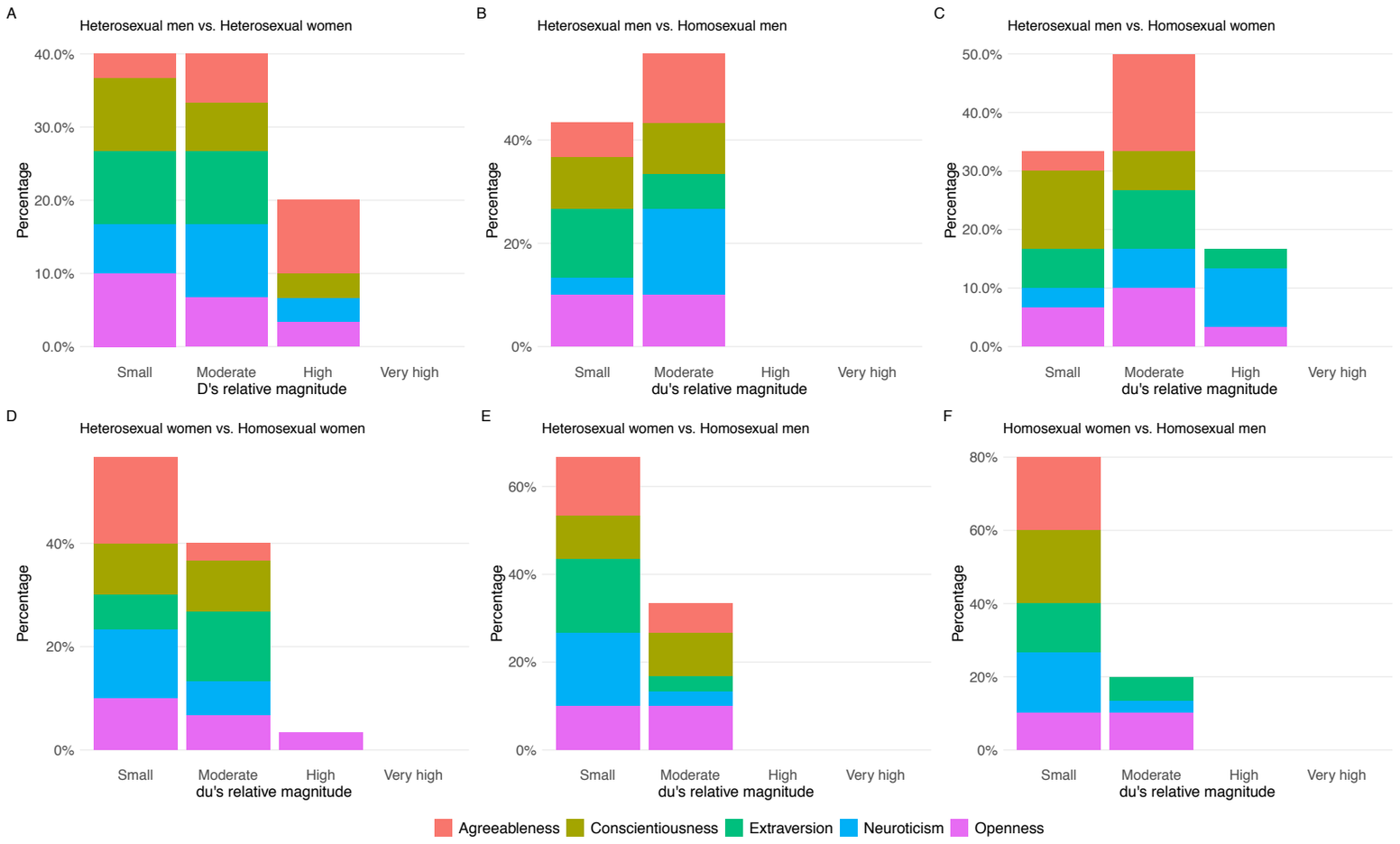

Figure 5.2

Percentage of facets according to their domains yielding small, moderate, high, and very high distances across group comparisons. 'Small' distance represents du values = (-.2, .2); 'Moderate' distance represents du values = [|.2, .5|); 'High' distance represents du values = [|.5, .8|); 'Very high' distance represents du values equal or higher than |.8|

Section summary

The personality distance (i.e d_u) between heterosexual men and heterosexual women was as large as the distance between heterosexual men and homosexual women. The distances between heterosexual women and the homosexual groups were smaller but not as small as the distance between both homosexual groups. Overall, the differences at facet level between heterosexual men and heterosexual women sourced from different domains but the largest differences came from Agreeableness facets. Interestingly, heterosexual men differed from homosexuals mainly on Agreeableness and Neuroticism. Heterosexual women differed from homosexual men on all domains similarly, but differences with homosexual women were mainly driven by Openness facets. From Figure 5.2 and Table 5.10 the higher differences across the four groups seemed to originate from Openness and Neuroticism facets.

5.5.2.8 The Shift Hypothesis vs. the Inversion Hypothesis at facet level

As mentioned in the introduction, the Shift Hypothesis would predict that the scores of homosexual people would shift away from their heterosexual counterparts. On the contrary, the Inversion Hypothesis predicts that there would be a complete cross-over, so homosexual men would score as heterosexual women and homosexual women as heterosexual men. This section compares the four groups across facets to determine if there was evidence for any of these hypotheses. Figure 5.3 serves as a rough summary, displaying in a single dimension the average raw score of each group, facilitating the identification of shifts in personality scores across sex/gender and sexual orientation groups. Note that the use of raw averages might distort the actual size of the distance between groups.

Replicating our findings at domain level there was strong support for the Shift Hypothesis (but not so much for the Inversion Hypothesis) on several Agreeableness facets. In "Trust","Straightforwardness" and "Altruism" facets, personality scores of homosexual participants shifted away from their heterosexual counterparts toward the center, as the Shift Hypothesis would predict. We also observed unilateral displacements of homosexuals man's scores towards heterosexual woman's scores in "Modesty". Similarly, homosexual man's

Table 5.10

Facets with the highest distance, d_u, across group comparisons

Domain	Facet	Distance size
Heterosexual men vs. Heterosexual women (M= .31, SD=.21)		
Agreeableness	Altruisim	-0.70
Neuroticism	Self-Consciousness	-0.68
Openness	Feelings	-0.67
Agreeableness	Straightforwardness	-0.60
Conscientiousness	Dutifulness	-0.60
Heterosexual men vs. Homosexual men (M= .21, SD=.11)		
Agreeableness	Altruisim	-0.42
Neuroticism	Anxiety	-0.41
Agreeableness	Tender-Mindedness	-0.39
Openness	Feelings	-0.33
Neuroticism	Depression	-0.32
Heterosexual men vs. Homosexual women (M= .31, SD=.18)		
Neuroticism	Self-Consciousness	-0.66
Openness	Feelings	-0.64
Neuroticism	Anxiety	-0.61
Extraversion	Gregariousness	0.56
Neuroticism	Depression	-0.51
Heterosexual women vs. Homosexual women (M= .19, SD=.15)		
Openness	Values	-0.52
Extraversion	Gregariousness	0.47
Agreeableness	Altruisim	0.39
Openness	Ideas	-0.38
Conscientiousness	Self-discipline	0.37
Heterosexual women vs. Homosexual men (M= .17, SD=.12)		
Conscientiousness	Dutifulness	0.41
Neuroticism	Self-Consciousness	0.40
Extraversion	Positive emotions	0.39
Openness	Feelings	0.32
Agreeableness	Straightforwardness	0.30
Homosexual women vs. Homosexual men (M= .14, SD=.10)		
Neuroticism	Self-Consciousness	0.42
Extraversion	Gregariousness	-0.38
Openness	Feelings	0.31
Openness	Values	0.30
Openness	Action	-0.26

Note:

Lables indicate the reference group. Positive values means that the first group was higer, conversely, negative values reflect that it was lower. M= Mean distance across all facets; SD= Standard desviation across all facets

scores displaced in "Tender-Mindedness" until they were almost identical to heterosexual women, hence there was evidence of a unilateral scores' inversion. All four groups scored similarly in the "Cooperation" facet, although there seemed to be a small group division according to sexual orientation.

It could be said that Neuroticism facets supported the Social Stress Theory because homosexual men scored higher than heterosexual men in all facets and homosexual women scored higher than the rest of the groups in "Depression", "Anxiety", and to a less extend, in "Impulsiveness". On the other hand, it could be considered that Neuroticism facets offered partial support for the Inversion Hypothesis, as homosexual men scored as heterosexual women in "Vulnerability", "Depression","Anxiety",and "Anger". In "Self-Consciousness" facets, homosexual man's scores unilaterally displaced towards the center but were still far from woman's scores. As previously found in the "Cooperation" facet, groups clustered according to sexual orientation in the "Impulsiveness" facet.

In Conscientiousness, the direction of the differences between heterosexual men and women varied across facets, potentially explaining why the sex/gender differences at domain level in this study were small. Moreover, this contradicting pattern of results at the facet level could also explain why sex/gender differences are not consistent across studies. Heterosexual men scored higher than heterosexual women in "Deliberation" and "Cooperation", conversely, heterosexual women scored higher in "Self-discipline" and "Dutifulness". Lastly, both scored similarly in "Achievement striving". Conscientiousness facets offered more support to the Inversion Hypothesis. Consistent with previous findings, there was a unilateral inversion of homosexual men scores in the "Self-discipline", "Order", "Deliberation", and "Competence" facets. These results deviate from what was found at domain level, where inversions were mostly on homosexual women's scores. The only evidence for the Shift Hypothesis was in the "Dutifulness" facet as the scores of homosexual men and women simultaneously displaced towards the center. Lastly, in the facet where heterosexuals scored similarly -i.e. "Achievement striving"- homosexual scores were displaced in the same direction and groups cluster

according to sexual orientation.

In Extraversion, as with Conscientiousness, the direction of the differences between heterosexual men and women varied across facets, which might explaining inconsistencies in the literature. Heterosexual men scored higher than heterosexual women in "Excitement-seeking" and "Assertiveness", while heterosexual women scored higher in "Warmth" and "Positive emotions". Heterosexuals scored similarly on "Gregariousness" and "Activity". Extraversion facets did not provide much support for the Shift Hypothesis, except for the "Excitement-seeking" facet, where homosexual men scored between heterosexual men and women. There was a unilateral inversion of homosexual man's scores in "Positive emotions", "Gregariousness", and "Assertiveness". In the "Activity" facet groups were divided again according to sexual orientation. Overall, as it happened at the domain level, Extraversion results did not offer evidence for both hypotheses.

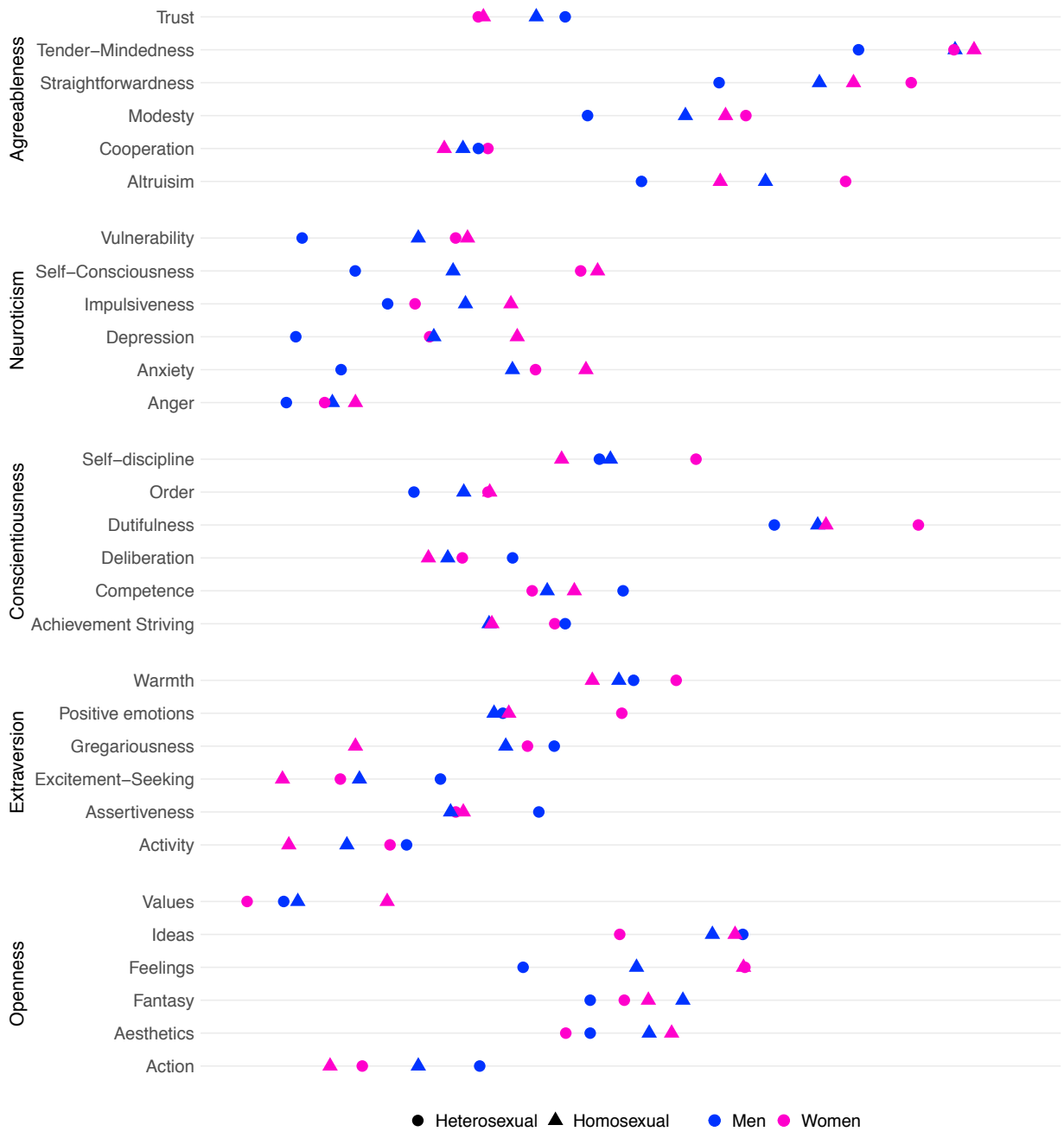

Figure 5.3

Average scores at each personality facet per sex/gender and sexual orientation group.

Notes: According to the Shift Hypothesis, personality scores of homosexual people should shift away from their heterosexual counterparts. The Inversion Hypothesis predicts that homosexuals should cross-over and score as the heterosexuals of their opposite sex. In this figure, the Shift Hypothesis is supported when the circles representing homosexual participants lay in between the triangles, which represent heterosexual participants. Conversely, the Inversion Hypothesis is suported when a circle and a triangle overlap, provided they are from different sexes. See full details on Table 7.2 in the Appendix.

Similarly to the domain level, Openness to experience offered evidence for both the Shift and the Inversion Hypotheses. As with Extraversion and Conscientiousness, the direction of differences between heterosexual men and women varied across facets. Heterosexual men scored higher in "Ideas" and "Action", while heterosexual women scored higher on "Feelings". Both heterosexual men and women scored similarly in "Aesthetics" and "Fantasy". Consistent with previous results, groups clustered according to sexual orientation in these two facets: "Aesthetics" and "Fantasy". Results also showed that homosexual man's scores were unilaterally displaced towards the center in "Action" and "Feeling" facets. Moreover, homosexual man's scores inverted in the "Ideas" facet. Lastly, homosexual woman's scores in "Values" were displaced extremely away from their heterosexual counterparts.

Taken together, these results support both the Shift and the Inversion Hypotheses. As happened at domain level, there was strong support for the Shift Hypothesis in Agreeableness and some more evidence was scattered across the rest of the domains, although it was often just unilateral shifts. There was strong support for the Inversion Hypothesis on Neuroticism and Conscientiousness facets, but only as referred to homosexual men. Finally, there was mixed evidence in the Extraversion and the Openness to experience facets. These results suggest that most evidence was partial, as there were few cases in which both homosexual groups changed their scores in the direction predicted by either of the hypotheses. Still, these results serve as evidence that sexual orientation is a relevant variable when studying personality as a continuum.

5.6 Overall summary: Comparing results at facet and domain level

It is interesting to compare the results at domain and facet level. Table 5.11 shows the individual facets and domains with the highest d_u across group comparisons, evidencing a reasonable amount of consistency across analyses. At both, facet and domain level, Neuroticism consistently yield the highest differences regardless of sexual orientation.

Table 5.11
Summary of facets and domains driving differences between groups

Group Comparison	Facet	Domain
Heterosexual Men vs. Heterosexual Women	Agreeableness and Neuroticism	Agreeableness and Neuroticism
Heterosexual Men vs. Homosexual Men	Agreeableness and Neuroticism	Neuroticism and Agreeableness
Heterosexual Men vs. Homosexual Women	Neuroticism and Openness	Neuroticism and Extraversion
Heterosexual Women vs. Homosexual Women	Openness and Extraversion	Extraversion and Openness
Heterosexual Women vs. Homosexual Men	Conscientiousness and Neuroticism	Conscientiousness and Agreeableness
Homosexual Women vs. Homosexual Men	Neuroticism and Extraversion	Neuroticism and Extraversion

Differences between heterosexual men and women were concentrated in Agreeableness and Neuroticism at facet and domain level. Interestingly, these two domains were also strongly influencing the differences between heterosexual and homosexual men at both facet and domain level. The differences between heterosexual men and homosexual women were also concentrated in Neuroticism, along with Openness at the facet level and Extraversion at the domain level. When comparing heterosexual women to homosexual men, Conscientioussness was relevant at the domain and facet level, as well as Neuroticism at the domain and Agreeableness at the facet level. However, differences between heterosexual women to homosexual women concentrated in Openness and Extraversion. Lastly, homosexual men and women differed mainly on Neuroticism and Extraversion.

5.7 Discussion

Personality differences between men and women have often been explored. However, few studies have considered personality differences across sexual orientation groups. Sexual orientation is a relevant variable, for it might reflect the continuum nature of personality traits better than sex/gender does. Two hypotheses – the Shift Hypothesis (Lippa, 2001) and the Inversion Hypothesis (Freud, 1953) – have been proposed to predict how sexual orientation would shape personality scores.

Around 500 participants from the UK were gathered through different social media platforms, as well as Prolific. The sample was divided into four groups according to participant's sex/gender and sexual orientation (i.e. heterosexual men, heterosexual women, homosexual men, and homosexual women). Evidence was found for multivariate normality (MVN) and variance homogeneity at domain but not at facet level. Hence personality differences were explored between the four sex/gender and sexual orientation groups, using Mahalanobis D for the former and Cohen's d for the latter.

With respect to personality differences between heterosexual men and women, the results supported the hypotheses and replicated previous research showing that women scored higher than men in Agreeableness and Neuroticism at facet and domain level (e.g.: Costa et al., 2001; Heineck, 2011; Kajonius and Mac Giolla, 2017; Lippa, 2005, 2008; Mac Giolla and Kajonius, 2019; Schmitt et al., 2008). On the contrary, our results did not replicated sex/gender differences in Extroversion at domain level (e.g.: Costa et al., 2001; Heineck, 2011; Kajonius and Mac Giolla, 2017; Lippa, 2005, 2008, 2010; Mac Giolla and Kajonius, 2019; Schmitt et al., 2008), although women scored higher than men in most of its facets. As was predicted in Hypothesis 1, there was a mix pattern of results at facet level in Openness to experience and Conscientiousness which, might explain why some researchers have found that women score higher (Costa et al., 2001; Kajonius and Mac Giolla, 2017; Lippa, 2005; Mac Giolla and Kajonius, 2019; Schmitt et al., 2008) while others have claimed the opposite (Heineck, 2011; Kajonius and Mac Giolla, 2017). Moreover, there was mixed evidence for Hypothesis 2 as

homosexual participants did not systematically score higher than heterosexual participants in Neuroticism and Openness. On the one hand, heterosexual women scored lower than homosexual men and women in Openness at the facet and domain level, but these results did not replicated with heterosexual men. On the other hand, homosexual men did not score higher than heterosexual women in Neuroticism at either domain or facet level. Homosexual women did score higher than heterosexual women but only in some Neuroticism facets. Nonetheless, results suggests that Neuroticism rose as an important trait at facet and domain level, distinguishing men and women regardless of their sexual orientation. These results offer certain support the Social Stress Theory, especially at the facet level.

Regarding the distance between groups, the biggest distance - the largest difference - at facet and domain level was between heterosexual men and women. Interestingly, the smallest distance, at domain and facet level, was between homosexual men and women. These results replicate previous research showing that differences between heterosexuals are larger than between homosexuals, suggesting that personality scores displace according to sexual orientation (Lippa, 2005). This finding could be taken as an indirect evidence for the Shift Hypothesis, as it shows that the distance between homosexual participants does not mirror that between heterosexual participants. However, this overall distance estimate could be biased by large differences on a few traits. Hence, we tested both hypotheses directly by comparing the four groups against each other.

In terms of the evidence supporting the Shift and Inversion Hypotheses, our results replicated Lippa (2005), showing that both may be found at facet and domain level. Similar to Lippa (2008), we found strong support for the Shift Hypothesis in Agreeableness at facet and domain level. Other researchers have found similar, but not as convincing, results (e.g.: Allen and Robson, 2020; Lippa, 2005). On the contrary, previous studies (e.g.: Lippa, 2005, 2008; Zheng et al., 2008) have observed solid evidence for the Shift Hypothesis in Neuroticism, but our results mostly supported the Inversion Hypothesis, replicating Greaves et al. (2017). We also replicated Greaves et al. (2017) results on Conscientiousness. Our Extraversion

results showed mix evidence at the facet level, but there was a clear tendency for homosexual participants' scores to often deviated from their heterosexual counterparts. Contrary to Lippa (2005) and Zheng et al. (2008), there were differences between heterosexual and homosexual groups at the facet and domain level. Also, unlike Allen and Robson (2020) and Lippa (2008), we did not find that homosexual woman's scores displaced more often than those of homosexual men. Consistent with previous research (e.g.: Allen and Robson, 2020; Lippa, 2005, 2008), the homosexuals participants scored higher in Openness to experience than either heterosexual group. Hypothesis 3 proposed that the extent to which homosexual scores shifted would be related to the degree to which heterosexual men and women differed. Although, this was not the case across facets and domains, we found that homosexual scores were displaced to extreme values when heterosexuals scored similarly. This was observed in Extraversion at the domain level, as well as in Openness at domain and facet level.

