

The Scales of General Well-Being (SGWB)

By

Ylenio Longo

Thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy in Applied Psychology

University of Nottingham, Division of Psychiatry and Applied Psychology
March 2017

Abstract

Well-being indicators are often conceptualized either as positive feeling (e.g. happiness), positive functioning (e.g. meaning, involvement, competence), or a combination of the two. Several measures of well-being have been developed based on this distinction, with lower-order factors (e.g. happiness, meaning) loading on separate higher-order factors (e.g. positive feeling and positive functioning). However, some limitations have been identified in the lower-order factor structure of current well-being questionnaires. Furthermore, recent evidence indicates that a single higher-order or general factor adequately explains most of the variation in well-being data. This dissertation presents the development and initial validation of a questionnaire, which attempts to overcome some limitations in the lower-order structure of previous questionnaires and assumes the presence of a single general factor of well-being. Following a review of current measures, commonalities among them were identified and, based on these points of agreement, the general factor was conceptualized as a set of subjective experiences and evaluations, which are relatively stable and interpreted as symptoms of good mental health. Fourteen constructs were identified as common lower-order indicators of well-being across measures: happiness, vitality, calmness, optimism, involvement, self-awareness, self-acceptance, self-worth, competence, development, purpose, significance, self-congruence and connection. The hypothesized properties of the higher- and lower-order factors were then tested in three studies. Study 1 developed an item pool and assessed the adequacy of its content by consulting a panel of six academics with expertise in well-being research. Studies 2 and 3 recruited two adult North American samples (total $N = 1,496$) to assess the questionnaire's psychometric properties, and specifically its dimensionality, reliability, measurement invariance, and relationships with external criteria. Each lower-order scale fit a unidimensional model and all fourteen scales fit a correlated-factors model as well as higher-order and bifactor models. Evidence of measurement invariance was found across gender, age and a longitudinal period of 5 weeks. Lower- and higher-order scale scores showed adequate internal and test-retest reliabilities. Finally, lower- and higher-order scale scores showed strong associations with previous measures of similar constructs and lower associations with measures of distinct constructs. The questionnaire aims to contribute to future research both as a global measure of well-being as well as a collection of fourteen individual health-related scales.

Keywords: scale development, scale validation, bifactor model, higher-order model, well-being, measurement, quality of life

Acknowledgements

Special thanks go to Iain Coyne (now at Loughborough University), not only for his encouragement and clear guidance, but also for allowing me to think for myself while solving some of the challenges along the way. He is a pleasure to work with and this thesis would not have been the same quality without his input.

I am also grateful for the support of other members of the supervision team who have contributed to the PhD process at different stages. Specifically, Stephen Joseph, Nick Manning and Hugh Middleton, with whom I shared several interesting conversations during the early stages of the PhD when I was based in the School of Sociology and Social Policy, and Nigel Hunt who provided support and career advice particularly through the final stages of the PhD.

I would also like to acknowledge the help of academics who were not directly involved in this project, but who nevertheless provided helpful input: Stavroula Leka, Guy Curtis, Tom Farsides, Petter Gustavsson, Veljko Jovanović, Ralph van den Bosch, Alexander Gunz, and Jessica Morgan.

The Economic and Social Research Council (ESRC) funded the research in this thesis (grant ES/J500100/1) and provided many additional resources, without which I would not have had the opportunity to spend three years researching such a fascinating topic. Similarly, I also want to thank all the people who took the time to participate in the studies.

Most importantly, I want to thank my mother, who worked hard so that I could have an education.

Table of contents

Abstract	i
Acknowledgements	ii
Table of contents	iii
List of tables	vii
List of figures	vii
List of Abbreviations and Symbols	viii
 Chapter 1: Introduction	 1
1.1. The importance of mental well-being	1
1.2. Hedonia and Eudaimonia	2
1.3. Current models	4
1.4. Structure of the thesis	7
 Chapter 2: The hierarchical structure of well-being	 9
2.1. The general factor of well-being	9
2.2. A conceptualization of the general factor	11
2.3. The lower-order factors in previous measures of well-being	15
2.4. The lower-order factors in the SGWB	20
2.4.1. Happiness, vitality and calmness	21
2.4.2. Optimism	21
2.4.3. Involvement	22
2.4.4. Self-awareness	22
2.4.5. Self-acceptance	22
2.4.6. Self-worth	23
2.4.7. Competence	23
2.4.8. Development	23
2.4.9. Purpose	23
2.4.10. Significance	24
2.4.11. Self-congruence	24
2.4.12. Connection	24
2.5. Summary	24

Chapter 3: Method.....	26
3.1. Literature review	26
3.2. Defining the construct.....	27
3.3. Level of measurement.....	28
3.4. Item generation	31
3.5. Principles of item writing	35
3.6. Descriptive statistics	38
3.7. Dimensionality: Factor analysis.....	40
3.7.1. Introduction to factor analysis.....	40
3.7.2. Exploratory and confirmatory factor analysis.....	43
3.7.3. A side note on communalities and uniquenesses	44
3.7.4. Model building and identification.....	45
3.7.5. Estimation	47
3.7.6. Model fit.....	49
3.7.7. Model respecification.....	53
3.7.8. Measurement invariance	55
3.7.9. Summary	56
3.8. Reliability.....	57
3.8.1. Internal consistency.....	57
3.8.2. Temporal stability	63
3.8.3. Parallel form.....	63
3.9. Construct validity.....	63
3.9.1. Content evidence	64
3.9.2. Criterion evidence	66
3.10. The relationship among dimensionality, reliability and validity	69
3.11. Sample size requirements	70
3.11.1. Simulation studies	71
3.11.2. Power analysis.....	73
3.11.3. Summary	74
3.12. Online sample recruitment and representativeness.....	75
3.13. Dealing with missing data.....	78
3.14. Summary	82
Chapter 4: Study 1 – Item development and content validation	83
4.1. Item development	83
4.2. Content validation.....	84

Chapter 5: Study 2 – Item selection and scale dimensionality.....	87
5.1. Method	87
5.1.1. Sample.....	87
5.1.2. Procedure.....	88
5.2. Results.....	88
5.2.1. Preliminary analyses	88
5.2.2. Dimensionality of each scale	89
5.2.3. Dimensionality of the entire questionnaire	91
5.3. Discussion	95
 Chapter 6: Study 3 – Further dimensionality and criterion validity testing.....	97
6.1. Method	98
6.1.1. Design	98
6.1.2. Sample and procedure	98
6.1.3. Measures	99
6.2. Results.....	101
6.2.1. Preliminary analyses	101
6.2.2. Dimensionality of each scale	102
6.2.3. Dimensionality of the entire questionnaire	103
6.2.4. Invariance across gender and age.....	104
6.2.5. Longitudinal invariance and test-retest reliability.....	107
6.2.6. Criterion validation	108
6.2.7. Construct value.....	111
6.3. Discussion.....	112
 Chapter 7: General discussion.....	114
7.1. Findings	114
7.2. Implications	117
7.2.1. A new measure of well-being	117
7.2.2. From a psychometric view, is there a general factor of well-being?	118
7.2.3. Which hierarchical model is more suited to study SGWB scores?... ..	119
7.3. Limitations and future directions	121
7.3.1. Subjective choice of models and constructs.....	121
7.3.2. The use of higher- and lower-order factor scores	122
7.3.3. The structure of well-being across cultures	123
7.3.4. Self-report measures.....	123
7.3.5. Cross-sectional design.....	125
7.3.6. Short form	125
7.4. Conclusion	125

References	127
Appendices.....	159
Appendix 1	159
Conceptual development of the 14 scales of GWB	159
Appendix 2	163
Past measures consulted during item development (study 1)	163
Appendix 3	165
Item pool submitted for content validation (Study 1)	165
Appendix 4	170
Content validation (Study 1)	170
Appendix 5	184
Item selection in individual CFAs (Study 2)	184
Appendix 6	195
Final scales.....	195
Appendix 7	198
Item loadings (Study 2).....	198
Appendix 8	199
Higher-order reliability analysis (Study 2)	199
Appendix 9	200
Survey information (Study 3)	200
Appendix 10	203
Item loadings (Study 3).....	203
Appendix 11	204
Higher-order reliability analysis (Study 3)	204
Appendix 12	205
Longitudinal invariance – Time 1 and 2 model fit.....	205

List of tables

Table 1 Types of measurements.....	30
Table 2 An example of items and content areas for competence.....	35
Table 3 Example of a reliability analysis	62
Table 4 Hypothesized convergent and discriminant correlations	67
Table 5 Number of items per scale at different stages of scale development	86
Table 6 Fit indices of individual scales after item selection (Study 2)	91
Table 7 Fit indices of the entire questionnaire (Study 2)	92
Table 8 Reliability and discriminant validity analysis (Study 2).....	93
Table 9 Competing models for self-worth and self-acceptance (Study 2).....	93
Table 10 Fit indices of individual scales (Study 3).....	103
Table 11 Fit indices of the entire questionnaire (Study 3)	103
Table 12 Reliability and discriminant validity analysis (Study 3).....	104
Table 13 CFA in each gender and age group (higher-order model)	105
Table 14 Multigroup CFA for gender and age groups (higher-order model)	107
Table 15 Test-retest reliability and longitudinal invariance	108
Table 16 Correlations between each WB scale and external criteria.....	110
Table 17 Correlations between well-being and ill-being scales.....	111
Table 18 Perceived value of each well-being construct.....	112

List of figures

Figure 1 Examples of hierarchical models.....	9
Figure 2 A one-factor model.....	42

List of Abbreviations and Symbols

α	Alpha (i.e. either reliability coefficient or type 1 error)
AVE	Average variance explained by a factor
β	Beta (i.e. either regression coefficient or type 2 error)
BIC	Bayesian Information Criterion
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
Δ	Delta (i.e. difference)
df	Degree of freedom
DWLS	Diagonally Weighted Least Squares
ECV	Explained common variance
EFA	Exploratory Factor Analysis
FIML	Full Information Maximum Likelihood
g	General factor
h^2	Communality
HTMT	Heterotrait-monotrait method
λ	lambda (i.e. factor loading)
ML	Maximum Likelihood
MLM	Robust Maximum Likelihood with Satorra-Bentler scaled chi-square and robust standard errors
MLR	Robust Maximum Likelihood with Yuan-Bentler chi-square and Huber-White robust standard errors
MVS	Maximum variance shared with other factors
N	Sample size
ϕ	Phi (factor correlations)
PUC	Percentage of uncontaminated correlations
PWB	Psychological well-being
R^2	Squared multiple correlation
RMSEA	Root Mean Square Error of Approximation
SRMR	Standardized Root Mean Square Residual
SGWB	Scales of General Well-Being
Σ	Sigma (i.e. sum)
SWB	Subjective well-being
ULSMV	Unweighted Least Squares Means and Variances
ULS	Unweighted Least Squares
ω	Omega (reliability index)
WLS	Weighted Least Squares
WLSMV	Weighted Least Squares Means and Variances
χ^2	Chi-square (fit index)

Chapter 1: Introduction

1.1. The importance of mental well-being

The study of mental well-being can have important consequences for society. A notable example of its value becomes apparent when observing the cost of low levels of well-being. Well-being is highly negatively related to the incidence of psychological health problems, such as stress, depression and anxiety. As one's levels of well-being reduce, it becomes over six times more likely to experience stress, depression and anxiety (Keyes, Dhingra & Simoes, 2010). These mental health problems cost the UK economy 70-100 billion GBP per year, amounting to 4.5% of its Gross Domestic Product (Davies, 2013). One way in which psychological health problems amount to such a high economic cost is through the loss of around 10 million working days per year (Health and Safety Executive, 2015). On the other hand, people who maintain a high level of well-being not only report the fewest instances of mental illness, but also the fewest missed days of work (Keyes, 2005).

Beyond these economic considerations, well-being is considered valuable by the populations of several countries (Diener, 2000). Therefore, it is worth researching not only because of its outcomes, but also because of its intrinsic importance to people's lives.

It has also been argued that measures of mental well-being can be used to assess individual and societal quality of life beyond current indicators, such as Gross Domestic Product (Layard, 2010). Therefore, the effective measurement of well-being can have a positive impact, not only on our understanding of the topic, but also by informing public policy, which can potentially affect the lives of entire populations.

This thesis presents the development and initial validation of a new measure of mental well-being: the Scales of General Well-Being (SGWB). It will review the current conceptualization and measurement of well-being, identify some of its limitations, and attempt to overcome these by developing and validating a new measure of well-being. Consistent with best practices in scale development (e.g. DeVellis, 2012), this chapter provides an overview of the historical background underlying some of the main current definitions of well-being in psychological research. The next chapter will focus on current issues in well-being measurement.

1.2. Hedonia and Eudaimonia

At a conceptual level, mental well-being is often defined as feeling good and functioning well (Huppert & So, 2013; Keyes, 2007; Ryan & Deci, 2001). This distinction is based on two perspectives rooted in ancient Greek philosophy: hedonia and eudaimonia (e.g. Ryan & Deci, 2001; Waterman, 2013). They can be defined, respectively, as the pursuit of pleasure and the pursuit of virtue. Hedonia is an ethical theory that attributes a higher moral significance to pleasant experiences (Waterman, 2013). Among its proponents, the philosopher Aristippus of Cyrene emphasized the importance of pursuing enjoyment, particularly physical pleasure, though without becoming overpowered by it (Diogenes Laërtius, trans. 1972; Tsouna-McKirahan, 2004). On the other hand, Epicurus favored a more balanced approach that prioritizes the avoidance of pain and the pursuit of long-term satisfaction (Epicurus, trans. 1994). In contrast, eudaimonia is an ethical theory that posits that one should pursue virtue, rather than pleasure. Among its proponents, Aristotle (trans. 2009) is perhaps the most influential and, in *Nichomachean Ethics*, he argued that we should pursue moral virtues such as balance, kindness, justice and excellence, and that only constant virtuous action can lead to well-being.

When philosophers, such as Epicurus and Aristotle, talked about hedonia and eudaimonia, their goal was not to define well-being, but rather to offer ethical guidelines for how to live (e.g. Ryff & Singer, 2008). As such, they focused on pursuits and behaviors (e.g. pursuing pleasure or excellence) that would help people live a blessed life. In other words, hedonia and eudaimonia, as described by Epicurus and Aristotle, are not definitions of well-being but rather behaviors and pursuits that should then lead to well-being.

On the other hand, modern researchers in psychology overwhelmingly define well-being as a set of experiences and evaluations. For example, while Diener's subjective well-being (SWB) is often linked to hedonia, his operationalization (or model) involves the *experience* of pleasant affect (e.g. happiness), rather than the *pursuit* of it. Furthermore, SWB is rooted in lay people's conceptions of happiness, rather than hedonic theory (Diener, 1984; Kesebir & Diener, 2008), and the link between hedonia and SWB was only made a posteriori (Kashdan, Biswas-Diener & King, 2008; Ryan & Deci, 2001). Similarly, current models of positive functioning and flourishing tend to define well-being predominantly in terms of subjective

experiences and evaluations, such as involvement and meaning (e.g. Keyes, 2013; Waterman, 2013; see also Kashdan et al., 2008).

Research has also been conducted on the behaviors, pursuits and events that lead to well-being, but these are typically presented as antecedents rather than components of well-being (Huta, 2013b; Huta & Ryan, 2010). For example, among the correlates and predictors of well-being in the literature, we can find behaviors such as volunteer work (McDougle, Handy, Konrath & Walk, 2014), pursuits that are in line with one's values and interests (i.e. self-concordant goals) (Sheldon, 2013). Theoretically, it is possible to conceptualize behaviors and pursuits as important correlates or antecedents, rather than defining elements of well-being, for several reasons. The range of pursuits and behaviors that may influence people's well-being is likely to be broad and vary from person to person (Waterman, 2008). According to Beck's cognitive model, behaviors and subsequent events do not define people's subjective experience. Rather, the same behavior or event is likely to produce different subjective evaluations and experiences in different people (e.g. Beck, 2011; Trower, Jones, Dryden & Casey, 2011). Furthermore, it is sound psychometric practice not to combine environmental conditions and behaviors with their experiential outcomes (De Vaus, 2002; Spector, 2013). While it is important to distinguish between the correlates of well-being and well-being itself, research on these correlates is of utmost importance, as they can be targeted in interventions that aim to improve people's psychological health (e.g. Huta, 2013b). In short, while the terms *hedonia* and *eudaimonia* describe pursuits and behaviors that lead to well-being, current conceptualizations and measures of well-being tend to focus on subjective experiences and evaluations.

An interesting difference can be found between the behaviors that lead to well-being and the subjective experiences that define well-being. Hedonic and eudaimonic behaviors, such as the pursuit of pleasant or productive activities, are weakly correlated. In several studies, the association between the two types of behaviors ranged approximately from $-.30$ (over a short timeframe) to $.30$ (over a long timeframe) (Huta, 2013b). That is, at any given moment, a person was likely to engage either in a hedonic or a eudaimonic activity, but not both. However, over several weeks, people who engaged in many eudaimonic pursuits were also somewhat more likely to engage in hedonic pursuits.

While hedonic and eudaimonic behaviors are weakly related, subjective experiences categorized as positive feelings and positive functioning are very strongly related. For example, the correlation between “hedonic” enjoyment and personal expressiveness (an index of positive functioning) has been found to range between .71 and .87 (Waterman, 1993; Waterman, Schwartz & Conti, 2008). Similarly, correlations between a latent SWB and a psychological well-being (PWB) factor ranged between .76 and .84 (Linley, Maltby, Wood, Osborne & Hurling, 2009; Keyes, Shmotkin & Ryff, 2002).

Like current well-being conceptualizations and measures, this thesis will focus not on ethical theories specifying pursuits or behaviors that may lead to well-being, but rather on subjective experiences and evaluations that define well-being.

If feeling good and functioning well are not independent constructs, like hedonic and eudaimonic behaviors, and correlate as high as .80, it may be useful to test whether positive feeling and positive functioning components can be explained by a single higher-order factor, rather than two separate but very strongly overlapping higher-order factors (Brown, 2015). Whenever two models of different complexity fit the data adequately, it is often appropriate to retain the simpler and more parsimonious of the two, even though its fit to data is often slightly lower (e.g. Bollen, 1989; Brown, 2015; Kim & Mueller, 1978). This possibility is further explored in the next chapter.

1.3. Current models

Over the last 30 years, the conceptualization and measurement of well-being has become increasingly more complex with new measurement models often integrating constructs from previous ones. During this time, SWB has been the most practically used measurement model of well-being in psychology (David, Boniwell & Ayers, 2013; Ryan & Deci, 2001). SWB focuses on positive feelings. It integrates earlier work on positive and negative affect and life satisfaction (e.g. Andrews & Withey, 1976; Bradburn, 1969; Campbell, Converse & Rodgers, 1976), and operationalizes well-being as frequent positive affect, infrequent negative affect, and a positive evaluation of one’s life (Diener, 1984; Diener, Colvin, Pavot & Allman, 1991). Despite its popularity as a focus for research, SWB has been criticized for not having a strong theoretical basis and for focusing too much on feelings of happiness,

thus neglecting important aspects of positive psychological functioning (e.g. Ryff, 1989).

To address this gap, Ryff (1989) synthesized earlier theories on personal growth (e.g. Rogers, 1961; Maslow, 1968; Jung, 1933; Allport, 1961), optimal lifespan development (e.g. Erikson, 1959; Buhler, 1935; Neugarten, 1973), and positive mental health (Jahoda, 1958), to create a six-factor model of PWB. This includes components such as self-acceptance, positive relations and purpose in life. Since then, Ryff's PWB has become one of the most widely utilized models of positive psychological functioning.

With the advent of the positive psychology movement in 1998, research interest on wellness soared. A PsycINFO search revealed that entries for the term *well-being* increased nearly six times during the following 14 years (29,879), compared to the 14 years preceding the advent of the movement (5,174). During that time, new models of positive functioning were introduced. For example, Ryan, Huta and Deci (2008) introduced a model of positive functioning based on their previous work on self-determination theory (e.g. Deci & Ryan, 2000) and mindfulness (Brown & Ryan, 2003). Waterman (2013; Waterman et al., 2010) presented a model based on eudaimonic philosophy and influenced by the work of philosophers such as Kraut (1979) and Norton (1976).

Some researchers have also integrated SWB and positive functioning components into more comprehensive models of well-being (e.g. Huppert & So, 2013; Keyes, 2002; Seligman, 2011). These integrative models have often used the term "flourishing" to describe well-being in a broader sense. The components included in these models often overlap, and differences among them reflect the variety of development approaches chosen by these authors. Specifically, Huppert and So's (2013) model of flourishing was developed based on symptoms of psychological ill-health; Keyes' (2002) model was based on Diener's SWB, Ryff's PWB and his own model of social well-being (Keyes, 1998); finally, Seligman's (2011) model was an expansion and revision of his previous work on authentic happiness (Seligman, 2002).

The models discussed here have been included in major recent reviews of the field (Huta, 2013a; Kashdan et al., 2008; Keyes, 2015; Ryan & Deci, 2001), and have generated large amounts of research. For example, an online search on Web of Science and Google Scholar shows that Keyes, Seligman and Ryan et al.'s

publications on well-being have been cited hundreds or thousands of times. Similarly, the previous models, on which Keyes, Seligman and Ryan et al.'s papers are based, have also been highly cited (e.g. Diener, 1984; Keyes, 1998; Peterson, Park, Seligman, 2005; Ryan & Deci, 2000; Ryff, 1989). The more recent models cannot obviously meet the same citation standards of the previous ones. Nevertheless, Diener et al. (2010) and Huppert and So's (2013) are among the top cited papers in the well-being journal *Social Indicators Research* since 2010, and Waterman et al.'s (2010) paper is among the top cited in *The Journal of Positive Psychology*.

Different scholars present a specific set of constructs to measure well-being, based on their research background as well as their perspective regarding what constitutes well-being. For example, Ryff (1989) chose to base her well-being model on previous work by Maslow, Jung, Neugarten, Jahoda and others. Similarly, Huppert and So (2013) chose to develop their model by identifying constructs that were conceptually opposite to symptoms of psychological ill-health, and provide a rationale for why that may be the case. Consequently, we can find a variety of perspectives about what constitutes well-being, which may be influenced by the authors values as well as their cultural and academic background.

While there is still disagreement regarding which model better captures well-being (e.g. David et al., 2013; Kashdan et al., 2008), a strong overlap can be found among these. Despite the variety of perspectives that influenced the models presented here, psychologists agree that mental health goes beyond the absence of mental illness, and that well-being can be measured through several lower-order constructs (e.g. Diener, 1984; Huppert & So, 2013; Keyes, 2007; Ryan & Deci, 2001; Ryff & Keyes, 1995; Seligman & Csikszentmihalyi, 2000). Furthermore, the constructs included in different models of well-being significantly overlap. The thesis will focus on this overlap to develop an integrative conceptual definition and measure of well-being.

As will be discussed in section 3.1, information regarding the well-being construct could be acquired qualitatively through interviews and focus groups. This is normally carried out in the initial steps of concept development, when a construct is less known and its development is in its infancy. On the other hand, when sufficient information about the construct can be found in the current literature, we may rely on this to conceptualize our construct. The assumption underlying the latter

approach is that different sources in the literature may closely capture the construct under focus. A fundamental assumption underlying the thesis is that different models complement, rather than compete with, each other.¹ They provide a variety of useful perspectives and, when these are considered in unison, a more comprehensive picture of well-being indicators emerges.

1.4. Structure of the thesis

Chapter 2 provides a psychometric review of previous well-being scales to facilitate the development of a new questionnaire. It evaluates well-being measurement at different hierarchical levels. Following this review, a new conceptualization of general and specific factors of well-being is presented and it is shown how these constructs are conceptually related to those presented in previous conceptualizations.

Chapter 3 outlines the methods adopted in the thesis and focuses on best practices in scale development. Specifically, it covers the following topics: scale review, concept development, item generation, factor and reliability analysis, measurement invariance and construct validity. These topics and practices inform the entire research endeavor of the thesis.

Chapter 4 presents the development of the item pool and the assessment of its content validity by six academics with expertise in well-being research. Following their feedback, the item pool is further revised (Study 1).

Chapter 5 covers the factor analysis of the item pool in a North American sample ($N = 507$). Since the questionnaire is designed to measure fourteen specific constructs as well as mental well-being as a whole, factor analyses test each lower-order scale, all fourteen scales as correlated factors, as well as higher-order and bifactor models (Study 2).

Chapter 6 outlines further analyses in a new North American sample ($N = 989$). The factor structure from Study 2 is replicated in this new sample. Furthermore, measurement invariance is tested across gender, age and a longitudinal period of 5 weeks. Finally, criterion validity is assessed by correlating lower- and higher-order scale scores with measures of conceptually similar and different constructs (Study 3).

¹ This is also acknowledged as a potential limitation in section 7.3.1.

Chapter 7 concludes the thesis by discussing and critically evaluating the findings and their relevance in the wider literature. It also covers the limitations of the studies and areas for future research.

Several appendices are included to provide further information on the analyses.

Chapter 2: The hierarchical structure of well-being

2.1. The general factor of well-being

Well-being indicators are often conceptualized as either positive feeling (e.g. happiness) or positive functioning (e.g. meaning, involvement, competence), and many of the well-being questionnaires and measurement models used in current research conform to this distinction (Kashdan et al., 2008).

Recently, however, several studies have shown that elements of positive feeling and positive functioning are highly correlated. For example, the two latent higher-order factors correlated at .76 in a British sample ($N = 2,593$) (Linley et al., 2009), .84 in an American sample ($N = 3,032$) (Keyes et al., 2002), and .96 in a large international sample ($N = 7,617$) (Disabato, Goodman, Kashdan, Short & Jarden, 2016). In the presence of such high correlations between positive feeling and positive functioning, it may not be necessary to conceptualize well-being as two separate factors, but as a single general factor (e.g. Kashdan et al., 2008; Disabato et al., 2016).

The adequacy of a single general factor of well-being has been tested using two hierarchical models: higher-order and bifactor models (see Figure 1). In higher-order models, the relationships among well-being items are explained by lower-order factors (e.g. happiness, meaning, involvement, competence, etc.) and, in turn, the relationships among lower-order factors are explained by a general factor (also labeled higher-order factor). On the other hand, bifactor models allow well-being items to load both on a general well-being factor as well as specific (or group) factors.

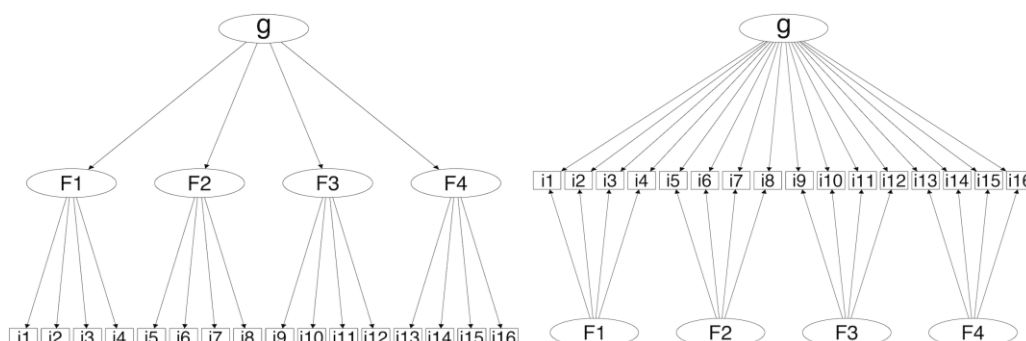


Figure 1 Examples of hierarchical models

Note: a higher-order model (left) and a bifactor model (right). In this higher-order model, lower- and higher-order factors can also be labeled first and second-order factors, respectively.

Both models assume that items are influenced by a general factor (g) as well as specific factors. If the main purpose of the model is simply to estimate g , then it makes little difference if this is estimated with a higher-order or a bifactor model, as the two general factors would be almost perfectly correlated (e.g. Jensen & Weng, 1994; Murray & Johnson, 2013). However, higher-order and bifactor models differ in several ways (Beaujean, 2015). First, they differ with regards to the relationships among factors. In higher-order models, g has a direct effect on lower-order factors. Therefore, changes in g are expected to produce changes in the lower-order factors. Conversely, in bifactor models, g and the specific factors are uncorrelated (e.g. Holzinger & Swineford, 1937). Changes in g should have no impact on the specific factors and vice versa. Second, in higher-order models g has an indirect causal effect on items through the lower-order factors while, in bifactor models, all factors have a direct causal effect on the items. Third, through bifactor models, we can explore the redundancy and predictive ability of specific factors after controlling for the general factor. For example, after including g in a bifactor model, the loadings on a specific factor may become nonsignificant, indicating that that specific factor is an artifact, as it explains no variance over and above g . However, it is not possible to test this in higher-order models. Fourth, bifactor models include more parameters. In other words, they are less parsimonious and because of this they tend to show better model fit even when the data has a higher-order structure (Murray & Johnson, 2013). Historically, hierarchical models have been used to study human intelligence (e.g. Jensen & Weng, 1994). However, recently, they have been shown to be useful in modeling well-being data (e.g. Chen, West & Sousa, 2006; Gallagher, Lopez & Preacher, 2009).

For example, Huppert and So (2013) proposed a measurement model of well-being with ten single-item indicators loading on two factors. However, a recent study showed that the two factors were highly correlated ($r = .76$), and after including a general factor, thus specifying a bifactor model, the two specific factors were unreliable (Longo, Coyne, Joseph & Gustavsson, 2016).