All in all, previous studies have found some evidence for the Shift Hypothesis in Extraversion, Agreeableness, and Neuroticism. Evidence for Conscientiousness has been less consistent and there has been no support in Openness to experience. I replicated these results in Agreeableness. However, results also supported the Inversion Hypothesis in Neuroticism and Conscientiousness, especially as it regards to homosexual men. This partly contradicts our Hypothesis 4 as, although the strict requisites of the of the Inversion Hypothesis should make it less likely to capture personality scores' shifts in the homosexual population, scores inversion were frequently observed.

This study suggests that there might be grounds to some stereotypes on behavioral differences between heterosexual and homosexual individuals. Besides this anecdotal connection, these personality differences across sex/gender and sexual orientation groups might prove relevant in other settings where personality plays a role. For instance, as we will further explore in the next Chapter (Chapter 6), personality sex/gender differences might be related to personality differences across occupations. Interestingly, some of Lippa's studies (e.g.: Lippa, 2005, 2008) support the idea that there is also a shift in the occupational preferences of

homosexual participants. According to Lippa (2005) "93% of lesbians have more male-typical interests than the average heterosexual woman does, and the effect size for gay versus heterosexual men implies that 90% of gay men have more female-typical interests than the average heterosexual man does" (p. 145). These results have been replicated in samples around the world like in China (e.g.: Zheng et al., 2011) or Samoa (e.g.: Semenyna, 2016). The question remains about the relationship between the shifts in personality and in occupational preferences.

On the question of whether homosexual personality scores are better predicted by the Shift or Inversion Hypotheses, and following this discussion, our results did not provide a definite answer. However, they did show that sexual orientation was a relevant variable distinguishing participant's personality scores at facet and domain level. As discussed in the introduction, to the extent that scores' inversion imply a shift, the Shift Hypothesis is more comprehensive. Therefore, researchers might want to expand the definition of the Shift Hypothesis to accommodate any degree of displacement in personality scores according to sexual orientation. On the other hand, these results suggest that inversions are frequently found, although might often be unidirectional. Further research would be needed to clarify this debate.

5.7.1 Limitations and strengths

The main limitation of this study was the small sample size (around 100 participants per sex/gender and sexual orientation group), which might have attenuated the power to detect small group differences. The scope of this study was also limited in terms of the number of sexual orientations we considered, as we did not include participants self-identifying as bisexuals or asexuals. Nevertheless, our sample has a reasonable size considering that, according to the latest UK census, only 1.4% of the respondents self-identified as homosexuals, and just 0.9% as bisexuals (Office for National Statistics, 2020).

The main strength of our study was that we measured group differences in terms of the multivariate distance between groups which provides a more accurate account of personality

group differences. Additionally, this still allowed us to explore if these group differences were driven homogeneously by all traits. Furthermore, we studied these differences at the facet and the domain level, which gives a more detailed account of what personality traits differentiate people across sex/gender and sexual orientation. Moreover, been able to compare results at the facet and domain level helped identifying why previous research results might seem contradictory. Finally, this study is among the few that has explored personality differences between heterosexuals and homosexuals participants. Note that Jonason and Luoto (2021) recently hypothesized that this type of research might be deliberately ignored to avoid misinterpretation of research results. However, this line of research allows us to explore the continuous nature of personality and to further understand sexual differentiation.

Altogether, our results provide important insights to the role of sexual orientation in personality differentiation. However, more research is needed to explore which facets and domains are more likely to show scores' shifts or inversions. Further work is needed to determine if the tendency for homosexual and heterosexual's scores to cluster around extreme values in certain traits is a universal phenomenon. Additionally, future research is required to explore why scores' shifts do not manifest uniformly across personality traits.

Chapter 6

Personality Differences Across Ocupations and Sexes/Genders

6.1 Introduction

6.1.1 Sex/gender differences in occupational preferences

Sex/gender differences in occupational aspirations have been found to be stable from an early age. For instance, girls in primary and secondary school tend to prefer subjects like literature or history, whilst boys like those related to Science, Technology, Engineering, and Mathematics, *STEM* (Croll, 2008; Dekhtyar et al., 2018; Gore et al., 2015, 2017; Holmes et al., 2018; Howard et al., 2011; Lubinski and Benbow, 1992). Biology related fields like Medicine seem to also be preferred by girls. Some fields like art, law, and music seem to be gender-neutral (Howard et al., 2011).

Su et al. (2009) meta-analysis showed that men had higher interest in STEM fields. For instance, the mean effect sizes (Cohen's d) for sex/gender differences in occupational preferences was .36 in science-related fields, .34 in mathematics, and 1.11 in engineering, with only slightly smaller effect sizes in older cohorts. Parker et al. (2014) found similar results in a sample of teenagers between 18-19 years old: 7% of females chose math, physics, and engineering (vs. 45% of males), 56% females preferred humanities and social sciences (vs. 26% males), 21% females inclined towards biological and medical sciences (vs. 15% males), and 16% females opted for law, economics, and business (vs. 14% males). Ellis et al. (2012) used a sample with a wider age range (from 18 to 56) and found consistent small-to-moderate sex/gender differences on specific jobs, women preferring beautician (d=-.42), daycare provider (d=-.40), and nurse (d=-.30) while men preferred being an athlete (d=.37), auto-mechanic (d=.34), test pilot (d=.32), or a police officer (d=.24). Richard A. Lippa has published several works showing that men and women incline towards different specific types of jobs and, although the effect sizes vary across countries, the results' patters are similar (Lippa, 2005, 2008, 2010, 2020).

6.1.2 Origins of sex/gender differences in occupational preferences

Recently, Stewart-Williams and Halsey (2021) published a comprehensive summary on sex/gender differences in occupational aspirations and the uneven sex/gender division across occupational fields – e.g.: men working in STEM while women work in non-STEM occupations. Stewart-Williams and Halsey (2021) conclusions suggested that occupational choices are influenced by stable personal characteristics[1]. Although this is a recent paper, this hypothesis is not new. Gottfredson and Holland had already proposed similar ideas in the mid-20th century (see Chapter 3). In fact, the relationship between personality and occupation has been of great interest for almost a century (e.g Bordin, 1943). Nonetheless, it is important to highlight that little research has focused on personality differences across actual occupations.

This study aims to tackle the limitations of previous research by studying the relation between personality traits and actual occupations. Based on the ideas put forward by Stewart-Williams and Halsey (2021), I hypothesize that if woman's unequal representation across occupation categories (e.g.: STEM vs. Non-STEM) was partly explained by enduring personal characteristics like personality traits, then the personality differences across fields should mirror personality sex/gender-differences. This study could have the potential of explaining uneven sex/gender-representation across fields (Ceci et al., 2014), improving vocational counseling, and enhancing job stability by better matching individuals to specific positions (Lubinski, 2000; Wilk et al., 1995).

6.1.3 Holland's RIASEC

As reviewed in Chapter 5 several pieces of research have found women scoring higher in Extraversion, Agreeableness, and Neuroticism. Sex/gender-differences in Conscientiousness and Openness to experience are still less clear. Arguably, these sex/gender-differences in

[1]The fact that these personal characteristics are stable does not necessarily imply that they are exclusively genetically driven. These characteristics might have been socialized at a very early age and, then, stabilized. We will discuss the degree to which genetic and environmental factors play a role in these personal characteristics, as well as its implications for Individual Differences research in Section 7.3.2.

personality could make women better suited for specific jobs – e.g.: higher Agreeableness might be related with less female representation in solitary jobs (Hunalp, 2020).

As previously discussed, the relationship between occupations and personality traits has been thoroughly studied, but actual occupational titles have rarely been used. Research on personality traits and occupations has been mostly based on Holland's personality-environment typology, which establishes a link between people's *interests* and occupations or *enviroments*. Holland created interests measures so they mirrored "personality inventories which reveal information such as the person's values, attitudes, needs, self-concept, preferred activities, and sources of threat and dissatisfaction" (Holland, 1959, p. 36). Holland's model includes six categories (i.e. realistic, investigative, artistic, social, enterprising, and conventional – RIASEC), which have been used to describe both occupation's and people's characteristics. According to Holland, *realistic* individuals like working with objects and machinery, *investigative* people have scientific pursuits, *artistic* people pursue self-expression opportunities and aesthetics, *social* individuals like being in contact with others, *enterprising* people like corporate environments, and *conventional* individuals prefer well-structured tasks (Holland, 1966). Men have been found to prefer realistic occupations whereas women tend to be more attracted by social, artistic (Hoff et al., 2020; Hunalp, 2020; Lippa, 1998; Su et al., 2009; Woods and Hampson, 2010), and conventional occupations (Lippa, 1998; Su et al., 2009; Woods and Hampson, 2010).

Larson et al. (2002) meta-analysis showed that interest in realistic occupations correlated negatively with Neuroticism (r=*-.08* for men and r=*-.10* for women, all correlations $p <$ .002), as well as with Openness to experience, positively for women (r=*.17*). Interest in investigative occupations correlated for men and women negatively with Neuroticism (r= *-.10*), positively with Openness to experience (r=*.26* for men and r=*.29* for women), and to Conscientiousness (r=*.05* for men and r=*.12* for women). Interest in artistic occupations correlated with Openness to experience (r=*.48*) and extroversion (r=*.09* for men and r=*.13* for women). Interest in social occupations correlated positively with all personality domains

except for Neuroticism (r=*-.08* for men and r=*-.14* for women). Interest in social occupations correlated especially high with Extraversion (r=*.32* for men and r=*.31* for women). Interest in enterprising occupations was negatively related to Neuroticism (r=*-.18* for men and r=*-.20* for women) but positively with Extraversion (r=*.44* for men and r=*.39* for women). Interest in conventional occupations correlated with Conscientiousness (r=*.30* for men and r=*.21* for women), Openness to experience (r=*.05* for men and r=*.16* for women), and Extraversion for men (r=*.16*), as well as negatively with Neuroticism (r=*-.12* for men and r=*-.07* for women). Larson et al. (2002) found five significant correlation differences between men and women: realistic occupations and Openness were only correlated for women, artistic occupations and Neuroticism were only correlated for men, conventional occupations and Extraversion were only correlated for men, conventional occupations and Openness were more highly correlated for women, and conventional occupations and Conscientiousness were more positively correlated for men. This results suggest a relation between personality traits and interests for women and men.

Soon after, Mount et al. (2005) carried another meta-analysis and replicated Larson et al. (2002) results. Mount et al. (2005) did not split their sample by sex/gender but, out of the 30 correlations between the Big Five and RIASEC, they found the same four correlations that surpassed .20 in Larson et al. (2002): extraversion and enterprising (r=*.40*, all correlations $p < .01$), openness to experience and artistic (r=*.41*), extraversion and social (r=*.29*), as well as, openness to experience and investigative (r=*.25*). Therefore, there are only a few moderate correlations between RIASEC categories and personality at the domain level.

Larson et al. (2002) and Mount et al. (2005) argue that personality domains and RIASEC categories must be different constructs, for none of the former explained more than 34% of the latter's variance. Other studies make similar claims on the basis that RIASEC categories are more stable during childhood and young adulthood than personality traits (e.g.: Hoff et al., 2020; Low et al., 2005). However, Holland himself defined RIASEC categories as personality inventories and research has shown a certain overlap between the two. Given

that the relationship between personality and RIASEC categories is likely to be confounded, it might be redundant to study personality differences across occupations when this are coded into RIASEC categories.

In conclusion, extensive research has shown that occupations attract people with specific personality traits. However, previous research has three main drawbacks. Firstly, most research has used RIASEC categories, which seem to overlap with personality measures. Secondly, regardless of this overlap, RIASEC categories cannot be directly translated into actual occupations. Lastly, previous research has only used the Big Five model at the domain level which, as discussed in Chapter 4, might not be comprehensive enough and can obscure patterns at lower levels of aggregation. This study aims to tackle these limitations as well as to expand this research by operationalizing occupations not only as occupational field (i.e. STEM vs. Non-STEM), but also as occupational orientation (i.e. Prediger's people-things dimension).

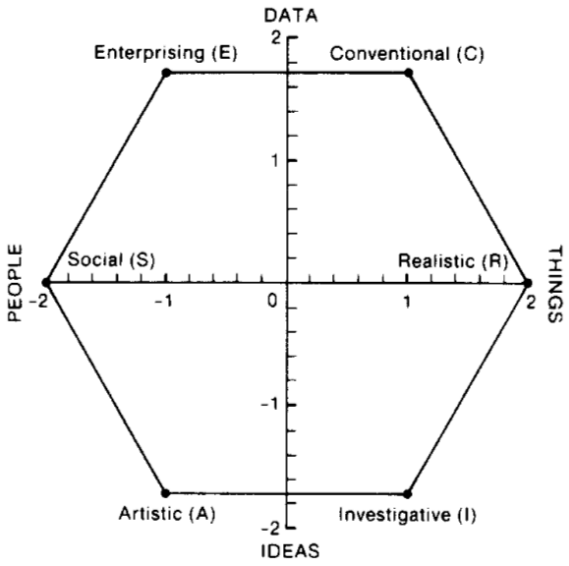

Figure 6.1
Representation of Holland's RIASEC and Prediger's dimensions.

Notes: Figure taken from "Dimensions underlying Holland's hexagon: Missing link between interests and occupations?" by D. J. Prediger, 1982, Journal of Vocational Behavior, 21(3), p.260 (http://doi.org/10.1016/0001-8791(82)90036-7).

6.1.4 *People vs. Things*

Prediger (1982) found that RIASEC categories could be factor-analyzed into two by-polar dimensions representing occupational orientation: people-things and data-ideas (Figure 6.1). The latter dimension informs about jobs that rely on pre-existing information as opposed to creating and developing new concepts. The people-things dimension reflects the degree to which socializing (e.g.: working with or directing teams, meeting clients, etc) is more relevant than working with objects or machinery. People-oriented jobs are related to Holland's enterprising, social and artistic occupations; while things-oriented jobs are closer to conventional, realistic, and investigative occupations. Given the previously reviewed correlations between personality and RIASEC categories, people-oriented occupations are likely to be positively correlated to Extraversion and Openness, while things-oriented occupations may be positively correlated to Conscientiousness and Openness. Both people-oriented and things-oriented occupations might correlate negatively with Neuroticism.

Subsequent research showed that, unlike the data-ideas dimension, people-things occupational orientation correlated strongly with sex/gender. Research has shown that women gravitate around people-oriented occupations and men towards things-oriented occupations (Lippa, 1998; Hustad et al., 2020; Su et al., 2009). Lippa (1998) found that the effect sizes for sex/gender differences in the people-things orientation across three of his studies were higher than 1.20. This is among the largest sex/gender-difference effect sizes ever found (Lubinski, 2000) but it might be inflated due to the aggregation of RIASEC scales which already present large sex/gender differences. In a recent meta-analysis, Hoff et al. (2018) showed that sex/gender differences in people-things dimensions widen during secondary school, precisely when students make relevant choices about their educational fields. However, this gap narrows across adulthood, when women become more attracted to things-oriented occupations and men increase their preference for people-oriented occupations (Hoff et al., 2018). Still, Lippa et al. (2014) argued that, in developed countries, sex/gender segregation between people- and things-oriented occupations is larger than across occupational status. Recently, Luoto (2020) showed that this tendency is even present in scientific publishing, where more

women publish in people-oriented fields and more men in things-oriented fields. Altogether, these research findings suggest that occupational orientation (i.e. Prediger's people-things dimension) is an important predictor of sex/gender differences in the labor market.

6.1.5 Occupational field vs. Occupational orientation

The unequal sex/gender-representation in people vs. things occupations, as well as in STEM vs. non-STEM positions, lead many researchers to hypothesize that the STEM vs. non-STEM divide mirrors the people-things dimension (e.g.: Lippa et al., 2014; Stewart-Williams and Halsey, 2021). To my knowledge, few studies have tested this hypothesis directly. Hence, we can only rely on the indirect information that some studies provide.

For instance, Hustad et al. (2020) reported that people-oriented occupations correlate strongly with verbal demands (r=.86) and weakly with mathematical-demands (r=-.16). Given that, from a young age, women tend to outperform men in verbal tasks (Benbow et al., 2000; Stoet and Geary, 2013, 2018) and that there is a strong link between abilities and interests (see Chapter 3), their results suggested that women are more suited to people-oriented occupations. Moreover, this could mean that there is a latent factor like cognitive abilities contributing to the sex/gender divide in both occupational field and orientation.

However, Hustad et al. (2020) also found that things-oriented occupations were strongly negatively related to verbal-demands (r=-.82), but did not significantly correlate to mathematical-demands. Hence, mathematically demanding occupations were compatible with both, things and people-oriented occupations. Additionally, Lippa et al. (2014) asked 78 college students to rate the orientation of 60 occupations and found that certain occupations, especially those related to biology, were fairly distributed across the people-things spectrum. Together, these two studies suggests that people- or things-oriented occupations might be found within the same occupational field. In other words, it is possible that the occupational orientation dimension is nested within the STEM vs. Non-STEM divide.

6.1.6 Evidence on personality differences between actual educational and occupational carreers

Few studies have compared personality across actual educational or occupational fields. Summarizing their results is not an easy task, for there is neither convergence in the personality inventories, nor on the operationalization of university studies or the occupational fields being used (De Fruyt and Mervielde, 1996; Vedel, 2016). As with occupations, some studies use bachelor titles (e.g.: Clariana, 2017) but others recategorize them into RIASEC (e.g.: Kaufman et al., 2013). Results also vary depending on which fields are included and compared – e.g.: social sciences students in De Fruyt and Mervielde (1996) scored the highest in openness but they might have scored lower than arts students had they been considered like in Kline and Lapham (1992).

Recently Vedel (2016) did a literature review on personality differences across university degrees. They found medium effect sizes in Extraversion, with students in Economics, Law, Political Sciences, and Medicine scoring higher than those in Arts, Humanities and Sciences. On the other hand, Natural Sciences tend to score higher in Agreeableness (De Fruyt and Mervielde, 1996; Rubinstein, 2005). Vedel (2016)'s literature review replicated this finding, in addition to few medium effect sizes showing that Law, Business, and Economics scored lower in Agreeableness compared to with Medicine, Psychology, Sciences, Arts, and Humanities. Regarding Neuroticism, students in Arts, Humanities, Psychology, Economics, and Business score higher than those in in Engineering, Law, and Sciences (Vedel, 2016). Evidence suggests engineers score higher in Conscientiousness (Clariana, 2017; De Fruyt and Mervielde, 1996; Kline and Lapham, 1992). Vedel (2016)'s also found medium effect sizes with Arts and Humanities scoring lower in Conscientiousness than Sciences, Law, Economics, Engineering, Medicine, and Psychology. Contrary to expectations, not all studies reported arts as the most Open to experience (De Fruyt and Mervielde, 1996; Rubinstein, 2005). Vedel (2016) found medium to large effect sizes with Humanities, Arts, Psychology, and Political Sciences scoring higher than Economics, Engineering, Law, and Sciences.

In accordance with these results, studies comparing people across occupations have also shown that different jobs attract specific personality profiles. Oh et al. (2018), for instance, compared the personality profiles of recently hired members at a pharmaceutical and a banking company and found that they were especially different in Neuroticism (d=.34), Openness (d=.14), and Conscientiousness (d=.10), but less so in Extraversion (d=.05) and Agreeableness (d=-.01). Lippa (2010), found that at an individual level and controlling for sex/gender, male-preferred occupations attracted individuals with lower levels of Extraversion, Neuroticism, and Agreeableness. Both pieces of research show that different occupational fields attract people with specific personality profiles. Additionally, King et al. (2017) found that employees working within the same occupational field tend to have more similar personality traits compared to other occupational fields.

6.1.7 Summary and hypotheses

From an early age, men and women incline toward different educational and occupational fields. The fact that women are overrepresented in Non-STEM and people-oriented occupations has lead experts to suggest that these differences could be due to some reasonably stable psychological characteristics. For instance, women may prefer occupations involving more socializing because they score higher than men in personality traits supporting social skills like Extraversion and Agreeableness.

Personality differences across occupations have been thoroughly studied. However, most studies have re-categorized occupations into Holland's RIASEC categories rather than using actual occupational titles. Actual occupations could be operationalized in terms of occupational field (e.g.: STEM vs. Non-STEM) or orientation (i.e. Prediger's people-things dimension). These categorization systems are often considered interchangeable (e.g.: Lippa et al., 2014; Stewart-Williams and Halsey, 2021) but, to my knowledge, no attempts have been made to understand the relation between the two.

Previous research has focused on personality differences at the domain level which might

obscure personality differences at lower levels of aggregation (see Chapter 4). Chapter 5 serves as an example of how sex differences at the domain level are averaged out and seem smaller than what they are. In this sense, using nuances could be considered the optimal level of granularity. However, nuances as a personality measure are not as established yet. Therefore, using both facets and nuances would provide a more clear and detailed account of the personality differences between occupations.

In this study we aim to test personality differences at facet and nuance level, using actual occupations categorized according to their field and their orientation. According to previous research, we hypothesize that:

- *Hypothesis 1 (H1):* If the uneven sex/gender representation in certain occupational fields was influenced by some stable characteristics like personality, the personality traits differentiating men and women should also show prominent differences between these occupational fields. In other words, personality differences across occupations should mirror sex/gender differences in personality.