Similarly, Keyes' (2002) proposed a model with fourteen lower-order factors loading on three higher-order factors: positive feeling, positive functioning and social well-being. However, after testing this model in a student sample ($N = 591$) and a large representative U.S. sample ($N = 4,032$), a single higher-order factor was found to adequately explain the variance in the lower-order factors (Gallagher et al.,

2009). Similar results were found when focusing only on positive feeling and positive functioning components in a large international sample ($N = 7,617$): the two higher-order factors had a strong latent correlation ($r = .96$), exhibited similar correlations with external criteria, their lower-order factors were adequately explained by a single higher-order factor, and these results were consistent across 7 geographical world regions (Disabato et al., 2016). Similar results were also found with Seligman's (2011) conceptualization of well-being. Two studies supported a higher-order model in which the lower-order factors loaded on a single higher-order well-being factor (Coffey, Wray-Lake, Mashek & Branand, 2016).

Additional studies showed support for a general factor of well-being using a bifactor model (Chen, Jing, Hayes & Lee, 2013; de Bruin & du Plessis, 2015; Jovanović, 2015). In the bifactor model, well-being items showed stronger loadings on the general factor compared to their specific factors, indicating that well-being items reflect primarily a single general factor of well-being. Consistent with the aforementioned simulation (Murray & Johnson, 2013), when both higher-order and bifactor models were tested, the higher-order model did not fit as well as the bifactor model (Chen et al., 2013; Jovanović, 2015).

Taken together, these results suggest that well-being data has a hierarchical structure, where items are influenced by a specific factor, as well as a single general factor of well-being, whether directly (in bifactor models) or indirectly (in higher-order models). In the present studies, both higher-order and bifactor models will be empirically tested. However, for simplicity and consistency, throughout the following sections, the higher-order model will be used as the hypothesized structure of well-being.

2.2. A conceptualization of the general factor

The literature has often defined at least two general factors of well-being, labeled positive feeling and positive functioning, or hedonia and eudaimonia, or pleasure and personal fulfillment, rather than a single general factor. Few researchers have attempted to conceptualize a single general factor. This section will draw from these efforts and will further attempt to conceptualize the general factor by identifying commonalities among positive feeling and positive functioning constructs.

Positive feeling is conceptualized as pleasant experience and positive evaluation of life and it is most often measured using Diener's (1984) tripartite measurement model of SWB, which includes three factors: life satisfaction, positive affect and negative affect.

There is less agreement regarding the conceptualization and measurement of positive functioning (Huta, 2013a). Nevertheless, this area of well-being research tends to conceptualize well-being as a set of subjective experiences, and few behaviors, which relate to personal fulfillment. Positive functioning is often measured by assessing people's perceptions of their purpose in life, competence, positive relationships, as well as behaviors and goals that are consistent with one's self-perception (e.g. Ryff, 1989; Waterman et al., 2010).

When questionnaire items load on a single factor, commonalities among the items are identified to label and define that factor (Kline, 2000). Similarly, if well-being data fit hierarchical models with a single general factor, it is necessary to develop a conceptual notion of well-being as a unidimensional hierarchical construct, by identifying commonalities among the lower-order factors of well-being.

The models considered here generally focus on well-being in one's life as a whole, rather than in specific domains. Of these models, four include elements of positive feeling and positive functioning (Diener et al., 2010; Huppert & So, 2013; Keyes, 2002; Seligman, 2011), while two focus on elements of positive functioning, or eudemonic well-being (Ryan et al., 2008; Waterman et al., 2010), which is a highly debated focus of research, thus deserving additional attention (Huta, 2013a). These operationalizations have been included in recent reviews of the field (Huta, 2013a; Kashdan et al., 2008; Keyes, 2015; Ryan & Deci, 2001), and have generated large amounts of research.

A total of 49 well-being constructs were identified in the six models, most of which share several common properties.

- Diener et al. (2010): Purpose and meaning, supportive relationships, engagement, contribution, competence, positive self-perception, optimism, feeling respected.
- Huppert and So (2013): Emotional stability, happiness, vitality, resilience, optimism, self-esteem, engagement, meaning, positive relationships and competence.

- Keyes (2002): Positive affect, low negative affect, life satisfaction (Diener, 1984); self-acceptance, positive relations, autonomy, mastery, purpose in life and personal growth (Ryff, 1989); social integration, social acceptance, social contribution, social actualization, social coherence (Keyes, 1998).
- Seligman (2011): Positive emotion, engagement, relationships, meaning and accomplishment.
- Ryan et al. (2008): Intrinsic goals, autonomous behavior, mindfulness, satisfying needs for autonomy, competence and relatedness.
- Waterman et al. (2010): Self-discovery, perceived development, purpose, effort, involvement and personal expressiveness.

As noted by Kashdan et al. (2008), with few exceptions, these constructs consist of the same experiential phenomena: subjective feelings and evaluations. Consistent with this notion, general well-being has been recently defined as “the experience of life going well” (Huppert & So, 2013, p. 838; see also Keyes, 2013). Only 4 of the 49 constructs did not share this commonality. Autonomous behavior, intrinsic goals, effort, and emotional resilience were not defined as subjective experiences or evaluations. Instead, intrinsic goals are defined as pursuits (Ryan et al., 2008); autonomous behavior (Ryan et al., 2008) and the “investment of effort in the pursuit of excellence” (Waterman et al., 2010, p. 45) are behaviors; and emotional resilience is defined as the “ability to manage anxiety and worry” (Huppert & So, 2013 p. 856). Beyond these exceptions, the first main point of convergence for well-being constructs is their subjective experiential nature.

Each of these subjective experiences is interpreted only as a symptom of good mental health. Well-being is not defined by any single symptom, but rather by a collection of symptoms (e.g. Huppert & So, 2013; Keyes, 2013; Seligman & Csikszentmihalyi, 2000). Just like each item loading on a factor is an imperfect measure of that factor, each well-being symptom, or lower-order factor, is merely an imperfect indicator of general well-being. A high score on a single lower-order factor does not guarantee high well-being. Consequently, it may be overly simplistic to measure well-being by focusing only on a single construct. For example, optimism is often used as an indicator of well-being but, on its own, it may also be related to negative outcomes, such as a lower likelihood to disengage from gambling even after losses (Gibson & Sanbonmatsu, 2004). However, when optimism is considered

together with many other indicators, its shortcomings as an indicator of well-being can be reduced. Therefore, this thesis will focus on multiple symptoms to provide a comprehensive account of well-being.

These symptoms are relatively stable, with test-retest reliabilities approximately ranging from .60 to .85 over a period of 4 to 8 weeks (e.g. Pavot & Diener, 2008; Ryff, 1989; Watson, Clark & Tellegen, 1988). They are not as stable as personality traits, with retest reliabilities often approximating a .70 to .90 range over a period of 4 to 8 weeks (Watson, 2004), or mental ability, with retest reliabilities around .90 over periods under 10 months (Schuerger & Witt, 1989). Nevertheless, around half or more of the variance in well-being tends to be stable over time. Furthermore, symptoms are typically rated over large timespans. For instance, when completing a well-being questionnaire respondents are often asked to focus on their life in general (e.g. Kern, Waters, Adler & White, 2015) or to report their short-term subjective experiences and evaluations over multiple instances (e.g. Kahneman, Krueger, Schkade, Schwarz & Stone, 2004; Csikszentmihalyi & Larson, 1987). In other words, well-being is usually conceptualized and measured at the trait level.

In all but one case, well-being symptoms involve the *presence* of a healthy experience (e.g. positive affect). Only one construct included in one operationalization involves the *absence* of an unhealthy experience: negative affect. Negative affect strongly overlaps with psychopathology constructs, such as depression and anxiety, with correlations as high as .77 (e.g. Brown, Chorpita & Barlow, 1998; see also Joiner, Catanzaro & Laurent, 1996; Larsen & Diener, 1992; Watson & Clark, 1992; Watson et al., 1988). Consistent with the majority of conceptualizations reviewed here, the current conceptualization will focus on the presence of healthy experiences. In other words, from an affective perspective (e.g. Russell, 1980), well-being conceptualizations seem to focus overwhelmingly on positively valenced (i.e. positive or pleasant) constructs.

In short, after identifying commonalities among lower-order factors of well-being, the general factor could be defined as a collection of relatively stable subjective feelings and evaluations, which represent symptoms of good mental health. So far, evidence supporting a single general factor was provided and commonalities among lower-order factors were used to conceptually define the general factor of well-being. The next section aims to identify some of the strengths

and limitations of previous lower-order measures of well-being, which will then help to summarize the constructs presented in previous models into a more parsimonious list of factors to include in the SGWB.

2.3. The lower-order factors in previous measures of well-being

Indeed, no conceptualization or psychometric scale is without limitations. The main purpose of this section is to explore some of the limitations regarding the lower-order factor structure of previous measures of well-being. This discussion will aid in the conceptualization and scale development process for the new measure, and will also further justify the need for a new measure. It is not the purpose of this chapter to provide an exhaustive account of the evidence supporting the validity of these measures. Previous reviews attempting to do so are referenced below (e.g. McDowell, 2010) and it has already been argued that these conceptualizations and measures provide useful information. Instead, limitations are explored in an attempt to avoid replicating these potential shortcomings when developing the new measure.

Keyes's conceptualization and measures of well-being integrate elements from previous models and measures of well-being focusing on positive feeling (Diener, 1984), positive functioning (Ryff, 1989) and social well-being (Keyes, 1998). The overall factor structure proposed by Keyes, with fourteen lower-order factors and three higher-order factors, showed good fit to the data in several cross-cultural studies (e.g. Keyes, 2013). However, as previously mentioned, when hierarchical models with only one general factor were tested, the constructs included in Keyes' conceptualization and measures were adequately explained by a single general factor (Gallagher et al., 2009; Jovanović, 2015). This indicates that it is not necessary to specify three higher-order factors corresponding to positive feeling, positive functioning and social well-being. Since Keyes' measures include elements of positive feeling, positive functioning and social well-being, these are discussed in greater depth below.

As previously mentioned, positive feeling is typically measured through questionnaires assessing affect and life satisfaction. Affect is defined as the most basic feelings that people can consciously access (Russell & Feldman Barrett, 1999). We can distinguish among different affective states based on their valence and arousal (e.g. Russell, 2003; Russell & Feldman Barrett, 1999). Valence distinguishes between pleasant and unpleasant feelings, while arousal refers to a sense of

behavioral activation or energy. Thus, vitality and energy can be categorized as *high*-arousal pleasant affect, happiness and pleasantness as *moderate*-arousal pleasant affect, and calmness and serenity as *low*-arousal pleasant affect (e.g. Diener, 2006; Ekkekakis, 2013; Longo, 2015; Yik, Russell & Steiger, 2011). Positive affect is usually measured with the Positive and Negative Affect Schedule (PANAS, Watson et al., 1988). This has been mainly criticized for focusing on high-arousal feelings (e.g. alertness and excitement) and neglecting low-arousal feelings (e.g. calmness and serenity) (e.g. Ekkekakis, 2013). On the other hand, all three arousal levels of positive affect can be found, for example, in measures used by Keyes et al. (2002), Kern et al. (2015) and Huppert and So (2013). Previous studies have successfully modeled different arousal levels of positive affect as one factor (e.g. Kern et al., 2015), two correlated factors (e.g. Longo, 2015) or three correlated factors (e.g. Tsai, Knutson & Fung, 2006). The potential cost of treating them as one factor is that the final measure may not fit a unidimensional model in some samples. In fact, in a study using an Australian sample, a unidimensional model including more than one positive arousal level was found not to fit the data adequately (Longo, 2015). Therefore, if different arousal levels are assumed to be unidimensional, the final measure may perform poorly and, if not enough items have been selected to measure all three arousal levels of positive affect, it may not be possible to measure these reliably. Therefore, as a conservative choice, it may be preferable to treat them as three factors to ensure that each aspect is adequately measured. If these three factors show poor discriminant validity, there will be no downside to merging them into a unidimensional model (Whetten, 1989).

Life satisfaction has been measured with single-item scales (e.g. Huppert & So, 2013) and multi-item scales such as the satisfaction with life scale (Diener, Emmons, Larsen & Griffin, 1985). This multi-item scale is supported by substantial evidence of dimensionality, reliability and validity (e.g. Pavot & Diener, 2008), and single-item scales have been shown to produce results consistent with multi-item scales (e.g. Cheung & Lucas, 2014). Life satisfaction is perhaps the broadest of all the constructs reviewed here. In fact, it has been used on its own as an indicator of overall well-being (e.g. Kern et al., 2015). Furthermore, it has very strong overlaps with many well-being constructs included here, such as moderate-arousal positive affect, positive relationships, accomplishment and meaning in life (e.g. Kern et al., 2015; Longo, 2015; Proctor, Tweed & Morris, 2015). The strongest overlaps have

been found between life satisfaction and positive affect. For example, in one study moderate-arousal positive affect was found to explain 68% of the variance ($\beta = .82$) in life satisfaction (Longo, 2015). Another study found that life satisfaction is almost perfectly correlated with positive affect ($r = .98$) when the two are measured at the trait level, with lower yet still strong state-level correlations (r range = .53 – .65) (Kaczmarek, Bujacz & Eid, 2015). It should also be noted that moderate-arousal positive affect terms (e.g. pleased, content and happy) are sometimes used in satisfaction scales (e.g. Alfonso, Allison, Rader & Gorman, 1996; Lavalle, Hatch, Michalos & McKinley, 2007; Sorthaix & Lönnqvist, 2014), and satisfaction scales are sometimes used as indicators of positive affect (e.g. Lyubomirsky, King & Diener, 2005). Similarly, composite life domain satisfaction has been strongly linked to positive affect. Specifically, general positive affect was found to explain around 66% of variance ($\beta = .81$) in composite domain satisfaction (e.g. Blore, Stokes, Mellor, Firth & Cummins, 2011). Additionally, some of the domains in a domain-satisfaction scale (e.g. satisfaction with one's relationships) conceptually overlap with constructs included in positive functioning models (e.g. positive relationships) (e.g. Kashdan et al., 2008). Taken together, it appears that life satisfaction is broader than other constructs reviewed here, current measures show adequate evidence of dimensionality, reliability and validity, and most of its variance can be explained by constructs such as positive affect, positive relationships, accomplishment and meaning in life. As such, it was decided not to include a separate measure of life satisfaction because this would be captured by other constructs assessed in the new questionnaire. Nevertheless, a life satisfaction scale was included during scale development to help validate the new questionnaire.

Beyond positive feeling, Keyes' conceptualization of well-being includes positive functioning, as conceptualized and measured in Ryff's (1989) scales of psychological well-being. Like other measures reviewed thus far, Ryff's scales of psychological well-being have been widely used and evidence supporting their validity has been reviewed by Ryff and Singer (2006). Nevertheless, two potential limitations have also been identified. First, each positive functioning construct attempts to capture a broad range of phenomena. This is apparent in the construct definitions, which are much longer than those of other well-being constructs, and encompass several subjective experiences and even some behaviors (see Appendix 1). As a consequence, several items are required to reliably measure each construct,

with shorter 3-item scales exhibiting extremely low levels of reliability (e.g. $\alpha < .56$) (Ryff & Keyes, 1995). Second, several studies have found very high correlations among the six scales, which indicates an excessive overlap and thus poor discriminant validity among the six scales. Specifically, four scales tend to be very highly correlated: self-acceptance and environmental mastery, purpose in life and personal growth (e.g. Abbott et al., 2006; Chen et al., 2013; Burns & Machin, 2009; Ryff, 1989; Ryff & Keyes, 1995; Springer & Hauser, 2006). Both of these limitations could potentially be overcome by narrowing the construct definitions, which should increase reliability by increasing the correlations among items measuring the same construct (e.g. Reise, Cook & Moore, 2015) and should reduce factor correlations by increasing the distinctiveness of each construct.

Beyond positive feeling and positive functioning, Keyes' conceptualization and measures of well-being include social well-being constructs, as measured by the 5-factor social well-being scale (Keyes, 1998). Items in the social well-being scale generally fit the hypothesized five-factor model (e.g. Keyes, 1998). However, when alternative models were explored, models with three or four factors have also been empirically supported (e.g. De Jager, Coetzee & Visser, 2008; Hashemi, Hekmati, Vahedi & Babapour, 2014; Keyes, 1998). In other words, it may be justified to merge some of the social well-being constructs into fewer than five factors. Additionally, as Keyes (2015) pointed out, some social well-being constructs are quite unique to the social well-being scale. In other words, they do not tend to exhibit shared meaning with constructs in other conceptualizations or measures of well-being. For example, social coherence involves understanding of how society works. While this may involve some degree of perceived competence and self-efficacy, which is included in many conceptualizations, no conceptualization or measure seems to include an equivalent construct. Nevertheless, the social well-being constructs that overlapped with constructs in other conceptualizations of well-being are included in Appendix 1.

Recently, more well-being conceptualizations and measures have been developed, which include elements of both positive feeling and positive functioning. Because of their recency, fewer studies have assessed their psychometric properties. Nevertheless, these have been useful in providing additional perspectives to develop a new conceptualization and measure of well-being. Using Seligman's (2011) five-factor conceptualization, Kern et al. (2015) developed a questionnaire to measure

well-being in adolescent students. In this study, ($N = 516$), only four factors were identified, with positive relationships and meaning in life being merged into one factor (Kern et al., 2015). This suggests that positive relationships and meaning in life as defined by Seligman (2011), may be conceptually too similar. Nevertheless, in other studies, including 831 U.S. adults and 322 Malaysian adults, the hypothesized five-factor structure was supported (Coffey et al., 2016; Khaw & Kern, 2015).

The next two measures also include elements of both positive feeling and positive functioning (Diener et al., 2010; Huppert & So, 2013). However, these use single items to measure each component of well-being. Single-item measures are particularly useful in large national surveys, where not many items can be implemented. However, single-item measures also exhibit an important shortcoming. Due to the lack of multiple indicators measuring the same construct, it is not possible to exhaustively test their lower-order factor structure or to adequately estimate their internal reliability and adequately control for a lack of reliability (Marsh, 2007). Since factor analyses of these scales are only relevant to discussions regarding the higher-order structure of well-being, no limitations relevant to the lower-order structure of well-being were identified.

So far, this section has focused on four broad well-being conceptualizations and measures proposed by Keyes, Seligman, Huppert and So, Diener, and their respective colleagues. While scholars tend to agree about the conceptualization of positive feeling constructs, there is greater disagreement about positive functioning (e.g. Huta, 2013a). Therefore, two more conceptualizations of positive functioning, or eudaimonic well-being, were considered. Waterman et al. (2010) proposed a conceptualization and multi-item measure of positive functioning focusing on six constructs. Conceptually the questionnaire was multidimensional, but empirically it was validated with a single-factor model. On the one hand, support for the one-factor model indicates that items in the questionnaire are measuring a single construct, similar to general well-being, though without its affective elements. On the other hand, support for a one-factor model negates any support for the discriminant validity of the questionnaire subscale scores. In other words, the one-factor model indicates that the items, which purportedly measure six constructs, are actually indistinguishable from each other. Some evidence of multidimensionality was presented in a later study, which found support for a three-factor model (Schutte, Wissing & Khumalo 2013). However, clearly these three factors did not reflect

Waterman et al.'s six-factor conceptualization. Therefore, like Ryff's scales of psychological well-being, if all the original factors are to be retained, it may be desirable to narrow the construct definitions proposed by Waterman et al., in order to increase the distinctiveness of each scale.

Finally, the conceptualization of eudaimonic well-being proposed by Ryan et al. (2008) can be operationalized through measures of basic needs and mindfulness. One of the first measures of the basic psychological needs for autonomy, competence and relatedness (Gagné, 2003) has received limited support for its factor structure. Negative items were found to be particularly detrimental to the factor structure of the questionnaire (Johnston & Finney, 2010). To overcome these limitations, new scales with adequate factor structures have been developed to measure need satisfaction in life in general (Sheldon & Hilpert, 2012) as well as specific domains (Longo, Gunz, Curtis & Farsides, 2016). In both cases, an adequate factor structure was identified after treating negatively-scored items as method effects or separate factors. Mindfulness was conceptualized as receptive and open awareness (Ryan et al., 2008). Several measures of mindfulness have been developed (e.g. Brown & Ryan, 2003; Cardaciotto, Herbert, Forman, Moitra & Farrow, 2008; Feldman et al., 2007) and, among these measures, we can find two key components, which are consistently measured: self-awareness and self-acceptance. Like positive affect, these two aspects may be successfully modeled by a single factor (e.g. Brown & Ryan, 2003). However, since two factors have often emerged (Cardaciotto et al., 2008; Feldman et al., 2007), the present studies will favor such a differentiation and, if these two factors show poor discriminant validity, there will be no downside to merging them into a unidimensional model.

The strengths and shortcomings identified in this section will help inform the conceptualization of the scales of general well-being, which is presented in the next section.

2.4. The lower-order factors in the SGWB

Of the 49 constructs presented in six operationalizations of well-being, the majority (44) share the commonalities presented in the conceptual definition of general well-being presented above. Furthermore, these 44 constructs often overlap with each other. For example, four of the measurement models included the same feeling of engagement and involvement in one's daily activities (Diener et al., 2010;

Huppert & So, 2013; Seligman, 2011; Waterman et al., 2010). In other words, there is a general consensus regarding some constructs chosen by different scholars to measure well-being. This section will focus on that consensus. Common features among constructs presented in different operationalizations were identified as the core of each construct (Jaccard & Jacoby, 2010). In short, the main focus was on points of agreement and shared meaning among the constructs.

During this process, some degree of subjectivity must be acknowledged, as it is engrained in the process of concept and scale development (Jaccard & Jacoby, 2010; Kline, 2000). Indeed, other researchers have included a variety of constructs to measure well-being. In this case, the choice of constructs was informed by the conceptual definition of general well-being presented above, apparent commonalities in the conceptual definitions of the lower-order constructs, as well as a review of the psychometric properties of previous measures of these constructs. After reviewing the conceptualizations and measures above, the following fourteen common constructs were identified to include in the SGWB. A summary of the following constructs and their conceptual relationships with constructs presented in other well-being models can be found in Appendix 1.

2.4.1. Happiness, vitality and calmness. All three arousal levels of positive affect can be found in the majority of the models or measures. As previously mentioned, all three arousal levels of positive affect can be found, for example, in measures used by Keyes et al. (2002), Kern et al. (2015) and Huppert & So (2013). These will be conceptualized and measured as three correlated factors, to ensure that each aspect is adequately measured with a sufficient number of items. If these three factors show poor discriminant validity, there will be no downside to merging them into a unidimensional model (Whetten, 1989). Thus, the first three constructs identified are the following: happiness – moderate-arousal pleasant feelings, such as feeling happy, cheerful and pleased; vitality – high-arousal pleasant feelings, such as feeling energetic and lively; and calmness – low-arousal pleasant feelings, like serenity and peacefulness.

2.4.2. Optimism is generally defined as a positive outlook on and expectations about the future (Scheier & Carver, 1985). Huppert and So (2013) and Diener et al. (2010) define it in broader terms, such as “feeling optimistic or hopeful about the future” (Huppert & So, 2013, p. 857). That definition is mirrored in more specific terms in Keyes’ concept of social actualization: being “hopeful about the condition

and future of society” (Keyes, 1998, p. 123). While the first definition focuses on the future in general, the second focuses more on the future of society. Another closely-related construct is social acceptance: the belief that people are good and trustworthy (Keyes, 1998). This also assumes an optimistic worldview. In fact, after correcting for unreliability, the correlations between social actualization and social acceptance ranged between .73 and .94 (Keyes, 1998), indicating that they tap into the same construct.

2.4.3. Involvement describes the flow state: an absorbing experience in which the individual is completely focused on the task at hand (Csikszentmihalyi, 1990). In well-being models, this construct has been described with terms such as engagement (Diener et al., 2010; Huppert & So, 2013; Seligman, 2011), involvement (Waterman et al., 2010), and it is ordinarily described by athletes as being “in the zone” (Martin & Jackson, 2008). Involvement or flow can be conceptualized as a multidimensional construct, comprising high concentration, lack of self-awareness, time distortion, etc., or as a unidimensional construct (i.e. core flow) defined as a subjective experience of absorption (Martin & Jackson, 2008). Consistent with the reviewed models of well-being that include the flow construct, this was conceptualized as a unidimensional experiential construct (i.e. core flow).

2.4.4. Self-awareness consists of knowing oneself (Waterman et al., 2010) and experiencing a state of mindful awareness: by noticing different aspects of one’s experience, such as thoughts, feelings, and physical sensations (Ryan et al., 2008). Both involvement and self-awareness entail a focus on the present moment, which can be directed towards the task at hand (involvement) or towards one’s inner experience (self-awareness). At the state level, they may seem antithetical, as one can either fully focus on the task at hand or focus on one’s internal state (e.g. Csikszentmihalyi, 1990). However, at the trait level, they were found to be positively related, as people who tend to experience mindful awareness also report more frequent flow experiences (Kee & Wang, 2008).

2.4.5. Self-acceptance. When people are aware of their internal states and dispositions, they may criticize them, deny them, or experience them in an accepting, non-judgmental way. Self-acceptance consists of experiencing different aspects of oneself (e.g. one’s past, personality, thoughts, and feelings) in a tolerant, receptive and non-judgmental way (Ryan et al., 2008, Ryff, 1989). Both self-acceptance and

self-awareness are commonly presented as two key components of mindfulness (e.g. Cardaciotto et al., 2008; Niemiec & Ryan, 2013; Ryan et al., 2008).

2.4.6. Self-worth consists of positive evaluations and feelings about oneself (Baumeister, Smart & Boden, 1996; Baumeister, Campbell, Krueger & Vohs, 2003). In other words, people with high self-worth like and feel good about who they are. While Huppert and So conceptualize self-worth as a separate construct, Ryff (1989) includes self-worth within self-acceptance. As Ryff and Keyes (1995) argue, constructs such as Ryff's self-acceptance may be too conceptually broad, which can lead to low reliability. Therefore, in this thesis these two constructs were initially separated. If sufficient evidence indicates that they measure a single factor, they can later be merged.

2.4.7. Competence consists of feeling and perceiving oneself as effective and able to overcome challenges and achieve desired outcomes (Deci & Ryan; 1985, 2000; Ryan & Deci, 2002). It is a felt sense of confidence in one's abilities, rather than an acquired skill (Ryan & Deci, 2002). All these facets have been shown to load strongly on a single factor (e.g. Longo, Gunz et al., 2016). Ryan and colleagues' definition seems to tap into the core of the construct that is shared across other models (e.g. Ryff, 1989; Seligman, 2011). This definition also conceptually overlaps with general self-efficacy, which focuses more on actual competence rather than perceived competence (e.g. Luszczynska, Gutiérrez-Doña & Schwarzer, 2005), and it also seems to underlie the more specific belief that one can make sense of the world and that life is predictable and controllable (Keyes, 1998).

2.4.8. Development consists in experiencing continuous growth and improvement. This is not a fixed state that can be attained at a single point in time. Instead, it represents a perceived improvement and expanding of the self that continues over time (Ryff, 1989). This is a more specific definition than Ryff's, which overlaps conceptually with significance and competence. For instance, according to Ryff, a person with high personal growth also has a "sense of realizing his or her potential, [...] is changing in ways that reflect more self-knowledge and effectiveness" (Ryff, 1989, p. 1072).

2.4.9. Purpose consists of having clear goals, a sense of direction and a larger aim in life. It should also be noted that purpose in life is often categorized as a key component of meaning in life: a broader concept, which also includes constructs like

significance (see Brandstätter, Baumann, Borasio & Fegg, 2012; Morgan & Farsides, 2009).

2.4.10. Significance is the feeling that what we do is worthwhile, rewarding and valuable (Huppert & So, 2013). Compared to some previous definitions, the current one focuses on the subjective experience of significance and does not focus on the beneficiary or target of one's activities, such as society (Keyes, 1998) or something larger than oneself (Seligman, 2011). Again, this was consistent with the goal to focus on the shared meaning among constructs in different conceptualizations of well-being.

2.4.11. Self-congruence is the perception that one's actions are compatible with one's interests, values, and beliefs. In other words, self-congruence is about being true to oneself: experiencing life in a way that is consistent with one's nature (Huta, 2013a; Niemiec & Ryan, 2013). Doing what other people ask of us does not undermine our need for self-congruence, as long as we endorse other people's requests and do not feel forced to act in a certain way (Ryan & Deci, 2002; Longo, Gunz et al., 2016). Perhaps, autonomy is the most widely used term for this construct. However, autonomy is often confused with being independent, which is a facet included in Ryff's definition (Ryan & Deci, 2001, 2002) and may be viewed as a set of behaviors involving a lack of connection with other people. On the other hand, the term self-congruence (Weinstein, Przybylski & Ryan, 2012) seems to better depict the subjective perception that one's behavior is in line with one's sense of self.

2.4.12. Connection describes a set of perceived positive links we can establish with other people. It involves a feeling of belonging, mutual caring, love, and closeness (e.g. Keyes, 1998; Ryan et al., 2008; Ryff, 1989). These facets have been shown to load strongly on a single factor (e.g. Longo, Gunz et al., 2016).

2.5. Summary

This chapter reviewed the higher- and lower-order factor structure of well-being measures that operationalize six conceptualizations of well-being (Diener et al., 2010; Huppert & So, 2013; Keyes, 2002; Ryan et al., 2008; Seligman, 2011; Waterman et al., 2010). It was concluded that a single general factor of well-being adequately explains covariances in well-being data. Based on similarities among well-being constructs, the general factor was conceptualized as a collection of

relatively stable subjective feelings and evaluations, which represent facets of good mental health. After examining the commonalities among the lower-order construct definitions, fourteen common constructs were identified and conceptually defined. These conceptual definitions will guide the scale development and validation process described in the following chapters. The next chapter will outline best practices in scale development and validation (e.g. DeVellis, 2012). Then, three studies will be conducted to develop and validate the questionnaire. Like previous questionnaires, the SGWB aim to measure well-being in the general adult population, and will be initially validated in North American samples.

Chapter 3: Method

In the social sciences, we often measure hypothetical constructs (i.e. latent variables): variables that cannot be directly measured. These constructs can be broad (e.g. general mental ability) or specific (e.g. working memory, numerical ability, etc.). We cannot directly measure someone's working memory in the same way we directly measure the length of a table. However, we can ask an individual to complete several memory tasks and based on their performance on the tasks, we infer their working memory ability (e.g. Englund, Decker, Woodlief & DiStefano, 2014). Similarly, we can measure personality traits, for example, by asking individuals the extent to which they find several statements about themselves accurate (e.g. Donnellan, Oswald, Baird & Lucas, 2006). This chapter will review the best practices and useful rules of thumb in scale development. It will focus particularly on well-being and affective constructs, although some examples will also involve mental ability and personality constructs. The following sections attempt to follow the chronological process of scale development.

3.1. Literature review

The first step in scale development consists in reviewing the literature on the topic (e.g. Clark & Watson, 1995). A sound knowledge of previous conceptual frameworks facilitates the later development and selection of items based not only on statistical grounds, but also on theoretical grounds. Most importantly, by reviewing previous work in the field we can assess whether the new measure is needed and is likely to make a contribution to the literature. Considering the time and resources necessary to develop a new questionnaire, this may be justified if there are no existing measures of a particular construct, or if they exhibit serious limitations. Identifying such limitations can also help develop a new measure that attempts to improve on past measures.

During the review process, it is useful to write down how other academics have defined the same or similar constructs, and broaden the search by identifying other related constructs (Clark & Watson, 1995; Jaccard & Jacoby, 2010). That information will be useful later on when testing convergent and discriminant validity (see Validity section).

If literature on a topic is lacking, it may be useful to consult with experts on the topic or conduct interviews and focus groups to learn more about the construct from the population of interest (Rattray & Jones, 2007).

This thesis introduced different conceptualizations of well-being in the introduction and presented a critical review of the measurement of well-being in the previous chapter.

3.2. Defining the construct

After reviewing the literature, we need to identify and describe the construct we are trying to measure. If a construct definition is too broad or ambiguous, the scale development and validation process becomes much more challenging. Without a clear notion of the target construct it is difficult to develop items to measure it, interpret the scores produced by the measure, or identify criteria to test its adequacy (Spector, 2013). The construct definition should be clear and thorough. Ideally, it should describe the population to whom the construct applies (e.g. age and/or gender groups), the type of information being assessed (e.g. attitudes, behaviors, etc.), the breadth of the construct (e.g. general anxiety or social anxiety), its location within a hierarchy of constructs (e.g. higher-order or lower-order), and its theoretical relation to other constructs already known in the literature (De Vaus, 2002; Reise, Waller & Comrey, 2000; Spector, 2013). It is good practice to write items that measure the same type of information (e.g. behaviors or attitudes) rather than mixing different types of information (De Vaus, 2002; Spector, 2013). The resulting definition is sometimes labeled a “conceptual definition”, as it describes a construct and its boundaries in more abstract terms.

On the other hand, operational definitions equate a construct with how it is measured (Jaccard & Jacoby, 2010). For example, a higher-order factor of well-being is operationalized (i.e. measured) through a combination of its lower-order latent factors, such as happiness, competence, etc. From this perspective, well-being is equivalent to its measurement.

It is important to define constructs both conceptually and operationally. In the absence of a clear conceptual definition of a construct (i.e. its properties and boundaries), it is more difficult to identify the most valid among several measures. For example, it would be difficult to identify a good involvement scale if we did not

know that involvement describes the subjective experience of immersion with the task at hand.

Conceptual and operational definitions can influence each other through deductive and inductive processes: we may start from an abstract idea and then identify different ways of measuring it, but we may also start from a set of specific variables to develop a broader conceptual definition (Jaccard & Jacoby, 2010). Regardless of our starting position, it is essential for our conceptualization and operationalization to be consistent with each other, as they both attempt to describe the same construct from different perspectives (DeVaus, 2002). In other words, in the process of scale development, we want our measure (operationalization) to perform in ways that are consistent with its conceptualization.

For example, in an attempt to conceptually define the general factor of well-being and its lower-order factors, we can use the following inductive and deductive approaches (Jaccard & Jacoby, 2010). In the inductive approach, the researcher identifies commonalities in the lower-order factors in order to define the general factor, thus proceeding from specifics to general. In the deductive approach, the researcher starts by formulating a conceptual definition for the general factor and then identifies the lower-order factors necessary to measure it, thus moving from general to specifics. Inductive and deductive approaches can be combined over several iterations. Thus, the conceptualization of g can be refined by identifying commonalities among its lower-order factors, and lower-order factors can be included or omitted based on the revised definition of g . This combination of inductive and deductive approaches was used in the previous chapter to define the higher- and lower-order factors of well-being.

3.3. Level of measurement

Measurement involves “the assignment of numerals to objects or events according to rules” (Stevens, 1946, p. 677). This section outlines four common types of measurement in order to place the psychometric scale development process in the wider context of scientific measurement. These measurement types are not mutually exclusive and were first outlined by Stevens (1946).

Nominal (or categorical) variables include two or more categories with no rank order. That is, the categories in each variable are simply different from each other and no category is “greater than” or “less than” any other category. For example, a

variable including different colors would be nominal. We may assign a number to each of these color categories (e.g. “1” for yellow, “2” for white, “3” for green, etc.), but these numbers are arbitrary and can be interpreted as names. In the color example, green would *not* be 3 times greater than yellow, as the numbers only serve the purpose of distinguishing among categories. Common examples of categorical variables in social research are gender and ethnicity. Analyses of these variables generally focus on the number of cases in a given category (i.e. frequency) (Stevens, 1946).

Ordinal (or ordered-categorical) variables include categories, which are ordered. Each category has a position (i.e. rank) on a scale from low to high. For example, in a sport tournament the participants can be ordered in 1st, 2nd, 3rd place, etc. However, the differences in score between each place may vary considerably. Perhaps, the person in 1st place performed only slightly better than the person in 2nd place, while the person in 2nd place may have performed much better than the person in 3rd place. In this case, the difference between the 1st and 2nd category would be very small, while the difference between the 2nd and 3rd category would be substantial. However, information about the differences between categories is absent in ordinal data. Therefore, based on ordinal data, we can state that the person in 1st place performed better than the person in 2nd place, but we cannot accurately quantify the difference between the two.

Like ordinal variables, interval variables provide information about ranking. Additionally, the differences between two adjacent points are always the same. A classic example is the measurement of temperature (in degrees Celsius or Fahrenheit), where the difference between 20° and 21° is the same as the difference between 25° and 26°. Interval variables can be transformed and analyzed using the linear model: $y = bx + a$. In other words, each value (x) on an interval variable can be transformed (into y) by both multiplying it by a constant (b) and adding it to a constant (a). Transforming interval variables only through a multiplication, without adding a constant, would not produce meaningful results (Stevens, 1946). This occurs due to a lack of an absolute “0” value. For example, both Celsius and Fahrenheit scales are somewhat arbitrary, in that neither scale includes an absolute “0” indicating the total absence of temperature. Consequently, 40° Celsius is not twice greater than 20° Celsius. In other words, simply multiplying 20° Celsius by a factor of two would not produce a temperature that is twice as great. If we

transformed 40° and 20° Celsius in Fahrenheit degrees, we would see that 104° is not twice greater than 68° Fahrenheit. This limitation of interval data could be overcome if the scales had an absolute “0” value indicating a total lack of the attribute. For example, when temperature is measured in Kelvin degrees, a value of 0 indicates that all thermal motion stops and no negative values are meaningful. Therefore, if we used a Kelvin scale, 40° Kelvin would indeed be twice greater than 20° Kelvin.

Ratio data meets the requirements of interval data. Additionally, ratio scales have an absolute “0”, below which no negative values are meaningful. This is the highest standard of scientific measurement, because ratio data can be subjected to a greater number of statistical analyses. Specifically, ratio data can be subjected to all the transformations that are applicable to the previous measures outlined.

Furthermore, each score in a ratio variable can be multiplied by a constant (e.g. Kelvin values shown above). Additionally, length (e.g. in meters) is another example of a ratio variable. Length is measured on a ratio scale because it has equal intervals, the interval between 1 meter and 2 meters is the same as the interval between 11 meters and 12 meters, and it has an absolute “0”: length can only be “0” or positive. The measurement types outlined hitherto and their properties are summarized in Table 1.

Table 1 Types of measurements

Scale type	Rank	Equal intervals	Absolute “0”
Nominal	No	No	No
Ordinal	Yes	No	No
Interval	Yes	Yes	No
Ratio	Yes	Yes	Yes

Note: The table was adapted from Kaplan and Saccuzzo (2012) (p. 29)

In psychological research, we may sometimes use ratio data. For example, in a psychological experiment, we may use reaction time, which includes an absolute “0” and equal intervals (e.g. a reaction time of 200ms is twice as long as a reaction time of 100ms). Additionally, we often try to measure hypothetical constructs (e.g. self-esteem) by asking several questions (e.g. Do you feel good about yourself?), and score these questions on a specific response format (e.g. a 7-point Likert format from strongly disagree to strongly agree). Strictly speaking, such a response format produces ordered-categorical (i.e. ordinal) data: responses are ordered, but the interval between “strongly disagree” and “disagree” may be somewhat different between the interval between “disagree” and “slightly disagree” (Jamieson, 2004; Stevens, 1946). However, it has been argued that it is sensible to assume interval

properties for Likert variables (Nunnally, 1978; Streiner, Norman & Cairney, 2015). Section 3.7.5 will reference evidence supporting this assumption.

3.4. Item generation

After identifying the constructs we need to measure and articulating a conceptual definition for each construct, we can start to write items to measure each construct. We need to determine the response format of our questionnaire. We may choose from several formats (e.g. vignettes, multiple choice, etc.). The Likert response format asks respondents to judge, for example, the extremity or frequency of an attribute on a scale with two or more ordered response options. This is one of the most common response formats for psychological constructs involving opinions, beliefs and attitudes (DeVellis, 2012). A well-developed questionnaire using a Likert format can be subjected to a wide range of analyses unlike, for instance, multiple choice (i.e. ipsative) formats, which are restricted to analyses such as Thurstonian item response theory (e.g. Brown & Maydeu-Olivares, 2012). Because of its wide applicability, this thesis will focus on the Likert format.

First, the number of response options to include in the response format are specified. These can vary considerably. At the lowest end of the spectrum, only two options may be included (e.g. false and true). However, this response format may be too simplistic and people may want to report more refined judgments that range between these two options. In that case, this restrictive response format may reduce participant cooperation (Kline, 2000; Rust & Golombok, 2009). At the highest end of the spectrum, we may include, say, as many as 100 response options. However, with so many response options, individuals may be unable to distinguish between a rating of 67 and 68. In fact, Miller (1956) has shown that it is difficult to retain more than seven categories in our short-term memory. Therefore, increasing the number of response options beyond 7 may simply increase random error variance, as people may not be able to meaningfully differentiate among more than 7 categories (Saris & Gallhofer, 2014). Because of this, psychometricians usually suggest using between 5 and 7 response options (Kline, 2000; Spector, 2013; Streiner et al., 2015). Considerations regarding the number of response options are also presented in a later section focusing on factor analysis estimation.

During this process, we must select the labels to assign to each response option. For example, the response labels could focus on frequency (e.g. never to

always), agreement (e.g. strongly disagree to strongly agree), degree or intensity (e.g. not at all to very much), similarity (e.g. unlike me to like me), etc. (Clark & Watson, 1995). There are no universal rules to choose a specific label (Rust & Golombok, 2009) but, generally, the nature of the constructs will help determine its response labels (Spector, 2013). For example, a frequency format may work well with clinical symptoms questionnaires, while an intensity format may work well with a mood questionnaire (Rust & Golombok, 2009). Response categories should be clearly ordered from low to high or from high to low. Additionally, it has been argued that the response format should include a category representing the equivalent of an absolute zero (e.g. not at all), in order to produce data that resembles ratio data as much as possible (Rossiter, 2002). Consistent with this reasoning, attempts should be made to choose response labels that produce equal intervals between adjacent categories (Rossiter, 2002). In practice, this is rarely tested, and requires the use of item response theory models (e.g. Wakita, Ueshima & Noguchi, 2012). However, as a rule of thumb, as long as the response format includes at least 5 categories and the distribution of the data closely approximates a normal distribution, the data can be factor analyzed under the assumption of interval measurement (Rhemtulla, Brosseau-Liard & Savalei, 2012; Streiner et al., 2015).

We may choose an odd or even number of response options. The main difference between the two is that, with an odd number of response options, we include a middle “non-committal” option (e.g. “don’t know” or “neither agree nor disagree”). With an even number, we force respondents to choose between reporting a lower or higher level of the construct. If we choose a scale consistent with Rossiter’s (2002) recommendations (e.g. from “not at all” to “very much”), the middle option will express a moderate level of the construct (e.g. “somewhat”) rather than a non-committal option (e.g. “don’t know”). The main issue with a middle option is that participants may select it out of inattention or laziness (De Vaus, 2002). However, not including the middle option may increase the error variance for those participants who really exhibit a moderate level of the construct (Clark & Watson, 1995; De Vaus, 2002). Since several measures were taken in the subsequent studies to control for inattention, it was deemed most appropriate to include a middle option.

After identifying an appropriate response format, we can generate several items to measure each construct. These items need to be consistent with the chosen response format (Sarlis & Gallhofer, 2014). For example, if we are using a frequency

format (e.g. never to always), including frequency terms in the items may confuse respondents, such as in the item “Do you always feel engaged at work?” In this case, we could either use this item with a different response format (e.g. not at all true to very true) or we may use the same response format (never to always), but rephrase the item to “How often do you feel engaged at work?” In general, frequency terms within items should be avoided because they may be interpreted in very different ways by different respondents (Rossiter, 2002). However, when used as different response labels to judge one’s subjective perception of frequency they tend to perform well (Streiner et al., 2015).

It is a good practice to develop multi-item scales whenever feasible, rather than single-item scales, because responses to each item are influenced by responses to previous items on the questionnaire. For example, in one study, respondents answered the following two items: “How happy are you with life in general?” and “How often do you go on dates?” (Strack, Martin & Schwarz, 1988). When the question about dating followed the question about general happiness, a negative and non-significant correlation was found between the two ($r = -.12, p > .05$). However, when the question about dating preceded the question about general happiness, the two were strongly and positively correlated ($r = .66, p < .001$). The difference can be explained by considering the consciously available information when answering each item. In the first scenario, one answers the happiness question by considering one’s life in general, without necessarily considering one’s dating life. Therefore the subsequent factual answer to the dating question will not necessarily correlate to the happiness scores. However, in the second scenario, after answering the dating question, that information, whether positive or negative, will be considered when evaluating one’s happiness. People who go on dates often may report a higher level of happiness, while people who rarely go on dates may report a lower level of happiness (Strack et al., 1988). When we measure a construct using only a single item, responses to that item are more likely to be biased than if we use a multi-item scale, where item order and other idiosyncratic effects influencing each item will be reduced by the use of multiple items (McCoach, Gable & Madura, 2013). Nevertheless, it must be noted that the use of single-item measures is somewhat justified in large national surveys, where it may be very costly to use many multi-item scales. Additionally, single-item scales may be used when measuring simple and straightforward concepts. For example, it has been argued that the concept of

“liking” is simple enough to be measured by a single item (e.g. “how much do you like this restaurant?”) (Hair, Black, Babin & Anderson, 2010).

How many items should we write in our initial item pool? We can expect several of the items in the initial item pool not to perform well in later analyses. In other words, during the scale development and validation process, several items will likely be omitted. Therefore, it is important to generate at least twice as many items as we are expecting to include in the final questionnaire (e.g. Kline, 2000). Three items per construct are typically an absolute minimum necessary to run a factor analysis. Including at least four items per construct is a typical recommendation (Brown, 2015). Therefore, the initial item pool should ideally include 8 or more items per construct.

A group of items measuring the same construct should exhibit the same level of specificity (Comrey & Lee, 1992). If two items are more highly correlated to each other than the rest of the items on the same scale, they will seem to measure a separate construct: a separate factor will emerge including only these two items. However, this factor would simply be an artifact due to the comparatively higher correlation between the two items. For example, Comrey and Lee (1992) present the following neuroticism items as an example (p. 210):

1. I am tense and nervous
2. I lack confidence
3. I am unable to cope with life’s problems
4. My moods fluctuate easily
5. I feel blue
6. I suffer from depression

While all six items tap into neuroticism, the last two are more redundant, as they are more similar to each other than the rest of the scale. Specifically, they both tap into depression, while other items tap into various aspects of neuroticism. If these items were included in our initial item pool, we may select only one of them. Alternatively, we may include both of them temporarily, with the purpose of later selecting the one that performs better in quantitative analyses (DeVellis, 2012).

The item writing process can be facilitated through the use of a grid (e.g. Rust & Golombok, 2009). For each construct, we need to identify several of its manifestations. For example, in the previous chapter, the construct competence was defined as “feeling and perceiving oneself as effective and able to overcome

challenges and achieve desired outcomes.” In other words, three content areas were specified: feeling effective, overcoming challenges and achieving outcomes. For each of these content areas, several items need to be generated (see Table 2). If some of these content areas were more important than others, we could write more items tapping into that area. In other words, we may assign greater weight to different content areas. In this case, although one extra item was written focusing on overcoming challenges, an equal weight was assigned to each content area for competence, and ultimately an equal number of items from each content area were selected. This procedure is further explained in the next chapter.

Table 2 An example of items and content areas for competence

Manifestation	Content Area
I am highly effective at what I do	General effectiveness
I tend to perform very well	General effectiveness
I am confident in my ability to master new things	General effectiveness
I can obtain the results I want	Ability to achieve
I am able fulfil my goals	Ability to achieve
I can accomplish what I attempt to do	Ability to achieve
I can successfully complete difficult tasks	Overcome challenges
I can handle difficult situations	Overcome challenges
I’m very good at overcoming challenges	Overcome challenges
I have a great ability to deal with problems	Overcome challenges

3.5. Principles of item writing

Several principles have been outlined to facilitate item writing. While an exhaustive account of these principles would typically require an entire chapter (e.g. Kline, 2000), a concise summary is presented here.

The most important principle is that items should be simple and clear. The terms included in each item should be frequently used in everyday speech, and as concrete as possible (e.g. De Vaus, 2002; Saris & Gallhofer, 2014). Simplicity can be achieved by writing items that are as short as possible (e.g. Spector, 2013). By focusing on simplicity and clarity, we attempt to write items that have the same meaning for all respondents (De Vaus, 2002). Some threats to clarity and simplicity include vague frequency terms (e.g. often, sometimes), quantity terms (e.g. a bit, a few) or feelings (e.g. do you enjoy music?) (Kline, 2000; Rust & Golombok, 2009). These terms may have different meanings for different people. For instance, in the last example, the terms “enjoy” and “music” are both vague. This item could be interpreted or rephrased as, “Do you play a musical instrument?”, “Do you like

playing that musical instrument?”, “What type of music do you listen to?” and “Do you like listening to that music?” Responses to the item “Do you enjoy music?” could be influenced by all these interpretations, but the single item score will not tell us the way in which each respondent interpreted the item. Therefore, the ambiguous item will produce ambiguous data.

A related principle is that each item should only measure a single construct. Double-barreled items clearly violate this principle as they ask several questions at the same time (Clark & Watson, 1995). As an extreme example, the item “do you like eating ice cream and reading books?” is clearly asking two questions: “do you like eating ice cream?” and “do you like reading books?” Once again, it is impossible to know the aspect of the item to which respondents are answering. Therefore, responses to the item will be ambiguous. As a rule of thumb, double-barreled items can often be avoided by avoiding the use of “and”, “but”, “by” and “because” (McCoach et al., 2013). For example, the item above included the word “and,” and it was made less ambiguous by rephrasing it as two items focusing on each element of the item. That said, when two or more aspects of an item correlate highly and clearly measure the same construct, the use of the conjunction “and” may be justified (Saris & Gallhofer, 2014), as in the hypothetical item “I feel happy and cheerful.”

Statements that apply to everyone or no one should be avoided (Clark & Watson, 1995). Items should produce a range of responses (e.g. from low to high). If an item only elicits a low (or high) response from every respondent, there will not be enough variation in responses for that item to be useful, as responses would not correlate to other items. An item that always produces the same response is as useful as a clock that always tells you the same time. For example, the items “do you sleep sometimes?” or “do you sleep less than 1 hour per month?” are not useful because virtually everyone would agree with the first one and disagree with the second one.

Items should generally not make assumptions about respondents (Saris & Gallhofer, 2014). For example, the item “what is the best movie you watched last week?” assumes that the respondent watched a movie, which may not be the case. Therefore, the item may be biased as participants may, perhaps, leave it blank or report a movie watched prior to last week. This could be revised by asking “Did you watch any movies last week?” and then, following an affirmative answer, “What was the best movie you watched last week?”

Items should not elicit a specific response from respondents (De Vaus, 2002). Such leading questions create the impression that there may be a wrong or correct answer to an item, which would clearly bias the data. For example, starting an item with the sentence “do you agree that ...” may invite respondents to endorse the item regardless of its content.

It is best to avoid using the word “not”, as it may lead to misunderstandings (De Vaus, 2002; Spector, 2013). Respondents may fail to see the “not”, and therefore agree with an item even though they hold the opposite attitude. For example, respondents who like their co-workers may erroneously agree with the item “my co-workers are not very nice.” People are used to thinking in terms of concrete objects rather than their absence (McCoach et al., 2013). Therefore, they may be confused by items including “not.”

Similarly, the use of reverse-scored items has been recently discouraged (e.g. McCoach et al., 2013). If we were measuring “happiness” we may also include few items such as “I feel sad.” These items would be scored in the opposite direction compared to the rest of the happiness items (e.g. “I feel happy”), and thus are labeled reverse-scored. Historically, the use of reverse-scored items was advocated to attempt to reduce acquiescent responding: the tendency to endorse all items regardless of their content (see Kline, 2000). However, reversed items can be confusing to respondents, particularly in long questionnaires (DeVellis, 2012; van Sonderen, Sanderman & Coyne, 2013); they can distort the dimensionality of a questionnaire (Schmitt & Stults, 1985; Schriesheim & Eisenbach, 1995; Woods, 2006; Wong, Rindfleisch & Burroughs, 2003); they have been found to be less effective in measuring well-being constructs like flow (Jackson & Eklund, 2004) and vitality (Bostic, Rubio & Hood, 2000); and, while they are normally included to prevent acquiescent responding, they have been found to be ineffective in doing so (van Sonderen et al., 2013).

So far, this chapter has focused on good practices in item writing. To summarize, items should be simple and clear; they should measure a single construct; they should produce a range of responses; they should not make assumptions about respondents or elicit a specific response from them; and they should avoid negative or reverse-scored statements. The next sections will focus on quantitatively and qualitatively analyzing the items in the item pool to select the most adequate measures of a construct.

3.6. Descriptive statistics

The previous section mentioned that items should produce scores that vary (e.g. DeVellis, 2012). For example, an individual with a low level of happiness would ideally report a low score on a happiness item, while an individual with a high level of happiness would ideally report a high score on a happiness item. If we assume that happiness varies across individuals, then responses to our items must also vary. The fact that item scores vary does not necessarily mean that the item is measuring the construct it aims to measure. However, if the scores do not vary at all, then we can conclude that the item is not very useful in measuring any construct.

To further emphasize the importance of score variability, it must be stated that this is influenced by several factors. In classical test theory, an individual's observed score on an item is influenced by two main components: the individuals' true score as well as error (e.g. Kline, 2000). In other words, score variability can be influenced by actual changes in the construct (i.e. true score variance) as well as changes in aspects unrelated to the construct (i.e. error variance). In a multi-item scale, true score variance and error variance can be estimated based on the inter-relationships among items measuring the same construct. The stronger the correlations among items, the higher is their proportion of true score variance. However, if scores do not vary, they cannot covary, which precludes any estimation of participants' true score. This topic will be further explored in the following sections on dimensionality (3.7) and reliability (3.8). For now, it is clear that, to assess the true score variance in our measures, it is necessary for them to produce a range of scores.

When variable scores vary, their distribution takes a specific shape. Theoretically, when variable scores approximate a normal distribution (i.e. a bell curve), the distances between scores are known. Therefore, these variables can be treated as interval measures (Gaito, 1980). In practice, factor analyses have been found to produce accurate results with items including five or more response options, as long as they approximate a normal distribution (Lei, 2009; Rhemtulla et al., 2012). When this is not the case, specific estimation techniques are required, and these are outlined in section 3.7.5 on estimation in factor analysis. Therefore, when item scores closely approximate a normal distribution, thereby exhibiting interval properties, we can analyze them with a wider range of statistical techniques. Several statistics and rules of thumb can be used to assess how closely score distributions approximate normality.

Score variability is typically assessed by observing the range and standard deviations of the scores for each item. The range of scores should ideally cover all possible responses to our items. For example, in a 5-point frequency scale ranging from never (1) to always (5), the overall responses would ideally range from one to five. During the development of a mindfulness scale, items that did not cover their full response range were omitted from further analyses (Brown & Ryan, 2003). Additionally, the standard deviation, which measures the spread of scores from the mean, should exhibit a non-zero value for all items. As a rule of thumb, in a scale with a 7-point response format, the standard deviation should not be lower than 1 (McCoach et al., 2013). Items that do not produce a range of scores (e.g. $SD = 0$) would typically be excluded from further analyses.

In addition to examining the spread of the item scores, it is useful to observe their mean values. This should be close to the middle response option (DeVellis, 2012). For example, an item with a 5-point response format ranging from 1 to 5 should ideally produce scores with a mean value of 3. If the mean were extremely high (e.g. 4.7) or extremely low (e.g. 1.3), together with a low standard deviation, it would indicate a restricted variance (McCoach et al., 2013). As a rule of thumb, in a 7-point scale, one may revise or omit items showing a mean with a distance of .50 or less from the lowest or highest response category (McCoach et al., 2013).

Two more statistics are useful in assessing the univariate shape of a score distribution: skewness and kurtosis. Skewness assesses the symmetry of the distribution relative to its mean. In a normal distribution, skewness would be 0, indicating that the distribution tails on both sides of the mean are equal. A positive skewness would indicate that most of the observed scores are below the mean. A negative skewness would indicate that most of the observed scores are above the mean. Extreme negative or positive skewness would indicate a lack of variability in responses. When absolute skewness values are above 2 the assumption of normality is not met and this may lead to bias in factor analysis (Curran, West & Finch, 1996; Ryu, 2011), unless specific estimators are used (see section 3.7.5).

Kurtosis assesses the degree to which the distribution of scores is peaked or flat. A kurtosis of 0 provides evidence for the normality of the distribution. A positive kurtosis indicates that the distribution has a higher peak than normal. A negative kurtosis indicates a relatively flat distribution. Absolute kurtosis values above 7 are generally considered problematic (Curran et al., 1996; Ryu, 2011).

In some circumstances z-scores are used to evaluate the significance of skewness and kurtosis values. These z-scores are calculated by dividing the skewness and kurtosis values by their standard errors. Z-scores larger than 1.96 indicate significant skewness ($p < .05$). However, this practice has been discouraged because, as the sample size increases, minimal skewness and kurtosis will appear significant, even though they are unlikely to bias factor analytical results (Fabrigar & Wegener, 2012).

Finally, in factor analysis using the maximum likelihood estimator, variable scores are often assumed to have a multivariate normal distribution. As the next section explains, confirmatory factor analyses are most often estimated with the maximum likelihood estimator (Brown, 2015). This estimator assumes the distribution among all variable combinations to be normal. This assumption can be assessed with Mardia's (1970) multivariate skewness and kurtosis statistic. As a rule of thumb, values above 5 or 6 indicate that the assumption of multivariate normality is not met (Bentler, 2006). When that is the case, a robust maximum likelihood estimator can be used (see section 3.7.5).

3.7. Dimensionality: Factor analysis

3.7.1. Introduction to factor analysis. After assessing the distribution of our data and establishing that there is sufficient variance in the item scores, we can conclude that our items may be measuring something. This variance may be influenced by one or more constructs as well as random error. By assessing the dimensionality of our data, we attempt to answer the following question: how many constructs (i.e. dimensions or latent variables) do the items measure? Dimensionality is sometimes labeled factor structure or structural validity, because it assesses the structure of the data (e.g. Burns & Machin, 2009).

Why is it important to assess the dimensionality of a questionnaire? It is essential for a measure of a single construct (e.g. job satisfaction) only to tap into a single construct or dimension. If it measures more than one construct any interpretation of its scores will be ambiguous (Kline, 2000). For example, if a purportedly unidimensional scale taps into both job satisfaction and perceived competence, a score of 5 out of 10 may be due to a high level of satisfaction but low competence, high satisfaction but low competence or moderate satisfaction and moderate competence. Therefore, if a purportedly unidimensional scale measures

clearly distinguishable constructs, items measuring each construct may be separated into different scales in order to provide less ambiguous scale scores.

The idea of construct or latent variable was introduced in a previous section. A latent variable is a hypothetical concept that cannot be directly measured (Brown, 2015). For example, happiness is an abstract concept. Unlike a table, whose length can be directly measured, happiness is measured indirectly, for instance, by asking several questions (i.e. items) about one's affective state. In factor analysis, latent variables are called "factors."