- *Hypothesis 2 (H2):* Given that the occupational orientation dimension (Prediger's people-things) could be nested within the occupational field (i.e. STEM vs. Non-STEM) divide, differences in personality traits across occupational orientations will be greater than trait differences between STEM vs Non-STEM.

These questions will be investigated across two studies using independent samples from Estonia (Study 1); as well as China and the UK (Study 2).

6.2 Study 1

6.2.1 Methods

6.2.1.1 Sample

Data were retrieved from the Estonian Genome Centre (EGC) of the University of Tartu including over 51 000 participants representative of the Estonian population (for details see www.biobank.ee). Most participants were randomly recruited by their physicians, although volunteers were also included. Participants signed a consent form (available at www.biobank.ee) and completed a computer-assisted personal interview. A subset of over 3 200 randomly selected participants were asked to additionally fill in a self-report personality questionnaire.

6.2.1.2 Materials

Personality. Participants completed the Estonian version of the NEO-PI-3 (De Fruyt et al., 2009; McCrae et al., 2005) which is a slightly modified and more readable version of the Revised NEO Personality Inventory (NEO-PI-R, Costa and McCrae, 1992b). The NEO-PI-3 has 240 items on a 5-point Likert scale ranging from "completely disagree" to "completely agree". Items can be grouped into 30 personality facets which can also be summarized into the Big Five domains, such that each domain score is a composite of six facet scores. In addition, informants close to the participants (e.g.: parents, partners, etc.) were asked to complete the Estonian version of the NEO-PI-3 as if they were the participant. Allik et al. (2010) found large convergence between self and informant personality scores across cultures. Moreover, McCrae (1994) adviced the combination of self and informant-reported personality scores to enhance measurement reliability. Therefore we used both self and informant-reported personality scores in our analyses.

Occupational field. Participants' self-reported job titles we re-coded to match the 2008 version of the International Standard Classification of Occupations (ISCO-08). ISCO-08 groups

jobs into ten broad categories according to occupational level which is generally determined by the education qualification needed. There were two requisites to select the professional and technical occupations that were used required formal education (i.e. university studies or vocational education) as long as they were easily identifiable as pertaining to either STEM or Non-STEM. leaving aside ambiguous groups like "university teachers", "sports science" (see Table 6.1). Additionally, occupations in which sex/gender-representation was related to confounding variables like physical strength were also excluded (e.g.: mining).

Table 6.1
Grouping of ISCO codes into STEM and Non-STEM occupations

Grouping	Professionals (Categoryies 1000 to 1999 in ISCO-08)	Technicians and Associate Professionals (Categoryies 2000 to 3999 in ISCO-08)
STEM		
Science, Technology, Engineering, and Mathematics	21 Science and Engineering Professionals	311 Physical and Engineering Science Technicians
	211 Physical and Earth Science Professionals	312 Mining, Manufacturing and Construction Supervisors
	212 Mathematicians, Actuaries and Statisticians	313 Process Control Technicians
	214 Engineering Professionals (excluding Electrotechnology)	315 Ship and Aircraft Controllers and Technicians
	215 Electrotechnology Engineers	351 Information and Communications Technology Operations and User Support Technicians
	216 Architects, Planners, Surveyors and Designers	352 Telecommunications and Broadcasting Technicians

Grouping	Professionals (Categoryies 1000 to 1999 in ISCO-08)	Technicians and Associate Professionals (Categoryies 2000 to 3999 in ISCO-08)
	251 Software and Applications Developers and Analysts 252 Database and Network Professionals	
Non-STEM		
Biology and Medical Sciences	213 Life Science Professionals 221 Medical Doctors	314 Life Science Technicians and Related Associate Professionals
	222 Nursing and Midwifery Professionals	321 Medical and Pharmaceutical Technicians
	223 Traditional and Complementary Medicine Professionals	322 Nursing and Midwifery Associate Professionals
	224 Paramedical Practitioners	323 Traditional and Complementary Medicine Associate Professionals
	225 Veterinarians	324 Veterinary Technicians and Assistants
	226 Other Health Professionals	325 Other Health Associate Professionals
Bussiness	241 Finance Professionals	331 Financial and Mathematical Associate Professionals
	242 Administration Professionals	332 Sales and Purchasing Agents and Brokers
	243 Sales, Marketing and Public Relations Professionals	333 Business Services Agents

Grouping	Professionals (Categoryies 1000 to 1999 in ISCO-08)	Technicians and Associate Professionals (Categoryies 2000 to 3999 in ISCO-08)
		334 Administrative and Specialized Secretaries 335 Government regulatory associate professionals
Law	261 Legal Professionals	341 Legal, Social and Religious Associate Professionals
Art	265 Creative and Performing Artists	343 Artistic, Cultural and Culinary Associate Professionals
Teaching	232 Vocational Education Teachers 233 Secondary Education Teachers 234 Primary School and Early Childhood Teachers 235 Other Teaching Professionals	
Other Social Sciences	262 Librarians, Archivists and Curators 263 Social and Religious Professionals 264 Authors, Journalists and Linguists	

6.2.2 Analysis

Study 1 used a correlated beta approach to measure the degree to which personality differences across fields mirrored personality sex/gender-differences. Specifically, personality scores were first regressed to occupation, sex/gender, and age. This regression model was run separately for personalty scores at the facet and nuance level (i.e. individual items). Occupation and sex/gender were dummy coded so that 1 represented STEM and men, respectively. Therefore, the regression estimate for occupation would reflect the difference in personality scores between STEM and non-STEM, controlling for age and sex/gender. Equivalently, the regression estimate for sex/gender would represent the difference in personality scores between males and females, controlling for occupation and age. Including covariates was crucial for two reasons: (1) to compensate for the potential implications of the uneven sex/gender-distribution in the STEM/non-STEM divide, and (2) to control for potential personality changes associated with the time an individual has been involved in the same field or position (e.g.: Oh et al., 2018). Finally, the sex/gender and occupation beta coefficients were extracted for each personality measure and correlated. This correlation would reflect the degree to which personality differences across fields mirror the personality differences between men and women.

These analyses were run at nuance and facet level for both self-reported and informant-reported personality scores. Lastly, these results were compared to those using the average between self and informant-reported personality scores at nuance and facet level. Aggregating both sources of information will reduce the measurement error, providing more reliable results.

6.2.3 Results and discussion

Table 6.2 shows the number of observations per sex/gender according to the occupational field and occupational status (i.e. professional or technical occupations). The sample size difference between non-STEM and STEM was noticeable, although this should be expected

Table 6.2
Frequency of Estonian occupations according to sex/gender, occupational status and occupational field.

	STEM		Non-STEM		Total
	Professionals	Technicians	Professionals	Technicians	
Men	123	48	196	59	426
Women	62	17	617	268	964
Total	185	65	813	327	1390

Notes. STEM= Science, Technology, Engineering, and Mathematics.

given the limited number of positions included in STEM.

First, I correlated self and informant-reported personality items at both, nuance (M=.29, SD=.09) and facet level (M=.46, SD=0.08). Correlations at the facet level tended to be higher than at nuance level, possibly because of higher reliability of facet-level data which results when aggregating nuances' common variance at the expense of their specific variance.

Table 6.3 shows the Spearman's correlation (ρ) between sex/gender and occupational field regression estimates when predicting self- or informant-reported personality scores at nuance level. As Table 6.3 shows, there was a small and non-significant correlation between the sex/gender and occupation regression estimates (ρ = .02, CI= [-.15, .19] for self-reported and ρ = .12, CI= [-.05, .28], for informant-reported). Similar results were found when using personality scores at the facet level as the outcome variable: ρ = .09, CI= [-.39, .53] for self-reported and .19, CI= [-.30, .60] for informant-reported personality scores.

Interestingly, there was a high - and significant - correlation between the self and informant-reported regression estimates for occupation (ρ =. 50 at nuance and ρ =. 75 at facet level) and sex/gender (ρ =.69 at nuances and ρ =. 75 at facet level), showing that personality differences between sexes/genders and STEM/Non-STEM replicated across both sources of information. This justified averaging the personality scores from both sources of information:

Table 6.3
Correlations between the sex/gender and occupational field regression estimates when predicting self- or informant-reported personality scores at nuance level.

		Self-report		Informant-report	
		Occupation R.E.	Sex/gender R.E.	Occupation R.E.	Sex/gender R.E.
Self-report	Occupation R.E.	1			
	Sex/gender R.E.	.02	1		
Informant-report	Occupation R.E.	.50	.08	1	
	Sex/gender R.E.	.22	.69	.12	1

Notes. R.E.= Regression Estimates.

self-reported and informant-reported personality scores. Using this average of personality scores as the outcome variable replicated previous results. The Spearman's correlation between the occupation and sex/gender regression estimates was .13, p = .04 at the nuance level and .21, p = .26 at facet level.

Overall and contrary to Hypothesis 1, the correlation between the occupation and sex/gender regression estimates was not strong enough to claim that personality differences between fields mirror personality sex/gender-differences, regardless of who reported the personality scores (i.e. self- or informant- reported) and of their aggregation level (i.e. nuance or facet level).

6.3 Study 2

6.3.1 Methods

6.3.1.1 Sample

A survey containing questions on various life outcomes together with a personality questionnaire was uploaded to Formr (Arslan et al., 2020) and was distributed through social media (e.g.: Twitter, facebook, Instagram, etc.) as well as in Prolific. A total of 1303 UK participants and 2489 Chinese participants started the survey. Out of those, only 936 UK and 1405 Chinese participants reached the personality items and provided an occupational title that could be codified.

All participants under 18 years-old (N= 60 UK and N= 78 Chinese participants) were eliminated. After examining data quality with R's *careless* package (Yentes and Wilhelm, 2018), another 12 UK participants and 48 Chinese participants were eliminated for giving the same answer 10 or more consecutive times in the personality questionnaire. This left a total of 864 UK participants (64% Women; age range between 19 and 74) and 1279 Chinese participants (73% Women; age range between 19 and 60). Table 6.4 shows the samples' demographics.

Table 6.4
Frequencies and descriptive statistics of the UK and Chinese samples.

Variable	N	Percent of the total sample	M	SD
UK Sample				
Sex/gender	864	100		
Women	556	65.2	N/A	N/A
Men	297	34.8	N/A	N/A
Age	864	100	35.64	11.15

Variable	N	Percent of the total sample	M	SD
Income	863	99.9	3.54	1.50
No income	101	11.7		
<10 000 pounds	146	16.9		
10 001 -18 000 pounds	149	17.3		
18 001 - 26 000 pounds	184	21.3		
26 001 - 44 000 pounds	215	24.9		
>44 000 pounds	68	7.9		
Education	864	100	4.68	1.00
No educational qualifications	1	.01		
Primary school	3	.03		
Secondary school	148	17.1		
Vocational Education or Apprenticeship	143	16.6		
Undergraduate Education	390	45.1		
Postgraduate Education	179	20.7		
Prediger's people-things dimension	853	98.7	50.51	33.08
Chinese Sample				
Sex/gender	864	100		
Women	931	74.1	N/A	N/A
Men	326	25.9	N/A	N/A
Age	864	100	28.75	7.34
Income	859	99.4	3.73	2.00
No income	251	19.7		

Variable	N	Percent of the total sample	M	SD
<10 000 yuan	211	16.6		
10 001 -18 000 yuan	181	14.2		
18 001 - 26 000 yuan	71	5.6		
26 001 - 44 000 yuan	109	8.6		
>44 000 yuan	451	35.4		
Education	861	99.7	5.24	.75
No educational qualifications	1	.01		
Primary school	4	.03		
Secondary school	26	.02		
Vocational Education or Apprenticeship	128	10.0		
Undergraduate Education	610	47.8		
Postgraduate Education	507	39.7		
Prediger's people-things dimension	853	71.6	38.23	20.42

6.3.1.2 Materials

Personality. A total of 90 items were selected from the pool of items created by Henry, Mõttus, & Vainik (in preparation). Items were mainly retrieved from the International Personality Item Pool (IPIP, Goldberg, 1999) and the Synthetic Aperture Personality Assessment (SAPA, Condon, 2018). Some items were written by the authors. Items were selected to maximize their re-test reliability (M=.71, SD=.09, Values range=.60-.85), their clarity, readability, as well as to minimize their redundancy. Following Condon et al. (2021)

advice, items were selected and categorized into facets according to: (1) their re-test reliability, (2) feedback from Dr McCrae, and (3) fit within the IPIP's facet definitions. Domain scores were computed as the average across six facet scores, each computed as the sum scores of three items.

Note that only self-reported personality scores were available for Study 2. Although Study 1 and Study 2 used different personality measures, both included nuances and facets of the Big Five.

Occupational Field. Study 2 classified occupations according to the same category system used in Study 1: the International Standard Classification of Occupations (ISCO-08). As both studies used ISCO-08, Study 2 closely replicated the occupations that were included in Study 1. Self-reported occupations were re-coded by two independent researchers per sample (cross-rater agreement was .50 for the UK sample and .55 for the Chinese sample), disagreements were discussed until consensus was reached. In the UK there were 101 participants working in STEM and 234 working in Non-STEM occupations. In the Chinese sample there were 74 participants working in STEM and 404 working in Non-STEM occupations (Table 6.5).

Occupational Orientation. Occupational orientation was measured by asking participants how much time they spend working with people/animals compared to the time they invest working with things (e.g.: computers, machines, etc.). The actual item label was:

"Thinking about your job, how much time do you spend working with people/animals and things? You can use the slider to mark the percentage of time you devote working with people or animals. The remaining percentage will be the time you use to work with things. For example, if you move the slider to 40, it would mean that you expend 40% of the time working with people and 60% working with things."

The answer options showed some examples of "working with people/animals" and "working

Table 6.5
Frequency of UK and Chinese occupations according to sex/gender, occupational status and occupational field.

		STEM		Non-STEM		Total
		Professionals	Technicians	Professionals	Technicians	
UK	Men	29	19	50	16	114
	Women	35	17	105	61	218
Total		64	36	155	77	332
China	Men	28	7	46	17	98
	Women	23	14	305	32	374
Total		51	21	351	49	472

Notes. STEM= Science, Technology, Engineering, and Mathematics.

with things". To facilitate the comparison of occupational field and orientation – and given that STEM occupations have often been considered things-oriented – this item was reversed so it reflected the percentage of time invested working with things.

6.3.2 Analysis

Study 2 had two aims: replicating Study 1 and expanding it by operationalizing occupation as occupational orientation (i.e. Prediger's people-things dimension). To replicate Study 1, each personality score (either at nuance or facet level) was regressed to occupation (1=STEM, 0= Non-STEM), sex/gender (1= men, 0= female), and age. The sex/gender and occupation regression estimates where extracted for each personality measure and correlated. Analyses were run separately for the UK and the Chinese samples. Afterwards, to expand this replication, occupation was operationalized as occupational orientation (i.e. high values corresponding to things-orientated occupations). Additionally, the correlation between occupational field and orientation was also computed (polyserial correlation) to empirically test whether occupational field and orientation were equivalent categories.

Table 6.6
Correlation between the regression betas of occupation (STEM vs. Non-STEM) and sex/gender when using personality scores at item and facet level

	Item level	Facet level
UK Spearman's rho	-0.10	-0.16
China Spearman's rho	0.15	0.22

6.3.3 Results and discussion

Table 6.6 shows the Spearman's correlation (ρ) between the occupation and sex/gender regression estimates when predicting personality nuances or facets for both the UK and the Chinese samples. These results largely replicated Study 1. On the one hand, correlations at the facet level (ρ_{UK}= -.16, CI = [-.58, .33] and ρ_{China}= .22, CI = [-.27, .62]) were stronger than at the nuance level (ρ_{UK}= -.10, CI = [-.36, .17] and ρ_{China}= .15, CI = [-.13 .40]). On the other hand, these correlations were small. Surprisingly, the direction of the regression estimates correlations was not consistent across samples (i.e. correlations were negative in the UK sample but not in the Chinese sample). These small correlations could be due to random variation but it could also reflect a consistent cultural difference. This will be further explored in the General Discussion. Overall, all correlations in Study 2 were slightly larger than in Study 1, but would still be classified as small effects.

Study 1 and 2 results showed a non-meaningful fluctuation around zero. In this sense, results of Study 2 replicate Study 1: the correlation between the occupation and sex/gender regression estimates was not strong enough to claim that personality differences between fields mirror personality sex/gender-differences at nuance or facet level.

Regarding the relationship between occupational field and orientation, the polyserial correlation in both samples was small (r=-0.01) in the UK sample and r=0.16 in the Chinese sample), meaning that occupational field and orientation are not interchangeable. In other words, many people in STEM occupations spend more time working with people than with things and, conversely, many people in Non-STEM occupations spend more time working

Table 6.7
Correlation between the regression betas of occupation (i.e. Prediger) and sex/gender when using personality scores at item and facet level

	Item level	Facet level
UK Spearman's rho	0.30	0.38
China Spearman's rho	0.03	0.08

with things than with people.

The results from the regression estimates correlations between sex/gender and occupational orientation are presented in Table 6.7 for both the UK and the Chinese samples. The strength of the correlation was notably higher for the UK sample (r_{range}=.30-.38, both non-significant) compared to the Chinese sample (r_{range}=.03-.08, both non-significant).

6.4 General Discussion

Holland's (1959) theory of personality-environment typology, as well as Gottfredson's (1981) theory of Circumscription and Compromise, suggest that people partly shape their occupational aspirations to match their interests and personality traits (see Chapter 3). Recently, Stewart-Williams and Halsey (2021) reviewed the evidence on sex/gender-segregation in the labor market and concluded that the psychological make up differentiating each sex/gender could be contributing to this divide.

This study address this idea by studying the degree to which personality differences across occupations mirrored personality sex/gender differences. Previous research testing this relationship has not used actual occupational titles but Holland's RIASEC categories. Additionally, personality differences have only been considered at domain level. To tackle these limitations, occupation was operationalized as field (STEM vs. Non-STEM) and as orientation (i.e. Prediger's people-things dimension). Additionally, personality scores at nuance and facet level were used. The analyses consisted in, first, regressing the personality scores to sex/gender, age, and occupation; and then, extracting and correlating the sex/gender and

occupation regression estimates for each personality measure. These analyses were replicated in three samples.

Study 1 used an Estonian sample and, contrary to Hypothesis 1, found no evidence that personality differences across occupational fields mirrored personality sex/gender differences either when using self-reported or informant-reported personality scores at nuance or at facet level. This result was replicated in Study 2, using a UK and Chinese sample of self-reported personality scores at nuance and facet level. This evidences that the way in which occupation had previously been operationalized – especially when using RIASEC categories (e.g.: Hoff et al., 2020; Hunalp, 2020; Larson et al., 2002; Mount et al., 2005)– could have inflated the correlation between occupations and personality traits.

Study 2 expanded this replication by (1) testing the correlation between occupational field and orientation; and (2) substituting occupational field by occupational orientation. Occupational field and orientation were not correlated implying that, contrary to popular belief, STEM occupations do not necessary involve working exclusively with computers, materials, data, etc. Conversely, Non-STEM occupations do not require constant interaction with other people. Hence, these results corroborated what some previous research had roughly suggested (Hustad et al., 2020).

Moreover, results from Study 2 indicate that personality sex/gender differences mirror personality differences across occupational orientations, at least, in the UK sample. Thus, these results partially supported Hypothesis 2. Given that there seemed to be no relationship between occupational field and occupational orientation, it could be that other psychological variables (e.g.: intelligence, spatial ability, etc.) drive people into different field but it is personality what pushes individuals toward more people- or things-oriented jobs. As previous research shows less sex/gender differences in collectivists countries (See Chapter 5), it could be that this trend was not as evident in the Chinese sample due to a range restriction in the personality sex/gender differences. On the other hand, the results from Study 2 might be reflecting consistent cultural differences between the UK and China labor markets. For

example, in China, personality may be more relevant for choosing the occupational field than occupational orientation.

Overall, our results highlight the need to contemplate how variables are operationalized, especially as it regards to occupations. For instance, using broader fields, rather than STEM vs. Non-STEM might reflect more nuanced personality differences across occupations. Furthermore, certain operationalizations might not be as relevant across contexts to study their relationship to the human psyche - e.g.: occupational orientation might provide more information than occupational field. Lastly, albeit further replications are needed, Study 2 results seem to echo Holland and Gottfredson theories on occupational choice, and how they could be shaped by the interplay between intelligence, personality, and the surrounding context. We will further explore the relationships between personality and intelligence in applied settings in Chapter 7.

6.4.1 Strengths and limitations

The Chinese and UK samples were small, so further replications are needed. In fact, the contradicting results on personality differences across sex/gender and occupational orientation might be due to the small sample sizes. Future studies would be necessary to confirm whether these results on occupational orientation reflected an actual cultural difference. Another limitation of this study is the recategorization of self-reported occupational titles into ISCO-08 categories. Although ISCO-08 is a comprehensive categorization system, it still involves a substantial amount of subjectivity. Following studies might obtain more accurate codes by allowing participants to select the codes that best represented their position.

Nonetheless, this study counts with numerous strengths. The most noticeable strength is the replication of results across three samples, across self and informant-reported personality scores at nuance and facet level. Hence, the evidence regarding personality differences across sex/gender and occupational field seems robust. Additionally, this study also tackled numerous limitations of previous research such as using actual occupations and exploring

personality differences at nuance and facet level. Moreover, this line of research has enormous potential to explain uneven sex/gender-representation across fields (Ceci et al., 2014), and even to improve vocational counseling and job stability (Lubinski, 2000; Wilk et al., 1995).

Chapter 7

Final Conclusion

7.1 Thesis overview

This thesis includes two studies on cognitive ability, two studies on personality, and two introductory chapters (Chapters 1 and Chapter 4) summarizing the main milestones in Intelligence and Personality research. This final chapter synthesizes each of the studies (Section 7.2) and addresses some of the cross-cutting complexities present across all of them (Section 7.3).

At this point I consider it relevant to make a personal disclaimer. This Chapter discusses sensitive historical events like race segregation. Note that the opinions expressed throughout this thesis, and especially in this Chapter, are mine and only mine, and by no means aim to represent the values or political/moral positions of the University of Edinburgh or any of its staff members. As such, I have taken the liberty to change the tone and to invite the reader into a more informal conversation.