For example, a happiness questionnaire may include items, such as "I feel happy", "I feel cheerful", etc. These items are assumed to be relatively imperfect measures of happiness. Specifically, they are assumed to correlate with each other and these inter-item correlations are explained by a single factor, which we may label "happiness." In path diagrams like the one in Figure 2, factors (or latent variables) are presented as circles or ellipses, and items (or measured variables) are presented as squares or rectangles. The arrows represent the causal direction of the relationships among variables. Specifically, the latent variable is assumed to cause the covariance among items. Therefore, arrows originate from the latent variable towards the measured variables (Hair et al., 2010). Any variance in the items that is not explained by the factor (i.e. the arrows to the right of the items) is assumed to be unique to each item and uncorrelated with any other variable. In Figure 2, the happiness factor is assumed to explain all the covariances among the four items and, after controlling for the factor, the item uniquenesses (or residuals) are assumed to be independent (i.e. uncorrelated). This type of model is known as a covariance-based model, where factors are assumed to cause the responses to items. It is also possible to use variance-based models, which assume that items cause latent variables, but there is currently little evidence supporting their use (Rönkkö, McIntosh & Antonakis, 2015) and they will not be discussed here.

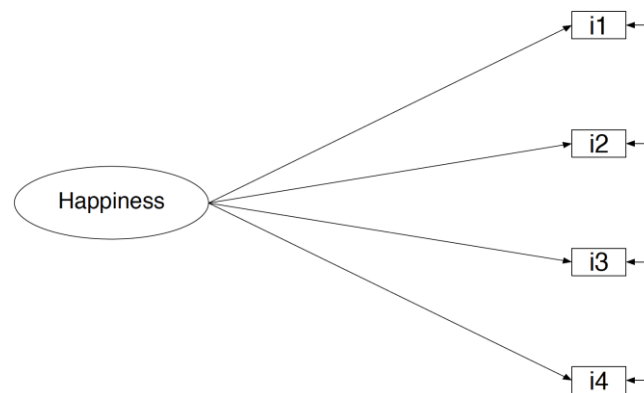


Figure 2 A one-factor model

Each item can be more or less related to other items on the same scale, and the strength of these relationships determines how well that item measures its latent variable. When an item measures a construct well, it will be highly related to other items measuring the same construct. In factor analysis, the strength of association between each item and the broader factor is assessed through factor loadings (λ), also called parameter estimates. Ideally, each item should only measure one construct and thus exhibit a high loading on only one factor. When an item only loads on a single factor (e.g. Figure 2), a factor loading is interpreted as a correlation coefficient or, more precisely, a β (beta) coefficient in a simple regression, where the factor is the predictor and the item is the outcome (Brown, 2015). When an item loads on more than one factor, it is predicted by more than one factor. Therefore, each factor loading is interpreted as a β coefficient in a multiple regression: the association of the item with a factor after controlling for the variance explained by the other factors on which the item loads (Brown, 2015). A loading of .3 or .4 is generally regarded as a minimum requirement to interpret an item as a measure of a factor. Loadings above .5 are preferable, as they indicate that the item shares at least 25% of its variance with the factor. Loadings above .7 are ideal, as they indicate that the item shares at least 50% of its variance with the factor (Hair et al., 2010). When each item exhibits a strong loading only on one factor with loadings near 0 on all other factors, the model is said to approximate a simple structure (Thurstone, 1947). Achieving a simple structure is important because it indicates that each item is a relatively pure measure of only one construct, the overall factor is easily interpretable and, most importantly, it is likely to replicate in a new sample (e.g. Fabrigar & Wegener, 2012).

3.7.2. Exploratory and confirmatory factor analysis. As previously stated, factor analysis is a statistical technique often used to assess the dimensionality of psychometric scales (Fabrigar & Wegener, 2012). Two major versions of this technique are generally used: exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Several differences can be identified between these two types of factor analysis.

First, while EFA is used to develop hypotheses, CFA is used to test hypotheses. EFA is used when the factor structure of the data is not well known: when we cannot anticipate the number of common factors and their relationship to the items (Fabrigar & Wegener, 2012). When it is difficult to specify a hypothetical factor structure, EFA is used to explore models with different numbers of factors and observe which items relate to these factors. In EFA, all items are allowed to load on all of the estimated factors. Results of such exploratory analyses can then be used to generate hypotheses about the structure of the data. These hypotheses can then be tested on a new sample using CFA. When we can formulate one or more hypothetical factor structures for the data, CFA is appropriate. In this case, the researcher specifies both the number of factors and the specific items loading on each factor. Typically, in a CFA we specify that each item will load only on one factor, with cross-loadings (i.e. loadings on further factors) fixed to “0”. This mirrors the expectation that each item should be a pure measure of only one latent variable (Fabrigar & Wegener, 2012). To specify a CFA model, the researcher can use a conceptual outline describing the factors and their respective items, or can refer back to prior exploratory work (Brown, 2015). Although rarely done in practice, we can also use EFA in a slightly more confirmatory way by specifying an a priori target matrix for the EFA loadings and observing whether the pattern of estimated loadings closely resembles our hypothesized target (e.g. Browne, 2001).

In both EFA and CFA it is possible to assess how well the model reproduces the correlations in the data, that is, how well the model fits the data. These procedures are further outlined in section 3.7.6 on model fit. Unfortunately, some well-known statistical packages (e.g. SPSS 21) only report limited information on model fit when conducting an EFA, but this is reported in more depth in specialized statistical packages (e.g. Mplus and the R packages psych and semTools). Model fit assessment in EFA can be used to support or reject the choice of a specific number of factors. For example, a model with two factors may not adequately portray the

relationships in the data, and thus show poor model fit, but a model with three factors may show better fit (e.g. Garrido, Abad & Ponsoda, 2016). In CFA, more sophisticated hypotheses can be tested regarding not only the number of factors, but also the relationships between each factor and the items (Osborne, 2014). As such, CFA is a more powerful tool for testing hypotheses. In CFA, we may free (i.e. estimate) some parameters, while fixing (i.e. not estimate) others to specific values, such as “0”. When parameters are fixed to 0, they are generally not represented in path diagrams. For example, the residual correlations among all items in Figure 2 are fixed to 0, and since they are not estimated, they are not portrayed in the path diagram. In a multi-factor CFA model, cross-loadings are generally fixed to 0. However, given a strong empirical and theoretical justification, we could free (i.e. estimate) some cross-loading or fix any parameter to a non-zero value (e.g. Brown, 2015). On the other hand, EFA allows a more limited set of modifications to the factor model, even though recent developments have expanded the number of modifications applicable to EFA models (Asparouhov & Muthén, 2009). As Paul Kline points out, “many psychologists, trained in the experimental tradition, prefer confirmatory analysis because it is a form of hypothesis testing. They condemn normal exploratory analysis as blind empiricism” (Kline, 2000, p. 145). Indeed, EFA results are more prone to interpretation errors due to sample-specific idiosyncrasies in the data (Fabrigar, Wegener, MacCallum & Strahan, 1999).

This thesis provides a conceptualization of well-being based on previous empirical work. In other words, it is possible to specify a confirmatory model based on specific hypotheses. While EFA plays an important role in research, due to the reasons outlined, as well as the limitations of exploratory approaches, the confirmatory approach was adopted. Additionally, CFA will allow the comparison of hypotheses, such as whether factors are correlated or not, which is not possible with EFA, where the presence or absence of factor correlations produce identical fit to the data. Confirmatory factor analysis involves several steps, which are outlined next.

3.7.3. A side note on communalities and uniquenesses. As a side note, EFA and CFA differ in how their communalities are calculated. The concept of communality is important not only with regards to factor analysis, but it will also be useful later on when discussing score reliability (section 3.8) and sample size requirements (section 3.11). When an item loads on a single factor, its communality (h^2) is simply equal to its squared loading (i.e. λ^2). This represents the percentage of variance in the item

explained by the factor. When an item loads on multiple independent (i.e. uncorrelated) factors, its communality equals the sum of its squared loadings on all factors. For example, if item 1 loads on factor 1 and factor 2, its communality will be the following (Brown, 2015):

$$h^2 = \lambda_{11}^2 + \lambda_{12}^2 \quad (1)$$

However, when an item loads on multiple correlated factors, loadings represent partial regression coefficients and the communality is a function of both loadings (λ) and factor correlation (ϕ) (Brown, 2015):

$$h^2 = \lambda_{11}^2 + \lambda_{12}^2 + 2\lambda_{11}\phi_{21}\lambda_{12} \quad (2)$$

If the communality (h^2) represents the percentage of variance in an item explained by all the factors on which the item loads, the uniqueness ($1 - h^2$) represents the percentage of variance in the item that is not explained by any factor. The concepts of communality and uniqueness will be further explored in the sections below.

3.7.4. Model building and identification. In CFA, the researcher must specify exactly which items load on which factors. Furthermore, for the CFA to produce admissible results, the specified model must meet several mathematical requirements.

First, there has to be a sufficient number of items for each factor. An absolute minimum of three items per factor has been recommended, with four items being an ideal minimum (Brown, 2015). This requirement is based on the fact that item variances and covariances provide the input to estimate the CFA parameters, such as loadings, residual variances, factor variances, factor correlations and any residual correlations.

An example will illustrate this procedure. Suppose we are testing a one-factor model with 4 items. The number of p items provides variances and covariances equal to $[p(p+1)]/2$ (Brown, 2015). Therefore, 4 items will produce 10 elements in the variance-covariance matrix that will serve as input for the CFA. The CFA can then estimate 1 factor variance, 3 unstandardized loadings and 4 residual variances. Therefore, 10 pieces of information in the variance-covariance matrix are used to estimate 8 CFA parameters. In short, there is a greater amount of input information relative to the parameters that need to be estimated in the CFA output. The difference between the pieces of information in the variance-covariance matrix and the

estimated parameters results in the model degrees of freedom (df). In this case, the model has a positive number of degrees of freedom ($df = 2$), and it is said to be overidentified, which is ideal in a CFA. Three unstandardized loadings are estimated instead of four, because one loading is usually fixed to 1 to set the scale for the factor. Factors are hypothetical variables, consequently they have no metric scale. By fixing either one item loading or the factor variance to 1, we are specifying that the standardized loadings can range from -1 to 1 (Hair et al., 2010).

With three items, the input variance-covariance matrix includes 6 input elements, and the CFA would estimate 6 elements: 2 loadings, 1 factor variance and 3 residual variances. When input and output parameters are equal, the degrees of freedom are 0, and the model is said to be just identified. In these cases, parameter estimates (e.g. loadings) will be estimated, but model fit cannot be tested because the model will fit the data perfectly.

With only two items, the input variance-covariance matrix contains fewer elements (3) than are estimated in the CFA (4). In this case, the degrees of freedom are negative, the model is said to be under-identified and the CFA will not produce admissible results. In short, CFA models should be overidentified, and to reach overidentification, a good rule of thumb is to include at least 4 items per factor.

Additionally, even when models are overidentified, it is necessary for the items to be adequately correlated with each other (Brown, 2015). When some variables exhibit correlations near 0 or near 1 with other variables, the CFA may not produce admissible results. Correlations near 0 will produce empirical underidentification, meaning that the CFA may not be able to produce parameter estimates or may provide out of range values (see section 3.7.7). Similarly, when correlations among items are close to 1, they indicate multicollinearity in the data. In other words, one or more variables are redundant and should be omitted from the analysis. Otherwise, the CFA may be unable to generate output and may indicate that the covariance matrix is not positive definite, which indicates that at least one of the variables is redundant, as it can be perfectly predicted by a linear combination of the other variables.

As a rule of thumb, before specifying the model, the researcher should ensure that items exhibit a minimum corrected item-total correlation of .30 (Spector, 2013). An item-total correlation is the correlation between the item and the whole scale including the item in question. The corrected (or adjusted) item-total correlation is the correlation between an item and all other items on the scale (McCoach et al.,

2013). A value below .30 indicates that an item is not sufficiently correlated with other items (Spector, 2013), and may therefore lead to empirical underidentification.

Additionally, factor analysis estimation (section 3.7.5) requires each variable not to be perfectly predicted by combinations of other variables. In other words, if we carried out an iterative regression analysis with each variable as the outcome and all other variables as predictors, the squared multiple correlations (i.e. R^2) should not be close to 1. If the R^2 were close to 1 it would indicate an extreme level of multicollinearity, which may hinder the estimation process. As a rule of thumb, squared multiple correlations should not exceed .90 (Kline, 2016).

3.7.5. Estimation. CFA output can be estimated using several methods. Most methods are iterative: parameter estimates are calculated over multiple cycles, each attempting to improve on previous iterations by producing estimates that fit the data more closely. When improvements from one iteration to the next become smaller than a set criterion, the output is considered to be stable, the factor analysis is said to have converged, and estimation stops (Kline, 2016). Depending on the data analyzed, the choice of estimation method can influence the fit indices and parameter estimates (e.g. loadings).

The most commonly used estimation method in confirmatory factor analysis is the maximum likelihood (ML) estimator (Brown, 2015). It is the default method in many statistical packages (e.g. AMOS, Lavaan, Mplus). ML estimation aims to find parameter estimates that most likely produced the data. It requires large samples and, as samples get larger, ML parameter estimates closely approximate the population parameters (Long, 1983). However, for the parameters to be unbiased, the data must be continuous and have a multivariate normal distribution (Brown, 2015).

If the assumption of multivariate normality is not met, a correction can be applied to the chi-square and standard errors produced by maximum likelihood estimation. Parameter estimates (e.g. loadings) will remain the same as those derived from ML, but the chi-square and its derived fit indices (e.g. RMSEA) as well as standard errors will be adjusted for non-normality (Brown, 2015). In this case, the estimation method is referred to as robust maximum likelihood. For example, methods of robust ML have been proposed by Satorra and Bentler (MLM; Satorra & Bentler 1994) and by Yuan and Bentler (MLR; Yuan & Bentler, 2000). The advantage of using MLR is that it can also estimate models with missing data, which is not possible with MLM (e.g. Beaujean, 2014).

The ML estimator and its derivations require data to be continuous. However, Likert response formats often produce ordinal data, rather than continuous data. Specific methods tailored to ordinal data also exist. These least squares approaches can be used to estimate factor loadings from polychoric or tetrachoric correlations (e.g. Flora & Curran, 2004). Polychoric correlations are used among polytomous items (i.e. with 3 or more response options). Tetrachoric correlations are a special case of polychoric correlations used among dichotomous items (i.e. with 2 response options). These estimate relationships, under the assumption that the observed ordinal variables represent unobserved continuous variables (Flora & Curran, 2004). The unweighted least squares (ULS) estimator identifies parameter estimates that minimize the sum of squared residuals. In other words, it attempts to produce estimates that reduce the difference between the sample covariances and the model-implied covariances (Saris & Gallhofer, 2014). Unlike ML, ULS makes no distributional assumptions, and can therefore work well with non-normal data. However, it requires all observed variables to be measured on the same scale (Long, 1983).

The weighted least squares (WLS) estimator is similar to the ULS, as it estimates loadings that minimize the sum of squared residuals. However, while ULS assigns an equal weight to all correlations, WLS assigns a different weight to different correlations based on the observed variances, covariances and multivariate kurtosis (Kaplan, 2000; Saris & Gallhofer, 2014 p196). The WLS also makes no distributional assumptions (e.g. Kline, 2016), but it is computationally demanding and requires very large samples (e.g. $N = 2,000$) (Hau & Marsh, 2004; Lei & Wu, 2012). WLS is sometimes labeled the asymptotically distribution-free (ADF) estimator (Brown, 2015). A less computationally demanding version of the WLS estimator is known as the diagonally weighted least squares (DWLS) estimator (Christofferson, 1977). This has been found to perform well in smaller samples and, in one simulation, it tended to converge more often than ULS, while ULS provided more accurate parameter estimates than DWLS (Forero, Maydeu-Olivares & Gallardo-Pujol, 2009).

Both the ULS and DWLS estimators have been further improved through chi-square adjustments (Muthén, 1993) similar to those presented by Satorra and Bentler (1994). These adjusted estimators are known as unweighted least squares means and

variances (ULSMV) and weighted least squares means and variances (WLSMV), respectively, and both have been found to perform well with ordinal data (Li, 2014).

Several simulation studies have shown that, when CFA is run on Likert scales with 2 to 4 response options, WLSMV estimation on polychoric correlations outperforms ML and MLR on Pearson correlations (Beauducel & Herzberg, 2006; Rhemtulla et al., 2012). In these cases, ML estimation on Pearson correlations tends to underestimate factor loadings and over reject well-fitting models. Polychoric correlations could be used with ML, but this may lead to higher rates of non-convergence and improper solutions (Lei, 2009; Rigdon & Fergusson, 1991). On the other hand, when using Likert scales with 5 response options or more, ML estimation performs as well as or better than WLSMV (Beauducel & Herzberg, 2006). Similarly, when responses across a 5-point scale slightly deviate from normality, robust ML performs as well as WLSMV (Lei, 2009; Rhemtulla et al., 2012).

In summary, when data are approximately normally distributed and the response format includes at least five categories, MLR can be used on Pearson correlations. When the response format includes four or fewer categories, WLSMV or ULSMV estimators should be used on polychoric or tetrachoric correlations. Since the new questionnaire will use a 5-point response format, if the distributions are approximately normal, the studies will use the MLR estimator.

3.7.6. Model fit. Factor analyses attempt to reproduce a large number of item correlations with a smaller number of factor loadings. In other words, the factor loadings summarize information about the relationships among items.

To evaluate our factor analytic model, we must first assess whether the model as a whole adequately reproduces the item correlations (Hair et al., 2010). If it does, we would conclude that the model accurately portrays (i.e. fits) the data. The goal is not for the model to reproduce the item correlations perfectly, but rather to approximate these relationships closely (Fabrigar & Wegener, 2012).

Model fit assessment can be explained through an example involving residual correlations. In factor analysis, item correlations or covariances are usually used to estimate factor loadings. In turn, factor loadings can be used to estimate correlations among the items, known as model-implied correlations. For example, in a one-factor model, we can estimate the model-implied correlation between item A and item B by multiplying their two factor loadings (e.g. Comrey & Lee, 1992). These model-implied correlations will often approximate item correlations, but the two will rarely

perfectly match. The difference between raw correlations and model-implied correlations are known as residual correlations. These indicate the extent to which the model does not fit the data. If the model fits the data perfectly, raw correlations and model-implied correlations will be identical and, therefore, the residual correlations will be equal to “0”. When the model does not fit the data perfectly, the factor loadings will not perfectly reproduce the item correlations, therefore raw and model-implied correlations will differ somewhat, producing non-zero residuals. Residual correlations are very useful in identifying specific areas where the model does not fit the data. As a rule of thumb, residual correlations above an absolute value of .10 indicate that some modifications may be necessary in a specific area of the model (Kline, 2016). For example, an item with several residual correlations above .10 may need to be omitted or a cross-loading may need to be added to the CFA model.

Before inspecting a residual correlation matrix, we need to assess whether the model as a whole fits the data reasonably well. To do this, we can use a number of model fit indices. Several indices have been proposed, but this section will present only the ones that have performed well in large simulations (e.g. Hu & Bentler, 1999) and have been consistently recommended (e.g. Brown, 2015; Kline, 2016). Each of the following fit indices assesses different aspects of model fit and they all have some shortcomings. Therefore, fit indices should be considered in unison, and together with considerations about sample size, average loadings and model complexity. Furthermore, several rules of thumb have been proposed regarding cut-off values for these fit indices. Indeed, these are approximate cut-off values and model fit should be treated as a continuum, rather than a categorical choice where we either accept or reject a model (Brown, 2015).

The chi-square (χ^2) fit index assesses the difference between the sample and model-implied covariance matrix. Therefore, it increases as residual covariances increase, but also as the sample size and model complexity increase (e.g. Hu & Bentler, 1999). When the model perfectly reproduces the relationships among items, the chi-square will be small and non-significant. However, this becomes less feasible when samples are large and when models become more complex (i.e. with many degrees of freedom). In these cases, even minimal misfit will cause the chi-square to be significant. For example, the chi-square will indicate worse fit as the number of variables per factor increases (Marsh, Hau, Balla, & Grayson, 1998). Unlike the fit

indices outlined next, the chi-square index tests whether the fit of a model is perfect, which is often considered an unrealistic expectation when analyzing social science data (e.g. Fabrigar & Wegener, 2012). Because of this limitation, its significance does not lead researchers to reject a model. Instead, additional fit indices are often reported and interpreted, which assess whether the model closely approximates the relationships among the items. Despite the limitations of the chi-square, this should always be reported as it can be used to estimate additional fit indices (Kline, 2016).

The root mean square error of approximation (RMSEA) is based on the chi-square test. Additionally, it adjusts for the sample size and model complexity. It favors parsimonious models, that is, models that can adequately explain the relationships among variables with as few estimated parameters as possible. Values above .10 indicate poor fit, values below .08 indicate acceptable fit and values below .06 indicate good fit (Brown, 2015). This index also provides a 90% confidence interval, whose width can help determine the precision of the fit estimate (e.g. Fabrigar & Wegener, 2012). When the width of the confidence interval is large, it indicates that the population RMSEA may vary substantially from the sample RMSEA. Confidence intervals tend to be wider when the sample size and the number of degrees of freedom are small. When the upper boundary of the confidence interval is above .10, it indicates that the population RMSEA may be above .10, potentially showing poor fit (Kline, 2016). In this case, we should acknowledge that the model may fit poorly in the population. The main shortcoming of the RMSEA is that, when a model has few degrees of freedom, the RMSEA will tend to over reject it even when it fits the data reasonably well (Kenny, Kaniskan & McCoach, 2015). Because of this, its use is not recommended in models with few degrees of freedom (Kenny et al., 2015).

The comparative fit index (CFI) assesses how well the estimated model fits the data compared to a null model, that is, a model in which all variables are specified as uncorrelated (Hair et al., 2010). Values above .90 indicate acceptable model fit, while values above .95 indicate good model fit (Brown, 2015; Hu & Bentler, 1999). Another closely related fit index is the Tucker-Lewis index (TLI). Its purpose is identical to the CFI. However, while the CFI is normed and thus its values can only range between 0 and 1, the TLI values are not normed. Therefore, the TLI may produce values above 1 (Hair et al., 2010).

The standardized root mean square residual (SRMR) is based on the residual matrix. Conceptually, it represents the average residual correlation (Brown, 2015). One shortcoming of the SRMR is that it does not perform well when analyzing categorical items (Yu, 2002). SRMR values below .08 generally indicate good fit (Hu & Bentler, 1999).

When comparing the fit of two competing nested models, we can use a chi-square difference test (e.g. Hair et al., 2010). A nested model includes a subset of the parameters of a competing model. For example, in a two-factor model where all cross-loadings are fixed to “0” is nested within a two-factor model including an estimated (i.e. free) cross-loading. The chi-square difference would calculate a p-value for the difference between the two models’ chi-squares and their degrees of freedom. If the chi-square difference is significant, we can conclude that adding a cross loading significantly improves the fit of the model. When robust estimators are used (e.g. MLR), the chi-square difference test must also take the adjustment for non-normality into account (Brown, 2015).

An alternative to the chi-square difference test is the Bayesian information criterion (BIC), which can be used to compare competing nested or non-nested models. It is not meaningful to interpret the BIC value of a single model. However, when comparing two models, the one with a smaller BIC is interpreted as the better fitting one (Brown, 2015). Another index with the same purpose as the BIC is the Akaike information criterion (AIC). The main difference between the two is that the BIC favors simpler models (i.e. with fewer parameter estimates) more than the AIC (Brown, 2015). Furthermore, some studies have found that the BIC is more robust to assumption violations than the AIC (Markon & Krueger, 2004; Schwarz, 1978). Both the BIC and AIC are only available with the ML or its related estimators (e.g. MLR).

After assessing global fit through these fit indices, it is also important to identify specific areas of the model, which may not fit the data well (i.e. local fit). This can be done by inspecting the residual correlations described above. Additionally, one can inspect the modification indices (Brown, 2015). These indices suggest model changes that would improve model fit, by reducing the chi-square value. Typically, we would focus on the largest modification indices, as these indicate the modifications that would produce the largest improvement in model fit. Like any exploratory approach, modification indices can be influenced by sampling error. Furthermore, modification indices may suggest the estimation of nonsensical

parameters, such as a covariance between an item residual and a factor, rather than suggesting a simple cross-loading (McCoach et al., 2013). Therefore, any actual model respecification should be guided, not only by the modification indices or correlated errors, but most importantly by a theoretical understanding of the constructs measured (Brown, 2015).

3.7.7. Model respecification. After specifying an overidentified model, this is estimated and we may evaluate its fit to the data, and its parameter estimates. Depending on these results we may need to respecify the model.

Measures of global or local fit may indicate that the model does not fit the data well. In that case, we may respecify the model based on alternative theories, residual correlations and modification indices, as explained above. These respecifications may depend on the purpose behind our specific factor analysis. For example, we may find that one item in our multi-item scale exhibits a cross-loading on another factor. If we are developing a new questionnaire and this includes many items on each factor, we may decide to omit the item with the cross-loading, as it is not a pure measure of its intended factor. However, if the factor analysis aims to examine the theoretical relationships among items and factors, finding a cross-loading may be theoretically interesting, and we may retain that item (Fabrigar & Wegener, 2012).

Items exhibiting low loadings (e.g. .20) or non-significant loadings on their hypothesized factor, may be omitted from further analyses, as they are not adequate measures of that factor. As previously mentioned, loadings should be above .50 and, ideally, above .70 (Hair et al., 2010).

It is also possible that standardized factor loadings or factor correlations may present values outside of their typical range. Like correlation coefficients, factor loadings and factor correlations range from -1 to 1 (Brown, 2015). Loadings that do not conform to this range are known as Heywood cases, and may indicate a number of potential issues: the population parameters may be close to the boundaries (e.g. 1), the model tested may need to be respecified (e.g. with fewer factors), more items per factor may be needed, the distribution of the data may require a different estimator, or the sample size may be too small (de Winter & Dodou, 2012; Fabrigar et al., 1999; Hair et al., 2010). Heywood cases can also appear in the form of negative variances, different loading signs for items scored in the same direction, or very large standard errors (Bentler & Chou, 1987; Hair et al., 2010). Most often, these can be remedied by respecifying the model. For example, one may reduce the number of

factors estimated, omit an item causing the Heywood case, or add more items if fewer than 3 items per factors are being used (de Winter & Dodou, 2012; Hair et al., 2010).

In some circumstances, parameter estimates and factor correlations may indicate that two factors are very highly related. As a rule of thumb, the scales should be somewhat related to each other without any inter-factor correlations above .80 or .85 (Brown, 2015). A very strong correlation between two factors indicates that they may be measuring the same construct, rather than two separate constructs.

This hypothesis can be further tested by comparing the average variance explained by each factor (AVE) relative to the variance shared by the two factors. The AVE is equivalent to the average factor loading squared. Since a loading represents the association between an item and its factor, a squared loading tells us the percentage of variance in the item, explained by the factor. Therefore, the mean of the squared loadings on a factor indicates how much variance in the items is explained by the factor (i.e. AVE). This is compared with the squared correlations between factors. If AVE is greater than all of the factor's squared factor correlations, the factor explains more variance in its items than it shares with other factors (Fornell & Larker, 1981; Hair et al., 2010). In that case, we can conclude that different factors measure distinct constructs.

An alternative method, which has been found to perform well in simulations, is the heterotrait-monotrait method (HTMT; Henseler, Ringle & Sarstedt, 2015). The method focuses on inter-item correlations: HTMT values show the ratio of the inter-item correlations between scales to the inter-item correlations within scales. For example, if the average correlation among items *across* two scales was .30 and the average inter-item correlation *within* two scales was .50, the HTMT value between the two would be $.30 / .50 = .60$. This can be interpreted in the same way as an inter-factor correlation, which takes into consideration the variance explained by each scale. Two cut-off points have been proposed for HTMT values (Henseler et al., 2015). Scales should exhibit HTMT values below 1, which indicates that a scale explains more variance in its items than it shares with other scales and, ideally, HTMT values should be below .85, which echoes the cut-off point for inter-factor correlations proposed by Brown (2015).

Both methods provide preliminary convergent and discriminant evidence of validity (further explained in the validity section 3.9.2). If the variance shared across

factors or scales is greater than the variance explained by a single factor or scale, we may respecify two factors as a single factor. If a single factor adequately explains the relationships among items from two or more scales, we may favor this one-factor model, as it is more parsimonious. In other words, whenever two models of different complexity fit the data adequately, we would often retain the simpler (i.e. more parsimonious) of the two (Brown, 2015).

If our model is altered substantially (e.g. omitting several items, combining factors, etc.), it is necessary to collect data from a new sample and evaluate the model once again in the new sample, to ensure that the model changes were not merely due to sampling variation (Brown, 2015).

3.7.8. Measurement invariance. When we assess the dimensionality of a measure, we usually focus on the entire sample available to us. However, within that sample we may find subgroups, such as different age or gender groups. If the questionnaire aims to compare different groups (e.g. males and females), the factor structure must be invariant across these groups. In other words, each item must measure the same construct across groups, and a given score on that item must reflect the same level of the construct across groups (Spector, 2013). Therefore, after assessing the factor structure of a questionnaire in the entire sample, it is important to test whether the factor and mean structures are consistent across groups within the entire sample. If measurement invariance does not hold, any group comparisons may be biased because a group difference may be due to true score change but also to a change in the structure of the measure.

For example, if we were using a measure of weight, its estimates should vary only when the weight of the objects measured vary. If the measure produced different weight estimates for objects identical in weight, but different in shape, the scale would not be invariant across shapes, and would therefore be biased (Millsap, 2011). Similarly, a measure of conscientiousness should produce the same result for males and females with an identical level of conscientiousness. If this were not the case, the measure would be non-invariant and, therefore, biased in relation to gender.

To assess measurement invariance, we must follow several steps (Brown, 2015). During this process, the researcher specifies a sequence of increasingly constrained models and tests whether model fit decreases substantially from one step to the next. Since several fit indices are influenced by sample size, it is preferable for the group sizes to be approximately equal (Brown, 2015).

First, as a preliminary step, the same CFA model is estimated in each group separately. Model fit should be acceptable in each group before proceeding to the next steps.

Second, the same model is estimated in the groups simultaneously. At this stage, we are evaluating whether the factor structure is consistent in all groups (configural invariance). If model fit is adequate, we can conclude that configural invariance is supported.

Third, each loading is constrained to be equal across the groups. If model fit does not significantly reduce after constraining loadings to equality, it indicates that the items have the same meaning across groups (metric invariance). This level of invariance only permits the comparison of latent variances and covariances across groups.

Fourth, each item intercept is constrained to be equal across the groups. If model fit does not significantly reduce after constraining intercepts to equality, it indicates that a given score on an item reflects the same level of the construct across groups (scalar invariance). This level of invariance allows the comparison of latent means across groups.

Additional steps and constraints can be subsequently added (e.g. equal residual variances). However, these levels of invariance are not necessary to compare scores across groups (Brown, 2015). Therefore, they will not be discussed.

A chi-square difference test may be used to assess changes in fit, but this has been found to reject invariance even when fit differences between models are trivial. Unlike the chi-square, the CFI has been found to accurately distinguish between invariant and non-invariant scales. Specifically, a CFI reduction greater than .002 indicates that the more constrained model fits substantially worse than the less constrained model, leading us to reject invariance (Meade, Johnson & Braddy, 2008).

3.7.9. Summary. Factor analysis can be used to explore (EFA) or test (CFA) the dimensionality of a questionnaire. CFA involves the specification of an overidentified model, the estimation of such a model, the assessment of its fit to the data, and model respecification when the model is not adequate. After identifying a model that fits the data well, we can proceed to the next step in scale development, which consists in examining the amount of random error present in our questionnaire responses.

3.8. Reliability

Reliability is defined as the “consistency of scores across replications of a testing procedure” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 33). Different items measuring the same construct, or the administration of the same scale on different occasions, should produce consistent (i.e. correlated) scores. When talking about score reliability, one may refer to the consistency of item scores at a single time point (i.e. internal consistency), the correlation between scores on the same scale at two different time points (test-retest), or the correlation between two versions of the same scale administered at a single time point (parallel form). These three types of reliability will be further explained next.

3.8.1. Internal consistency. After using factor analysis to identify the specific number of constructs measured by the questionnaire, we can assess the amount of random error present in the measurement of each construct. According to classical test theory, the scores on a questionnaire are influenced by respondents’ latent score (or true score) as well as unique variance (e.g. Kline, 2000). The latent or true score is the participant’s hypothetical score on a measure if this was administered over an infinite number of replications (AERA et al., 2014). Unique variance consists of score variations that are not due to changes in one’s latent score (e.g. McCoach et al., 2013). It consists of variance that is not shared among items, and this includes specific and random variance (e.g. Kline, 2000). Specific variance is variance that an item may not share with other items on the same scale, but may share with other scales and variables. Random variance is variance that is uncorrelated with any variable on any scale. Unique variance, which includes both specific and random variance, is often assumed to be random and is generally labeled error variance, uniqueness, or residual variance (e.g. Kline, 2000). Unique variance may be caused by fatigue, lack of attention or motivation, vague questionnaire instructions, time of day, or distractions (AERA et al., 2014). In summary, the total variance in a scale is the sum of its latent variance and its unique variance.

If we repeatedly measured the length of a table with a standard ruler, repeated measurements would be highly consistent with each other, because each estimate would be primarily influenced by the actual length of the table (i.e. true score variance). Therefore reliability would be high. However, if we tried to measure the

length of a table using an elastic or a ruler with no unit labels, repeated measures of the same table will be more inconsistent because each measurement would be influenced both by the length of the table as well as the imprecision of our measure, which would increase the amount of error variance in our measurement. Therefore, reliability would be low.

Mathematically, internal consistency is defined as the ratio of latent score variance to the total variance, which includes latent variance and error variance (e.g. Kline, 2000). Therefore, the lower the amount of error variance in our data, the higher the reliability of the scores produced by our measure.

Reliability increases as the item covariances and the number of items in the questionnaire increase (e.g. Kline, 2000). High inter-item correlations indicate that items on a scale measure the same construct. In an extreme scenario, where all items correlate perfectly, they would all perfectly measure the same construct. Therefore, the latent variance would equal the total variance in the scale. In this case, the scale scores would include no error variance, and would be perfectly reliable. However, such an extreme scenario is rare and may not be desirable, as explained in section 3.10 on the relationship between reliability and validity.

Scales with more items tend to be more reliable because, when multiple items are used to measure a construct, the idiosyncrasies inherent to each item will cancel each other out. Similarly, if we were measuring time or temperature with a single instrument, any idiosyncrasy or flaw in that instrument would bias our measurement. Conversely, calculating an average estimate using multiple instruments would reduce the amount of error variance, and thus increase reliability. However, as the length of the questionnaire increases, respondents may be unwilling to complete all the questions, or fatigue may increase the error variance in the responses (AERA et al., 2014). Therefore, it is necessary to find a balance between the scale length necessary for reliability and respondent motivation and fatigue.

The number of items needed to achieve adequate reliability will also depend on the breadth of the construct we are measuring. Items measuring broader constructs (e.g. well-being or intelligence) will be less similar to each other and will therefore exhibit lower inter-item correlations. In that case, more items will be needed to achieve adequate reliability (e.g. Clark & Watson, 1995; Reise et al., 2015). On the other hand, items measuring a narrower construct (e.g. happiness, numerical ability)

will be more similar to each other and will therefore exhibit higher inter-item correlations. In that case, fewer items will be needed to achieve adequate reliability.

How high should reliability be? Reliability values can vary from 0 (lowest) to 1 (highest). As a rule of thumb, values of .70 or higher are considered acceptable for research purposes (Kline, 2000), while values of .80 or higher are considered good (e.g. DeVellis, 2012). Additionally, when the interpretation of a measured score carries important consequences, a greater level of precision is required (AERA et al., 2014). For example, when results on a test are used to determine the admission or rejection in a university or the hiring decisions of a company, a higher level of reliability is necessary, and values above .90 are warranted (e.g. DeVellis, 2012). Such a high reliability would indicate that random error has little influence on the scores. As error is reduced, the confidence intervals of our estimates (e.g. correlations) become narrower. Thus, high reliability, like a large sample size, increases statistical power: the ability of a scale to detect significant relationships (DeVellis, 2012).

As previously mentioned, reliability is influenced by the amount of random error in our scores, and random error is influenced by item characteristics (e.g. ambiguity), respondent characteristics (e.g. fatigue), as well as contextual characteristics (e.g. distractions). Therefore, reliability is clearly a property of the scores, rather than a property of the measure. That is, the same scale can produce different reliability estimates depending on the sample and the context of administration. In other words, we may state that reliability is a property of the scale administered to a specific population in a specific context and, as such, it should be estimated whenever the scale is administered to a new sample (e.g. AERA et al., 2014; Osborne, 2014).

Before describing current reliability estimates, it must be reiterated that dimensionality should be assessed before testing reliability (Osborne, 2014). Reliability coefficients such as Cronbach's alpha do not assess dimensionality (e.g. Schmitt, 1996), and a lack of unidimensionality can bias these reliability coefficients (DeVellis, 2012).

Cronbach's alpha (α) is perhaps the most frequently used measure of reliability (Graham, 2006). It is estimated based on item variances and covariances. For example, the following equation shows one way of estimating Cronbach's α , based on formula 16 presented by Cronbach (1951, p. 304):

$$\alpha = \frac{k^2 \overline{Cov_{ij}}}{Var_t} \quad (3)$$

where, k^2 is the number of items squared, $\overline{Cov_{ij}}$ is the mean inter-item covariance and Var_t is variance of the composite test score. In other words, the coefficient α will increase as the number of items increases and as the average covariance among items increases. One limitation of Cronbach's α is that the coefficient may increase even when adding items, which are only moderately related to other items on the scale, and this can give a false appearance of reliability (Schmitt, 1996). Because of this, when the number of items exceeds 40, Cronbach's α is not very informative (Clark & Watson, 1995).

Cronbach's α accurately estimates reliability when the scale is unidimensional and all items measure the target construct equally well. From a factor analytic perspective, the second assumption implies that all items must have equal loadings on their factor, a condition known as essential tau-equivalence (Graham, 2006). Calculating a single alpha for multidimensional scales (i.e. violating the unidimensionality assumption) would not be appropriate and may lead to the overestimation of reliability (e.g. Cortina, 1993). On the other hand, large variation in item loadings (i.e. violating the essential tau-equivalence assumption) can lead to the underestimation of reliability, which is particularly problematic because this assumption is often not met in real world data (Raykov, 1997; Graham, 2006).

One reliability estimate that does not assume essential tau-equivalence, is McDonald's omega total (ω_t) (McDonald, 1999), which has been found to produce more accurate reliability estimates than α (e.g. Revelle & Zinbarg, 2009). This estimates reliability through the common factor model, presented in the dimensionality section above. For example, in a one-factor model, the scale's latent variance is equal to the sum of item loadings squared, while the unique variance is equal to the sum of uniquenesses (i.e. $1 - h^2$). The sum of latent and unique variances is equal to the total variance. Therefore, the reliability estimate (ω_t) is once again equal to the ratio of latent variance to the total variance (e.g. McDonald, 1999; Rodriguez, Reise & Haviland, 2016):

$$\omega_t = \frac{(\sum \lambda_j)^2}{[(\sum \lambda_j)^2 + \sum (1 - h^2)]} \quad (4)$$

where $(\sum \lambda_j)^2$ is the sum of loadings squared and $\sum(1 - h^2)$ is the sum of uniquenesses. However, latent variance may arise from a unidimensional as well as a multidimensional construct. Even scores on narrow scales (e.g. numerical ability) may exhibit some degree of multidimensionality (e.g. Revelle & Condon, in press). In that case, McDonald's ω_t would consider such multidimensionality as part of the scores' latent variance. Therefore, multidimensionality would increase the total reliable variance. For example, in a bifactor model with one general and several group factors (e.g. Rodriguez et al., 2016):

$$\omega_t = \frac{(\sum \lambda_{gen})^2 + \sum_{k=1}^K (\sum \lambda_{grp k})^2}{(\sum \lambda_{gen})^2 + \sum_{k=1}^K (\sum \lambda_{grp k})^2 + \sum(1 - h^2)} \quad (5)$$

where $(\sum \lambda_{gen})^2$ is the sum of loadings squared for the general factor and $\sum_{k=1}^K (\sum \lambda_{grp k})^2$ is the sum of loadings squared for the each group (or specific) factor. This would not be as informative as a reliability index, which accounts for such multidimensionality. In other words, rather than focusing on the total reliable variance, we may focus on the percentage of variance that reflects only one construct (McDonald, 1999; Revelle & Zinbarg, 2009). McDonald's omega hierarchical (ω_h) coefficient meets precisely this need, as it estimates the proportion of variance due only to a single factor. That is, it estimates general factor saturation: within a bifactor model, it estimates the amount of variance due only to the general factor (Revelle & Condon, in press; Rodriguez et al., 2016).

$$\omega_h = \frac{(\sum \lambda_{gen})^2}{(\sum \lambda_{gen})^2 + \sum_{k=1}^K (\sum \lambda_{grp k})^2 + \sum(1 - h^2)} \quad (6)$$

In other words, the coefficients alpha (α) and omega total (ω_t) estimate the reliable variance in a scale due to all sources of common variance, and could potentially be inflated when a scale exhibits some degree of multidimensionality. On the other hand, omega hierarchical (ω_h) estimates the proportion of variance due only to a single factor. Therefore, high values of ω_h would require the scale to closely approximate a unidimensional model: where the correlations among items are due primarily to a single general factor (e.g. Rodriguez et al., 2016). While adding poor items (i.e. items loading strongly on factors other than the general factor) may inflate the coefficients alpha and omega total, this should not inflate omega hierarchical.

For example, let us consider two multidimensional questionnaires with 15 items each. For each questionnaire, the items reflect both a general factor as well as a

specific factor. The two hypothetical matrices with factor loadings are presented in Table 3. On the left, we find a matrix with a strong general factor and three specific factors. On the right, we find a matrix with a weak general factor and the same three specific factors.

Table 3 Example of a reliability analysis

Variables	Example 1: Strong g				Example 2: Weak g			
	g	F1	F2	F3	g	F1	F2	F3
V1	0.70	0.60	0.00	0.00	0.20	0.60	0.00	0.00
V2	0.70	0.60	0.00	0.00	0.20	0.60	0.00	0.00
V3	0.70	0.60	0.00	0.00	0.20	0.60	0.00	0.00
V4	0.70	0.60	0.00	0.00	0.20	0.60	0.00	0.00
V5	0.70	0.60	0.00	0.00	0.20	0.60	0.00	0.00
V6	0.70	0.00	0.50	0.00	0.20	0.00	0.50	0.00
V7	0.70	0.00	0.50	0.00	0.20	0.00	0.50	0.00
V8	0.70	0.00	0.50	0.00	0.20	0.00	0.50	0.00
V9	0.70	0.00	0.50	0.00	0.20	0.00	0.50	0.00
V10	0.70	0.00	0.50	0.00	0.20	0.00	0.50	0.00
V11	0.70	0.00	0.00	0.40	0.20	0.00	0.00	0.40
V12	0.70	0.00	0.00	0.40	0.20	0.00	0.00	0.40
V13	0.70	0.00	0.00	0.40	0.20	0.00	0.00	0.40
V14	0.70	0.00	0.00	0.40	0.20	0.00	0.00	0.40
V15	0.70	0.00	0.00	0.40	0.20	0.00	0.00	0.40

Note: V= variable; g= general factor; F= specific (or group) factor; non-zero loadings in bold.

In the first case (i.e. strong g), Cronbach's alpha and McDonald's omega coefficients suggest adequate reliability ($\alpha = .95$; $\omega_t = .97$, $\omega_h = .83$). This scale is not perfectly unidimensional, which inflates the α and ω_t coefficients. However, the ω_h shows that the general factor accounts for most of the variance in the items. In other words, scores on the first scale reflect primarily one construct. In the second case (i.e. weak g), Cronbach's alpha and McDonald's omega coefficients suggest nearly adequate reliability ($\alpha = .66$; $\omega_t = .73$, $\omega_h = .23$). If we did not assess the dimensionality of this scale or its ω_h , we may even interpret the α and ω_t values as suggesting approximately adequate reliability. However, the ω_h clearly indicates that a single score would not adequately capture the variance among these fifteen items. Instead, scale scores corresponding to the three specific factors should be used. In short, the difference between the two scale reliabilities is clearly captured by ω_h , but not as much by the other reliability estimates. In this thesis, ω_h was estimated using the omega function in the psych package in R.

3.8.2. Temporal stability. Temporal stability (or test-retest reliability) is an estimate of the consistency of a scale score under the same conditions at different points in time. In other words, it assesses whether a construct is stable or fluctuates over time. Dispositional constructs (e.g. personality traits) are typically expected to be stable over time. Therefore, scores on a personality measure should exhibit high test-retest reliability. On the other hand, testing temporal stability may be inappropriate with constructs that vary over time. For example, we do not need to measure test-retest reliability for transient emotions, which are by definition bound to have low temporal stability (Kline, 2000).

Temporal stability is important when researching changes in scores across time. For example, in an intervention study, changes in scores before and after the intervention should be due to changes in the actual construct. Such changes can be inferred if the construct is stable over time. Otherwise, changes in scores may be due to the lack of stability in our measurement (McCoach et al., 2013).

Test-retest reliability is assessed by correlating scale responses from the same sample once (i.e. Time 1) and then again after a few weeks, months or years (i.e. Time 2). While cognitive measures tend to exhibit very high temporal stability (e.g. around .90), affective measures often produce lower reliabilities (e.g. around .70) (McCoach et al., 2013). Additionally, test-retest reliability will be lower as the time gap between administrations gets larger and when changes in the latent score take place between time 1 and time 2 (Kline, 2000).

3.8.3. Parallel form. Sometimes, there may be two versions or forms of the same test with slight differences in wording or order of the questions. In this case, we can assess the equivalence between the two forms. In other words, we can test the parallel form reliability. This is assessed through a correlation analysis between the two forms, with correlations around .90 being indicative of adequate reliability (Kline, 2000). In practice, parallel forms are rarely used because their development and validation requires twice the amount of work than creating a single scale and it is often preferable to create a better test by combining the best performing items of both forms (Rust & Golombok, 2009).

3.9. Construct validity

Construct validity is generally defined as the degree to which a questionnaire measures what it claims to measure (e.g. Rust & Golombok, 2009). More

specifically, it can be defined as “the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests” (AERA et al., 2014, p. 11). Construct validity is a unitary concept. Classic psychometric texts (e.g. Kline, 2000) often talk about different types of validity (e.g. content validity, criterion validity, etc.). However, recently, these have been conceptualized, not as different types of validity, but rather as different types of evidence supporting construct validity, which is a single concept (e.g. Brown, 2015; Spector, 2013).

Like dimensionality and reliability, construct validity is not a property of the measure, but rather of its use and interpretation in a specific context. Strictly speaking, validation studies do not establish the validity of a questionnaire. Rather, they provide evidence supporting its use and interpretation in a specific context (AERA et al., 2014; McCoach et al., 2013). Therefore, construct validity cannot be proven and it is never truly attained. Rather, we can only find evidence for or against the use of a scale (e.g. Kline, 2000). Validation is an ongoing process of testing the performance of a questionnaire and collecting evidence supporting or undermining its use in specific contexts.

A statement supporting the use of a psychometric measure should include evidence from a variety of sources (Spector, 2013). These sources can be grouped into two main categories: content evidence and criterion evidence (Furr, 2011).

3.9.1. Content evidence. Evidence based on the instrument’s content can be gathered even before we test its dimensionality and reliability (e.g. McCoach et al., 2013). Content evidence indicates whether the items tap into the entire range of the construct, and whether they are unambiguous (e.g. Spector, 2013). It is important to avoid construct underrepresentation, that is focusing only on a very specific aspect of the construct, and construct irrelevant variance, that is items tapping into constructs beyond the one under focus (Messick, 1995). Quantitative and qualitative information is gathered from individuals with expertise on the subject. Specifically, experts are usually asked to rate how well each item matches the definition of the construct. Therefore, the assessment of the content of an instrument requires clear definitions in order to compare those definitions to the items (McCoach et al., 2013). A quantitative scale can be used to assess the correspondence between the items and the construct definition. Additionally, qualitative questions can then gather suggestions regarding item ambiguity, construct underrepresentation or construct irrelevant variance (McCoach et al., 2013).

In extremely exploratory research, where construct definitions are not available, it may not be possible to gather content evidence of validity (Kline, 2000). Additionally, it must be noted that expert judgments may be wrong from time to time (McCoach et al., 2013), and these interpretations may be different from those of the respondents who will ultimately use the test. Therefore, we normally look for an approximate consensus among expert judges. For example, as a rule of thumb, a 75% agreement about the correspondence between the item and its construct definition is considered a minimum cutoff point to retain an item (Hardesty & Bearden, 2004). Following the assessment from content experts, the scale developer can use this feedback to revise the questionnaire and its conceptualization, where necessary. For example, the scale developer may revise construct definitions, modify items or omit them from further analyses.

“Face evidence” of validity is a strand of evidence also related to the questionnaire’s content. It tests whether the construct being measured is apparent to the respondents (Spector, 2013). While content evidence is gathered from expert judgments, “face evidence” is gathered from the respondent population. For example, during the validation of a student motivation scale, we may ask several students to assess whether the items reflect the proposed definition of student motivation. In other words, they evaluate whether the items are transparent in revealing what they measure. The benefit of a high level of transparency is that it may increase respondents’ cooperation with the study (Rust & Golombok, 2009). However, transparency may not be desirable in some context, such as employee selection, where respondents would likely attempt to cheat or distort their attitudes (Kline, 2000). In any case, this type of evidence is more relevant to attitude and personality measures, rather than ability measures, where only one correct answer can be provided (Rust & Golombok, 2009).

While it is not necessary for respondents to know exactly which constructs are being measured, it is important for them to understand the meaning of the items clearly. Expert feedback can provide some information regarding the clarity of the items. Additionally, it is possible to interview a small group of individuals from the respondent population to explore their interpretations of each item. This is known as cognitive interviewing (e.g. DeVellis, 2012). It is generally carried out before the questionnaire is administered to a large sample, and it involves asking respondents to verbally express their thought process while interpreting and answering each

question. It is particularly useful when dealing with new constructs and new fields of research, or when a questionnaire is being adapted in a population for whom the meaning may not be clear. For example, a questionnaire developed for an adult population may include items that are difficult to understand in an elementary school sample. In this case, cognitive interviewing could help identify problematic items and help develop simpler ones. When cognitive interviewing is not necessary or feasible, it can still be helpful to ask respondents to indicate whether any items are ambiguous or unclear, during the first questionnaire administration (DeVellis, 2012). For example, one study replaced or rewarded items if 2% or more respondents rated them as confusing (Greco, Baer & Smith, 2011).

In summary, it is a good practice to assess the adequacy of questionnaire content through expert review and to evaluate item clarity for the respondents either through cognitive interviewing or through probing questions in the first administration of the questionnaire. Despite the importance of content evidence, it is generally not considered to be sufficient to support the interpretations drawn from questionnaire scores. To strengthen the validation argument, criterion evidence is also necessary (Kline, 2000).

3.9.2. Criterion evidence. Criterion evidence involves correlating each construct measured by the questionnaire with measures of similar and different constructs (McCoach et al., 2013). These comparison measures (i.e. criteria) should already be supported by previous validation studies.

Criterion evidence is typically gathered after dimensionality and reliability have been assessed. If dimensionality is not satisfactory, the meaning of each construct measured is in question. Therefore, it makes little sense to proceed to gathering criterion evidence. Rather, further dimensionality work is typically carried out before proceeding to this stage of test validation (McCoach et al., 2013). Additionally, once a stable and unambiguous factor structure is identified, criterion evidence can help further assess the dimensionality of the scale.

Criterion evidence supporting the use of a questionnaire would show a pattern of correlations where the questionnaire is highly related to measures of similar constructs (i.e. convergent evidence) and it is poorly related to measures of different constructs (i.e. discriminant evidence). For example, the short flow scale and the core flow scale both measure the flow experience (i.e. a state of deep involvement and absorption with the task at hand), but they do so from different perspectives: the

short flow scale includes several *correlates* of flow while the core flow scale focuses strictly on the *experience* of flow. The two scales were found to correlate around .70 (Martin & Jackson, 2008). If both scales focused on experiential flow, we may expect them to exhibit higher correlations, perhaps greater .80. As a rule of thumb, correlations above .80 provide compelling evidence suggesting that two questionnaires measure the same construct (Brown, 2015). Conversely, a modest correlation, say of .20, would provide strong evidence against the interpretation of the two questionnaires as measuring the same construct.

Discriminant evidence can be gathered by correlating the questionnaire with measures that it should not be strongly correlated with. For example, a measure of well-being is not expected to correlate highly with a measure of the dark triad personality traits (i.e. narcissism, psychopathy and Machiavellianism) (e.g. Aghababaei & Blachnio, 2015). Therefore, a small correlation with dark triad traits (e.g. .10 or .20) would provide discriminant evidence of validity.

In a multidimensional scale, both convergent and discriminant evidence could also be gathered in the following way. If we are researching three constructs (A, B, and C), we may develop three scales to measure them (A₁, B₁, C₁) and identify three previous scales measuring similar constructs (A₂, B₂, C₂). If participants complete the new scales and the previous ones (A₁, B₁, C₁, A₂, B₂, C₂), the correlation between each respective scale (e.g. A₁ and A₂) should be higher than the correlation between different constructs (e.g. B₁ and A₂). These results would provide evidence of both convergent and discriminant validity (see Table 4). In this scenario, a significance test of the differences between correlations would be particularly useful. For example, Steiger's (1980) test assesses whether the correlation between similar measures (A₁ and A₂) is significantly higher than the correlation between different measures (e.g. B₁ and A₂), while taking into account the correlation between the two measures in the same questionnaire (e.g. A₁ and B₁).

Table 4 Hypothesized convergent and discriminant correlations

	A₁	B₁	C₁
A₂	Higher	Lower	Lower
B₂	Lower	Higher	Lower
C₂	Lower	Lower	Higher

Criterion evidence can also be gathered by testing differences in scores among known groups (McCoach et al., 2013). For example, if we are validating a depression

questionnaire, and we administer it to a sample of the general population and a sample of individuals diagnosed with clinical depression, the latter group would be expected to exhibit a significantly higher depression score. This may be tested, for example, through an independent t-test, and such a result would provide criterion evidence of validity.

Criterion evidence is usually gathered at a single time point. That is, the new questionnaire and its criteria are measured concurrently, thus providing concurrent criterion evidence. However, if the questionnaire aims to predict future events, these criteria may be measured at a later time from the administration of the new questionnaire. Then, predictor and outcome variables are assessed as usual through correlation or regression analyses (McCoach et al., 2013).

If the new questionnaire predicts external criteria, after controlling for previous measures of the same construct, this would provide incremental evidence of validity (Kline, 2000). For example, if a new mental ability measure predicts more variance in work performance than a previous mental ability measure, it may be favored over its predecessor. Regression analysis may be used in this case. However, if the two predictors are highly related, that is if the new questionnaire is strongly correlated to the alternative measure of the same construct, this high collinearity can bias the regression estimates. For example, when predictors correlate at .70 or more, beta coefficients will be biased (Hair et al., 2010). In this case, a correlation analysis may be more adequate (Hair et al., 2010). Once again, Steiger's test may be used to assess whether the correlation between the new measure and the outcome is significantly higher than the correlation between previous measure and the outcome.

It has been argued that criterion evidence suffers from a potential shortcoming. It is possible to inflate relationships between predictor and outcome by increasing the similarity between the two. If the new predictor measure is more similar to the outcome measure than the previous predictor measure, then it will show stronger criterion evidence of validity (Rossiter, 2002). For example, we may be conducting research on a theory stating that self-efficacy predicts happiness. After reviewing the literature, we encounter an existing self-efficacy scale, with items such as "I feel competent", and a happiness scale, with items such as "I feel happy." We can then develop a new self-efficacy scale with apparently stronger criterion evidence than the original one, simply by writing items such as "I'm happy about my level of competence." While the new scale would predict the theorized outcome better than

the previous one, it would indeed be less valid because it would likely be a less precise measure of self-efficacy. Therefore, it has been argued that criterion evidence of validity would be more adequate if we already knew the correlation between the two constructs in the population (Rossiter, 2002). For instance, if we knew that self-efficacy correlated with happiness at .50 in the population, and we found a correlation that is much higher (e.g. .80) or lower (.20) we could question the strength of the criterion evidence of validity. However, that information is almost never available (Rossiter, 2002), which is why criterion evidence only provides a limited amount of information that needs to be considered together with other validation evidence (e.g. content evidence).

3.10. The relationship among dimensionality, reliability and validity

Within the context of psychometrics, the aim is to develop questionnaires and provide evidence supporting the dimensionality, reliability and validity of their scores within a specific context. When the scores on the instrument show good dimensionality, we may infer that it likely measures a certain number of constructs. When these factors show high reliability, we may infer that the constructs are measured with small amounts of unique (i.e. residual) variance. When the items are deemed adequate by an expert panel, and their scale scores correlate to external variables in a way that is consistent with their hypothesized relationships, we may infer that the scales likely measure the constructs they are supposed to measure, within circumstances consistent with the current study.

A reliability analysis, using Cronbach's α or McDonald's ω^2 , bears little meaning if the dimensionality of the scale is inadequate (e.g. Furr, 2011). Similarly, factor and reliability analyses would not provide information regarding what constructs we are measuring but, in an ideal scenario, they would indicate that we are adequately measuring a certain *number* of constructs. As such, adequate dimensionality is necessary to assess reliability. If we do not know the factor structure of the data, we cannot adequately assess the presence of random error within that structure. In short, adequate dimensionality is necessary but not sufficient for reliability.

Similarly, adequate reliability is necessary to assess construct validity (e.g. Kline, 2000). If most of the variance in a scale is random variance, any correlation to external criteria (i.e. validity coefficient) will be highly imprecise and attenuated by

the random error. In other words error variance will reduce correlations between scales, thus undermining criterion evidence of validity. However, high reliability does not guarantee that a scale will show a correlation pattern to external criteria that is consistent with the hypothesized correlation pattern. Even if a scale shows good dimensionality and reliability, it may be measuring a different construct than what it is supposed to measure. Therefore, adequate reliability is necessary but not sufficient for validity.