7.2 Summary of the findings and contributions to the field

Chapter 2 and Chapter 3 used data from a British longitudinal study, the Millennium Cohort Study (MCS). Both research pieces contributed to the integration of intelligence and educational research by showing how intelligence measures could complement classical theories from the Educational field. Promoting the cross-over between both fields is relevant because, beyond the historical and ethical reasons that contributed to their division (see Section 7.3.3), there is enough evidence showing that intelligence is a strong predictor of educational outcomes. Moreover, these two studies show how the combination of both fields lead us to explore new and innovative research avenues.

The goal of Chapter 2 was to expand Eccles's expectancy-value model so it included measures of cognitive ability. This study explored the relationship between verbal ability (as a proxy

for cognitive ability) and motivation changes after the primary to secondary school transition. Although effects were not strong, they showed (a) that Verbal Ability was more related to Self-Competence than to Task-Values, and (b) that students with higher Self-Competence and verbal ability at age 11 experienced less change in Self-Competence. Similarly, Chapter 3 integrated educational and intelligence research by exploring how feedback experiences like change in grades might mediate the relationship between verbal ability and occupational aspirations' change. As in Chapter 2, change was small and could not be measured reliably, so the mediation was not significant. Nonetheless, this study highlights the similarities between theories based on Individual Differences (e.g.: Holland's theory of vocational interests or Gottfredson's theory of Circumscription and Compromise) and educational theories (Lent and colleagues' Social Cognitive Career Theory), evidencing the possibilities and benefits of combining them.

Chapter 5 and Chapter 6 explore personality differences between sex/gender and sexual orientation groups, as well as, across occupations. Chapter 5 showed that personality scores varied across sex/gender and sexual orientation groups. In this sense, there seemed to be slightly more evidence favoring the Inversion than the Shift Hypothesis, but none was sufficient to explain the results. In Chapter 6 there was evidence that personality plays a role in occupational choices, especially when operationalizing jobs according to their occupational orientation (i.e. Prediger's people-things dimension) rather than their occupational fields (i.e. STEM vs. Non-STEM). Both Chapter 5 and Chapter 6 contribute to the field of Individual Differences in two ways. On the one hand, they expanded the frontiers of knowledge, as these research questions remain largely unexplored. Additionally, both studies serve as evidence of the pervasiveness of personality in daily life choices. On the other hand, both studies exemplified the usefulness of personality scores at lower levels of aggregation: not only did they allowed us to contextualize inconsistencies in previous research results, but they also allowed us to explore personality differences in more detail. In this sense, these two studies support a change in research focus from theorizing around personality traits at the domain level to prioritizing research on the predictive power of facets and nuances. This

new methodological approach has multiple advantages, mainly that it ensures higher quality (in terms of inter-rater and retest reliability) and higher predictive power of personality measures. Moreover, focusing on the basic building blocks helps setting the foundations for an easier to update personality model (for more details see Section 4.4.2).

7.3 Cross-cutting themes

Beyond the contributions of each of these chapters to the field, there are recurring themes throught the thesis. This section will focus on three key themes: (1) the use of primary and secondary data, (2) the relevance of theories on Individual Differences, and (3) the importance of using Individual Differences in applied settings.

7.3.1 Primary vs. Secondary data

Psychologist use two types of data: Primary and Secondary data. The former refers to any data collected and used by the same researcher. For instance, the data used in the personality studies included in this thesis – i.e. Chapters 5 and 6. The latter refers to data collected by an individual – or a public/private agency – unrelated to the researcher that might eventually use it (e.g.: the case of the MCS data used in the intelligence studies). I will now summarize the advantages and disadvantages of each data type concerning their use in this thesis.

The results obtained in studies using Secondary data (Chapters 2 and 3) were not as interesting as the other two. This is partly due to the nature of the research questions, but also because of the measures used. Unlike with Secondary data, Primary data tends to incorporate more focused and in-depth questions. Secondary data, especially in the case of longitudinal studies, has a broad focus so it is unlikely to find a dataset that satisfies your particular researchers interests. Additionally, these longitudinal studies tend to follow a "patchy" structure because measures are not systematically used across waves. This is precisely what happened in the MCS, where motivation and intelligence measures were

not consistently used across waves. Moreover, these measures were not as comprehensive as it would have been ideal. For instance, it would have been preferable to have a broad measure of intelligence rather than a single verbal ability test. These reasons evidence that Chapters 2 and 3 might have benefited from using Primary data. However the collection of this data would have been prohibitive in terms of time and resources. Note that some data-sharing initiatives like *the Open Science Framework* might provide a third solution to this issue. These sites could help gathering Primary data from researchers in a given area. This would be especially useful for popular research questions (e.g.: personality differences across groups) but less so for more specific cross-field research.

Additionally, research is costly in terms of time and/or money. Monetary costs are mainly associated to participant payment and survey distribution. In this sense, it might take more time to collect data on a voluntary basis compared to collecting data through platforms like Prolific, which selects participants according to the research criteria, distributes the survey, and organizes participant payment. Researchers collecting Primary data have to cover those expenses by themselves hence, they tend to have smaller sample sizes than what Secondary data can offer. For instance, the MCS used in Chapters 2 and 3 had data from over 3 000 children and their families. However, due to material constraints, I only managed to collect around 1000 participants for Chapters 5 and 6. In this sense, Secondary data offer a higher power to detect small effects and differences that would go otherwise unnoticed.

Moreover, Primary data is always up to date, as it is collected *ex-profeso* to address a specific research question. Sometimes it is difficult to find up-to-date or relevant secondary data because its costs reduce the number of people and occasions when these studies could be carried out. In this thesis, I was able to use the MCS which follows children and teenagers born in 2001 making it an extremely useful data set for the study of current trends in educational and occupational outcomes.

Secondary data tends to save researchers the time and effort required to prepare ethical approval applications and data protection protocols. Secondary data was preferred for Chapters

2 and 3 because using vulnerable populations involves dealing with multiple gatekeepers and requires high data protection standards. These reasons also explain why using Primary data in this type of research would yield small sample sizes. Additionally, the MCS was a longitudinal study, which allowed the study of change in the outcome variables. Longitudinal studies are very valuable to understand the interaction between different variables and to follow the developmental stages of certain psychological processes (e.g.: personality changes during childhood or in the workplace). Of course, change can also be evaluated cross-sectionaly by testing children of different ages and comparing their results, but this approach does not control for unmeasured variables which might be affecting change. For this reason, longitudinal studies preferred over primary data when trying to establish the direction of causality, as I aimed in the mediation model of Chapter 3.

Using a longitudinal study would have enriched the discussion of Chapter 6, as it would have allowed to test whether personality traits lead people into different occupations, or if it was previous experiences in the labor market what lead to changes in personality. Establishing the direction of causality is relevant in psychology, especially when considering genetic and environmental influences. We will further discuss this issue in the next section.

Section summary

Psychologists use Primary or Secondary data depending on their resources and the research question at hand. Each data type has strengths and weaknesses but both face challenges when we try to derive theories from them. This is mainly due to the myriad of factors involved in every aspect of human life and psyche.

7.3.2 Theories in Individual Differences

Theories are helpful to guide research but, as this thesis shows, some factors like cognitive ability or personality facets/nuances are valuable by themselves even if they are not integrated into a larger framework. As already mentioned in Section 4.5.4, theories are helpful because they help us describe and/or predict phenomena. In this sense, theories do not need

to be perfectly accurate to be useful as long as they have high predicting power. Researchers often focus on predictive power because it allows us to control our environment and take actions that minimize the costs and maximize the gains. For instance, the better we predict tornadoes, the lower the costs of human, wildlife, and material losses.

As I will now exemplify, the problem faced by Individual Differences theories and, more generally by psychology, is the plethora of intertwine environmental and genetic factors influencing each observable behavior:

> *"[T]here is no direct way of manipulating psychological variables such as thoughts or affects (Chiesa, 1992; Hughes et al., 2016). Instead, they have to be manipulated indirectly via verbal instruction or other external stimuli, and such techniques are typically not precise enough to change just one variable. For example, it is (at least currently) impossible to manipulate feelings of loss of control without changing any other psychological states, such as motivation, attention, or feelings of anxiety […] This creates a problem for finding psychological causes because when interventions are fat-handed, we cannot assume that they are unconfounded manipulations that license causal inferences. More specifically, we cannot assume that they change putative effect Y only via a route that goes through the putative cause X."*
>
> — Eronen and Bringmann (2021, p. 6)

For these reasons, it is not only difficult for psychologists to create accurate descriptions of each phenomenon, but also to find which factors predict them better. Thus, building psychology theories is arduous. As I will further explain, this has led some experts to abandon theorizing altogether to focus on those building blocks with high predictive power. This thesis is an example of this approach as it focuses on using elements with high predictive power to complement previous research or theories. On the one hand, this thesis shows the relevance of intelligence in educational theories. On the other hand, this thesis uses personality facets

and nuances to explain inconsistencies in previous research and to provide a more detailed account of what aspects of personality contribute to sex/gender differentiation.

This section further explains the benefits and drawbacks of using theories in Individual Differences – or psychology – research. Additionally, it also examines the alternative method of tackling research followed in this thesis.

Why are theories important in Individual Differences or, more broadly, in psychology?

John et al. (1988) and, more recently R. J. Stemberg in Deary and Sternberg (2021), argued that theories are necessary to guide research, for they help to identify and to detangle the environmental and genetic causes, providing a framework of reference so research findings are critically evaluated. According to these authors, theories also enable us to communicate and accumulate knowledge so that, eventually, we build a comprehensive theory of human psychology. In this sense, unifying models or theories can reduce redundancy while improving predictive power. However, this emphasis on theories has also been criticized (e.g.: Deary and Sternberg, 2021; Mõttus et al., 2017a). The main criticism revolves around how theorizing (in the sense of merging and creating a narrative connecting multiple constructs) has been prioritized over studying their basic components, especially given recent evidence showing that these components individually can be as useful as full complex models. This might sound counterintuitive, but the following paragraphs will clarify it.

How easy is it to describe psychological phenomena?

Describing a psychological phenomenon is not easy for it requires identifying the key variables influencing it, as well as the causal mechanisms connecting them. As an example, we can take the model we used in Chapter 3 where we predicted occupational aspirations change from verbal ability and change in grades.

Figure 7.1 serves as a broader representation of variables affecting a student's occupational

aspirations at a given time. As Figure 7.1 shows, higher SES has been linked to higher educational and occupational aspirations (Croll, 2008; Dubow et al., 2009; Gottfredson, 1981; Haller et al., 1974; Lee and Rojewski, 2012; Schoon and Parsons, 2002; Schoon and Polek, 2011), regardless of students' abilities (Cochran et al., 2011). However, this is not a straightforward relationship, as many other variables might moderate this effect – e.g.: SES has been found to interact with race (Howard et al., 2011) and school achievement (Lee and Rojewski, 2012). Additionally, students from higher SES families have been found to be more certain of their occupational aspirations and are more likely to justify them based on interest and personal development rather than money (Ball et al., 2002; Galliott and Graham, 2015; Gore et al., 2015). As a consequence, SES could be more relevant when predicting certain types of aspirations involving higher economical risk in the long term like art (Gore et al., 2017).

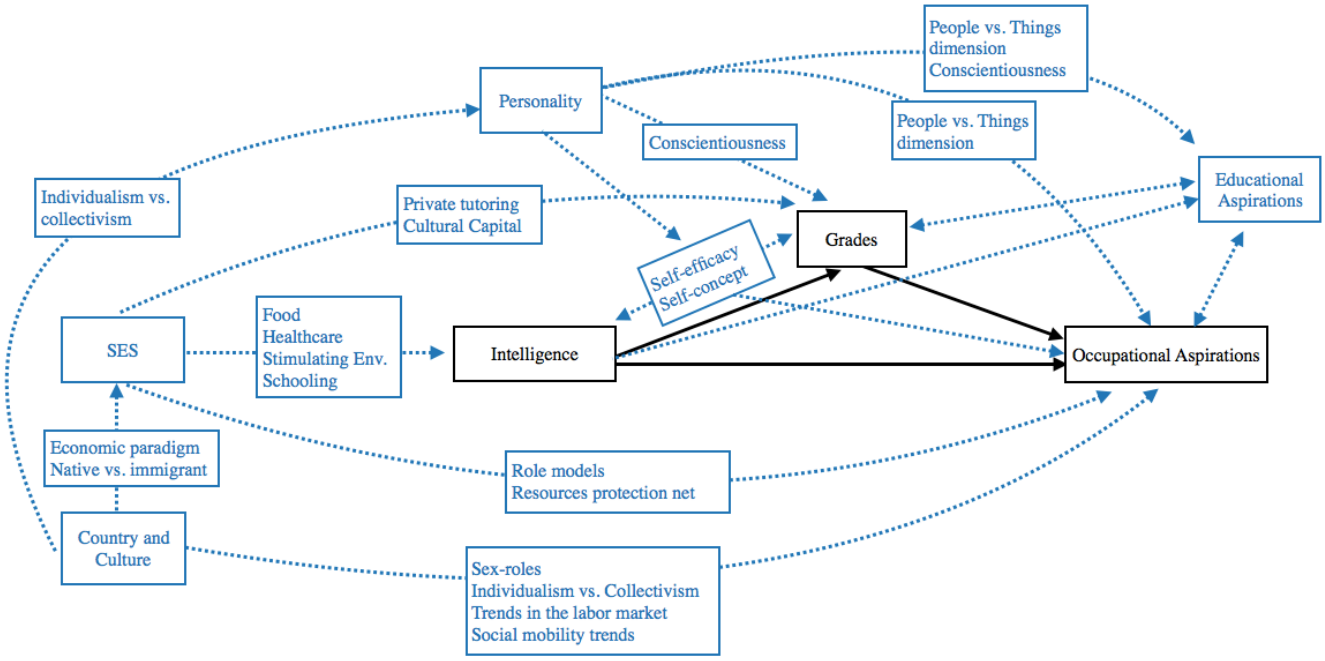

Figure 7.1
Potential causal mechanism influencing students' occupational aspirations.

Notes: This hypothetical model shows some of the factors affecting directly and indirectly (blue) each of the variables included in a simplified version of the model used in Chapter 3 (black).

Parents' expectations have also been related to children's occupational and educational aspirations (Mau and Bikos, 2000). Parents might also foster different characteristics in their children influencing them to be better suited for specific types of jobs (Schulenberg et al., 1984). Additionally, high SES has been related to high parental educational aspirations and expectations (Ditton et al., 2019; Schoon and Parsons, 2002). In fact, the relationship between SES and students' school achievement has been found to be mediated by cultural resources and parents' educational expectations (Ditton et al., 2019).

According to Bourdieu, high SES families would help students to navigate the labor market because they have the knowledge and the personal connections, in other words, "they have not just been given the map; they are the cartographers themselves" (Gale, 2015, p. 90). Lastly, high SES background is related to high intelligence (Cochran et al., 2011; Deary et al., 2005; Dubow et al., 2009; Schoon and Polek, 2011). Research suggests that SES confounds the heritability or expression of intelligence so, SES can widen school achievement differences between students of similar abilities (Turkheimer and Harden, 2013; von Stumm, 2017). Higher-SES has also been directly related to higher school achievement (Bergman et al., 2014; Gore et al., 2017; Schoon, 2001; Schoon and Parsons, 2002).

Research shows that student have sex-stereotypical occupational aspirations as females tend to choose Humanities and boys, STEM (Croll, 2008; Gore et al., 2015, 2017; Holmes et al., 2018; Howard et al., 2011). This seems to be the case regardless of students' actual abilities. Some research shows that verbally talented men, would rather choose to work in STEM-related than in Humanities-related occupations and, although to less extreme, mathematically talented women would choose the opposite (Baird, 2012; Coyle et al., 2015). Potentially, it is possible that sex/gender stereotypes affect occupational aspirations through self-competence beliefs (Shapka et al., 2006). Additionally, the country and the culture in which students are embedded can also affect their occupational aspirations. For instance, traditional beliefs held by individuals might influence their adherence to sex-stereotyped occupational aspirations (Baird, 2012). Indeed, attending to a single- or mixed-sex/gender

school has been found to moderate sex-differences in occupational aspirations (Schoon, 2001). Moreover, the impact of the culture in which students are raised might depend on whether they are natives from that culture or immigrants (Marjoribanks, 2002).

There can also be an interplay between culture and actual sex/gender differences. As we have previously discussed in Chapter 6, personality sex/gender differences are larger in developed countries which could influence the sex/gender segregation of the labor market. On the other hand, individuals might have less economic pressure in developed countries, which might prompt individuals to transgress their sex/gender roles in terms of their occupational aspirations. Additionally, personality might also influence occupational aspirations through educational achievement and educational aspirations, especially given the degree of diligence required for studying (Dumfart and Neubauer, 2016).

The historical context in which individuals are embedded also mediates the effects that background SES, occupational aspirations, and school achievement have on future occupational status (Schoon and Parsons, 2002). Individuals' perceptions of social mobility patterns, for instance, can make some generations more ambitious, although it might not be enough to alter the general tendency of students to follow their parents' careers (Croll, 2008) or what they think their parents expect (Helwig, 2001). This discussion shows how difficult it is to flag every factor influencing students' occupational aspirations at a given time, let alone to determine which is the direction of causality: does SES increase intelligence or does intelligence lead to high SES? Does personality influence self-competence through grades or does self-competence shape personality through grades?

Note that this is just one example to showcase the complexities involved in a single study. But, the same logic could be applied to psychology studies across the board. For instance, thinking about other studies in this thesis we could ask: does personality lead people into different jobs, or do specific experiences in each occupation shape people's personalities? Could the personality sex/gender differences caused by biological factors become more salient due to specific parenting choices or cultural values?

These questions could be easier to address if we directly controlled environmental or genetic factors. Current advances in genetic research (e.g.: CRISPR technology), could allow us to control certain aspects of an individuals' genome, but such studies would raise ethical concerns. Similarly, not all environmental factors could be controlled ethically. In fact, past attempts to control genetic or environmental factors lead to terrible consequences. For instance, the psychiatrist Peter Bela Neubauer lead a study in collaboration with the adoption agency Louise Wise Services, in which twins and triplets were separated at birth and placed in families of different SES without the consent of either the biological or the adoptive parents. These children were tested multiple times across their life span, without most of them ever knowing the actual goal of those follow-ups. This study was documented in the 2018 film *Three Identical Strangers*, through the dramatic story of a set of triplets reunited as adults who, as a consequence, experienced mental health issues which eventually led one of them to suicide.

These practices are unthinkable nowadays. Methods have been developed to study the relationship between genetic and environmental factors indirectly. Longitudinal studies, for instance, help to control environmental and genetic factors that might affect a phenomenon across time (e.g.: the influence of therapy on personality changes, Roberts et al., 2017). Another example is the technique of Genome-wide Association Study (GWAS) which allows researchers to parse the effects of environment and genes through genetic data of thousands of participants (for an extensive review on this technique and its implications on Intelligence research see Plomin, 2019).

Can inaccurate theories still be useful?

As we have just discussed, there is a myriad of factors involved in psychological phenomena, hampering our ability to build accurate theories, especially in research fields like Individual Differences. Unavoidably, the next question is: can an inaccurate theory still be useful? This is a complex philosophical question, but it could be synthesized into two opposing views. On the one hand, theories have an intrinsic value for conveying information about the world,

even if it does not perfectly match reality. Thus, theories *do not need* to be *useful* in the sense of having high predictive power. On the other hand, most people *want* theories that are useful to predict phenomena, regardless of the degree to which they mirror reality[1]. In other words, under this view, *wrong*[2] theories can still be useful.

This last point is important to our debate on the value of theories in the field of Individual Differences, so I hope the reader allows detouring slightly to fully explain why *wrong* theories can still be useful and how this might impact Individual Differences research.

For a moment, let's imagine we were standing in the middle of a forest on a pitch-black night. Now we turn on a flashlight and we can see what lies in front of us. We could think of ourselves as researchers, the forest as the world of knowledge, and the torch as our research methods. The light allows us to see what is ahead of us, but not everything around us. Scientists would interpret this situation differently depending on their ontology[3]. To some scientists the torch shows the world as it is, knowing the surroundings would just be a matter of moving the torch. Then, theories would help us weave all evidence together. Theories could further guide research by making predictions about specific phenomena. If these predictions are confirmed, these scientists would agree that their theories are closer to the *truth*. Through this continuous cycle of mutual refinement between methods and theories, researchers hope to eventually create a single theory able to explain a specific phenomenon

[1]It is important to highlight that this "predictive power fever" is not intrinsic to scientific research. However, in many fields, it allows researchers to attract research funding. Hence, given the economic system where mainstream research develops, having tools with high predictive power usually translates into higher research funds which, conversely, leads to more research opportunities.

[2]Note that I have used the words "wrong" and "truth" deliberately through this section to simplify the discussion. Words like "right", "truth", and "wrong" are loaded with ontological and metaphysical assumptions. Such discussion is out of the scope of this thesis thus, I have used them following their popular acceptation: "wrong" is used in opposition to "right" or to "the truth", which I would define as "an accurate reflection of something".

[3]Ontology refers to the specific definition of knowledge to which people, theories, research, etc. ascribe to. Each philosophical current follows a different ontology, but they could be roughly divided into two. On the one hand, we have ontologies derived from Descartes's Dualism – e.g.: empiricism, positivism, etc. – for which knowledge is independent of the observer and, as such, "the truth of the world" is cognoscible. Quantitative research is largely based on these philosophical currents. On the other hand, we find the ontologies derived from Idealism – e.g.: interpretivism, constructivism, etc. – for which knowledge is context and subject-dependent hence, generalizations are neither viable nor useful. These are the ontologies on which most qualitative research is founded (for more details see Smith, 1983).

(or, even, the entire universe).