Furthermore, as mentioned in the reliability section above, we must consider the breadth of the construct we attempt to measure (Reise et al., 2015). A very specific construct (e.g. high-arousal positive affect) may be adequately measured with as few as four items, and this scale may be highly reliable (e.g. ω_t or $\omega_h > .80$). On the other hand, questionnaires measuring broad constructs may need to include several items and several subscales to reach adequate reliability. For example, well-being and general mental ability are relatively broad constructs. If a short four-item scale aimed to measure general well-being and this scale had a reliability of .95, perhaps we may question its validity. Such a high reliability would suggest that the items do not probably cover the entire range of the construct and, instead, focus on a very specific aspect of well-being (e.g. high-arousal positive affect). Similarly, it is possible to achieve high reliability by writing items that are nearly identical to each other (Clark & Watson, 1995). These items will correlate very highly with each other, thus producing high reliability estimate, but they will also result in a highly redundant scale, known as a bloated specific (Cattell, 1978). Including these redundant items will increase reliability, but it will not improve the correlations with external criteria. In other words, additional redundant items will not increase the predictive power of the scale, which will be approximately equivalent to using a single item (Clark & Watson, 1995).

In summary, adequate dimensionality is necessary but not sufficient for reliability and reliability is necessary but not sufficient for validity.

3.11. Sample size requirements

After developing an overidentified factor model, one may start to collect data to assess the dimensionality, reliability and validity evidence supporting the use of the questionnaire. Factor analysis is considered to be a large-sample technique (e.g. Kline, 2016). While a sample of 100 participants may sometimes be adequate for a

regression analysis (e.g. Green, 1991), in factor analysis this would usually be considered inadequate or an absolute minimum (Kline, 2016). As sample size increases, sampling error is reduced. In other words, with larger samples, the difference between the sample parameter estimates and the population parameters becomes smaller. Therefore, as sample size increases, factor analysis is more likely to produce stable estimates that closely approximate the population parameters (Jackson, 2001). However, how large should a sample be? The answer to this question is complex and needs to be explored from a variety of perspectives.

As a rule of thumb, a sample of 200 is regarded as a minimum requirement for CFA, with larger samples often recommended (Barrett, 2007; McCoach et al., 2013). While rules of thumb such as this one may be arbitrary, simulation studies provide the following evidence-based guidelines.

3.11.1. Simulation studies. First, sample size must increase as the size of the communalities or factor loadings decrease. Of course, communalities are based on the factor loadings, which are based on the inter-item correlations. Therefore, as the relationships among the items become stronger, factor loadings and communalities increase, and as they increase, smaller samples can be used in factor analysis. For example, one EFA simulation found that in a model with 6 factors, 4-items per factor, and high communalities (range $h^2 = .60 - .80$), a sample of 170 observations was sufficient to closely approximate the population loadings. However, if the same model had low communalities (range $h^2 = .20 - .40$), a sample of 400 observations was required (Mundfrom, Shaw & Ke, 2005). These results are consistent with findings from Hogarty, Hines, Kromrey, Ferron & Mumford, (2005) and MacCallum, Widaman, Zhang & Hong (1999). Similarly, in CFA simulations, population parameters were more closely approximated when loadings were higher (Jackson, 2001). Furthermore, when factor loadings were low ($\lambda = .40$), and sample size was small ($N = 100$), the convergence rates were very low (e.g. below 30%). With such low loadings, increasing the sample size to 250 or 500 observations increased convergence rates. However, when factor loadings were high ($\lambda = .80$) convergence rates were high even with small samples (Enders & Bandalos, 2001). In short, with high communalities, smaller samples are needed to achieve convergence and produce estimates that closely approximate the population values.

We can draw the same conclusion if we look at the ratio of the number of observations per item or the number of observation per parameter estimated. For

example, in an EFA simulation, when average communalities were high ($h^2 \geq .60$), accurate parameter estimates were produced with as few as 3 observations per item. However, when communalities were low ($h^2 \leq .40$), 40 or more observations per item were required (Hogarty et al., 2005; MacCallum et al., 1999). In CFA, increasing the ratio of observations to estimated parameters was found to slightly improve the stability of parameter estimates and fit indices (Jackson, 2003). However, with large enough samples ($N = 400$) and high loadings ($\lambda = .80$), further increases in the observations to estimates ratio made little difference (Jackson, 2003). It is difficult to predict the average item loadings and communalities. However, as mentioned in the reliability section, inter-item correlations, and therefore communalities, tend to increase when measuring relatively narrow constructs, like numerical ability, rather than broad constructs, like general mental ability (e.g. Reise et al., 2015).

Second, sample size must increase as the number of variables per factor decreases. For example, in an EFA simulation presented above, a six-factor model with high communalities (range $h^2 = .60 - .80$) and 12 items per factor required a sample size of 75 observations to adequately approximate population loadings; but, with 4 items per factor it required 170 observations; and with only 3 items per factor it required 350 observations (Mundfrom et al., 2005). In short, specifying a model with fewer than 4 indicators seems to require particularly large samples. In CFA, increasing the number of indicators per factor was also found to improve the stability of the parameter estimates (Jackson, 2001). Additionally, including many indicators per factor was found to increase convergence rates (Marsh et al., 1998). However, when the sample size reached 400 or more observations, convergence rates were almost identical when including 3, 6, or 9 variables per factor (Marsh et al., 1998). In short, a higher number of variables per factor improves parameter estimation and convergence rates. Yet, in a large enough sample (e.g. $N = 400$), the number of variables per factor does not seem to have as much of an impact.

Third, sample size requirements may differ depending on the chosen estimator. In the CFA studies presented above, maximum likelihood estimation was generally used (e.g. Jackson, 2001). With moderate and severe non-normality, the Satorra-Bentler chi-square (MLM estimation) was found to produce lower type I errors than the default ML chi-square in small ($N = 100$) to large ($N = 1,000$) samples (Curran et al., 1996; see also Hu, Bentler, & Kano, 1992). The Yuan-Bentler chi-square (MLR estimation) has also been found to have acceptable rejection rates in moderate

sample ($N > 200$) (Savalei, 2010). Since other estimators (e.g. WLS, WLSMV) are not used in the thesis, considerations regarding the sample size necessary for these estimators are not discussed here.

3.11.2. Power analysis. In addition to the guidelines from simulation studies above, power analysis procedures have been developed to determine the sample size necessary to test the fit of a model (e.g. MacCallum, Browne & Sugawara, 1996). The goal is to identify the sample size necessary for a fit index to distinguish between a model exhibiting poor fit and a model exhibiting good fit.

The RMSEA fit index was proposed to conduct a power analysis because its distributional properties are known, and these permit the use of confidence intervals and hypothesis testing, which are used in power analyses (MacCallum et al., 1996). Specifically, two opposite and competing hypotheses are considered: the null hypothesis and the alternative hypothesis. With real-world data, the null hypothesis generally states that the model fits the data closely (close fit), while the alternative hypothesis states that the model fits the data poorly (poor fit).

An $RMSEA \leq .05$ generally indicates close fit, while an $RMSEA \geq .08$ or $.10$ generally indicates mediocre fit (MacCallum et al., 1996). With that in mind, the research question during the power analysis is the following: if the model fits the data poorly ($RMSEA = .08$) and we test the null-hypothesis of close fit ($RMSEA \leq .05$), what is the probability of correctly rejecting the null-hypothesis? (MacCallum et al., 1996, pp. 138-139). This probability is referred to as the power of the analysis, and it should generally be 80% or higher ($power \geq .80$) (Cohen, 1992). With a power of $.80$, there is an 80% probability of correctly rejecting the good fit hypothesis when model fit is poor, and there is a 20% probability of incorrectly rejecting the good fit hypothesis when model fit is actually good (Type II error, or β). It must also be noted that the probability of accepting the good fit hypothesis when this is false (Type I error, or α), is generally fixed at $.05$ (Cohen, 1992). In other words, a common practice in power analysis is to identify a sample size that allows only a 5% probability of wrongly accepting the good fit hypothesis when model fit is actually poor.

As the sample size increases, the RMSEA confidence intervals become narrower. Therefore, larger samples increase the power to distinguish between poor fit and close fit. Additionally, as the model degrees of freedom increase, indicating greater model complexity, a power of $.80$ is achieved with smaller and smaller

samples. Therefore, we aim to find the sample that achieves adequate power ($\geq .80$), given a specific model complexity. This can be tested using the semTools package (semTools Contributors, 2016) in R (R Core Team, 2015).

For example, a one-factor model with 10 items would have 35 degrees of freedom and would require a sample of 279 participants to achieve a power of .80. The same one-factor model with 20 items would have 170 degrees of freedom and would achieve the same power with 94 participants. If we were testing a more complex model (i.e. with more than 170 degrees of freedom), a lower sample size would be needed for the RMSEA to achieve a power of .80.

The main limitation of this approach is linked to the limitations of the RMSEA fit index. This approach is not as useful in very simple or very complex models. Simple models have very low degrees of freedom and, in these circumstances, the RMSEA will tend to over-reject models with good fit (Kenny et al., 2015). For example, a sample of 500 participants would generally be considered more than adequate to test a one-factor model with 5 items and strong loadings (e.g. Hair et al., 2010). However, such a simple model would only have 5 degrees of freedom. In this case, RMSEA would reach adequate power (.80) with a sample closer to 1,500 participants. Therefore, with 500 participants, its confidence interval will likely be broad and may include values greater than .10, which are considered indicative of poor fit. On the other hand, when analyzing a very complex model with, say, 1,000 degrees of freedom, the RMSEA may achieve adequate power with as few as 34 observations. Based on the simulation studies presented above, a sample of 34 observations would be inadequate to test virtually any CFA model.

3.11.3. Summary. In short, larger samples are not only desirable to produce sample estimates that closely approximate population estimates, but they are also required to achieve convergence when inter-item correlations (and consequently item loadings and communalities) are low. From the simulations studies presented here we can conclude that it is highly desirable to factor analyze models with high communalities ($h^2 > .60$), with at least 4 items per factor and with large samples (e.g. $N \geq 400$). In the previous chapter, a well-being model with 14 hypothesized factors was proposed. If a minimum of 4 items were used to measure each of the 14 correlated factors, this model would have more than 1,000 degrees of freedom. Therefore, based on the RMSEA power analysis a large sample ($N \geq 400$) would produce an RMSEA power of 1, thus exceeding the typical (.80) criterion (Cohen, 1992). As a conservative

choice, a target sample of 500 or more participants was chosen in the following studies.

Beyond factor analysis, this thesis will also use correlation analyses to test the relationships between different scales. Given an α of .05 and a power of .80, with a sample of 500 observations, the studies will be able to identify even relatively small two-tailed correlations (e.g. $r = .13$).

3.12. Online sample recruitment and representativeness

While the previous section focused on sample size, this section will focus on sample recruitment and representativeness. The advent of crowdsourcing platforms has provided researchers with convenient ways to recruit large samples. The current thesis will use the Amazon Mechanical Turk website (mturk.com) to recruit samples for the studies. Therefore, this section will outline some of the advantages and limitations of this platform.

The MTurk website provides access to a diverse population of individuals who complete online tasks. According to the website information pages, the platform was originally designed to recruit a large temporary workforce to carry out tasks that are better performed by human beings than computers, such as transcribing audio recordings or identifying objects inside images or videos. However, the variety of tasks and workers available on the site has increased substantially since its release.

It has been estimated that the platform includes over 500,000 individuals, the majority of whom reside in the U.S. (Paolacci & Chandler, 2014). The platform allows researchers to recruit large samples for relatively low compensation. Specifically, compensation is generally at or below the minimum wage in the U.S. (Litman, Robinson & Rosenzweig, 2015). However, increasing compensation closer to the minimum wage tends to improve the speed of recruitment (Buhrmester, Kwang & Gosling, 2011). The ability to recruit large samples can benefit psychological research by improving the statistical power of the analyses. The main consequence of low statistical power is the inability to detect small effects. For example, a sample of 100 individuals would be underpowered to identify a correlation of .20 between two variables. If such a relationship was present in the population, it may be undetected in the study, leading to a type II error. Furthermore, factor analysis requires large samples. From this perspective, using MTurk may be beneficial.

More important than sample size is sample diversity and representativeness. It is a common practice in psychological research to recruit student samples. For example, over 70% of the studies published in the *Journal of Personality and Social Psychology* in 2002 included student samples (Gosling, Vazire, Srivastava & John, 2004). Compared to the general adult population, student samples tend to exhibit a lower age, and a higher level of education and socio-economic family background (Gosling et al., 2004). On the other hand, the MTurk samples have been found to be more diverse and to approximate the general population demographics better than typical internet samples or student samples (Buhrmester et al., 2011; Casler, Bickel & Hackett, 2013). Compared to national probability samples, MTurk respondents tend to be more liberal and younger, although still older than student samples (Berinsky, Huber, Lenz, 2012). While MTurk samples may not be as representative as national probability samples, the respondents within the U.S. show demographics comparable to a large U.S. stratified survey (Huff & Tingley, 2015).

Several studies were able to replicate previous findings using MTurk samples: a behavioral categorization task produced the same choice patterns in MTurk and in-person samples (e.g. Casler et al., 2013); the prevalence of psychiatric disorders (e.g. generalized anxiety, depressive symptoms) in an MTurk sample were very similar to those previously reported in representative samples (Shapiro, Chandler & Mueller, 2013); established personality and health-related scales were found to exhibit similarly adequate internal and test-retest reliabilities in MTurk samples compared to traditional methods (Buhrmester et al., 2011; Shapiro et al., 2013).

Despite the advantages of online recruitment through MTurk, some concerns have also been raised. Initial criticisms about internet research argued that only individuals from a high socio-economic background had access to the internet. Therefore, internet samples may not be representative of a large proportion of the population that cannot afford internet access. Indeed, U.S. Census data estimated that, in the year 2000, only 51% of American households had at least one computer and overall 42% were able to access the internet (Newburger, 2001). However, computer and internet use has been steadily increasing, with the U.S. Census data reporting over 83% of households owning a desktop, laptop or handheld computer, and over 74% reporting internet use, in 2013 (File & Ryan, 2014). In short, internet access is becoming virtually ubiquitous, thus reducing the selection bias arising from internet research.

Another criticism involves the researcher's lack of control over the experiment or questionnaire administration process. In other words, by creating an online task that is completed by participants whenever and wherever they like, the researcher loses control over the environment in which the task is carried out (Kraut et al., 2004). For example, different respondents may complete a task using different devices (e.g. PC, tablet, smartphones). This may be particularly problematic for experiments using subtle visual stimuli, which may be presented in different ways depending on the device used. This issue may be somewhat prevented by including instructions that direct respondents to adjust the size (Woods, Velasco, Levitan, Wan & Spence, 2015) or brightness of the screen (Yung, Cardoso-Leite, Dale, Bavelier & Green, 2015). However, in survey studies, differences in the devices used across participants seem less problematic, as participants simply need to be able to read and respond to questions, and common survey platforms (e.g. qualtrics.com) tend to be compatible across devices. The results of experiments may also be biased by the fact that some MTurk respondents have already completed the same or a similar task, and are therefore not naïve to the purposes of the experiment (Berinsky et al., 2012). However, if this is a concern, the researcher can select workers who have previously only completed no more than a certain amount of tasks on the online platform.

More troubling is the possibility that MTurk workers may be impersonating someone else, or sending multiple submissions (Kraut et al., 2004). However, MTurk requires workers to verify their account through personal identification (e.g. social security number), which would likely reduce such attempts. Furthermore, settings on both MTurk and separate online survey platforms (e.g. qualtrics.com) can prevent respondents from completing a questionnaire more than once, based on their IP address and MTurk ID. Fake or malicious submissions may, perhaps, be motivated by an economic gain. For example, a worker may not pay sufficient attention while completing the task and complete several tasks quickly to obtain a higher compensation. However, the economic gains are generally small, as tasks tend to involve payments below the minimum wage (Litman et al., 2015). Most importantly, MTurk users cite task enjoyment as their primary motivation for completing a task, while economic gains play only a marginal role: only 12% of workers use MTurk as a main source of income (Ipeirotis, Provost & Wang, 2010). Finally, the gains of quick and inattentive work are diminished by the fact that a researcher or a business, receiving poor work from an MTurk user, is entitled to reject that work and deny any

payment, which would also undermine that worker's future ability to qualify for tasks advertised on MTurk. Indeed, participation on one's study can be limited only to those MTurk users whose previous work received a high approval rating. This has been shown to increase the quality of the data, by including only individuals who tend to pay attention to a task (Peer, Vosgerau & Acquisti, 2014). Finally, researchers can include screening questions in their tasks (e.g. "please leave this answer blank if you're paying attention"), which have been shown to be effective in screening out inattentive participants and increase the statistical power of a study (Maniaci & Rogge, 2014).

Online research presents both advantages and disadvantages over traditional survey methods. However, the evidence provided hitherto suggests that the advantages of online research using MTurk outweigh many of the disadvantages (e.g. Shaughnessy, Zechmeister & Zechmeister, 2012).

3.13. Dealing with missing data

Missing values are a virtually unavoidable part of research (Newman, 2014). When dealing with missing data, we must first assess the amount and patterns of missing data. The amount of missing data can be examined by reporting the total percentage of missing values, that is, the percentage of cells in the data set that are missing; the percentage of observations (i.e. cases) that have no missing values (i.e. observations valid listwise); and the percentage of available cases available for pairwise correlations (i.e. covariance coverage or observations valid pairwise); and the number of missing values for each observation and each item (e.g. Newman, 2014; Lang & Little, in press). In a multi-item questionnaire, an item that many people refuse to answer may be problematic and could be removed from further analyses (e.g. Kline, 2000). Similarly, if an individual fails to complete half or more of the items, they may be removed from further analyses (Graham, 2009; Hair et al., 2010). As a rule of thumb, missing data is not considered problematic when fewer than 10% of the observations have missing values (Newman, 2014), or when fewer than 10% of the values in each observation or item are missing (Hair et al., 2010).

We must also assess whether missing values occur in a systematic or random way. This can be partly ascertained by looking at the number of missing values of each observation and item. Additionally, observing missing value patterns also helps to identify any systematic influence on the results. From this perspective, the data

can be classified into one of three categories. These categories are partly overlapping and are often considered to be part of a continuum (Graham, 2009; Beaujean, 2014).

First, when no systematic patterns are present, the data are considered to be missing completely at random (MCAR). In this case, the presence or absence of a value is due entirely to chance, and there is no systematic reason why the data are missing across respondents.

Second, when missingness on a variable is related to another variable in the dataset, the data are considered to be missing at random (MAR). In other words, some systematic patterns are identified, but we know the cause behind this missingness. For example, if females tend to avoid responding to an item more than males, we can predict the item missingness from the gender variable. Therefore, our analyses will be unbiased as long as we include gender in our model (e.g. as a predictor) (Beaujean, 2014).

Third, when missingness on a variable is related to another variable *not* in the dataset, the data is considered to be missing not at random (MNAR). For example, if missingness is related to gender, but we did not include gender in our questionnaire, we cannot include it in our model. Therefore, parameter estimates may be biased.

Differentiating between these categories can be complex. This can be attempted, for example, by looking at nonresponse patterns. A random pattern (i.e. MCAR) can be found when values are missing in a seemingly arbitrary arrangement, in which no specific variable or groups of variables seem to cause a large portion of nonresponse (Lang & Little, in press). One common nonrandom pattern is the univariate nonresponse pattern, where values are missing only in one variable (Lang & Little, in press). As previously stated, an excessive and unusual amount of missing values in only one variable may be problematic and may lead to the exclusion of that variable from scale development and validation (e.g. Kline, 2000). Another systematic pattern is the monotone nonresponse pattern, where missing data appears in an increasing order similar to a staircase pattern. This may be common in multi-wave longitudinal studies where, at each wave, more participants permanently drop off (Lang & Little, in press). Additionally, some statistical tests have been proposed to assess whether data meet the MCAR assumption (e.g. Little, 1988). However, it has been argued that we “cannot be sure whether data really are missing at random, or whether the missingness depends on unobserved predictors or the missing data themselves.” (Gelman & Hill, 2007, p. 531). Most importantly, rather than focusing

on whether these assumptions have been violated, “we should answer the question of whether the violation is big enough to matter to any practical extent” (Graham, 2009, p. 567). For instance, if 3 out of 400 respondents exhibit a nonrandom missing pattern, this seems unlikely to bias the parameter estimates.

Common approaches to deal with missing data include listwise and pairwise deletion. Listwise deletion removes all observations (i.e. cases) with missing data. The advantage of listwise deletion is that it is not prone to type 1 errors. However, the reduction in sample size can lead to a loss of power and therefore a higher probability of making a type 2 error (Graham, 2009). In other words, with listwise deletion it is more likely to find non-significant results even when there is a significant effect in the population. Furthermore, it can lead to biased estimates when systematic nonresponse is present (Graham, 2009).

Pairwise deletion may be slightly better than listwise deletion because more data are preserved in the analyses, which can help maintain higher power (e.g. Brown, 2015). However, like listwise deletion, it can produce biased estimates with systematic nonresponse patterns (Newman, 2014). Furthermore, since different sample sizes are used for different pairwise estimates, the overall sample size for the analysis remains unclear (e.g. Beaujean, 2014). This variability in sample size can also lead to non-positive definite correlation matrices and correlation values that exceed their boundaries from -1 to 1 (Lang & Little, in press).

Mean replacement (or mean imputation) consists in replacing missing values for each variable with that variable’s overall mean for the entire sample (Lang & Little, in press). This method has been criticized because, by increasing the number of values equal to the mean, variances and covariances are reduced and response distributions are distorted (e.g. Brown, 2015; Graham, 2009).

Another type of single imputation is regression imputation. This method carries out a regression analysis treating each variable with missing values as an outcome variable and all other variables as predictors. In this way, missing values are replaced by values predicted by the regression (e.g. Tabachnick & Fidell, 2013). Even when the data are missing completely at random, this method will produce inflated correlations, and underestimated standard errors (e.g. Beaujean, 2014). This is a result of the fact that this method produces imputed values that lie exactly on the regression line. This type of imputation may be improved by adding a random error

term (e.g. Lang & Little, in press). However, a better alternative would be to create multiple datasets with imputed values.

Multiple imputation creates several (e.g. 20 or more) datasets where imputed values are drawn randomly from their predicted distribution. Analyses are then conducted separately on each imputed dataset and results are then pooled together (Graham, 2009; Lang & Little, in press). This procedure produces parameter estimates and standard errors that accurately reproduce the population parameters even when non-random missingness patterns are present (Allison, 2009). When data are not missing at random and auxiliary variables are available to account for the missingness mechanism, these variables can be included in the multiple imputation model to reduce the potential bias from the systematic missingness. For example, if 20% of the responses to an income item are missing, this missingness may not be random, as people who are very wealthy or very poor may be more likely to avoid answering this item. However, if we have data on other socio-economic variables, which are highly correlated with income, we can include these auxiliary variables during multiple imputation to better estimate the missing values on income (e.g. Lang & Little, in press).

Another excellent method is full-information maximum likelihood (FIML). This does not impute missing values. Instead, it is a model-based approach that uses all the information in the response patterns to estimate the parameter estimates (e.g. correlations, factor loadings) and standard errors. It has been found to be one of the most effective methods to deal with missing data, by producing estimates that closely approximate those generated from complete data (Graham, 2009; Newman, 2014). It is regarded as the best method of dealing with missing data in CFA and SEM, and can be used in conjunction with MLR estimation when the data exhibit non-normal distributions (Brown, 2015). Like multiple imputation, FIML can use auxiliary variables when necessary, but it is simpler to implement (Lang & Little, in press).

Overall, the two recommended methods for dealing with missing data are multiple imputation and FIML, and either of these methods should be used as default solutions (Beaujean, 2014). This thesis will focus on correlation and factor analysis, and will use FIML to estimate missing data.

3.14. Summary

To summarize, several steps are necessary to develop and validate a questionnaire (e.g. McCoach et al., 2013):

1. Review the literature to understand the conceptual foundations of the questionnaire and evaluate its necessity (see Chapters 1 and 2).
2. Define the construct(s) conceptually (see Chapter 2).
3. Develop items based on those construct definitions as well as best practices in item writing (see Chapter 4).
4. Assess the content evidence of validity by asking a group of content experts to evaluate the items. Based on these evaluations, make changes to the items where necessary (see Chapter 4).
5. Collect data by administering the questionnaire to a relatively large sample, and assess its response distributions, item clarity, dimensionality and reliability. Based on these findings, select well-performing items (see Chapter 5).
6. Replicate these findings in a new sample. This is particularly important if changes to the questionnaire were made in the previous step (see Chapter 6).
7. Finally, assess the criterion evidence of validity by correlating the questionnaire scores with other measures of conceptually similar and different constructs or testing mean differences across demographic groups (see Chapter 6).

While the previous chapters presented a literature review of the current conceptualization and measurement of well-being (steps 1 and 2), the following chapters will put the next steps into practice, by describing the development and initial validation of the scales of general well-being.

Chapter 4: Study 1 – Item development and content validation

4.1. Item development

The item development process was adapted from DeVellis (2012). Items were written based on the conceptual definitions of well-being presented in Chapter 2. Specifically, items were designed to measure fourteen constructs: happiness, vitality, calmness, optimism, self-awareness, self-acceptance, self-worth, competence, involvement, development, purpose, significance, self-congruence and connection.

The main goal during item development was to generate items that only tapped into the single construct under focus, and were as clear and concise as possible. To accomplish this, the first items generated closely resembled the definition of the construct they were designed to measure (e.g. “I feel happy” for happiness). Then, previous scales attempting to measure each construct, or closely related constructs, were examined (see Appendix 2), and key terms were extrapolated from these scales to develop new items. Additionally, a thesaurus (Oxford Dictionaries, 2014) was used to identify related terms describing each construct, while omitting terms that were too close in meaning to avoid any redundancy (e.g. bloated specifics) (Cattell, 1978). In few cases, closely related items (e.g. energetic and full of energy) were included in the initial item pool, so that the one exhibiting better psychometric properties may later be retained (DeVellis, 2012).

The initial item pool contained a minimum of 10 items for each construct (see Table 5). The aim for the final scale was to include approximately 5 items to measure each construct, with 4 items as a minimum. This target was chosen so that the final scales would minimize participant burden while maintaining adequate internal reliability. Whenever a single construct (e.g. self-awareness) comprised several facets (e.g. awareness of one's thoughts, feelings, bodily sensations, etc.), a grid was used to ensure at least two items were generated to tap into each facet (Kline, 2000).

A 5-point response format was used: 1 = not at all true, 2 = a bit true, 3 = somewhat true, 4 = mostly true, 5 = very true (adapted from Liao, 2014; Weinstein et al., 2012). Well-being data tends to be negatively skewed: most people report average to high levels of well-being (e.g. Springer & Hauser, 2006). Consequently, a scale that provides more options covering moderate to high levels of well-being has been recommended (Saris & Gallhofer, 2014) and has been found to produce data with a near-normal distribution (Liao, 2014).

Items were developed over several drafts through collaboration between the PhD student and the supervisor. Specifically, the PhD student would write a draft of the item pool and revise it based on feedback from the supervisor. During this process several items were revised and omitted. For example, the item “I avoid criticizing myself” was originally written to measure self-acceptance. However, during the review process, it became clear that the term criticism can also be interpreted as constructive, and avoiding criticism altogether could suggest an element of narcissism. Similarly, the item “I have a sense of direction” was written to tap into purpose. Indeed, this item perfectly matches the outlined definition of purpose in life (see also Ryff, 1989). However, on its own, the item could also be interpreted as spatial awareness (e.g. being able to orient oneself). Following such discussions, both items were omitted from the item pool. The final draft is presented in Appendix 3.

Items were grouped based on the construct they measured, so that each construct would be measured in a single section of the questionnaire. The grouped format was chosen for three reasons. First, it allows future researchers to use single scales (e.g. happiness) without compromising their validity due to item-order effects (e.g. Schwarz, 1999). On the other hand, if happiness items were randomly distributed across the questionnaire, responses to each happiness item would likely be influenced by the responses to its preceding item. Second, grouped formats have been shown to improve convergent and discriminant validity (Harrison & McLaughlin, 1996; Schriesheim, Solomon & Kopelman, 1989). Third, it has been argued that participants may find it easier to understand the meaning of a construct when its items are presented in a grouped format (De Vaus, 2002; Green & Yang, 2009).

All items were phrased as declarative statements. Theoretically, reversed well-being items were expected to tap into ill-being constructs (Huppert & So, 2013). From a psychometric perspective, reversed items exhibit several limitations that were outlined in section 3.5.

4.2. Content validation

Six academics with expertise in well-being research evaluated whether the items were clear and relevant to the construct they attempted to measure. Each academic had published several articles on the subject and was contacted through

their university email address. For each component of well-being, the experts were presented with a definition of the construct, followed by a list of items attempting to measure that construct. Then, they assessed whether each item was a good measure of the construct on a Likert scale (1 = inadequate, 2 = needs revision, 3 = adequate). Additionally, they were asked to provide comments if they thought any item was inadequate or needed revision. Based on these recommendations, of the 147 items submitted for content validation, 18 items were excluded from further analyses and 6 items were edited.

Items were generally retained if they had a mean rating of at least 2.30 / 3, which corresponds approximately to a rating of 7.5 / 10 (Hardesty & Bearden, 2004). The comments were also taken into consideration (see Appendix 4). For instance, two vitality items received a rating below 2.30, yet the justifications for such a low rating did not seem relevant: an academic found these two items to be very similar to two items in the PANAS. Since, that similarity does not hinder the items from performing well in the new scale, the items were retained. Conversely, other items received an adequate rating, but the comments indicated potential issues that could not be overlooked. For instance, the development item “I feel I am growing” had an adequate rating, but one reviewer argued that it may be interpreted too literally. Indeed, when presented on its own, the item may be misinterpreted as physical growth. Therefore, it was excluded from further analyses.

Furthermore, based on the comments, a few items were edited to improve clarity. For example, according to one academic, the self-acceptance item “I accept the choices I’ve made” did not seem to fit with the rest of the scale, which focused more generally on accepting different aspects of oneself and one’s flaws. Therefore, it was edited as “I accept my shortcomings.”

Finally, the academics were asked whether all aspects of each construct were represented in the item pool. Nearly all responses to this question were positive. One exception mentioned that the competence scale may also tap into constructs like self-efficacy and resilience. Indeed, the chosen definition of competence closely resembles that of self-efficacy and resilience, although it focuses more on one’s perception of confidence (e.g. Deci & Ryan, 2000). Therefore, this was not deemed problematic, especially after considering that all competence items received high ratings. After content validation, each scale included 8 or more items and the total item pool comprised 129 items (see Table 5).

Table 5 Number of items per scale at different stages of scale development

Scales	Item development	Content validation	CFA (study 2)
Happiness	10	8	4
Vitality	11	8	4
Calmness	10	8	4
Optimism	10	9	5
Involvement	10	9	4
Self-awareness	10	9	4
Self-acceptance	12	11	4
Self-worth	10	8	5
Competence	10	10	6
Development	10	9	5
Purpose	10	9	5
Significance	10	8	5
Congruence	12	11	5
Connection	12	12	5
Total	147	129	65

Chapter 5: Study 2 – Item selection and scale dimensionality

Study 2 examined the factor structure of the questionnaire and the extent that selected items exhibited desirable psychometric properties. The aim was to develop a questionnaire that could be used to measure specific constructs (e.g. self-acceptance) as well as well-being as a whole. Therefore, the study assessed the dimensionality of each individual scale as well as the entire well-being questionnaire. Items were selected so that each scale would include at least 4 items, to maintain an adequate level of reliability and facilitate dimensionality testing (e.g. Kline, 2016, Chapter 3). For each scale to have adequate dimensionality, it needed to fit a unidimensional model. That is, it had to measure only one construct. Otherwise, the scores generated from it would have an ambiguous meaning. Similarly, responses to the entire questionnaire were expected to fit the hypothesized factor structure with 14 correlated factors (Furr, 2011; Chapter 2). Finally, the hierarchical factor structure of well-being was examined through higher-order and bifactor models. Based on the definition of well-being presented, a single higher-order or general factor was expected to adequately explain the relationships among the fourteen factors. That is, scores on the questionnaire were expected to be adequately captured by a single overall score.

5.1. Method

5.1.1. Sample. Data were collected from the Amazon Mechanical Turk panel. In total, 573 responses from U.S. residents were recorded. Since few responses were unfinished, the final sample included 560 responses. Ages ranged from 19 to 77 years ($M = 37.18$, $SD = 13.10$), and 60.9% were female. A small monetary incentive was used to encourage participation (approx. \$0.50), and respondents spent on average 14 minutes to complete the survey.

To ensure high quality of the data, only people whose previous work had been rated as adequate at least 95% of the time were able to access the questionnaire. This cutoff was based on a recent study showing that participants with such a high approval rating maintain a higher focus throughout the questionnaire than those with a lower approval rating (Peer et al., 2014). Additionally, a screening question was included toward the end of the survey: “I read instructions carefully. To show that you are reading these instructions, please leave this question blank.” (Maniaci & Rogge, 2014). Around 8% of participants failed this screening question and were

therefore omitted from further analyses. Finally, seven cases were omitted, due to a lack of variation in their responses (Cummins, 2013). These were identified by computing an inter-item standard deviation for each respondent, and omitting those with a value of 0, indicating no response variation. The final sample included 507 participants.

5.1.2. Procedure. The questionnaire focused on well-being in life in general (see Chapter 2). Therefore item instructions stated the following: “Below you’ll find several statements about your experiences. Please indicate how true each statement is regarding the experiences in your life overall. There are no right or wrong answers. Please, choose the answer that best reflects your experience rather than what you think your experience should be.” Items measuring the same construct (e.g. calmness) were grouped together, and each group of items was presented on a separate page to help structure the questionnaire and make it less cognitively taxing. Furthermore, within each page, item order was randomized to minimize the effect of question order (De Vaus, 2002; Saris & Gallhofer, 2014). At the end of each page, respondents were asked to indicate whether any items on that page were confusing or ambiguous, and a blank space was provided to report such items.

5.2. Results

5.2.1. Preliminary analyses. Overall, a very small total amount of missing values was found (below 1%). The highest amount of missing values for a single item was less than 1%. The highest amount of missing values for a single observation (i.e. case) was 5 out of 129 items. Additionally, 84% of cases were complete listwise and, 98% of cases were complete pairwise. In other words, while listwise deletion would considerably reduce the sample size, all pairwise correlations in the data could be estimated using virtually the entire sample. The few cases with missing values did not seem to exhibit problematic nonresponse patterns. For example, the most frequent nonresponse pattern occurred in only four cases and consisted of nonresponse only to a single item, rather than pairs of items. During confirmatory factor analysis, the few missing values in the data were estimated using full information maximum likelihood (see section 3.13).

At the end of each page of the questionnaire, participants were given a blank space to indicate whether any of the items seemed confusing or ambiguous. Items were generally found to be clear. In fact, no item was reported as confusing by 2% or

more participants. Therefore, no items were excluded at this stage. Nevertheless, when an item was found to be confusing, albeit by few participants, this was taken into consideration during subsequent item selection.

Response distributions and item correlations were then examined using the psych package version 1.5.8 (Revelle, 2015) in R version 3.2.3 (R Core Team, 2015). All items produced responses covering the entire range of the 5-point response format. On average, means approximated the middle value of the 5-point format (i.e. 3), although there was a tendency to report slightly higher values ($M = 3.62$, range = 2.85 – 4.24). Standard deviations indicated sufficient variability in the responses (mean $SD = 1.11$, range = 0.88 – 1.30). Absolute values of univariate skewness ($M = -.61$, range = -1.38 and 0.02) and kurtosis ($M = -.29$, range = -1.04 and 1.55) were relatively low, compared to the cutoff points of 2 and 7, respectively (Fabrigar & Wegener, 2012). Again, skewness values indicated a tendency to report values slightly above the middle point. While univariate skewness and kurtosis were relatively low, Mardia's multivariate skewness and kurtosis values were above five, indicating a significant deviation from multivariate normality (Bentler, 2006). Therefore, robust maximum likelihood (MLR) estimation was used during factor analyses (Beaujean, 2014).

An average of the corrected item-total correlations exceeded .50 across the whole questionnaire, indicating that items were adequately related to each other (Hair et al., 2010). Furthermore, multicollinearity was generally not present: only one item had a squared multiple correlation above .90 (Kline, 2016), which was subsequently removed during factor analyses.

5.2.2. Dimensionality of each scale. Analyses were conducted using the Lavaan package (version .5–20, Rosseel, 2012). First, the dimensionality of each scale was assessed. Since each scale was based on a conceptual framework and previous measures of the same construct, a confirmatory factor analysis (CFA) was conducted (Fabrigar & Wegener, 2012).

Items were selected iteratively based on a combination of statistical and theoretical considerations. From a statistical perspective, items exhibiting the highest modification indices, expected parameter change and residual correlations were candidates for exclusion. From a theoretical and psychometric perspective, when choosing between two items, the simpler, clearer, less redundant item, which better reflected the definition of the construct was selected (DeVellis, 2012; see also

Appendix 5). For example, the items “I’m optimistic” and “I look on the bright side” aimed to measure optimism and exhibited strong factor loadings, above .80. However, a residual correlation between the two items significantly reduced model fit, indicating that one of them should be omitted from the scale. In this case, the item “I look on the bright side” was omitted, because it was lengthier: it conveyed the same idea as “I’m optimistic” but with greater complexity. In another example, the items “I feel calm” and “I feel leisurely” aimed to measure calmness. Once again, a residual correlation between the two indicated that one should be omitted. In this case, “I feel leisurely” was omitted because it is an infrequent expression (books.google.com/ngrams/), and thus less likely to be understood by respondents, as three participants also pointed out. In both examples, omitting these items produced a statistical improvement, in terms of model fit, as well as a psychometric improvement, in terms of item length and clarity. Thus, items were selected until each scale exhibited good fit and included items covering the entire content range of each construct’s definition.

Several fit indices were used to evaluate model fit of the final scales: the Yuan-Bentler chi-square, the Comparative Fit Index (CFI), the Standardized Root Mean Square Residual (SRMR), and the Root Mean Square Error of Approximation (RMSEA). CFI and RMSEA were based on the Yuan-Bentler chi-square. A non-significant chi-square indicates excellent fit, which becomes rarer to achieve as sample size and model complexity increase; CFI values above .90 indicate acceptable fit and values above .95 indicate good fit; SRMR values below .06 indicate good fit; RMSEA values below .08 indicate acceptable fit and values below .06 indicate good fit, and its upper confidence interval should not be above .10 (Brown, 2015; Kline, 2016). However, RMSEA was found to perform poorly in models with small degrees of freedom, as it over-rejects well-fitting models (Kenny et al., 2015). Therefore, RMSEA was not taken as much into consideration. In addition to fit indices, correlation residuals were examined to identify specific areas of misfit and these should not exceed a value of .10 (Kline, 2016).

Before item selection, fit indices showed generally acceptable or marginally acceptable fit. After item selection fit substantially improved (see Table 6). All indices showed good fit and residual correlations never exceeded .10. Interestingly, despite the large sample size χ^2 was not significant. Such good fit may be due to several factors: the initial item pool was derived from previously validated measures,

the items were then subjected to content validation and further refined, well-fitting items were then selected during CFA, and the final models were relatively simple, as they included only 4 to 6 items.

Table 6 Fit indices of individual scales after item selection (Study 2)

Scale	YB χ^2	df	p	CFI	SRMR	RMSEA (90% CI)
Happiness	4.58	2	.101	.998	.006	.050 (.000-.100)
Vitality	0.22	2	.897	1.00	.002	.000 (.000-.018)
Calmness	1.89	2	.389	1.00	.004	.000 (.000-.071)
Optimism	7.98	5	.157	.998	.008	.034 (.000-.068)
Involvement	0.03	2	.983	1.00	.001	.000 (.000-.000)
Self-awareness	1.28	2	.528	1.00	.008	.000 (.000-.060)
Self-acceptance	0.37	2	.830	1.00	.004	.000 (.000-.000)
Self-worth	2.77	5	.735	1.00	.004	.000 (.000-.028)
Competence	6.31	9	.708	1.00	.007	.000 (.000-.029)
Development	3.17	5	.674	1.00	.006	.000 (.000-.031)
Purpose	3.39	5	.641	1.00	.006	.000 (.000-.039)
Significance	2.37	5	.796	1.00	.004	.000 (.000-.016)
Congruence	1.25	5	.941	1.00	.005	.000 (.000-.000)
Connection	6.19	5	.288	.999	.009	.022 (.000-.062)

5.2.3. Dimensionality of the entire questionnaire. After selecting items in each scale, a CFA was conducted on the entire well-being questionnaire. The final questionnaire is presented in Appendix 6. Based on the conceptual definition of well-being used during scale development, the well-being items selected thus far should fit a measurement model with 14 correlated factors (Chapter 2).

A variety of models were tested, including alternatives to the hypothesized correlated 14-factor model (see Table 7). In the first model all items loaded on a single factor. If this model exhibited good fit, it would indicate that different scales have no discriminant validity from each other. The model exhibited poor fit and high residual correlations (some of which reached the .40 range). This suggests that the fourteen factors are likely not indistinguishable.

The second model specified 14 independent (i.e. uncorrelated) factors: all items loaded on their hypothesized factor, but all factor correlations were fixed to 0. If the model adequately fit the data, it would indicate that the 14 scales are independent of each other. This would suggest that, unlike previous well-being measures, the scales do not measure a common construct (Disabato et al., 2016), which would threaten the validity of the questionnaire. However, the model showed poor fit and very high residual correlations, which would particularly affect the

SRMR, as shown in Table 7. This suggests that the factors are likely not independent.

The third model was the same as the second model, but the factors were correlated. Model fit was satisfactory, thus supporting the hypothesized factor structure. Considering the complexity of the model, it was highly unlikely for the χ^2 to be significant. Nevertheless, all other fit indices indicated good fit. Furthermore, residual correlations were very low, the highest being .14. This was slightly higher than .10, but considering the large number of items and factors it was not deemed problematic. In short, the hypothesized model, with 14 correlated factors, fit the data best.

Table 7 Fit indices of the entire questionnaire (Study 2)

Model	YB χ^2	df	p	CFI	SRMR	RMSEA (90% CI)	BIC
One factor	13,198.05	2,015	.000	.581	.081	.105 (.103-.106)	82,156
14 independent factors	6,819.42	2,015	.000	.820	.419	.069 (.067-.070)	74,361
14 correlated factors	2,751.51	1,924	.000	.969	.033	.029 (.027-.031)	70,053
Higher-order	3,218.26	2,001	.000	.954	.051	.035 (.033-.037)	70,141
Bifactor	3,005.88	1,950	.000	.960	.044	.033 (.031-.035)	70,214

Factor loadings were strong ($M = .86$, range = .61 - .94, see also Appendix 7). All fourteen scales exhibited adequate reliability, with McDonald's omega hierarchical (ω_h) coefficients ranging from .81 to .91 (see Table 8). McDonald's ω_h estimates reliability based on exploratory factor analysis. In the present studies, reliability was estimated using minres extraction and oblique promax rotation. McDonald's ω_h was used to measure reliability because this was found to outperform Cronbach's α (Zinbarg, Revelle, Yovel & Li, 2005). Factor correlations were moderate to strong and all of them were positive. This indicates that the covariance among the 14 factors may be explained by a higher-order or general factor, a possibility that is explored below.

To further assess the dimensionality of the scale, the average variance explained (AVE) by each factor was compared with the variance shared between factors (see Table 8). AVE corresponds to the average factor loading squared for each scale, while variance shared corresponds to the squared correlations between factors. AVE should ideally be .50 or greater, indicating that at least 50% of the variance in the items is reliable variance captured by the factor (Fornell & Larker, 1981; Hair et al., 2010), thus supporting convergent validity. Indeed, this was the case for every scale. The variance explained by each factor should also be greater than the variance shared with other factors. In other words, the AVE of a factor

should exceed the squared correlations between that factor and all other factors (Fornell & Larker, 1981; Hair et al., 2010), which would indicate that different factors are not measuring the same construct, thus supporting discriminant validity. This was true in all cases except one: the self-acceptance factor explained 63% of the variance in its items, but shared 66% of variance with the self-worth factor.

Table 8 Reliability and discriminant validity analysis (Study 2)

	Hap	Vit	Cal	Opt	Inv	Aw	Acc	Wor	Cop	Dev	Pur	Sig	Cog	Con
ω_h	.90	.89	.90	.88	.89	.81	.84	.90	.88	.88	.89	.91	.85	.87
AVE	.81	.77	.81	.77	.74	.56	.63	.81	.74	.76	.75	.81	.66	.71
MVS	.61	.61	.52	.60	.36	.39	.66	.66	.50	.45	.56	.56	.40	.39

Note: ω_h = McDonald's omega hierarchical; AVE = average variance explained by the factor; MVS = maximum variance shared with other factors

An alternative to Fornell and Larker's method, is the recently developed heterotrait-monotrait method (HTMT; Henseler et al., 2015). This method is applied to a matrix of inter-item correlations. Based on these inter-item correlations, inter-scale correlations are estimated. HTMT values show the ratio of the inter-item correlations between scales to the inter-item correlations within scales. In other words, HTMT values are interpreted in the same way as factor correlations. HTMT values below 1 indicate that two scales are not perfectly correlated: stronger correlations are present within each scale compared to the correlations between scales. HTMT values below .85 provide even stronger evidence of discriminant validity (Henseler et al., 2015). In this sample, the highest HTMT value was .80 between self-acceptance and self-worth. In other words, based on the HTMT method, the two scales did not exhibit problematic discriminant validity. However, consistent with the Fornell and Larker method, they strongly overlapped.

Therefore, a further analysis was carried out to investigate whether the two factors could be merged together. A one-factor model of self-acceptance and self-worth fit the data poorly compared to a two-factor model, which showed good fit (see Table 9). Therefore, despite the overlap between self-acceptance and self-worth they were maintained separate.

Table 9 Competing models for self-worth and self-acceptance (Study 2)

Model (no. factors)	YB χ^2	df	p	CFI	SRMR	RMSEA (90% CI)	BIC
Worth-acceptance (one)	239.01	27	.000	.919	.066	.124 (.113-.136)	10,579
Worth-acceptance (two)	32.00	26	.193	.998	.017	.021 (.000-.039)	10,268

Two hierarchical models were then tested: a higher-order model and a bifactor model. These two models were specified to assess whether it would be appropriate to

produce a single overall well-being score from the questionnaire. The higher-order model was similar to the independent factors model, with the addition of a single higher-order factor on which the fourteen lower-order factors loaded. That is, the higher-order factor was expected to explain the correlations among the fourteen lower-order factors. Model fit was slightly worse compared to the correlated factors model, as evidenced by a higher BIC (see Table 7). However, this was expected, as fewer parameters were estimated, and more parsimonious models tend to have lower fit indices. Indeed, 91 factor correlations were now explained by the more parsimonious higher-order loadings. Despite this, the higher-order model fit the data adequately. Loadings on the higher-order factor were high ($M = .76$, range = .57 - .88). Additionally, factor correlations were used to compute McDonald's omega hierarchical ($\omega_h = .86$), which supported the reliability of the higher-order factor. The omega analysis also indicated that, after controlling for a single general factor, additional factors had unacceptable reliability estimates, below .20 (see Appendix 8).

The bifactor model was also similar to the independent factors model, with the addition of a general factor on which all items loaded. In other words, in the bifactor model, each item loaded both on its specific factor (e.g. happiness) as well as the general factor (i.e. g). Again, this model did not fit as well as the correlated factors model, perhaps due to its increased parsimony. Nevertheless, model fit was adequate. Loadings on the specific factors remained significant, indicating that, even after controlling for the general factor, each scale provided significant additional variance. Furthermore, the general factor saturation was high ($\omega_h = .94$), again supporting its reliability. The general factor in the bifactor model was expected to be more reliable than the higher-order factor due to the increase in number of items, which tends to increase reliability.

Two additional indices were recently found to be useful when evaluating bifactor models: explained common variance (ECV) and the percent of uncontaminated correlations (PUC). These indices indicate the extent to which a single score from g is unbiased in representing the variance in the items (Rodriguez et al., 2016). ECV represents the proportion of common variance explained by a factor (e.g. g). The higher the common variance explained by g , the more accurately a g score will represent the variance in the items. Since eigenvalues represent the amount of common variance explained by each factor, ECV is estimated by dividing the eigenvalue of one factor by the sum of all eigenvalues (Ten Berge & Sočan,

2004). In the present study, eigenvalues were 28.83 for the general factor, and 19.46 for all specific factors (total = 48.29). Therefore, the general factor explained the majority (i.e. 60%) of the common variance in the items ($ECV = 28.83 / 48.29 = .60$).

The ECV is moderated by the percentage of uncontaminated correlations (PUC), which represents the percentage of correlations among items in the model that can be attributed solely to the general factor. The PUC depends on the model specification, and specifically on the overall number of items, the number of specific factors, and the number of items in each specific factor. Models with many items, and many specific factors, with few loadings in each specific factor will have a high PUC. In this case, the correlations among items in different specific factors will solely be explained by the general factor and will, thus, be uncontaminated by specific factors. Conversely, models with few items, and few specific factors, with many loadings in each factor, will have a low PUC because a larger proportion of correlations among items will be explained (i.e. “contaminated”) by the specific factors. When PUC is high, an overall g score will be unbiased even when ECV is relatively low (e.g. .50) (Bonifay, Reise, Scheines & Meijer, 2015). In the present bifactor model, there were 65 items producing 2,080 correlations. The specific factors together explained 121 of those correlations. Therefore, 1,959 correlations were uncontaminated by specific factors. Thus, the PUC was .94, indicating that 94% of the correlations reflected only the general factor ($PUC = 1,959 / 2,080 = .94$). In short, the higher-order and bifactor models supported the use of a single general factor of well-being.

5.3. Discussion

In study 2, the item pool was subjected to factor analyses to select items that fit the hypothesized factor structure. CFAs were conducted on each scale individually as well as the entire well-being questionnaire. After item selection, good fit was found for each scale as well as the entire questionnaire. Thus, results in the present sample supported the dimensionality of scores in each individual scale as well as the questionnaire as a whole. Different factors were related to but distinct from each other, as evidenced by each factor’s higher AVE relative to its MVS, and HTMT values below .85. One exception was found when using only the Fornell and Larker method. While the self-acceptance factor seemed to share slightly more variance

with the self-worth factor than it explained in its items, the two scales fit the data adequately only when they were specified as two separate factors, which was consistent with the chosen conceptualization of well-being. Furthermore, each scale exhibited adequate internal consistency. Taken together the results show support for the hypothesized 14 correlated factors, thus providing preliminary support for the conceptual definition of the lower-order well-being factors presented in Chapter 2.

With regards to the hierarchical structure of the questionnaire, higher-order and bifactor models with a single general factor fit the data adequately. Furthermore, the ω_h , ECV and PUC indicated that the general factor explained a substantial amount of variance and an overall well-being score would adequately summarize the variation in the responses. In other words, support was found for the hypothesized general factor of well-being.

To ensure that the current results are replicable, these findings need to be supported in a new sample, which was recruited in the next study (Chapter 6). The next study will also investigate whether the dimensionality of the questionnaire is invariant across gender and age, and whether it is stable over time. In addition to assessing the dimensionality of the questionnaire, criterion evidence of validity will be examined. In other words, the next study will also assess the relationships between the new scales, measures of similar well-being constructs, and measures of different constructs (e.g. social desirability).

Chapter 6: Study 3 – Further dimensionality and criterion validity testing

The purpose of study 3 was to further test the dimensionality of the new well-being questionnaire. It was expected that the factor structure supported in study 2 would also be supported in a new sample. Test-retest reliability and longitudinal invariance were also assessed. Furthermore, since differences in well-being have been found across gender and age (e.g. Meade & Dowswell, 2015; Meisenberg & Woodley, 2015; Ryff, 1989), the measurement invariance of the questionnaire across gender and age was assessed.

Additionally, criterion validation was conducted by correlating scale and general well-being scores with external measures. The correlation between each scale and previous measures of the same (or similar) constructs was examined. Strong correlations were expected between each scale of GWB and its respective comparison measures (e.g. between the new and previous measures of happiness or pleasant affect). Comparatively weaker relationships were expected between each scale and measures of different constructs (e.g. between the new happiness scale and a comparison measure of self-awareness).

A general well-being score was expected to be very strongly correlated with scores of SWB, PWB, social well-being and flourishing. Furthermore, the relationship between general well-being, social desirability and personality traits was assessed. If the new questionnaire taps into the same general construct as previous measures of well-being, it should exhibit similar relationships to external variables. Based on previous studies, well-being was expected to exhibit a moderate correlation with social desirability (Fastame, Penna & Hitchcott, 2015; McCrae & Costa, 1983), moderate to strong correlations with neuroticism and extraversion, moderate correlations with conscientiousness and agreeableness and a small correlation with imagination (Donnellan et al., 2006; Steel, Schmidt & Shultz, 2008). Additionally, general well-being was expected to have a small correlation with dark triad traits: narcissism, psychopathy and Machiavellianism (Aghababaei & Błachnio, 2015).

Finally, based on the conceptual definition of well-being (Chapter 2), it was expected that each well-being construct would be negatively correlated to measures of psychological ill-being (e.g. depression, anxiety) (e.g. Watson et al., 1988), and would be perceived as having a high subjective value. Subjective value has been previously studied with regards to well-being constructs, like positive affect (Diener, 2000), and to identify the most relevant among several components of well-being

(e.g. Delle Fave, Brdar, Freire, Vella-Brodrick & Wissing, 2011; Hogan et al., 2015). Symptoms of good mental health are assumed to be intrinsically valuable. In other words, they are worth experiencing for their own sake, rather than as means to an end (Alexandrova, 2012; Seligman, 2011). Therefore, each construct measured in the scales of general well-being was expected to be rated with a high perceived value score.

6.1. Method

6.1.1. Design. The study was divided into 3 sections. In the first section, all participants completed the SGWB. Due to the large number of measures included in the survey, the second section was divided into two parts (2a and 2b), and participants were randomly assigned to one of these two conditions. Half of the sample responded only to section 2a while the other half responded only to section 2b. This design allowed participants to complete the survey in approximately 20 minutes, rather than an estimated 30-40 minutes. In section 2a, participants completed comparison measures tapping into the fourteen constructs assessed by the new well-being questionnaire. Specifically, for each of the fourteen constructs, two comparison scales were included to assess convergent and discriminant validity. Additionally, participants completed a measure of social desirable responding and a personality questionnaire. In section 2b, participants completed measures of SWB, PWB and social well-being, a psychological ill-being scale, and a measure of the perceived value of each well-being construct (see below).

Five weeks after completing the survey, all participants were asked to complete the questionnaire again. The goal was to receive responses from approximately half of the sample to assess the test-retest reliability of the questionnaire. This short follow-up also included a measure of the dark triad traits to further assess discriminant validity.

6.1.2. Sample and procedure. Data were collected from Amazon Mechanical Turk. In total, 1,101 responses from U.S. residents were recorded. Ages ranged from 17 to 83 years ($M = 36.93$, $SD = 12.13$), and 56.6% were female. A small monetary incentive was used to encourage participation (\$0.77), and respondents spent on average 20 minutes to complete the survey. Once again, only individuals whose previous work had been rated as adequate more than 95% of the time were able to

access the survey (Peer et al., 2014). Furthermore, the survey could only be accessed by workers who had not previously participated in study 2.

Screening questions were embedded in the survey. Section 1 included the item “please leave this item blank”, while sections 2a and 2b included the item “I have seventeen fingers on my left hand,” which required a negative response (DeSimone, Harms & DeSimone, 2015). Failing any of these screening questions resulted in exclusion from further analyses. Around 8% of participants failed these screening questions and were therefore omitted from further analyses. Furthermore, 18 cases showed no variation in their responses to the SGWB, as evidenced by an inter-item standard deviation of 0, and were therefore omitted (Cummins, 2013). Finally, 3 participants were excluded because they reported not being fluent in English. The final sample included 989 participants, of which 486 completed section 2a and 503 completed section 2b.

In the follow-up conducted five weeks later, a total of 446 responses were recorded. The time gap between the main study (time 1) and the follow-up (time 2) was consistent with previous studies, whose time gap usually ranged between 4 and 8 weeks (Pavot & Diener, 2008; Ryff, 1989; Watson et al., 1988). Ages ranged from 19 to 83 ($M = 39.72$, $SD = 12.65$), and 61% were female. Following the same screening procedure as above, 12 cases failed the attention check and 3 showed no response variation. Thus, the final sample included 431 cases.

6.1.3. Measures. In section 1, participants completed the SGWB. Each of the fourteen scales was presented on a separate page. Furthermore, scale order and item order were randomized. Omega hierarchical coefficients are presented in Table 12.

After completing the new well-being questionnaire, participants were randomly assigned to one of two conditions. Further information regarding each scale, the number of items, scoring and omega coefficient can be found in Appendix 9.

In section 2a, participants completed measures tapping into the fourteen constructs assessed by the new well-being questionnaire. Specifically, for each of the fourteen constructs, two comparison scales were included to assess convergent and discriminant validity. Happiness was measured with the pleasant affect scale (Feldman Barrett & Russell, 1998) and the happiness scale (Huppert & So, 2013). Vitality was measured with the subjective vitality scale (Ryan & Frederick, 1997; Bostic et al., 2000) and the energy / general activation scale (Thayer, 1986). Calmness was measured with the pleasant deactivated (Feldman Barrett & Russell,

1998) and the calmness / general deactivation scale (Thayer, 1986). Optimism was measured with the life orientation test-revised (LOT-R, Scheier, Carver & Bridges, 1994) and the optimism scale (Huppert & So, 2013). Involvement was measured with the Core flow scale (Martin & Jackson, 2008) and the total concentration subscale of the Dispositional Flow Scale 2 (DFS2; Jackson & Eklund, 2004). Self-awareness was measured with the awareness subscale of the Cognitive and Affective Mindfulness Scale-Revised (Feldman et al., 2007) and the awareness subscale of the Philadelphia Mindfulness Scale (Cardaciotto et al., 2008). Self-acceptance was measured with the acceptance subscale of the Cognitive and Affective Mindfulness Scale-Revised (Feldman et al., 2007) and the self-kindness subscale of the self-compassion scale (Neff, 2003). Self-worth was measured with the self-esteem scale (Sheldon et al., 2001) and the Rosenberg self-esteem scale (Rosenberg, 1965). Competence was measured with the competence satisfaction subscale of the balanced measure of psychological needs (Sheldon & Hilpert, 2012) and the new general self-efficacy scale (Chen, Gully & Eden, 2001). Development was measured with the intentional behavior subscale of the Personal Growth Initiative Scale – II (PGIS II; Robitscheck et al., 2012) and the love of learning subscale of the VIA-IS (Peterson & Seligman, 2004). Purpose was measured with the purpose in life test-short form (Schulenberg, Schnetzer & Buchanan, 2011) and the purposeful life subscale of the meaningful life measure (Morgan & Farsides, 2009). Significance was measured with an adapted version² of the valued life subscale of the meaningful life measure (Morgan & Farsides, 2009) and the meaning in life scale (Huppert & So, 2013). Self-congruence was measured with the Authorship subscale of the index of autonomous functioning (Weinstein et al., 2012) and the Autonomy satisfaction subscale of the balanced measure of psychological needs (Sheldon & Hilpert, 2012). Connection was measured with the relatedness satisfaction subscale of the balanced measure of psychological needs (Sheldon & Hilpert, 2012) and the relatedness satisfaction subscale of the Need Satisfaction and Frustration Scale adapted to one's life in general (Longo, Gunz et al., 2016). Consistent with the SGWB, instructions asked the participants to focus on how they felt in their life in general. Additionally, participants completed a measure of social desirable responding, the short form of the Marlowe-Crowne Social Desirability Scale (MC-A; Reynolds, 1982; Crowne &

² Adapted to one's activities rather than one's life.