For many other scientists, it is a leap of faith to think that good predictions immediately translate into accurate descriptions of the world. After all, we are blind to the *actual reality of the world*, we are in a pitch-black forest with a single torch lighting our way. Hence, we could now ask, is it possible to make accurate predictions from *wrong* theories? If you ask this question putting any psychological theory as an example, people will answer – probably without doubting – "yes". People may start hesitating when applying this question to physics. Can we make accurate predictions from *wrong* physics theories? "Probably not, I mean, what are the chances?". There is a reason why people will answer differently depending on what research example they are given but we will come to this later. Coming back to our question: can we make accurate predictions from *wrong* physics theories? Newton described gravity as a force between two objects, this force would depend on their masses and the distance between them. If the reader remembers their school years, you probably recall thinking about gravity as "something that pulls you toward the Earth's center". Some two hundred years after Newton, Einstein put forward a completely different description of gravity, proposing something like: "planets sit on an invisible mesh which stretches as planets travel through it. Thus, pulling whatever objects around them (e.g.: the moon), towards them, creating what we know as gravity". Newton's and Einstein's theories are quite different, yet they are both used because they are great at predicting how objects move. Another example of how a *wrong* physics theory can make accurate predictions is the fact that quantum mechanics theories (the branch of physics aiming to explain the behavior of atoms and subatomic particles) do not seem to match the theory of general relativity (which is more related to astrophysics and larger body masses). These theories are the foundations of current physics research even though some of their proposals do not converge. Now, it is possible that we are missing a link explaining the relationship between both or, it could be that both theories are *equally off* from the *truth*. It is possible that our methods have biased us into those theories. It is also plausible that the things we have been interested in predicting are not the things that would help us find out *the truth*. These questions and hypotheses are part of a broader

philosophical debate that goes beyond this thesis' scope, what it is important for us is to acknowledge that (1) theories create a narrative connecting scientific evidence, (2) theories might be *wrong*, and (3) *wrong* theories can still be useful if they have a strong predictive power. As we will later see, this third point is especially important in the theory-debate within psychology and, especially, in Individual Differences.

Before, I mentioned that people tend to assume that psychological theories are wrong more often than physics theories. This is partly due to how psychological research designs experiments and builds theories. Theories' predictive power partly depends on the degree to which confounders are identified and causal mechanisms are specified. As we saw in the previous example on occupational aspirations, identifying confounders and causal mechanisms is extremely difficult in most psychological research. Hence, it is more likely that psychological theories have a larger marging of error compared to physics theories. So, if (1) it is – at best – extremely difficult to gauge how accurate a theory is, and (2) inaccurate theories can still be useful if they have strong predictive power, but if (3) psychological theories with strong predictive power are hard to devise due to the complexities of the human psyche; one might ask: why do we even care about theories?

Experts in the field of Individual Differences have also recently voiced concerns about theorizing *too much* (e.g.: Deary and Sternberg, 2021; Mõttus et al., 2017a, 2020). The critiques of these experts around theories are not so much about the theories themselves but about them detracting the focus from the careful examination of their building blocks or, what is more, disregarding the study of these basic components arguing that they are *less useful*. For instance, there are many theories about intelligence, but there is still research needed to understand how each component interacts or how they manifest on the brain. There is also extensive research on the Big Five but, until recently, few efforts have been devoted to producing high quality measures of their lowest levels of aggregation – i.e. nuances and facets. Chapter 5 in this thesis exemplifies how there is intrinsic value in studying personality at lower levels of aggregation to fully understand results patterns at the domain level.

There are other issues related to theories beyond this main criticism. On the one hand, Deary and Sternberg (2021) argued that theories might constrain researchers' creativity, holding back the development of new theories or the update of old ones. For instance, the Big Five model assumes that all domains are independent but this does not allow to study how they might co-develop as network models suggests. Moreover, the over-emphasis on the Big Five has disregarded any research exploring other personality traits. The studies in this thesis used nuances and facets related to the Big Five, but it cannot be discarded that other personality traits could create wider personality differences between sex/gender and sexual orientation groups or across occupations. Personality studies like those included in this thesis show that using facets and nuances not only provides us with a deeper understanding of personality, it also allows us to examine new personality traits before incorporating them into a broader theoretical framework. On the other hand, Chapters 2 and 3 show how educational theories are effective at describing some factors influencing educational outcomes but, since evidence consistently suggests that cognitive ability is a strong predictor, it should be more often considered. In this sense, cognitive ability measures could be used as independent predictors, although existing theories could also be updated to integrate cognitive ability or, even, new theories could be developed to fully understand how intelligence and other contextual components interact in the educational sphere.

Besides these ulterior consequences, theories might also have practical consequences. I. Deary pointed out that theories might promote confirmation bias[4] which could also lead to publication bias[5]. Another issue arises when theories are proposed so fast that there is no time to fully test them. This tendency of developing new concepts to re-brand old constructs hampers – rather than allows – our ability to accumulate the knowledge needed to build sound theories. The previous section partly explains this re-branding tendency in Individual Differences: environmental and genetic factors intertwine in psychological traits

[4]Confirmation bias refers to indiviuals' tendency to focus on the evidence that confirms the values or theories they already hold, disregarding any other evidence, especially that contradicting it.

[5]Publication bias is a phenomenon by research showing specific results are more likely to be published. Publication bias is a pervasive practice in psychology, leading to multiple unreplicable findings and flawed theories.

to such an extent that any theory including one more element can still add information – regardless of how little it is – to previous ones. This circles back to one of the quotes from Chapter 4): "How does one conclude that a description is 'reasonably sufficient' or 'comprehensive' or 'fully adequate' ?" (Block, 1995, p. 187).

Focusing on the building blocks

The solution that I. Deary, R. Mõttus, and other experts have proposed is to invest more resources in the basic aspects of theories. In some sense, they propose to change the focus from designing theories to guide experiments, to gathering large amounts of evidence before building theories (or not building theories altogether). In this way, theories about specific psychological phenomenon would emerge from collecting enough evidence about how a it relates to behaviors or life outcomes.

This thesis aimed to provide some examples by (a) taking elements from different theories and testing if they could work together (e.g.: using verbal abilities and change in grades to predict change in occupational aspirations) and (b) putting the focus on model's basic elements (e.g.: using personality facets and nuances to study personality differences across occupations).

Section summary

In certain fields like psychology, identifying and controlling all factors involved in a specific phenomenon is remarkably complex. For this reason and, although theories on Individuals Differences (and psychology as a whole) have an intrinsic value, it is unlikely that their predictive power will ever compete with that of physics theories. This, among other reasons, has led experts to question the current focus of Individual Differences research on theorizing. Hence, some experts are advocating for research centered on more basic psychological constructs so we can gather enough evidence to improve existing theories and/or develop more solid ones.

7.3.3 The measure of Individual Differences across applied settings

The aim of this thesis has never been to design intervention programs using Individual Differences measures for the educational system or the labor market. However, as previously mentioned, some consider that research is useful to the degree that it can be applied to real-life settings. For this reason, this section explores whether research on Individual Differences – especially intelligence research – should inform educational policy (or other applied fields, for that matter). It is, however, a complex discussion with plenty of nuances and layers which would be impossible to fully cover in this section. Hence, I will summarize three key conundrums that this liaison could entail and the main arguments that are usually proposed.

Conundrum 1: Who would benefit?

One of the biggest concerns about integrating intelligence research into educational policies lies in the degree to which students across the board would benefit equally.

Some people argue that not all students would perceive the same benefits. According to them, intelligence testing would likely lead to a higher stratification of the student body: the more able[6] students would be praised and attended while the less able students would be stigmatized and ignored. However, this statement is based on classist[7] values imposing a hierarchical view of educational attainment. In this sense, normalizing intelligence testing could challenge these classist values. It might even replace them with an educational system that not only values diversity but also mirrors and accommodates the full spectrum of ability.

An educational system that considers ability might highlight differences between students, but it is hard to imagine that ignoring these differences would make them disappear. Ad-

[6]The word "more able" in this Chapter refers to students that would score notably higher than the average across intelligence tests. On the contrary, "less able" refers to students that would score lower than the average across intelligence tests.

[7]Classism refers to "a belief that a person's social or economic station in society determines their value in that society" (Classism, 2021).

ditionally, for something to be funded it has to be labeled: students in either extreme may not be adequately attended unless it is evident that their needs differ. Note that supporting students according to their needs might not involve dramatic changes in the curriculum, it might be as simple as investing more time in the students that, given their abilities, are more likely to lose their motivation. Chapter 2 in this thesis aimed to tackle precisely the importance of identifying whether more verbally able students are at a higher or lower risk of disengaging from specific school subjects or, even, from the school system as a whole during the primary to secondary school transition.

Nonetheless, funding could be a double-edged sword. Given the positive relationship between intelligence and family SES (plus the positive correlation between SES and intelligence heritability), more able students could count with the genes for high intelligence, in addition to the economic resources to pay culturally enriching activities. These individual resources, coupled with the extra resources provided by the education system, could enlarge the social stratification when factoring in SES. The gaps between each social stratum could become even larger in countries where there is a weak system of social welfare, unable to guarantee free access to healthcare and/or education. Moreover, regardless of how big these gaps were, brighter students already have more chances of succeeding, this hypothetical system could be sacrificing the future of less able students so that more able students could do better. Some people deny that this might be the case. They argue that a system incorporating intelligence testing would remove the focus from the average student – who are currently the ones benefiting the most – to those on either extreme; hence, the more able and the less able students would benefit equally.

Conundrum 2: Social cost

It is possible that incorporating intelligence and personality testing into the educational sphere would lead to a more diversified system, this is, more optional activities or educational tracks (e.g.: a university track, vocal education track, science track, etc.). This could be undesirable in societies where educational tracks might be valued differently. This ex-

pectation is based on the prevalence of classist values by which the hierarchical division of society often depends on educational attainment and wealth.

Contrary to what it could be expected, research evidence shows that sex/gender segregation in STEM and non-STEM fields is higher in more gender-equal countries (see Chapter 6). This suggests that there will not be a homogeneous distribution of students in terms of sex/gender (and, possibly, race) across educational tracks. In turn, this could unnecessarily perpetuate stereotypes on the abilities and interests of specific student groups, potentially setting the scene for a less harmonious school atmosphere. However, others argue that there is no reason to assume that the disconnect between societal values on gender equality and actual social dynamics on the workforce would disappear in a different educational system. And even if it did, it might be replaced by societal values recognizing diversity as an asset.

On the other hand, it could be argued that if more gender equal countries have a highly segregated labor force, people are already – *willingly* – gravitating towards occupations that fulfill their abilities and skills. Thus, the cost of a more diversified educational system might not be compensated if the segregation of the labor force is unlikely to increase beyond the current standards. Nonetheless, some people argue that a more diversified educational system could be more efficient in grouping and preparing students for the labor market. All in all, this debate could be boiled down to: unrealistic but seemingly just values (e.g.: everybody has the same abilities and skills, thus they can access any job) vs. realistic but seemingly unjust values (e.g.: people have abilities and skills that make them suitable only to specific positions). The social cost of each of these hypothetical scenarios is difficult to estimate.

Lastly, some people argue that if educational tracks reinforced stereotypes and this permeated the real world, it could destroy previous efforts aimed to deconstruct those stereotypes and to ensure equality of opportunity. Consequently, if these stereotypes were reinstated, individuals wanting to defy them would experience more pressure and stigma. Conversely, some people argue that society is already discriminating certain students, specifically, those

who do not pursue university studies. The stigma around vocational education could be counteracted in a system that acknowledges that ability (1) is a spectrum and (2) might influence career options but not personal worth. Additionally, this system could provide all students with better-tailored career counseling so they make more informed career choices. Altogether, it is hard to gauge which would lead to higher social costs: limiting individuals' freedom to choose a career path or fostering unrealistic expectations on individuals' career paths.

Some of the research questions connected to these issues were explored in Chapter 3 and Chapter 6. Chapter 3 in this thesis exemplifies how could this conundrum be explored empirically by reflecting on the connection between cognitive ability, feedback experiences (e.g.: grades), and occupational aspirations. Such research show how relevant it could be to provide students with a wide range of learning experiences where they can discover and exploit their strengths. These experiences could lead (organically) to a diversification of the school system which could be, potentially, better suited for the labor market, reducing student frustration and increasing employability. Moreover, Chapter 3 also explored how relevant it could be to consider personality traits when assessing students on their occupational orientation.

Conundrum 3: Economic cost

Any change in the educational system involves investing resources. Integrating intelligence or personality testing in the school system could imply anything from small changes in the curriculum to a reconceptualization of the whole educational system. The economic cost of these changes should be put in balance with the benefit they might yield.

According to some, diverting school funds benefiting the majority of the students to smaller groups would be economically less efficient. However, others argue that the cost could be compensated in the long run. Even if changes in the educational system involved the creation of several tracks used by few students, these students would be provided with the

education needed for them to achieve their maximum potential: (1) more able students could be potentiated to become future political leaders, researchers, writers, artists, etc. and (2) less able students could be given more opportunities to find suitable occupations that grant their access to basic services. In this sense, a more diversified system could be especially beneficial in countries with a weak welfare system, as it would reduce the overall social cost. Nonetheless, it could still be the case that not all groups would need those extra resources to succeed in the labor market. In particular, high-SES families could make the economic effort themselves without the system providing for them.

No studies in this thesis would allow to estimate the economic cost of any change in the educational system regarding the use of personality or intelligence measures. This research would involve an extraordinary effort to find comparable groups and to track them across their childhood and adulthood before being able to estimate what educational policies could be more cost-effective. This is probably one of the reasons why this type of research has not been carried out at all. However, the most relevant reason is that the use of Individual Differences measures in the educational system it is still questioned due to its record of controversial research outputs.

7.3.3.1 The burden of history

As the previous discussion suggests, there are arguments for and against the use of Individual Differences research in applied settings. However, arguments in favor are rarely publicly discussed. In this sense, it is important to understand some historical events to fully comprehend why it is so controversial to use Personality and Intelligence measures in the educational system. The reader might want to refer to Fass (1980) for a further discussion on the social impact of intelligence testing in the US in the 20th century.

Firstly, we should start by acknowledging that no research occurs in a vacuum. Research outcomes depend on who the researchers are, who funds the research, what purpose that research has, what are the researcher's values, what are the funding bodies' values, etc.

Research does not only depend on its immediate surrounding, it also depends on the cultural values of the society as a whole. An example is the drastic increase in funding in the US for all science and technology-related activities, from school programs to the creation of NASA, during the space race in the 50s.

Secondly, it is crucial to understand how were the first large scale intelligence measures developed and used. As mentioned in Chapter 1, the Binet-Simon Scale to assess children's cognitive ability was developed in France after the French revolution. By the beginning of the 20th century, the Binet-Simon Scale was brought to the US to help accommodate the increasing number of immigrant children into the educational system. Intelligence testing quickly became an important tool in the educational system (Fass, 1980). L. Terman and other US researchers used the Binet-Simon Scale to develop questionnaires suited for large-scale testing (Boake, 2002; Fass, 1980; Terman, 1916). These large data sets prompted researchers to study intelligence differences across sex/gender and races (Fass, 1980). It is important to remember that at that time, there was a high degree of racial segregation in the US and other countries around the world. Therefore, these research results unavoidably lead to numerous controversies. On the one hand, researchers at the time lacked the sensibility to present their results and, on the other hand, many people held the legitimate fear that these results would serve to justify - and, even further increase - the segregation of the population. The debate over the *rightness* of such research settled, mainly, due to a strong discrediting campaign that undermined Intelligence research and limited its use in the educational context. Even UNESCO published a report in 1950 discouraging the use of Individual Differences measures (Intelligence measures in particular) in schools. These events set precedence in dividing Intelligence and Educational research and even nowadays, the cross-over between both fields is scarce.

The debate over these research controversies has cooled down but it has never extinguished. Intelligence research still carries the burden of the social trauma that certain individuals feed. Until these social tensions are dissolved either through an overt confrontation or an

indirect by-pass, it is unlikely that Intelligence research is used systematically in certain applied fields like education. However, intelligence research has proven to be essential to understanding educational achievement hence, efforts to foster this partnership should be made. This thesis aims to be a humble contribution to the reunification of these fields.

Section summary

In this thesis I have showed the connection between Individual Differences and educational or occupational outcomes. The evidence in this thesis shows that intelligence and personality might influence students' school motivation and occupational aspirations, as well as their occupational careers. However, the evidence on this thesis is not enough to ensure that an educational system integrating individual differences testing would be better.

Nonetheless, previous research suggests that Intelligence and Personality are important for multiple life outcomes because they have strong predictive power – regardless of the overall theories in which they are framed. However, given the history of Individual Differences research, several other aspects must be put in balance before introducing measures of personality and cognitive ability in the educational system (e.g.: whether the student population would benefit uniformly, as well as the social and economic costs of such policies). In this sense, it is possible that more research is needed to address (a) how – rather than why – individual differences testing could be incorporated into the educational system; and (b) what could be the consequences of each approach.

7.4 Concluding remarks

This thesis combines the use of Primary and Secondary data to explore how Individual Differences – i.e. Intelligence and Personality – influence educational and occupational outcomes. This thesis supports the increasingly popular idea that theories in Individual Differences (and in psychology more broadly) might not yield as many benefits as focusing on the basic elements having high predictive power. In this sense, Personality research would benefit

from setting the focus on its building blocks so theories can be grounded on solid evidence. Similarly, Educational research would profit from using intelligence measures because, even if they are not part of classical theories, they are strong predictors of educational outcomes. Eventually, these lines of research might help smooth the social tensions around the use of Individual Differences – and especially intelligence – measures in applied settings like the educational system or the labor market.

A wise emperor, therefore, knows how to choose the right person for the right task. He is like a skillful carpenter who knows to use straight timber to make shafts, curved timber to make wheels, long timber to make beams, and short timber to make posts. Wood of all shapes and lengths is thus fully utilized. The emperor should make use of personnel in the same way, using the wise for their resourcefulness, the ignorant for their strength, the brave for their daring, and the timid for their prudence. As a good carpenter does not discard any timber, so a wise emperor does not discard any gentleman. A mistake should not lead the emperor to ignore a gentleman's virtues, nor should a flaw overshadow his merits.

– Tang Taizon Emperor, 648 a.C.

Retrived from Ebrey (2009, p. 113)

Appendix

Table 7.1
Distance, d_u, across factes and group comparisons

Domain	Facet	Cohen's du	Distance size
Heterosexual man vs.	Heterosexual woman		
Agreeableness	Modesty	-0.52	High
Agreeableness	Tender-Mindedness	-0.39	Moderate
Agreeableness	Trust	0.26	Moderate
Agreeableness	Straightforwardness	-0.60	High
Agreeableness	Cooperation	-0.04	Small
Agreeableness	Altruisim	-0.70	High
Conscientiousness	Competence	0.32	Moderate
Conscientiousness	Achievement Striving	0.04	Small
Conscientiousness	Order	-0.19	Small
Conscientiousness	Self-discipline	-0.28	Moderate
Conscientiousness	Dutifulness	-0.60	High
Conscientiousness	Deliberation	0.14	Small
Extraversion	Positive emotions	-0.38	Moderate
Extraversion	Gregariousness	0.08	Small
Extraversion	Warmth	-0.13	Small
Extraversion	Assertiveness	0.25	Moderate

Table 7.1
Distance, d_u, across factes and group comparisons (continued)

Domain	Facet	Cohen's du	Distance size
Extraversion	Activity	0.06	Small
Extraversion	Excitement-Seeking	0.32	Moderate
Neuroticism	Anxiety	-0.50	Moderate
Neuroticism	Anger	-0.12	Small
Neuroticism	Depression	-0.32	Moderate
Neuroticism	Self-Consciousness	-0.68	High
Neuroticism	Impulsiveness	-0.08	Small
Neuroticism	Vulnerability	-0.49	Moderate
Openness	Action	0.40	Moderate
Openness	Values	0.13	Small
Openness	Ideas	0.42	Moderate
Openness	Aesthetics	0.06	Small
Openness	Fantasy	-0.10	Small
Openness	Feelings	-0.67	High
Heterosexual man vs.	Homosexual man		
Agreeableness	Modesty	-0.30	Moderate
Agreeableness	Tender-Mindedness	-0.39	Moderate
Agreeableness	Trust	0.08	Small
Agreeableness	Straightforwardness	-0.31	Moderate
Agreeableness	Cooperation	0.06	Small
Agreeableness	Altruisim	-0.42	Moderate
Conscientiousness	Competence	0.23	Moderate
Conscientiousness	Achievement Striving	0.24	Moderate
Conscientiousness	Order	-0.12	Small
Conscientiousness	Self-discipline	-0.03	Small

Table 7.1
Distance, d_u, across factes and group comparisons (continued)

Domain	Facet	Cohen's du	Distance size
Conscientiousness	Dutifulness	-0.16	Small
Conscientiousness	Deliberation	0.20	Moderate
Extraversion	Positive emotions	0.03	Small
Extraversion	Gregariousness	0.13	Small
Extraversion	Warmth	0.04	Small
Extraversion	Assertiveness	0.24	Moderate
Extraversion	Activity	0.18	Small
Extraversion	Excitement-Seeking	0.24	Moderate
Neuroticism	Anxiety	-0.41	Moderate
Neuroticism	Anger	-0.13	Small
Neuroticism	Depression	-0.32	Moderate
Neuroticism	Self-Consciousness	-0.29	Moderate
Neuroticism	Impulsiveness	-0.22	Moderate
Neuroticism	Vulnerability	-0.32	Moderate
Openness	Action	0.20	Moderate
Openness	Values	-0.04	Small
Openness	Ideas	0.10	Small
Openness	Aesthetics	-0.16	Small
Openness	Fantasy	-0.27	Moderate
Openness	Feelings	-0.33	Moderate
Heterosexual man vs.	Homosexual woman		
Agreeableness	Modesty	-0.43	Moderate
Agreeableness	Tender-Mindedness	-0.48	Moderate
Agreeableness	Trust	0.25	Moderate
Agreeableness	Straightforwardness	-0.40	Moderate

Table 7.1
Distance, d_u, across factes and group comparisons (continued)

Domain	Facet	Cohen's du	Distance size
Agreeableness	Cooperation	0.13	Small
Agreeableness	Altruisim	-0.23	Moderate
Conscientiousness	Competence	0.16	Small
Conscientiousness	Achievement Striving	0.25	Moderate
Conscientiousness	Order	-0.19	Small
Conscientiousness	Self-discipline	0.10	Small
Conscientiousness	Dutifulness	-0.19	Small
Conscientiousness	Deliberation	0.23	Moderate
Extraversion	Positive emotions	-0.02	Small
Extraversion	Gregariousness	0.56	High
Extraversion	Warmth	0.11	Small
Extraversion	Assertiveness	0.20	Moderate
Extraversion	Activity	0.37	Moderate
Extraversion	Excitement-Seeking	0.42	Moderate
Neuroticism	Anxiety	-0.61	High
Neuroticism	Anger	-0.19	Small
Neuroticism	Depression	-0.51	High
Neuroticism	Self-Consciousness	-0.66	High
Neuroticism	Impulsiveness	-0.35	Moderate
Neuroticism	Vulnerability	-0.47	Moderate
Openness	Action	0.45	Moderate
Openness	Values	-0.35	Moderate
Openness	Ideas	0.03	Small
Openness	Aesthetics	-0.20	Moderate
Openness	Fantasy	-0.16	Small

Table 7.1
Distance, d_u, across factes and group comparisons (continued)

Domain	Facet	Cohen's du	Distance size
Openness	Feelings	-0.64	High
Heterosexual woman vs. Homosexual woman			
Agreeableness	Modesty	0.07	Small
Agreeableness	Tender-Mindedness	-0.08	Small
Agreeableness	Trust	-0.01	Small
Agreeableness	Straightforwardness	0.19	Small
Agreeableness	Cooperation	0.16	Small
Agreeableness	Altruisim	0.39	Moderate
Conscientiousness	Competence	-0.14	Small
Conscientiousness	Achievement Striving	0.22	Moderate
Conscientiousness	Order	0.00	Small
Conscientiousness	Self-discipline	0.37	Moderate
Conscientiousness	Dutifulness	0.37	Moderate
Conscientiousness	Deliberation	0.09	Small
Extraversion	Positive emotions	0.34	Moderate
Extraversion	Gregariousness	0.47	Moderate
Extraversion	Warmth	0.24	Moderate
Extraversion	Assertiveness	-0.02	Small
Extraversion	Activity	0.34	Moderate
Extraversion	Excitement-Seeking	0.18	Small
Neuroticism	Anxiety	-0.13	Small
Neuroticism	Anger	-0.09	Small
Neuroticism	Depression	-0.21	Moderate
Neuroticism	Self-Consciousness	-0.05	Small
Neuroticism	Impulsiveness	-0.28	Moderate

Table 7.1
Distance, d_u, across factes and group comparisons (continued)

Domain	Facet	Cohen's du	Distance size
Neuroticism	Vulnerability	-0.04	Small
Openness	Action	0.10	Small
Openness	Values	-0.52	High
Openness	Ideas	-0.38	Moderate
Openness	Aesthetics	-0.27	Moderate
Openness	Fantasy	-0.06	Small
Openness	Feelings	0.00	Small
Heterosexual woman vs. Homosexual man			
Agreeableness	Modesty	0.20	Small
Agreeableness	Tender-Mindedness	0.00	Small
Agreeableness	Trust	-0.16	Small
Agreeableness	Straightforwardness	0.30	Moderate
Agreeableness	Cooperation	0.10	Small
Agreeableness	Altruisim	0.27	Moderate
Conscientiousness	Competence	-0.05	Small
Conscientiousness	Achievement Striving	0.22	Moderate
Conscientiousness	Order	0.06	Small
Conscientiousness	Self-discipline	0.26	Moderate
Conscientiousness	Dutifulness	0.41	Moderate
Conscientiousness	Deliberation	0.04	Small
Extraversion	Positive emotions	0.39	Moderate
Extraversion	Gregariousness	0.06	Small
Extraversion	Warmth	0.17	Small
Extraversion	Assertiveness	0.01	Small
Extraversion	Activity	0.14	Small

Table 7.1
Distance, d_u, across factes and group comparisons (continued)

Domain	Facet	Cohen's du	Distance size
Extraversion	Excitement-Seeking	-0.06	Small
Neuroticism	Anxiety	0.06	Small
Neuroticism	Anger	-0.02	Small
Neuroticism	Depression	-0.01	Small
Neuroticism	Self-Consciousness	0.40	Moderate
Neuroticism	Impulsiveness	-0.15	Small
Neuroticism	Vulnerability	0.11	Small
Openness	Action	-0.18	Small
Openness	Values	-0.18	Small
Openness	Ideas	-0.30	Moderate
Openness	Aesthetics	-0.22	Moderate
Openness	Fantasy	-0.16	Small
Openness	Feelings	0.32	Moderate
Homosexual woman vs. Homosexual man			
Agreeableness	Modesty	0.12	Small
Agreeableness	Tender-Mindedness	0.08	Small
Agreeableness	Trust	-0.15	Small
Agreeableness	Straightforwardness	0.11	Small
Agreeableness	Cooperation	-0.07	Small
Agreeableness	Altruisim	-0.13	Small
Conscientiousness	Competence	0.08	Small
Conscientiousness	Achievement Striving	0.01	Small
Conscientiousness	Order	0.06	Small
Conscientiousness	Self-discipline	-0.13	Small
Conscientiousness	Dutifulness	0.03	Small

Table 7.1
Distance, d_u, across factes and group comparisons (continued)

Domain	Facet	Cohen's du	Distance size
Conscientiousness	Deliberation	-0.06	Small
Extraversion	Positive emotions	0.04	Small
Extraversion	Gregariousness	-0.38	Moderate
Extraversion	Warmth	-0.07	Small
Extraversion	Assertiveness	0.03	Small
Extraversion	Activity	-0.17	Small
Extraversion	Excitement-Seeking	-0.23	Moderate
Neuroticism	Anxiety	0.18	Small
Neuroticism	Anger	0.07	Small
Neuroticism	Depression	0.19	Small
Neuroticism	Self-Consciousness	0.42	Moderate
Neuroticism	Impulsiveness	0.12	Small
Neuroticism	Vulnerability	0.13	Small
Openness	Action	-0.26	Moderate
Openness	Values	0.30	Moderate
Openness	Ideas	0.07	Small
Openness	Aesthetics	0.06	Small
Openness	Fantasy	-0.09	Small
Openness	Feelings	0.31	Moderate

Note:

Lables indicate the reference group. Positive values means that the first group was higer, conversely, negative values reflect that it was lower

Table 7.2
Raw facets' average per sex and sexual orientation group

Group	Facet	Raw facet average
Heterosexual men	Modesty	11.98
Homosexual women	Modesty	13.25
Heterosexual women	Modesty	13.44
Homosexual men	Modesty	12.88
Heterosexual women	Tender-Mindedness	15.37
Heterosexual men	Tender-Mindedness	14.49
Homosexual women	Tender-Mindedness	15.56
Homosexual men	Tender-Mindedness	15.38
Homosexual men	Trust	11.50
Heterosexual women	Trust	10.96
Heterosexual men	Trust	11.77
Homosexual women	Trust	11.01
Homosexual men	Straightforwardness	14.12
Homosexual women	Straightforwardness	14.44
Heterosexual women	Straightforwardness	14.98
Heterosexual men	Straightforwardness	13.20
Homosexual men	Cooperation	10.82
Heterosexual men	Cooperation	10.96
Homosexual women	Cooperation	10.65
Heterosexual women	Cooperation	11.05
Homosexual men	Altruisim	13.62
Heterosexual women	Altruisim	14.37
Heterosexual men	Altruisim	12.48
Homosexual women	Altruisim	13.21

Table 7.2
Raw facets' average per sex and sexual orientation group (continued)

Group	Facet	Raw facet average
Heterosexual women	Competence	11.46
Homosexual men	Competence	11.60
Heterosexual men	Competence	12.30
Homosexual women	Competence	11.85
Homosexual men	Achievement Striving	11.06
Heterosexual women	Achievement Striving	11.67
Homosexual women	Achievement Striving	11.09
Heterosexual men	Achievement Striving	11.77
Homosexual men	Order	10.83
Heterosexual women	Order	11.05
Heterosexual men	Order	10.37
Homosexual women	Order	11.07
Heterosexual men	Self-discipline	12.09
Homosexual men	Self-discipline	12.19
Heterosexual women	Self-discipline	12.98
Homosexual women	Self-discipline	11.74
Heterosexual women	Dutifulness	15.04
Heterosexual men	Dutifulness	13.71
Homosexual men	Dutifulness	14.11
Homosexual women	Dutifulness	14.19
Homosexual men	Deliberation	10.68
Heterosexual women	Deliberation	10.81
Heterosexual men	Deliberation	11.28
Homosexual women	Deliberation	10.50

Table 7.2
Raw facets' average per sex and sexual orientation group (continued)

Group	Facet	Raw facet average
Heterosexual men	Positive emotions	11.19
Homosexual women	Positive emotions	11.25
Homosexual men	Positive emotions	11.11
Heterosexual women	Positive emotions	12.29
Homosexual men	Gregariousness	11.22
Heterosexual men	Gregariousness	11.67
Homosexual women	Gregariousness	9.82
Heterosexual women	Gregariousness	11.42
Homosexual women	Warmth	12.02
Homosexual men	Warmth	12.27
Heterosexual men	Warmth	12.40
Heterosexual women	Warmth	12.80
Heterosexual men	Assertiveness	11.52
Homosexual women	Assertiveness	10.82
Homosexual men	Assertiveness	10.71
Heterosexual women	Assertiveness	10.75
Heterosexual women	Activity	10.14
Heterosexual men	Activity	10.30
Homosexual women	Activity	9.21
Homosexual men	Activity	9.74
Heterosexual women	Excitement-Seeking	9.68
Homosexual men	Excitement-Seeking	9.86
Heterosexual men	Excitement-Seeking	10.61
Homosexual women	Excitement-Seeking	9.15

Table 7.2
Raw facets' average per sex and sexual orientation group (continued)

Group	Facet	Raw facet average
Heterosexual men	Anxiety	9.69
Homosexual women	Anxiety	11.96
Homosexual men	Anxiety	11.28
Heterosexual women	Anxiety	11.49
Heterosexual men	Anger	9.18
Heterosexual women	Anger	9.54
Homosexual women	Anger	9.82
Homosexual men	Anger	9.61
Heterosexual men	Depression	9.27
Homosexual men	Depression	10.55
Heterosexual women	Depression	10.51
Homosexual women	Depression	11.32
Heterosexual men	Self-Consciousness	9.82
Homosexual women	Self-Consciousness	12.07
Homosexual men	Self-Consciousness	10.73
Heterosexual women	Self-Consciousness	11.91
Homosexual women	Impulsiveness	11.26
Heterosexual men	Impulsiveness	10.12
Homosexual men	Impulsiveness	10.84
Heterosexual women	Impulsiveness	10.38
Heterosexual men	Vulnerability	9.33
Homosexual women	Vulnerability	10.86
Heterosexual women	Vulnerability	10.75
Homosexual men	Vulnerability	10.41

Table 7.2
Raw facets' average per sex and sexual orientation group (continued)

Group	Facet	Raw facet average
Heterosexual women	Action	9.89
Homosexual women	Action	9.59
Homosexual men	Action	10.41
Heterosexual men	Action	10.98
Heterosexual women	Values	8.82
Heterosexual men	Values	9.16
Homosexual women	Values	10.12
Homosexual men	Values	9.29
Heterosexual men	Ideas	13.41
Heterosexual women	Ideas	12.27
Homosexual women	Ideas	13.34
Homosexual men	Ideas	13.13
Heterosexual women	Aesthetics	11.77
Heterosexual men	Aesthetics	12.00
Homosexual men	Aesthetics	12.55
Homosexual women	Aesthetics	12.75
Heterosexual women	Fantasy	12.32
Heterosexual men	Fantasy	12.00
Homosexual women	Fantasy	12.54
Homosexual men	Fantasy	12.86
Homosexual women	Feelings	13.42
Heterosexual women	Feelings	13.43
Heterosexual men	Feelings	11.38
Homosexual men	Feelings	12.43

References

Achaa-Amankwaa, P., Olaru, G., & Schroeders, U. (2020). Coffee or Tea? Examining Cross-Cultural Differences in Personality Nuances Across Former Colonies of the British Empire. *European Journal of Personality*, *35*(3), 383–397. http://doi.org/10.31234/osf.io/dpqrx

Alexander, G. M., & Wilcox, T. (2012). Sex Differences in Early Infancy. *Child Development Perspectives*, *6*(4), 400–406. http://doi.org/10.1111/j.1750-8606.2012.00247.x

Allen, M. S., & Robson, D. A. (2020). Personality and Sexual Orientation: New Data and Meta-analysis. *The Journal of Sex Research*, *57*(8), 953–965. http://doi.org/10.1080/00224499.2020.1768204

Allik, J., Realo, A., Mõttus, R., Borkenau, P., Kuppens, P., & Hřebíčková, M. (2010). How people see others is different from how people see themselves: A replicable pattern across cultures. *Journal of Personality and Social Psychology*, *99*(5), 870–882. http://doi.org/10.1037/a0020963

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), i–171. http://doi.org/10.1037/h0093360

Alpu, O., Yuksek, D., & Department of Statistics, Eskisehir Osmangazi University, Eskisehir 26480, Turkey. (2016). Comparison of some multivariate normality tests: A simulation study. *International Journal of Advanced and Applied Sciences*, *3*(12), 73–85. http://doi.org/10.21833/ijaas.2016.12.011

American Psychological Association. (2012). Guidelines for psychological practice with lesbian, gay, and bisexual clients. *American Psychologist*, *67*(1), 10–42. http://doi.org/10.1037/a0024659

Andersen, L., & Cross, T. L. (2014). Are Students With High Ability in Math More Motivated in Math and Science Than Other Students? *Roeper Review*, *36*(4), 221–234. http://doi.org/10.1080/02783193.2014.945221

Andersson, H., Lovén, J., & Bergman, L. R. (2014). The Importance of High Competence in Adolescence for Career Outcomes in Midlife. *Research in Human Development*, *11*(3), 204–216. http://doi.org/10.1080/15427609.2014.936180

Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, *52*(1), 376–387. http://doi.org/10.3758/s13428-019-01236-y

Bachman, J. G., & O'Malley, P. M. (1986). Self-concepts, self-esteem, and educational experiences: The frog pond revisited (again). *Journal of Personality and Social Psychology*, *50*(1), 35–46. http://doi.org/10.1037/0022-3514.50.1.35

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by Supervised Principal Components. *Journal of the American Statistical Association*, *101*(473), 119–137. http://doi.org/10.1198/016214505000000628

Baird, C. L. (2012). Going against the flow: A longitudinal study of the effects of cognitive skills and gender beliefs on occupational aspirations and Outcomes1. *Sociological Forum*, *27*(4), 986–1009. http://doi.org/10.1111/j.1573-7861.2012.01365.x

Ball, S. J., Davies, J., David, M., & Reay, D. (2002). 'Classification' and 'Judgement': Social class and the 'cognitive structures' of choice of Higher Education. *British Journal of Sociology of Education*, *23*(1), 51–72. http://doi.org/10.1080/01425690120102854

Bao, A.-M., & Swaab, D. F. (2011). Sexual differentiation of the human brain: Relation to gender identity, sexual orientation and neuropsychiatric disorders. *Frontiers in Neuroendocrinology*, *32*(2), 214–226. http://doi.org/10.1016/j.yfrne.2011.02.007

Basler, A., & Kriesi, I. (2019). Adolescents' development of occupational aspirations in a tracked and vocation-oriented educational system. *Journal of Vocational Behavior*, *115*, 103330. http://doi.org/10.1016/j.jvb.2019.103330

Benbow, C. P., Lubinski, D., Shea, D. L., & Eftekhari-Sanjani, H. (2000). Sex Differences in Mathematical Reasoning Ability at Age 13: Their Status 20 Years Later. *Psychological Science*, *11*(6), 474–480. http://doi.org/10.1111/1467-9280.00291

Bergman, L. R., Corovic, J., Ferrer-Wreder, L., & Modig, K. (2014). High IQ in Early Adolescence and Career Success in Adulthood: Findings from a Swedish Longitudinal Study. *Research in Human Development*, *11*(3), 165–185. http://doi.org/10.1080/15427609.2014.936261

Bergman, L. R., Ferrer-Wreder, L., & Žukauskienė, R. (2015). Career outcomes of adolescents with below average IQ: Who succeeded against the odds? *Intelligence*, *52*, 9–17. http://doi.org/10.1016/j.intell.2015.06.003

Binet, A., & Simon, T. (1905-1916). New methods for the diagnosis of the intellectual level of subnormals. In H. H. Goddard (Ed.), *Development of intelligence in children (the Binet-Simon Scale)* (pp. 37–90). Williams & Wilkins Co. http://doi.org/10.1037/11069-002

Block, J. (1995). A Contrarian View of the Five-Factor Approach to Personality Description. *Psychological Bulletin*, *117*(2), 187–215. http://doi.org/10.1037/0033-2909.117.2.187

Boake, C. (2002). From the Binet–Simon to the Wechsler–Bellevue: Tracing the History of Intelligence Testing. *Journal of Clinical and Experimental Neuropsychology*, *24*(3), 383–405. http://doi.org/10.1076/jcen.24.3.383.981

Booth, T., & Irwing, P. (2011). Sex differences in the 16PF5, test of measurement invariance and mean differences in the US standardisation sample. *Personality and Individual Differences*, *50*(5), 553–558. http://doi.org/10.1016/j.paid.2010.11.026

Booth, T., & Murray, A. L. (2018). How Factor Analysis Has Shaped Personality Trait Psychology. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 933–951). Chichester, UK: John Wiley & Sons, Ltd. http://doi.org/10.1002/9781118489772.ch29

Bordin, E. S. (1943). A Theory of Vocational Interests as Dynamic Phenomena'. *Educational and Psychological Measurement*, *3*(1), 46–62. http://doi.org/10.1177/001316444300300105

Bourdieu, P., & Passeron, J.-C. (1977). Cultural capital and pedagogical communication. In *Education, society and culture* (pp. 71–106). SAGE Publications.

Briggs, S. R. (1989). The Optimal Level of Measurement for Personality Constructs. In D. M. Buss & N. Cantor (Eds.), *Personality Psychology* (pp. 246–260). New York, NY: Springer US. http://doi.org/10.1007/978-1-4684-0634-4_19

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships*, *35*(1), 32–58. http://doi.org/10.1177/0265407517718387

Cattell, R. B. (1943). The description of personality. I. Foundations of trait measurement. *Psychological Review*, *50*(6), 559–594. http://doi.org/10.1037/h0057276

Cattell, R. B. (1947). Confirmation and clarification of primary personality factors. *Psychometrika*, *12*(3), 197–220. http://doi.org/10.1007/BF02289253

Ceci, S. J., Ginther, D. K., Kahn, S., & Williams, W. M. (2014). Women in Academic Science: A Changing Landscape. *Psychological Science in the Public Interest*, *15*(3), 75–141. http://doi.org/10.1177/1529100614541236

Chamorro-Premuzic, T., Harlaar, N., Greven, C. U., & Plomin, R. (2010). More than just IQ: A longitudinal examination of self-perceived abilities as predictors of academic performance in a large sample of UK twins. *Intelligence*, *38*(4), 385–392. http://doi.org/10.1016/j.intell.2010.05.002

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods*, *21*(4), 603–620. http://doi.org/10.1037/met0000088

Charles, M., & Bradley, K. (2009). Indulging Our Gendered Selves? Sex Segregation by Field of Study in 44 Countries. *American Journal of Sociology*, *114*(4), 924–976. http://doi.org/10.1086/595942

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. http://doi.org/10.1207/S15328007SEM0902_5

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy*, *17*(3), 305–308. http://doi.org/10.1016/S0005-7894(86)80060-0

Clariana, M. (2017). Personalidad, Procrastinación y Conducta Deshonesta en Alumnado de distintos Grados Universitarios. *Electronic Journal of Research in Education Psychology*, *11*(30). http://doi.org/10.14204/ejrep.30.13030

Classism. (2021). Merriam-Webster online dictionary. Houghton Mifflin Harcourt Publishing Company. Retrieved from https://www.merriam-webster.com/dictionary/classism

Cochran, D. B., Wang, E. W., Stevenson, S. J., Johnson, L. E., & Crews, C. (2011). Adolescent Occupational Aspirations: Test of Gottfredson's Theory of Circumscription and Compromise. *The Career Development Quarterly*, *59*(5), 412–427. http://doi.org/10.1002/j.2161-0045.2011.tb00968.x

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. http://doi.org/10.1201/b18676-7

Condon, D. M. (2018). The SAPA Personality Inventory: An empirically-derived, hierarchically-organized self-report personality assessment model. http://doi.org/10.31234/osf.io/sc4p9

Condon, D. M., Wood, D., Mõttus, R., Booth, T., Costantini, G., Greiff, S., … Zimmermann, J. (2021). Bottom Up Construction of a Personality Taxonomy. *European Journal of Psychological Assessment*, *36*(6), 923–934. http://doi.org/10.1027/1015-5759/a000626

Corder-Bolz, C. R. (1978). The Evaluation of Change: New Evidence. *Educational and Psychological Measurement*, *38*(4), 959–976. http://doi.org/10.1177/001316447803800416

Costa, P. T., & McCrae, R. R. (1976). Age Differences in Personality Structure: A Cluster Analytic Approach. *Journal of Gerontology*, *31*(5), 564–570.

http://doi.org/10.1093/geronj/31.5.564

Costa, Paul T., & McCrae, R. R. (1978). Objective personality assessment. In *The clinical psychology of aging* (pp. 119–143). New York: Plenum: Springer. http://doi.org/10.1007/978-1-4684-3342-5_5

Costa, Paul T., & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653–665. http://doi.org/10.1016/0191-8869(92)90236-I

Costa, Paul T., & McCrae, R. R. (1992b). Professional manual: Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI). *Odessa, FL: Psychological Assessment Resources*, *61*.

Costa, Paul T., Robert, M. R., & Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences*, *12*(6), 887–898. http://doi.org/10.1016/0191-8869(91)90177-D

Costa, Paul T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *American Psychological Association*, *81*(2), 322–331. http://doi.org/10.1037/0022-3514.81.2.322

Coyle, T. R., Snyder, A. C., & Richmond, M. C. (2015). Sex differences in ability tilt: Support for investment theory. *Intelligence*, *50*, 209–220. http://doi.org/10.1016/j.intell.2015.04.012

Cramer, A. O. J., Van Der Sluis, S., Noordhof, A., Wichers, M., Geschwind, N., Aggen, S. H., ... Borsboom, D. (2012). Dimensions of Normal Personality as Networks in Search of Equilibrium: You Can't like Parties if you Don't like People. *European Journal of Personality*, *26*(4), 414–431. http://doi.org/10.1002/per.1866

Croll, P. (2008). Occupational choice, socio-economic status and educational attainment: A study of the occupational choices and destinations of young people in the British Household Panel Survey. *Research Papers in Education*, *23*(3), 243–268. http://doi.org/10.1080/02671520701755424

Davis, J. A. (1966). The Campus as a Frog Pond: An Application of the Theory of Relative Deprivation to Career Decisions of College Men. *American Journal of Sociology*, *72*(1), 17–31. http://doi.org/10.1086/224257

De Fruyt, F., De Bolle, M., McCrae, R. R., Terracciano, A., Costa, P. T., & Collaborators of the Adolescent Personality Profiles of Cultures Project. (2009). Assessing the Universal Structure of Personality in Early Adolescence: The NEO-PI-R and NEO-PI-3 in 24 Cultures. *Assessment*, *16*(3), 301–311. http://doi.org/10.1177/1073191109333760

De Fruyt, F., & Mervielde, I. (1996). Personality and interests as predictors of educational streaming and achievement. *Wiley Online Library*, *10*(5), 405–425. http://doi.org/10.1002/(sici)1099-0984(199612)10:5<405::aid-per255>3.0.co;2-m

Deary, I. J., Cox, S. R., & Ritchie, S. J. (2016). Getting Spearman off the Skyhook: One More in a Century (Since Thomson, 1916) of Attempts to Vanquish *g*. *Psychological Inquiry*, *27*(3), 192–199. http://doi.org/10.1080/1047840X.2016.1186525

Deary, I. J., & Sternberg, R. J. (2021). Answer five self-inflicted questions about human intelligence. *Intelligence*, *86*, 101539. http://doi.org/10.1016/j.intell.2021.101539

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13–21. http://doi.org/10.1016/j.intell.2006.02.001

Deary, I. J., Taylor, M. D., Hart, C. L., Wilson, V., Smith, G. D., Blane, D., & Starr, J. M. (2005). Intergenerational social mobility and mid-life status attainment: Influences of childhood intelligence, childhood social factors, and education. *Intelligence*, *33*(5), 455–472. http://doi.org/10.1016/j.intell.2005.06.003

Deary, I. J., Weiss, A., & Batty, G. D. (2010). Intelligence and Personality as Predictors of Illness and Death: How Researchers in Differential Psychology and Chronic Disease Epidemiology Are Collaborating to Understand and Address Health Inequalities. *Psychological Science in the Public Interest*, *11*(2), 53–79. http://doi.org/10.1177/1529100610387081

Dekhtyar, S., Weber, D., Helgertz, J., & Herlitz, A. (2018). Sex differences in academic strengths contribute to gender segregation in education and occupation: A longitudinal examination of 167,776 individuals. *Intelligence*, *67*, 84–92.

http://doi.org/10.1016/j.intell.2017.11.007

Del Giudice, M. (2017). Heterogeneity Coefficients for Mahalanobis' *D* as a Multivariate Effect Size. *Multivariate Behavioral Research*, *52*(2), 216–221. http://doi.org/10.1080/00273171.2016.1262237

Del Giudice, M. (2018). Maha_v8c.R.

Del Giudice, M. (2019). Measuring Sex Differences and Similarities. In *Gender and sexuality development: Contemporary theory and research* (D. P. VanderLaan & W. I. Wong). New York, NY: Springer.

Del Giudice, M., Booth, T., & Irwing, P. (2012). The Distance Between Mars and Venus: Measuring Global Sex Differences in Personality. *PLoS ONE*, *7*(1), e29265. http://doi.org/10.1371/journal.pone.0029265

DesJardins, S. L., Toutkoushian, R. K., Hossler, D., & Chen, J. (2019). Time may change me: Examining how aspirations for college evolve during high school. *The Review of Higher Education*, *43*(1), 263–294. http://doi.org/10.1353/rhe.2019.0096

Ditton, H., Bayer, M., & Wohlkinger, F. (2019). Structural and motivational mechanisms of academic achievement: A mediation model of social-background effects on academic achievement. *The British Journal of Sociology*, *70*(4), 1276–1296. http://doi.org/10.1111/1468-4446.12506

Dubow, E. F., Boxer, P., & Huesmann, L. R. (2009). Long-term Effects of Parents' Education on Children's Educational and Occupational Success: Mediation by Family Interactions, Child Aggression, and Teenage Aspirations. *Merrill-Palmer Quarterly*, *55*(3), 224–249. http://doi.org/10.1353/mpq.0.0030

Dumfart, B., & Neubauer, A. C. (2016). Conscientiousness Is the Most Powerful Noncognitive Predictor of School Achievement in Adolescents. *Journal of Individual Differences*, *37*(1), 8–15. http://doi.org/10.1027/1614-0001/a000182

Eccles, J. S. (1983). Expectancies, values and academic behaviors. In *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–146). Spence, J. T.

Elliott, C. D. (1986). The factorial structure and specificity of the British Ability Scales. *British Journal of Psychology*, *77*(2), 175–185. http://doi.org/10.1111/j.2044-8295.1986.tb01992.x

Ellis, L., Ratnasingam, M., & Wheeler, M. (2012). Gender, sexual orientation, and occupational interests: Evidence of their interrelatedness. *Personality and Individual Differences*, *53*(1), 64–69. http://doi.org/10.1016/j.paid.2012.02.008

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127. http://doi.org/10.1037/a0018053

Epskamp, S., Cramer, A., Waldorp, L., Schmittmann, V., & Borsboom, D. (2011). Qgraph: Network representations of relationships in data. *R Package Version 0.4*, *10*. http://doi.org/10.18637/jss.v048.i04

Eronen, M. I., & Bringmann, L. F. (2021). The Theory Crisis in Psychology: How to Move Forward. *Perspectives on Psychological Science*, 174569162097058. http://doi.org/10.1177/1745691620970586

Eysenck, H. J. (1947). *Dimensions of personality*. New Brunswick, N.J: Transaction Publishers. http://doi.org/10.4324/9780203793268

Eysenck, H. J. (1991). Dimensions of personality: 16, 5 or 3?Criteria for a taxonomic paradigm. *Personality and Individual Differences*, *12*(8), 773–790. http://doi.org/10.1016/0191-8869(91)90144-Z

Eysenck, H. J. (1992). A reply to Costa and McCrae. P or A and Cthe role of theory. *Personality and Individual Differences*, *13*(8), 867–868. http://doi.org/10.1016/0191-8869(92)90003-8

Eysenck, M. W. (2016). Hans Eysenck: A research evaluation. *Personality and Individual Differences*, *103*, 209–219. http://doi.org/10.1016/j.paid.2016.04.039

Farrell, P. J., Salibian-Barrera, M., & Naczk, K. (2007). On tests for multivariate normality and associated simulation studies. *Journal of Statistical Computation and Simulation*, *77*(12), 1065–1080. http://doi.org/10.1080/10629360600878449

Fass, P. S. (1980). The IQ: A Cultural and Historical Framework. *American Journal of Education*, *88*(4), 431–458. http://doi.org/10.1086/443541

Feingold, A. (1994). Gender Differences in Personality: A Meta-Analysis. *American Psychological Association*, *116*(3), 429–456. http://doi.org//10.1016/j.paid.2007.05.004

Felson, R. B., & Reed, M. D. (1986). Reference Groups and Self-Appraisals of Academic Ability and Performance. *Social Psychology Quarterly*, *49*(2), 103–109. http://doi.org/10.2307/2786722

Fischer, F., Schult, J., & Hell, B. (2013). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education*, *28*(2), 529–543. http://doi.org/10.1007/s10212-012-0127-4

Fitzsimons, E. (2017). Millennium cohort study sixth survey 2015-2016: User guide.

Freud, S. (1953). Three essays on the theory of sexuality (1905). In *The standard edition of the complete psychological works of Sigmund Freud, volume VII (1901-1905): A case of hysteria, three essays on sexuality and other works* (pp. 123–246). http://doi.org/10.1037/e417472005-189

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. http://doi.org/10.1177/2515245919847202

Gale, T. (2015). Widening and expanding participation in Australian higher education: In the absence of sociological imagination. *The Australian Educational Researcher*, *42*(2), 257–271. http://doi.org/10.1007/s13384-014-0167-7

Galliott, N., & Graham, L. J. (2015). School based experiences as contributors to career decision-making: Findings from a cross-sectional survey of high-school students. *The Australian Educational Researcher*, *42*(2), 179–199. http://doi.org/10.1007/s13384-015-0175-2

Gaspard, H., Häfner, I., Parrisius, C., Trautwein, U., & Nagengast, B. (2017). Assessing task values in five subjects during secondary school: Measurement structure and mean level differences across grade level, gender, and academic subject. *Contemporary Educational Psychology*, *48*, 67–84.

http://doi.org/10.1016/j.cedpsych.2016.09.003

Giangrande, E., & Turkheimer, E. (2021). Race, ethnicity, and the scarr-rowe hypothesis: A cautionary example of fringe science entering the mainstream. http://doi.org/10.31234/osf.io/md3qw

Giudice, M. D. (2021). Individual and group differences in multivariate domains: What happens when the number of traits increases? http://doi.org/10.31234/osf.io/rgzd2

Goldberg, Lewis R. (1981). Unconfounding Situational Attributions From Uncertain, Neutral, and Ambiguous Ones: A Psychometric Analysis of Descriptions of Oneself and Various Types of Others. *Journal of Personality and Social Psychology*, *41*(3), 517–552. http://doi.org/10.1037/0022-3514.41.3.517

Goldberg, Lewis R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality Psychology in Europe*, *7*(1), 7–28.

Gollwitzer, M., Christ, O., & Lemmer, G. (2014). Individual differences make a difference: On the use and the psychometric properties of difference scores in social psychology: Reliability of difference scores. *European Journal of Social Psychology*, *44*(7), 673–682. http://doi.org/10.1002/ejsp.2042

Gore, J., Holmes, K., Smith, M., Fray, L., McElduff, P., Weaver, N., & Wallington, C. (2017). Unpacking the career aspirations of Australian school students: Towards an evidence base for university equity initiatives in schools. *Higher Education Research & Development*, *36*(7), 1383–1400. http://doi.org/10.1080/07294360.2017.1325847

Gore, J., Holmes, K., Smith, M., Southgate, E., & Albright, J. (2015). Socioeconomic status and the career aspirations of Australian school students: Testing enduring assumptions. *The Australian Educational Researcher*, *42*(2), 155–177. http://doi.org/10.1007/s13384-015-0172-5

Gottfredson, Linda S. (1979). Aspiration-job match: Age trends in a large, nationally representative sample of young White men. *American Psychological Association*, *26*(4), 319–328. http://doi.org//10.1037/0022-0167.26.4.319

Gottfredson, Linda S. (1981). Circumscription and Compromise: A Developmental Theory of Occupational Aspirations. *Journal of Counseling Psychology*, *28*(6), 545–579. http://doi.org/10.1037/0022-0167.28.6.545

Gottfredson, Linda S., & Becker, H. J. (1981). A challenge to vocational psychology: How important are aspirations in determining male career development? *Journal of Vocational Behavior*, *18*(2), 121–137. http://doi.org/10.1016/0001-8791(81)90001-4

Gottfredson, Linda S., & Lapan, R. T. (1997). Assessing Gender-Based Circumscription of Occupational Aspirations. *Journal of Career Assessment*, *5*(4), 419–441. http://doi.org/10.1177/106907279700500404

Government Digital Service. (2020). What qualification level mean.

Great Britain, Office for National Statistics, Great Britain, & Government Statistical Service. (2010). *Standard occupational classification 2010*. Basingstoke, Hampshire: Palgrave Macmillan.

Greaves, L. M., Barlow, F. K., Huang, Y., Stronge, S., & Sibley, C. G. (2017). Personality across Sexual Identity and Gender in a National Probability Sample in New Zealand. *Sex Roles*, *77*(9-10), 653–662. http://doi.org/10.1007/s11199-017-0752-0

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. http://doi.org/10.1037/1082-989x.6.4.430

Guo, J., Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2015). Directionality of the Associations of High School Expectancy-Value, Aspirations, and Attainment: A Longitudinal Study. *American Educational Research Journal*, *52*(2), 371–402. http://doi.org/10.3102/0002831214565786

Haller, A. O., Otto, L. B., Meier, R. F., & Ohlendorf, G. W. (1974). Level of Occupational Aspiration: An Empirical Analysis. *American Sociological Review*, *39*(1), 113–121. http://doi.org/10.2307/2094280

Hegna, K. (2014). Changing educational aspirations in the choice of and transition to post-compulsory schooling: A three-wave longitudinal study of oslo youth. *Jour-*

nal of Youth Studies, *17*(5), 592–613. http://doi.org/10.1080/13676261.2013.853870

Heineck, G. (2011). Does it pay to be nice? Personality and earnings in the united kingdom. *ILR Review*, *64*(5), 1020–1038. http://doi.org/10.1177/001979391106400509

Helwig, A. A. (2001). A test of Gottfredson's theory using a ten-year longitudinal study. *Journal of Career Development*, *28*(2), 77–95. http://doi.org/10.1177/0894845301028002

Heternormativity. (2021). Merriam-Webster online dictionary. Houghton Mifflin Harcourt Publishing Company. Retrieved from https://www.merriam-webster.com/dictionary/heteronormativity

Hoff, K. A., Briley, D. A., Wee, C. J. M., & Rounds, J. (2018). Normative changes in interests from adolescence to adulthood: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *144*(4), 426–451. http://doi.org/10.1037/bul0000140

Hoff, K. A., Song, Q. C., Einarsdóttir, S., Briley, D. A., & Rounds, J. (2020). Developmental structure of personality and interests: A four-wave, 8-year longitudinal study. *Journal of Personality and Social Psychology*, *118*(5), 1044–1064. http://doi.org/10.1037/pspp0000228

Holland, John L. (1959). A theory of vocational choice. *Journal of Counseling Psychology*, *6*(1), 35–45. http://doi.org/10.1037/h0040767

Holland, John L. (1966). Exploring careers with a typology: What we have learned and some new directions. *American Psychological Association*, *51*(4), 397–406.

Holmes, K., Gore, J., Smith, M., & Lloyd, A. (2018). An Integrated Analysis of School Students' Aspirations for STEM Careers: Which Student and School Factors Are Most Predictive? *International Journal of Science and Mathematics Education*, *16*(4), 655–675. http://doi.org/10.1007/s10763-016-9793-z

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270. http://doi.org/10.1037/h0023816

Horn, J. L., & Mcardle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144.

http://doi.org/10.1080/03610739208253916

Howard, K. A. S., Carlstrom, A. H., Katz, A. D., Chew, A. Y., Ray, G. C., Laine, L., & Caulum, D. (2011). Career aspirations of youth: Untangling race/ethnicity, SES, and gender. *Journal of Vocational Behavior*, *79*(1), 98–109. http://doi.org/10.1016/j.jvb.2010.12.002

Hunalp, B. I. (2020). What happens in rich and gender egalitarian countries: A systematic review of gender differences across 6 features in order to explain STEM gender disparity. http://doi.org/10.13140/RG.2.2.31970.35521

Hunt, E. (2010). *Human intelligence*. Cambridge University Press.

Hustad, I. B., Bandholtz, J., Herlitz, A., & Dekhtyar, S. (2020). Occupational Attributes and Occupational Gender Segregation in Sweden: Does It Change Over Time? *Frontiers in Psychology*, *11*. http://doi.org/10.3389/fpsyg.2020.00554

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581–592. http://doi.org/10.1037/0003-066X.60.6.581

Hyde, J. S. (2014). Gender Similarities and Differences. *Annual Review of Psychology*, *65*(1), 373–398. http://doi.org/10.1146/annurev-psych-010213-115057

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, *73*(2), 509–527. http://doi.org/10.1111/1467-8624.00421

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*(4), 336–352. http://doi.org/10.1016/S0005-7894(84)80002-7

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport: CT: Praeger.

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

Personality, *2*(3), 171–203. http://doi.org/10.1002/per.2410020302

Johnson, W., & Bouchard, T. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*(4), 393–416. http://doi.org/10.1016/j.intell.2004.12.002

Johnson, Wendy, Bouchard, T. J., Krueger, R. F., McGue, M., & Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence*, *32*(1), 95–107. http://doi.org/10.1016/S0160-2896(03)00062-X

Johnson, Wendy, McGue, M., & Iacono, W. G. (2007). Socioeconomic status and school grades: Placing their association in broader context in a sample of biological and adoptive families. *Intelligence*, *35*(6), 526–541. http://doi.org/10.1016/j.intell.2006.09.006

Jonason, P. K., & Luoto, S. (2021). The dark side of the rainbow: Homosexuals and bisexuals have higher Dark Triad traits than heterosexuals. *Personality and Individual Differences*, *181*, 111040. http://doi.org/10.1016/j.paid.2021.111040

Kaiser, D. T. (2019). Nature and evoked culture: Sex differences in personality are uniquely correlated with ecological stress. *Personality and Individual Differences*, *148*, 67–72. http://doi.org/10.31234/osf.io/u73gh

Kajonius, P., & Mac Giolla, E. (2017). Personality traits across countries: Support for similarities rather than differences. *PLOS ONE*, *12*(6), e0179646. http://doi.org/10.1371/journal.pone.0179646

Kan, K.-J., de Jonge, H., van der Maas, H. L. J., Levine, S. Z., & Epskamp, S. (2020). How to Compare Psychometric Factor and Network Models. *Journal of Intelligence*, *8*(4), 35. http://doi.org/10.3390/jintelligence8040035

Kaufman, J. C., Pumaccahua, T. T., & Holt, R. E. (2013). Personality and creativity in realistic, investigative, artistic, social, and enterprising college majors. *Personality and Individual Differences*, *54*(8), 913–917. http://doi.org/10.1016/j.paid.2013.01.013

King, D. D., Ott-Holland, C. J., Ryan, A. M., Huang, J. L., Wadlington, P. L., & Elizondo, F. (2017). Personality Homogeneity in Organizations and Occupations:

Considering Similarity Sources. *Journal of Business and Psychology*, *32*(6), 641–653. http://doi.org/10.1007/s10869-016-9459-4

Kline, P., & Lapham, S. L. (1992). Personality and faculty in British universities. *Personality and Individual Differences*, *13*(7), 855–857. http://doi.org/10.1016/0191-8869(92)90061-S

Kovacs, K., & Conway, A. R. A. (2016). Process Overlap Theory: A Unified Account of the General Factor of Intelligence. *Psychological Inquiry*, *27*(3), 151–177. http://doi.org/10.1080/1047840X.2016.1153946

Krieger, N., Chen, J. T., & Selby, J. V. (1999). Comparing individual-based and household-based measures of social class to assess class inequalities in women's health: A methodological study of 684 US women. *Journal of Epidemiology & Community Health*, *53*(10), 612–623. http://doi.org/10.1136/jech.53.10.612

Larson, L. M., Rottinghaus, P. J., & Borgen, F. H. (2002). Meta-analyses of Big Six Interests and Big Five Personality Factors. *Journal of Vocational Behavior*, *61*(2), 217–239. http://doi.org/10.1006/jvbe.2001.1854

Lee, I. H., & Rojewski, J. W. (2012). Development of occupational aspirations in early Korean adolescents: A multiple-group latent curve model analysis. *International Journal for Educational and Vocational Guidance*, *12*(3), 189–210. http://doi.org/10.1007/s10775-012-9227-6

Lent, R. W., Brown, S. D., & Hackett, G. (2002). Social cognitive career theory. *Career Choice and Development*, *4*(1), 255–311. http://doi.org/10.4135/9781412952675

Lippa, R. A. (1998). Gender-Related Individual Differences and the Structure of Vocational Interests: The Importance of the People-Things Dimension. *Journal of Personality and Social Psychology*, *74*(4), 996–1009. http://doi.org/10.1037/0022-3514.74.4.996

Lippa, R. A. (2001). On Deconstructing and Reconstructing MasculinityFemininity. *Journal of Research in Personality*, *35*(2), 168–207. http://doi.org/10.1006/jrpe.2000.2307

Lippa, R. A. (2005). Sexual orientation and personality. *Annual Review of Sex Research*, *16*(1), 119–153. http://doi.org/10.1080/10532528.2005.10559831

Lippa, R. A. (2008). Sex Differences and Sexual Orientation Differences in Personality: Findings from the BBC Internet Survey. *Archives of Sexual Behavior*, *37*(1), 173–187. http://doi.org/10.1007/s10508-007-9267-z

Lippa, R. A. (2010). Sex Differences in Personality Traits and Gender-Related Occupational Preferences across 53 Nations: Testing Evolutionary and Social-Environmental Theories. *Archives of Sexual Behavior*, *39*(3), 619–636. http://doi.org/10.1007/s10508-008-9380-7

Lippa, R. A. (2020). Interest, Personality, and Sexual Traits That Distinguish Heterosexual, Bisexual, and Homosexual Individuals: Are There Two Dimensions That Underlie Variations in Sexual Orientation? *Archives of Sexual Behavior*, *49*(2), 607–622. http://doi.org/10.1007/s10508-020-01643-9

Lippa, R. A., Preston, K., & Penner, J. (2014). Women's Representation in 60 Occupations from 1972 to 2010: More Women in High-Status Jobs, Few Women in Things-Oriented Jobs. *PLoS ONE*, *9*(5), e95960. http://doi.org/10.1371/journal.pone.0095960

Loehlin, J. C., & Goldberg, L. R. (2014). Do personality traits conform to lists or hierarchies? *Personality and Individual Differences*, *70*, 51–56. http://doi.org/10.1016/j.paid.2014.06.018

Low, K. S. D., Yoon, M., Roberts, B. W., & Rounds, J. (2005). The Stability of Vocational Interests From Early Adolescence to Middle Adulthood: A Quantitative Review of Longitudinal Studies. *Psychological Bulletin*, *131*(5), 713–737. http://doi.org/10.1037/0033-2909.131.5.713

Lubinski, D. (2000). Scientific and Social Significance of Assessing Individual Differences: "Sinking Shafts at a Few Critical Points." *Annual Review of Psychology*, *51*(1), 405–444. http://doi.org/10.1146/annurev.psych.51.1.405

Lubinski, D., & Benbow, C. P. (1992). Gender Differences in Abilities and Preferences Among the Gifted: Implications for the Math-Science Pipeline. *Current Directions in Psychological Science*, *1*(2), 61–66. http://doi.org/10.1111/1467-8721.ep11509746

Luoto, S. (2020). Sex differences in people and things orientation are reflected in sex differences in academic publishing. *Journal of Informetrics*, *14*(2), 101021. http://doi.org/10.1016/j.joi.2020.101021

Mac Giolla, E., & Kajonius, P. J. (2019). Sex differences in personality are larger in gender equal countries: Replicating and extending a surprising finding. *International Journal of Psychology*, *54*(6), 705–711. http://doi.org/10.1002/ijop.12529

Machin, S., & Pekkarinen, T. (2008). ASSESSMENT: Global Sex Differences in Test Score Variability. *Science*, *322*(5906), 1331–1332. http://doi.org/10.1126/science.1162573

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Retrieved from http://library.isical.ac.in:8080/jspui/bitstream/10263/6765/1/Vol02_1936_1_Art05-pcm.pdf

Marjoribanks, K. (2002). Family background, individual and environmental influences on adolescents' aspirations. *Educational Studies*, *28*(1), 33–46. http://doi.org/10.1080/03055690120090361

Marsh, Herbert W. (1984). Self-Concept, Social Comparison, and Ability Grouping: A Reply to Kulik and Kulik. *American Educational Research Journal*, *21*(4), 799–806. http://doi.org/10.3102/00028312021004799

Marsh, Herbert W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295. http://doi.org/10.1037/0022-0663.79.3.280

Marsh, Herbert W., Chessor, D., Craven, R., & Roche, L. (1995). The Effects of Gifted and Talented Programs on Academic Self-Concept: The Big Fish Strikes Again, *32*(2), 285–319. http://doi.org/10.3102/00028312032002285

Marsh, Herbert W., & Hau, K.-T. (2003). Big-Fish–Little-Pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, *58*(5), 364–376. http://doi.org/10.1037/0003-066X.58.5.364

Matthews, G., Deary, I. J., & Whiteman, M. C. (2009). *Personality Traits* (Third). Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511812743

Mau, W.-C., & Bikos, L. H. (2000). Educational and Vocational Aspirations of Minority and Female Students: A Longitudinal Study. *Journal of Counseling & Development*, *78*(2), 186–194. http://doi.org/10.1002/j.1556-6676.2000.tb02577.x

McCabe, K. O., Lubinski, D., & Benbow, C. P. (2020). Who shines most among the brightest?: A 25-year longitudinal study of elite STEM graduate students. *Journal of Personality and Social Psychology*, *119*(2), 390–416. http://doi.org/10.1037/pspp0000239

McCrae, Robert R. (1994). The Counterpoint of Personality Assessment: Self Reports and Observer Ratings. *Assessment*, *1*(2), 159–172. http://doi.org/10.1177/107319119400100200

McCrae, Robert R. (2015). A More Nuanced View of Reliability: Specificity in the Trait Hierarchy. *Personality and Social Psychology Review*, *19*, 97–112. http://doi.org/10.1177/1088868314541857

McCrae, Robert R., & Costa, P. T. (1985a). Comparison of EPI and psychoticism scales with measures of the five-factor model of personality. *Personality and Individual Differences*, *6*(5), 587–597. http://doi.org/10.1016/0191-8869(85)90008-X

McCrae, Robert R., & Costa, P. T. (1985b). Updating Norman's "Adequate Taxonomy": Intelligence and Personality Dimensions in Natural Language and in Questionnaires. *Journal of Personality and Social Psychology*, *49*(3), 710–721. http://doi.org/10.1037/0022-3514.49.3.710

McCrae, Robert R., Costa, P. T., Jr., & Martin, T. A. (2005). The NEOPI: A More Readable Revised NEO Personality Inventory. *Journal of Personality Assessment*, *84*(3), 261–270. http://doi.org/10.1207/s15327752jpa8403_05

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. http://doi.org/10.1016/j.intell.2008.08.004

Mecklin, Christopher J., & Mundfrom, D. J. (2003). On Using Asymptotic Critical Values in Testing for Multivariate Normality. *InterStat*, *1*(1-12), 152–164.

Mecklin, Christopher J., & Mundfrom, D. J. (2005). A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate nor-

mality. *Journal of Statistical Computation and Simulation*, *75*(2), 93–107. http://doi.org/10.1080/00949650420001932 33

Meyer, I. H. (2010). The Right Comparisons in Testing the Minority Stress Hypothesis: Comment on Savin-Williams, Cohen, Joyner, and Rieger (2010). *Archives of Sexual Behavior*, *39*(6), 1217–1219. http://doi.org/10.1007/s10508-010-9670-8

Molenaar, D., & Borsboom, D. (2013). The formalization of fairness: Issues in testing for measurement invariance using subtest scores. *Educational Research and Evaluation*, *19*(2-3), 223—244. http://doi.org/10.1080/13803611.2013.767628

Mostafa, T., & Ploubidis, G. (2017). *Millennium Cohort Study:Sixth Survey 2015-2016, technical report on response (Age 14)* (pp. 1–10). Centre for Longitudinal Studies.

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the Big Five personality traits and the Big Six vocational interest types. *Personnel Psychology*, *58*(2), 447–478. http://doi.org/10.1111/j.1744-6570.2005.00468.x

Mõttus, R., Bates, T. C., Condon, D. M., Mroczek, D., & Revelle, W. (2017). Leveraging a more nuanced view of personality: Narrow characteristics predict and explain variance in life outcomes. http://doi.org/10.31234/osf.io/4q9gv

Mõttus, R., Kandler, C., Bleidorn, W., Riemann, R., & McCrae, R. R. (2017). Personality traits below facets: The consensual validity, longitudinal stability, heritability, and utility of personality nuances. *Journal of Personality and Social Psychology*, *112*(3), 474–490. http://doi.org/10.1037/pspp0000100

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. http://doi.org/10.1016/j.jrp.2014.07.005

Mõttus, R., Wood, D., Condon, D. M., Back, M., Baumert, A., Costantini, G., ... Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the Big Few traits. http://doi.org/10.31234/osf.io/hvk5p

Muthén, L. K., & Muthén, B. O. (1998). *Mplus User's Guide. Sixth Edition.* Los Angeles, CA.

Nicolas, S., Andrieu, B., Croizet, J.-C., Sanitioso, R. B., & Burman, J. T. (2013). Sick? Or slow? On the origins of intelligence as a psychological object. *Intelligence*, *41*(5), 699–711. http://doi.org/10.1016/j.intell.2013.08.006

Noble, M., Wright, G., Smith, G., & Dibben, C. (2006). Measuring Multiple Deprivation at the Small-Area Level. *Environment and Planning A: Economy and Space*, *38*(1), 169–185. http://doi.org/10.1068/a37168

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes. *Journal of Abnormal and Social Psychology*, *66*(6), 574–583. http://doi.org/10.1037/h0040291

OECD. (2007). *PISA 2006: Science Competencies for Tomorrow's World: Volume 1: Analysis*. OECD. http://doi.org/10.1787/9789264040014-en

Office for National Statistics. (2015). *Chapter 3: Equivalised income*.

Office for National Statistics. (2020). *Sexual orientation, UK: Experimental statistics on sexual orientation in the UK in 2017 by region, sex, age, marital status, ethnicity and socio-economic classification.* (Statistical Bulletin) (p. 14). UK.

Office for National Statistics. (2021). The National Statistics Socio-economic Classification (NS-SEC).

Oh, I.-S., Han, J. H., Holtz, B., Kim, Y. J., & Kim, S. (2018). Do birds of a feather flock, fly, and continue to fly together? The differential and cumulative effects of attraction, selection, and attrition on personality-based within-organization homogeneity and between-organization heterogeneity progression over ti. *Journal of Organizational Behavior*, *39*(10), 1347–1366. http://doi.org/10.1002/job.2304

Pace, V. L., & Brannick, M. T. (2010). How similar are personality scales of the "same" construct? A meta-analytic investigation. *Personality and Individual Differences*, *49*(7), 669–676. http://doi.org/10.1016/j.paid.2010.06.014

Parker, P., Nagy, G., Trautwein, U., & Lüdtke, O. (2014). Predicting career aspirations and university majors from academic ability and self-concept. In I. Schoon & J. S. Eccles (Eds.), *Gender Differences in Aspirations and Attainment* (pp. 224–246). Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9781139128933.015

Patton, W., & Creed, P. (2007a). Occupational aspirations and expectations of australian adolescents. *Australian Journal of Career Development*, *16*(1), 46–59. http://doi.org/10.1177/103841620701600108

Patton, W., & Creed, P. (2007b). The relationship between career variables and occupational aspirations and expectations for australian high school adolescents. *Journal of Career Development*, *34*(2), 127–148. http://doi.org/10.1177/0894845307307471

Pervin, L. A. (1994). A Critical Analysis of Current Trait Theory. *Psychological Inquiry*, *5*(2), 103–113. http://doi.org/10.1207/s15327965pli0502_1

Plewis, I., Calderwood, L., Hawkes, D., Hughes, G., & Joshi, H. (2007). Millennium cohort study: Technical report on sampling. London: Centre for Longitudinal Study, Institute of Education.

Plomin, R. (2019). *Blueprint: How DNA makes us who we are*. Mit Press.

Preckel, F., Goetz, T., Pekrun, R., & Kleine, M. (2008). Gender Differences in Gifted and Average-Ability Students: Comparing Girls' and Boys' Achievement, Self-Concept, Interest, and Motivation in Mathematics. *Gifted Child Quarterly*, *52*(2), 146–159. http://doi.org/10.1177/0016986208315834

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior*, *21*(3), 259–287. http://doi.org/10.1016/0001-8791(82)90036-7

Revelle, C. L., & Bonaccio, S. (2013). The Nature and Structure of"Intelligence." In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *The Wiley-Blackwell Handbook of Individual Differences* (pp. 187–216). Oxford, UK: Wiley-Blackwell. http://doi.org/10.1002/9781444343120.ch7

Revelle, W., Wilt, J., & Condon, D. M. (2013). Individual Differences and Differential Psychology: A Brief History and Prospect. In T. Chamorro-Premuzic, S. von Stumm, & A. Furnham (Eds.), *The Wiley-Blackwell Handbook of Individual Differences* (pp. 1–38). Oxford, UK: Wiley-Blackwell. http://doi.org/10.1002/9781444343120.ch1

Rigby, J. (2015). The Life and Evolution of Early Intelligence Theorists: Darwin, Galton, and Charcot. In S. Goldstein, D. Princiotta, & J. A. Naglieri (Eds.), *Handbook of Intelligence* (pp. 93–104). New York, NY: Springer New York. http://doi.org/10.1007/978-1-4939-1562-0_7

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, *143*(2), 117–141. http://doi.org/10.1037/bul0000088

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin*, *92*(3), 726—748. http://doi.org/10.1037/0033-2909.92.3.726

Rojewski, Jay W. (1996). Educational and Occupational Aspirations of High School Seniors with Learning Disabilities. *Exceptional Children*, *62*(5), 463–476. http://doi.org/10.1177/001440299606200506

Rojewski, Jay W., & Yang, B. (1997). Longitudinal Analysis of Select Influences on Adolescents' Occupational Aspirations, *51*(3), 375–410. http://doi.org/10.1006/jvbe.1996.1561

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015). Intelligence and school grades: A meta-analysis. *Intelligence*, *53*, 118–137. http://doi.org/10.1016/j.intell.2015.09.002

Rubinstein, G. (2005). The big five among male and female students of different faculties. *Personality and Individual Differences*, *38*(7), 1495–1503. http://doi.org/10.1016/j.paid.2004.09.012

Saucier, G., & Iurino, K. (2020). High-dimensionality personality structure in the natural language: Further analyses of classic sets of English-language trait-adjectives. *Journal of Personality and Social Psychology*, *119*(5), 1188–1219.

http://doi.org/10.1037/pspp0000273

Schmitt, D. P., Realo, A., Voracek, M., & Allik, J. (2008). Why can't a man be more like a woman? Sex differences in Big Five personality traits across 55 cultures. *Journal of Personality and Social Psychology*, *94*(1), 168–182. http://doi.org/10.1037/0022-3514.94.1.168

Schoon, I. (2001). Teenage job aspirations and career attainment in adulthood: A 17-year follow-up study of teenagers who aspired to become scientists, health professionals, or engineers. *International Journal of Behavioral Development*, *25*(2), 124–132. http://doi.org/10.1080/01650250042000186

Schoon, I., & Parsons, S. (2002). Teenage Aspirations for Future Careers and Occupational Outcomes. *Journal of Vocational Behavior*, *60*(2), 262–288. http://doi.org/10.1006/jvbe.2001.1867

Schoon, I., & Polek, E. (2011). Teenage career aspirations and adult career attainment: The role of gender, social background and general cognitive ability. *International Journal of Behavioral Development*, *35*(3), 210–217. http://doi.org/10.1177/0165025411398183

Schulenberg, J. E., Vondracek, F. W., & Crouter, A. C. (1984). The Influence of the Family on Vocational Development. *Journal of Marriage and the Family*, *46*(1), 129–143. http://doi.org/10.2307/351871

Seeboth, A., & Mõttus, R. (2018). Successful Explanations Start with Accurate Descriptions: Questionnaire Items as Personality Markers for More Accurate Predictions: Items as personality markers. *European Journal of Personality*, *32*(3), 186–201. http://doi.org/10.1002/per.2147

Semenyna, S. W. (2016). The Relationship between Adult Occupational Preferences and Childhood Gender Nonconformity among Samoan Women, Men, and Fa'afafine. *Human Nature*, *27*(3), 13. http://doi.org/10.1007/s12110-016-9258-7

Shapka, J. D., Domene, J. F., & Keating, D. P. (2006). Trajectories of career aspirations through adolescence and young adulthood: Early math achievement as a critical filter. *Educational Research and Evaluation*, *12*(4), 347–358. http://doi.org/10.1080/13803610600765752

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, *75*(3), 417–453. http://doi.org/10.3102/00346543075003417

Smith, J. K. (1983). Quantitative versus qualitative research: An attempt to clarify the issue. *Educational Researcher*, *12*(3), 6–13. http://doi.org/10.3102/0013189x012003006

Spearman, C. (1904). "General Intelligence" objectively determined and measured, *15*(2), 201–292. http://doi.org/10.2307/1412107

Spinath, B., Eckert, C., & Steinmayr, R. (2014). Gender differences in school success: What are the roles of students' intelligence, personality and motivation? *Educational Research*, *56*(2), 230–243. http://doi.org/10.1080/00131881.2014.898917

Spinath, B., Harald Freudenthaler, H., & Neubauer, A. C. (2010). Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation. *Personality and Individual Differences*, *48*(4), 481–486. http://doi.org/10.1016/j.paid.2009.11.028

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, *34*(4), 363–374. http://doi.org/10.1016/j.intell.2005.11.004

Spinath, F. M., Spinath, B., & Plomin, R. (2008). The nature and nurture of intelligence and motivation in the origins of sex differences in elementary school achievement. *European Journal of Personality*, *22*(3), 211–229. http://doi.org/10.1002/per.677

Steinmayr, R., & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality*, *22*(3), 185–209. http://doi.org/10.1002/per.676

Stewart-Williams, S., & Halsey, L. G. (2021). Men, women and STEM: Why the differences and what should be done? *European Journal of Personality*, *35*(1), 3–39. http://doi.org/10.1177/0890207020962326

Stipek, D., & Iver, D. M. (1989). Developmental Change in Children's Assessment of Intellectual Competence. *Child Development*, *60*(3), 521–538.

http://doi.org//10.2307/1130719

Stoet, G., & Geary, D. C. (2013). Sex Differences in Mathematics and Reading Achievement Are Inversely Related: Within- and Across-Nation Assessment of 10 Years of PISA Data. *PLoS ONE*, *8*(3), e57988. http://doi.org/10.1371/journal.pone.0057988

Stoet, G., & Geary, D. C. (2018). The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science*, *29*(4), 581–593. http://doi.org/10.1177/0956797617741719

Stoet, G., & Geary, D. C. (2020). Gender differences in the pathways to higher education. *Proceedings of the National Academy of Sciences*, *117*(25), 14073–14076. http://doi.org/10.1073/pnas.2002861117

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*(6), 859–884. http://doi.org/10.1037/a0017364

Terman, L. M. (1916). The uses of intelligence tests. In *The measurement of intelligence.* (pp. 3–21). Houghton, Mifflin; Company.

Thomas, D. R., & Zumbo, B. D. (2012). Difference Scores From the Point of View of Reliability and Repeated-Measures ANOVA: In Defense of Difference Scores for Data Analysis. *Educational and Psychological Measurement*, *72*(1), 37–43. http://doi.org/10.1177/0013164411409929

Thorndike, R. L. (1987). Stability of factor loadings. *Personality and Individual Differences*, *8*(4), 585–586. http://doi.org/10.1016/0191-8869(87)90224-8

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, *33*(204), 753. http://doi.org/10.2307/2279078

Tupes, E. C., & Christal, R. E. (1961). *Recurrent Personality Factors Based on Trait Ratings* (No. 61-97) (pp. 225–251). United States Air Force.

Turkheimer, E., & Harden, K. P. (2013). Behavior Genetic Research Methods. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (Second, pp. 159–187). New York: Cambridge University

Press. http://doi.org/10.1017/CBO9780511996481.012

UNESCO. (1950). The Race question. In *UNESCO and its programme* (Vol. 3, pp. 30–39). Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000128291

Van Der Maas, H., Kan, K.-J., Marsman, M., & Stevenson, C. E. (2017). Network Models for Cognitive Development and Intelligence. *Journal of Intelligence*, *5*(16). http://doi.org/10.3390/jintelligence5020016

Vedel, A. (2016). Big Five personality group differences across academic majors: A systematic review. *Personality and Individual Differences*, *92*, 1–10. http://doi.org/10.1016/j.paid.2015.12.011

Vernon, P. E. (1947). Research on personnel selection in the Royal Navy and the British Army. *American Psychologist*, *2*(2), 35–51. http://doi.org/10.1037/h0056920

von Stumm, S. (2017). Socioeconomic status amplifies the achievement gap throughout compulsory education independent of intelligence. *Intelligence*, *60*, 57–62. http://doi.org/10.1016/j.intell.2016.11.006

Wach, F.-S., Spengler, M., Gottschling, J., & Spinath, F. M. (2015). Sex differences in secondary school achievement The contribution of self-perceived abilities and fear of failure. *Learning and Instruction*, *36*, 104–112. http://doi.org/10.1016/j.learninstruc.2015.01.005

Watt, H. M. G., Shapka, J. D., Morris, Z. A., Durik, A. M., Keating, D. P., & Eccles, J. S. (2012). Gendered motivational processes affecting high school mathematics participation, educational aspirations, and career plans: A comparison of samples from Australia, Canada, and the United States. *Developmental Psychology*, *48*(6), 1594–1611. http://doi.org/10.1037/a0027838

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*(3), 461–481. http://doi.org/10.1037/0033-2909.91.3.461

Wilk, S. L., Desmarais, L. B., & Sackett, P. R. (1995). Gravitation to Jobs Commensurate With Ability: Longitudinal and Cross-Sectional Tests. *American Psycho-*

logical Association, *80*(1), 79–85. http://doi.org/10.1037/0021-9010.80.1.79

Willett, J. B. (1988). Chapter 9: Questions and Answers in the Measurement of Change. *Review of Research in Education*, *15*(1), 345–422. http://doi.org/10.3102/0091732X015001345

Woods, S. A., & Hampson, S. E. (2010). Predicting adult occupational environments from gender and childhood personality traits. *Journal of Applied Psychology*, *95*(6), 1045–1057. http://doi.org/10.1037/a0020600

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from http://yihui.org/knitr/

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. http://doi.org/10.1177/1745691617693393

Yentes, R. D., & Wilhelm, F. (2018). Careless: Procedures for computing indices of careless responding.

Zheng, L., Goldberg, L. R., Zheng, Y., Zhao, Y., Tang, Y., & Liu, L. (2008). Reliability and concurrent validation of the IPIP Big-Five factor markers in China: Consistencies in factor structure between Internet-obtained heterosexual and homosexual samples. *Personality and Individual Differences*, *45*(7), 649–654. http://doi.org/10.1016/j.paid.2008.07.009

Zheng, L., Lippa, R. A., & Zheng, Y. (2011). Sex and sexual orientation differences in personality in china. *Archives of Sexual Behavior*, *40*(3), 533–541. http://doi.org/10.1007/s10508-010-9700-6