Marlow, 1960), and a personality questionnaire, the mini-IPIP measure of the big five personality factors (Donnellan et al., 2006).

In section 2b, SWB was measured with the 5-item satisfaction with life scale (Diener et al., 1985) and the 20-item positive and negative affect schedule (PANAS, Watson et al., 1988). PWB was measured with Ryff's 42-item psychological well-being scales (PWB, Ryff, 1989). Social well-being was measured with Keyes' 15-item social well-being scale (Keyes, 1998). A score of flourishing was then generated by combining scores of SWB, PWB and social-well-being (Keyes, 2002). Before combining scores from questionnaires using different Likert scale lengths (e.g. a 5-point scale and a 7-point scale), they were transformed to z-scores (DeVaus, 2002). Psychological ill-being was measured with the 29-item patient-reported outcomes measurement system (PROMIS-29 v2) (Cella et al., 2010). Finally, to test the perceived value of each well-being construct, participants were asked to what extent they find each construct important in their life (see Table 18).

The follow-up study also included a 12-item concise measure of the dark triad: narcissism, psychopathy and Machiavellianism (Jonason & Webster, 2010). The three traits were found to be well explained by a general factor (Jonason, Kaufman, Webster & Geher, 2013). Therefore, an overall dark triad score was used.

Scale proximity has been found to influence the strength of the correlation between scales. That is, when two groups of items are presented in close proximity to each other, their correlation tends to be higher than when they are presented, say, at the beginning and end of a questionnaire (Weijters, De Beuckelaer & Baumgartner, 2014). To attempt to control for the influence of proximity, the order of the scales in each section was randomized.

6.2. Results

6.2.1. Preliminary analyses. A small amount of missing values was found in each section (less than 1%). The highest amount of missing values for a single item in each section was approximately 1%. The highest amount of missing values for a single case was 7 in section 1, 4 in section 2a, and 15 in section 2b. In the last section, the higher amount was due only to two cases who accidentally skipped a couple of pages on their online survey and later reported doing so to the survey administrator. Most cases were complete listwise: 95% in section 1, 84% in section 2a, and 82% in section 2b. Furthermore, 98% or more cases were complete pairwise

in each section. No problematic nonresponse patterns were identified. For example, the same pattern occurred in no more than 2 observations in section 1 and 2a, and no more than 4 observations in section 2b. Consistent with study 2, full information maximum likelihood was used during confirmatory factor analyses. Furthermore, in correlation analyses using scale scores, full-information maximum likelihood correlation matrices were estimated using the psych package. P-values in these correlation analyses used Holm (1979) adjustment for multiple tests.

All items in the SGWB produced responses covering the entire range of the 5-point Likert scale. On average, means approximated the middle value of the 5-point scale (i.e. 3), although there was a tendency to report slightly higher values (average $M = 3.67$, range = 2.86 – 4.21). Standard deviations indicated sufficient variability in the responses (average $SD = 1.08$, range = 0.83 – 1.26). Absolute values of univariate skewness ($M = -.60$, range = -1.26 and 0.02) and kurtosis ($M = -.24$, range = -0.85 and 0.95) were relatively low. Mardia's multivariate skewness and kurtosis values were above five, indicating a significant deviation from multivariate normality (Bentler, 2006). Therefore, robust maximum likelihood (MLR) estimation was used during factor analyses (Beaujean, 2014). Finally, no problematic multicollinearity was detected, as squared multiple correlations among the 65 items were below .90 (Kline, 2016).

6.2.2. Dimensionality of each scale. The study used the same statistical package, model specifications and fit indices used in the previous study. First, each scale was factor analyzed. Overall, results were consistent with the previous study, indicating that items in each scale fit a unidimensional model (see Table 10). Interestingly, in the purpose scale, CFI and SRMR suggested good fit, while RMSEA was slightly higher than desired, with an upper boundary of the 90% confidence interval slightly above .1. Nevertheless, as previously mentioned, RMSEA tends to over reject well-fitting models when few items are included (Kenny et al., 2015). Therefore, the overall findings supported the dimensionality of the scales.

Table 10 Fit indices of individual scales (Study 3)

Scale	YB χ^2	df	p	CFI	SRMR	RMSEA (90% CI)
Happiness	1.37	2	.504	1.000	.002	.000 (.000-.042)
Vitality	0.37	2	.833	1.000	.001	.000 (.000-.008)
Calmness	1.04	2	.596	1.000	.002	.000 (.000-.037)
Optimism	13.93	5	.016	.996	.008	.043 (.022-.064)
Involvement	4.89	2	.087	.998	.007	.038 (.005-.070)
Self-awareness	5.51	2	.064	.996	.010	.042 (.000-.082)
Self-acceptance	4.40	2	.111	.998	.008	.035 (.000-.071)
Self-worth	13.16	5	.022	.997	.007	.041 (.022-.060)
Competence	24.88	9	.003	.995	.011	.042 (.027-.058)
Development	15.36	5	.009	.996	.008	.046 (.026-.067)
Purpose	42.27	5	.000	.982	.016	.087 (.070-.105)
Significance	18.61	5	.002	.994	.010	.052 (.037-.069)
Congruence	11.05	5	.050	.996	.012	.035 (.013-.056)
Connection	34.42	5	.000	.986	.017	.077 (.058-.098)

N = 989

6.2.3. Dimensionality of the entire questionnaire. The dimensionality of the entire questionnaire was also consistent with the previous study (see Table 11).

Specifically, the one-factor model and the independent-factors model did not fit the data, but the hypothesized correlated factors model fit the data well. Standardized factor loadings were generally high ($M = .86$, range = .59 - .94, see also Appendix 10) indicating that each item adequately contributed to its factor (Hair et al., 2010).

Table 11 Fit indices of the entire questionnaire (Study 3)

Model	YB χ^2	df	p	CFI	SRMR	RMSEA (90% CI)	BIC
One factor	23,798.32	2015	.000	.578	.081	.105 (.103-.106)	154,240
14 independent factors	10,954.42	2015	.000	.827	.425	.067 (.066-.068)	137,737
14 correlated factors	3,305.95	1924	.000	.973	.032	.027 (.026-.028)	128,792
Higher-order	4,091.09	2001	.000	.959	.049	.032 (.031-.034)	129,253
Bifactor	3,725.51	1950	.000	.966	.042	.030 (.029-.032)	129,148

N = 989

All fourteen scales exhibited adequate reliability, with α coefficients ranging from .82 to .92 (see Table 12). The AVE of each factor exceeded .50, supporting internal convergent validity. Squared correlations among factors did not exceed the AVE of each factor, supporting discriminant validity. Furthermore, the highest HTMT value was .78, which was below the .85 cut-off criterion. In other words, both the Fornell and Larker method and the HTMT method supported the discriminant validity of the scale scores.

Table 12 Reliability and discriminant validity analysis (Study 3)

	Hap	Vit	Cal	Opt	Inv	Aw	Acc	Wor	Cop	Dev	Pur	Sig	Cog	Con
ω_h	.92	.91	.92	.86	.89	.82	.84	.91	.88	.91	.88	.90	.85	.86
AVE	.85	.80	.86	.75	.78	.59	.64	.84	.71	.79	.74	.80	.62	.68
MVS	.61	.57	.56	.61	.37	.30	.54	.59	.46	.47	.53	.53	.45	.47

Note: ω_h = McDonald's omega hierarchical; AVE = average variance explained by the factor; MVS = maximum variance shared with other factors.

Consistent with study 2, hierarchical models showed adequate fit, which was slightly worse than the correlated factors model due to an increase in model parsimony. The higher-order model exhibited a strong general factor with loadings ranging from .53 (self-awareness) to .87 (optimism) ($M = .76$). A single general factor was highly reliable ($\omega_h = .86$) and, after controlling for this, additional factors had unacceptable reliability estimates, below .18 (see Appendix 11).

In the bifactor model, loadings on the specific factors remained significant, indicating that, after controlling for the general factor, each item in each scale measures additional construct-specific variance. Furthermore, the general factor was reliable ($\omega_h = .94$). Eigenvalues were 29.18 for the general factor and 19.53 for all specific factors (total = 48.71). Therefore, the general factor explained the majority (i.e. 60%) of the common variance in the items ($ECV = 29.18/48.71 = .60$). Furthermore, since the PUC is purely based on the model structure, and the bifactor model in study 3 was identical to the bifactor model in study 2, PUC remained .94. Overall, the dimensionality of the questionnaire was successfully replicated in the new sample.

6.2.4. Invariance across gender and age. Measurement invariance is met when individuals with the same level of the latent variable, belonging to different groups, answer each item in the scale in same way (e.g. Brown, 2015). For example, males and females with a high overall competence rating should respond to a specific competence item in the same way. When this is the case, measurement invariance is achieved, and the questionnaire can be used to compare latent scores across groups. However, when specific items produce different response patterns across groups with the same latent score, group comparisons based on those latent scores may be biased.

Multigroup CFA was conducted on the entire sample ($N = 989$) to assess whether the questionnaire was invariant across groups that differ in gender and age. Age groups were derived from splitting the continuous age variable into three categories of approximately equal sizes: group 1 ($n = 353$, $M = 25.86$, range = 17-30, $SD = 3.12$), group 2 ($n = 310$, $M = 35.08$, range = 31-40, $SD = 2.88$), and group 3 (n

= 329, $M = 53.20$, range = 41-83, $SD = 7.92$). These three groups did not attempt to represent young, middle-aged and older adults. Instead, the categorization simply aimed to produce three groups of increasing age and approximately equal sample sizes. Multigroup CFA involves several steps (e.g. Beaujean, 2014; Brown, 2015). As a preliminary step, the measurement model was tested separately in each group, and adequate fit was expected in all groups. The goal was to assess the measurement invariance of both the lower-order and the general well-being factor. Therefore, the higher-order model was used for measurement invariance tests. The higher-order model fit the data adequately in all groups (see Table 13). Therefore, it was possible to proceed with invariance testing.

Table 13 CFA in each gender and age group (higher-order model)

Group	YB χ^2	df	p	CFI	SRMR	RMSEA (90% CI)
<i>Gender</i>						
Male	3080.21	2001	.000	.950	.037	.037 (.034-.039)
Female	3487.40	2001	.000	.954	.036	.036 (.034-.038)
<i>Age</i>						
Group 1	3119.44	2001	.000	.941	.040	.040 (.037-.042)
Group 2	3158.85	2001	.000	.939	.043	.043 (.041-.046)
Group 3	3120.64	2001	.000	.939	.041	.041 (.039-.044)

Invariance testing proceeds through the following three steps. In step 1, the model is tested in all groups at the same time and the questionnaire is expected to fit the same factor structure across different groups (e.g. males and females). When this hypothesis is supported, the measure is said to meet configural invariance. In step 2, the factor loadings are constrained to be equal in both groups. That is, each item is expected to measure its factor equally well in both groups (metric invariance). In step 3, in addition to the previous constraints, intercepts are constrained to be equal in both groups. In other words, for each level of the latent variable, the observed scores in each item are expected to be the same in both groups, which is known as scalar invariance. When scalar invariance is achieved, latent scores can be meaningfully compared across groups. Consistent with best practices using higher-order models, each invariance test was conducted sequentially first on the lower- and then higher-order sections of the model (Byrne & Stewart, 2006).

During measurement invariance tests, identification was achieved by standardizing factor variances to 1 rather than choosing an arbitrary marker indicator (Beaujean, 2014). Generally, the two identification methods produce similar results. A limitation of the marker indicator method is that it does not allow to assess the

invariance of the chosen marker indicators while, by fixing the factor variances one, we can examine the invariance of all indicators. However, during the next steps of invariance testing, keeping the variances constrained to equality can also introduce bias, as the latent variances may vary across groups (e.g. Yoon & Millsap, 2007). Therefore, when the loadings were constrained to equality (metric invariance), the latent variances of the factors were fixed to 1 in the first group and freed in the comparison groups (e.g. Little, Slegers & Card, 2006; Millsap, 2011). In other words, when the item loadings were constrained to equality, the lower-order factor variances were freed in the comparison groups. When the higher-order factor loadings were constrained to equality, the higher-order factor variance was freed in the comparison groups. This allowed the factor variances to differ across groups while ensuring the model was identified. Similarly, when the item intercepts were constrained to equality, lower-order factor intercepts (i.e. means) were freed in the comparison groups, and when higher-order intercepts were constrained (i.e. fixed to 0 in all groups), the higher-order factor mean was freed in the comparison groups.

Invariance is determined by assessing changes in model fit. Specifically, invariance is supported when there is no significant difference in fit between one step and the next (Hair et al., 2010). For instance, the metric invariance model should fit as well as the configural invariance model for us to conclude that metric invariance has been achieved. Changes in CFI were used to assess measurement invariance (see section 3.7.8). Specifically, based on results from these simulation studies, it has been recommended that reductions in CFI should not be greater than .01 (Cheung & Rensvold, 2002) and ideally should not be greater than .002 (Meade et al., 2008).

Results of multigroup CFA show that CFI changes (i.e. Δ CFI) never exceeded the value of .002 (see Table 14). In other words, scalar invariance was supported, indicating that scores for each scale and the questionnaire as a whole can be meaningfully compared across gender and age groups. The fact that the chi-square difference was significant during some of the steps, indicates that some significant differences were present, specifically with regards to item and lower-order factor intercepts. However, based on the Δ CFI these differences are minor and do not bias the comparison of overall well-being scores.

Table 14 Multigroup CFA for gender and age groups (higher-order model)

Model	YB χ^2	df	p	SB $\Delta\chi^2$	Δdf	Δp	CFI	ΔCFI
<i>Gender</i>								
Configural invariance	6567.59	4002	.000	-	-	-	.953	-
Metric invariance (lower)	6627.64	4053	.000	52.76	51	.406	.952	.001
Metric invariance (higher)	6648.01	4066	.000	20.25	13	.089	.952	.000
Scalar invariance (lower)	6760.63	4117	.000	118.00	51	.000	.951	.001
Scalar invariance (higher)	6829.93	4130	.000	78.31	13	.000	.950	.001
<i>Age</i>								
Configural invariance	9400.11	6003	.000	-	-	-	.940	-
Metric invariance (lower)	9532.81	6105	.000	126.11	102	.053	.939	.001
Metric invariance (higher)	9580.11	6131	.000	48.21	26	.005	.939	.000
Scalar invariance (lower)	9758.47	6233	.000	181.38	102	.000	.938	.001
Scalar invariance (higher)	9846.67	6259	.000	95.87	26	.000	.936	.002

N = 989

6.2.5. Longitudinal invariance and test-retest reliability. Next, the study examined whether the questionnaire performed consistently over time. In other words, temporal stability (i.e. test-retest reliability) and longitudinal invariance were examined.

First, test-retest reliability was examined for each scale score (i.e. a mean score for each of the 14 scales) as well as the *g* score (i.e. the mean of the 14 scale scores). If scores on the same measure are consistent over time, this would produce a high correlation between the scores on the same scale at time 1 and time 2. Based on other dispositional measures of well-being, the scales were expected to be fairly stable with test-retest correlations approximately ranging between .60 and .80 (Pavot & Diener, 2008; Watson et al., 1988). Pearson's *r* was first estimated first by correlating mean time 1 and time 2 scales scores (see Table 15). Then, a CFA was conducted by specifying two correlated factors corresponding to the scale at time 1 and time 2. Residuals for each item at time 1 were correlated with residuals of the same item at time 2 (Brown, 2015). Individual scales were analyzed, instead of a larger higher-order model, because of the reduced sample size and increased model complexity from the inclusion of correlated residuals among time 1 and time 2 items. Since CFA takes random error into account when estimating factor correlations, the test-retest correlation was expected to be higher when correlating CFA factors compared to average scales scores. However, since each scale achieved adequate internal consistency, only a small amount of random error variance was present. Therefore, the correlations based CFA are only slightly higher than correlations based on average scale scores. As hypothesized, test-retest correlations based on

CFA on each of the fourteen scales ranged from .58 to .86 ($M = .74$), while the general factor was more stable with a correlation of .88.

Additionally, the metric and scalar longitudinal invariance of each scale was tested. If the scale exhibits longitudinal invariance it would indicate that changes in responses over time reflect actual changes in the construct rather than measurement bias (Brown, 2015). If each construct is measured well across time, the questionnaire could be used, for example, in short interventions that aim to improve people's well-being. As a preliminary step, the scales fit the data adequately both at time 1 and time 2, as evidenced by adequate CFI, SRMR and RMSEA values (see Appendix 12). This suggested that invariance testing was appropriate. First, the CFA model used to assess test-retest reliability was used as the configural invariance model. Second, the loading of each item at time 1 was constrained to be equal to its respective loading at time 2, and the factor variance at time 2 was freed (metric invariance). Third, the intercepts were constrained to be equal, and the factor mean at time 2 was freed (scalar invariance) (McArdle & Nesselroade, 2014).

To facilitate the interpretation of the results, only CFI values are presented in Table 15. Reductions in model fit were generally smaller than .002 and, in the only case slightly greater than .002 (i.e. self-awareness), ΔCFI never exceeded .01. Taken together, the results supported the longitudinal invariance of the SGWB.

Table 15 Test-retest reliability and longitudinal invariance

	<i>r</i>	<i>r</i> (CFA)	CFI Configural	CFI Metric	CFI Scalar
Happiness	.74	.77	.998	.998	.998
Vitality	.72	.75	.999	1.00	.999
Calmness	.66	.68	.999	.999	1.00
Optimism	.82	.86	.991	.991	.991
Involvement	.59	.62	.999	.999	1.00
Awareness	.54	.58	.997	.998	.994
Acceptance	.68	.74	.986	.985	.985
Self-worth	.84	.86	.995	.994	.993
Competence	.73	.76	.989	.989	.988
Development	.62	.65	.998	.998	.998
Purpose	.79	.83	.983	.984	.984
Significance	.73	.75	.989	.989	.989
Congruence	.63	.68	.988	.989	.989
Connection	.76	.79	.995	.995	.994
GWB	.84	.88	.930	.930	.928

$n = 431$

6.2.6. Criterion validation. Study 3 also examined the correlations between each scale of GWB, and measures of similar constructs. Each comparison measure was

expected to correlate most strongly with its respective new measure. These hypothesized correlations are presented in bold in Table 16. For example, in the first row, a measure of pleasant affect (or moderate arousal positive affect) was expected to be strongly correlated to the SGWB happiness scale. Indeed, the comparison measure correlated most strongly with happiness ($r = .83$), and its second highest correlation was with calmness ($r = .74$). This provided some support for the convergent and discriminant validity of the happiness scale. In addition, a stricter test was used: Steiger's test of significant difference between the two dependent correlations was carried out with the psych package. This test compared the highest correlation between criterion and SGWB scale (r_{12}) with the second highest correlation between the criterion and another SGWB scale (r_{13}), while taking into account the correlation between two scales in the SGWB (r_{23}) (Steiger, 1980). Results of this test are presented in the last column of Table 16 and show that, for example, the correlation between pleasant affect and happiness was significantly higher than the correlation between pleasant affect and calmness. In all cases but two, the correlation between each well-being scale and its respective comparison measure was the highest. Furthermore, the difference between this high correlation and the second highest for each row was significant in the majority of cases. Overall, the hypothesized correlations were supported.

Table 16 Correlations between each WB scale and external criteria

	Hap	Vit	Cal	Opt	Inv	Aw	Acc	Wor	Cop	Dev	Pur	Sig	Cog	Con	<i>p</i>
Pleasant affect	.83	.67	<u>.74</u>	<u>.72</u>	.49	.34	.54	.71	.52	.56	.53	.61	.54	.65	.000
ESS-happy	.80	.66	<u>.67</u>	<u>.72</u>	.46	.34	.56	.74	.57	.56	.55	.63	.54	.63	.002
Subjective vitality	<u>.75</u>	.82	.67	<u>.71</u>	.54	.33	.52	.71	.60	.62	.62	.65	.51	.55	.001
Energy	<u>.71</u>	.85	.63	.63	.48	.35	.45	.65	.54	.61	.56	.58	.48	.52	.000
Pl. deactivation	<u>.66</u>	.56	.83	.57	.42	.31	.52	.59	.43	.46	.37	.48	.43	.48	.000
Calmness	<u>.42</u>	.34	.59	.38	.30	.27	.38	.37	.24	.30	.25	.30	.29	.33	.000
LOTR	.65	.55	.54	.76	.35	.30	.48	<u>.67</u>	.48	.47	.50	.52	.41	.53	.000
ESS-Optimism	<u>.66</u>	.59	.56	.75	.36	.28	.46	.65	.49	.53	.56	.55	.42	.55	.000
Core flow	.53	<u>.55</u>	.51	.51	.64	.35	.43	.50	.53	.51	.50	.55	.47	.46	.012
Concentration	.44	.48	.44	.44	.62	.36	.40	.45	.50	.40	.42	.41	<u>.55</u>	.41	.030
CAMSR-Aw	.38	.43	.35	.38	<u>.48</u>	.49	.45	.43	.40	.39	.41	.39	.44	.40	.697
PHLMS-Aw	.27	.32	.25	.38	<u>.42</u>	.60	.37	.31	.41	.34	.38	.33	.40	.38	.000
CAMSR-Ac	.48	.43	.45	.49	.39	.41	.63	<u>.53</u>	.47	.35	.38	.41	.44	.45	.000
Self-kindness	.52	.42	.49	.53	.36	.36	.66	<u>.61</u>	.40	.39	.38	.42	.40	.42	.060
Sheldon SE	.72	.61	.62	<u>.75</u>	.47	.40	.67	.87	.64	.52	.57	.63	.58	.59	.000
Rosenberg SE	<u>.72</u>	.59	.61	.71	.48	.35	.60	.83	.62	.50	.54	.62	.56	.59	.000
BMPN-CS	.53	<u>.56</u>	.43	.53	.51	.33	.42	.54	.75	.56	.50	.51	.48	.40	.000
NGSE	.64	.59	.53	<u>.70</u>	.49	.37	.48	.68	.78	.57	.59	.58	.48	.51	.000
PGIS II-IB	.47	.56	.38	.51	.49	.40	.32	.46	<u>.59</u>	.63	.57	.56	.49	.42	.235
VIA-Learn	.22	.22	.16	.22	.26	.29	.18	.25	.41	.34	.28	.25	.35	.25	N/A
PILSF	.65	.61	.52	.68	.47	.36	.47	.67	.59	.60	.77	<u>.72</u>	.53	.55	.010
MLM-Purpose	.47	.46	.37	.53	.39	.26	.35	.50	.48	.53	.73	<u>.58</u>	.37	.41	.000
MLM-Valued	.60	.56	.52	.61	.56	.36	.50	.62	.53	.55	<u>.63</u>	.71	.54	.55	.001
ESS-Significant	.65	.56	.53	<u>.68</u>	.47	.30	.47	.67	.49	.55	.63	.76	.47	.55	.000
Authorship	.53	.50	.47	.53	.53	.51	.52	.55	<u>.56</u>	.49	.51	.54	.77	.53	.000
BMPN-AS	.54	.53	.52	.55	.55	.39	.51	.57	.52	.50	.47	.54	.55	.51	N/A
BMPN-RS	.55	.44	.44	.54	.42	.32	.46	.53	.44	.39	.49	<u>.55</u>	.46	.73	.000
NSFS-RS	<u>.59</u>	.54	.51	.57	.44	.35	.53	.57	.46	.42	.47	.53	.48	.82	.000

Note: Correlations hypothesized to be the highest are in bold; 2nd highest correlations underlined. All correlations were significant ($p < .01$). The names of criterion measures are shortened here but are presented in exactly the same order and their complete names in the Measures section (6.1.3). $n = 486$.

Scale scores were then averaged into a general well-being composite score.

The relationship between general well-being and previous measures of well-being was strong: .77 with SWB, .81 with PWB, .61 with social well-being, and .86 with flourishing. All correlations were significant ($p < .001$, $n = 503$). Additionally, the relationship between general well-being, social desirability and personality traits was explored. As hypothesized, well-being exhibited a moderate correlation with social desirability ($r = .29$), strong correlations with extraversion ($r = .46$) and neuroticism ($r = -.57$), moderate correlations with agreeableness ($r = .32$) and conscientiousness ($r = .37$), and a small correlation with imagination ($r = .16$). All correlations were significant ($p < .001$, $n = 486$). Additionally, a small correlation was found with the general dark triad score ($r = -.22$, $p < .001$, $n = 431$). In other words, the general well-being score was related to external measures as predicted, further supporting its validity as a well-being measure.

To measure psychological ill-being, five subscale scores from the patient-reported outcomes measurement system (PROMIS-29 v2) (Cella et al., 2010) were used. The three remaining PROMIS subscales were omitted because they focused on physical aspects (e.g. physical pain), which were not the focus of the study. Generally, correlations were moderate to strong in magnitude (see Table 17). All correlations were significant ($p < .05$), with only one exception between self-awareness and sleep disturbance ($r = -.07$, $p = .112$). In short, well-being scales related to ill-being scales as predicted, although self-awareness seemed to have more modest correlations to ill-being.

Table 17 Correlations between well-being and ill-being scales

	Hap	Vit	Cal	Opt	Inv	Aw	Acc	Wor	Cop	Dev	Pur	Sig	Cog	Con
Anxiety	-.52	-.44	-.59	-.46	-.38	-.15	-.46	-.52	-.38	-.28	-.30	-.41	-.33	-.34
Depression	-.67	-.55	-.60	-.59	-.45	-.18	-.49	-.64	-.47	-.40	-.43	-.54	-.43	-.47
Fatigue	-.52	-.51	-.51	-.43	-.39	-.16	-.40	-.47	-.34	-.25	-.29	-.35	-.30	-.39
Sleep disturbance	-.39	-.37	-.42	-.34	-.27	-.07	-.32	-.37	-.25	-.18	-.19	-.30	-.21	-.32
Ability to participate	-.45	-.43	-.44	-.41	-.35	-.15	-.34	-.41	-.31	-.27	-.29	-.38	-.30	-.34

$n = 503$

6.2.7. Construct value. Finally, the perceived value of each well-being construct was assessed. Symptoms of good mental health are assumed to be worth experiencing for their own sake (Alexandrova, 2012; Seligman, 2011). Therefore, each construct measured in the SGWB was expected to exhibit a high subjective value rating (see Chapter 2). Specifically, on a 7-point scale, a score above the middle point (i.e. 3.50) was expected. Participants reported mean scores around 6 (see Table 18), indicating that every construct included in the SGWB had a high perceived value in this sample, which further supported the inclusion of these constructs in the questionnaire.

Table 18 Perceived value of each well-being construct

Construct	Item	M	95% CI	
Happiness	Feeling happy and cheerful	6.04	5.94	6.14
Vitality	Feeling energetic / full of energy	5.80	5.69	5.91
Calmness	Feeling calm / relaxed	6.02	5.92	6.11
Optimism	Being optimistic and hopeful	6.12	6.02	6.22
Involvement	Feeling completely involved and engaged in what you do	5.89	5.79	5.98
Awareness	Being in touch with how you feel	5.75	5.64	5.86
Acceptance	Accepting yourself the way you are	6.10	6.00	6.21
Self-worth	Liking yourself a lot	5.99	5.88	6.10
Competence	Feeling highly effective at what you do	6.05	5.96	6.15
Development	Feeling you're consistently improving, developing and advancing	6.04	5.94	6.14
Purpose	Having a purpose and a mission in life	6.02	5.91	6.14
Significance	Feeling that what you do is important and worthwhile	6.13	6.03	6.23
Congruence	Feeling that what you do is consistent with how you see yourself	5.89	5.78	6.00
Connection	Feeling close and connected with the people around you	5.94	5.83	6.06

Instructions: rate how important each of the following things are in your life. Do not focus on whether these are currently present in your life. Instead, think whether you would want these things in your ideal life.

Rating scale: (1) no importance whatsoever, (7) extraordinarily important and valuable (Diener, 2000).

6.3. Discussion

In study 3, the scales of general well-being were subjected to dimensionality, invariance, reliability and criterion validity analyses in a new North American sample. Once again, CFAs were conducted on each scale individually as well as the entire well-being questionnaire. Furthermore, the stability of the hierarchical structure of the questionnaire was tested over a 5-week period. Finally, correlation analyses were used to examine the relationships between the questionnaire's scores, theoretically-similar health constructs and theoretically-distinct constructs.

Tests of dimensionality in study 3 were consistent with previous results from study 2. In study 3, the factor structure of the questionnaire was supported: individual scales fit unidimensional models while the entire questionnaire fit a 14 correlated factors model and both higher-order and bifactor models. The 14-factor model was also supported by the fact that each factor exhibited a higher explained variance relative to its shared variance with other factors, using both the Fornell and Larker method and the HTMT method. While, in study 2, self-acceptance and self-worth exhibited a slightly higher shared variance, this was not the case in study 3, where all factors were related to but distinct from each other. The hierarchical factor structure of the questionnaire was also found to be invariant across gender, age and a 5-week time gap. Internal consistencies for each scale as well as the general factor of

well-being were adequate. Test-retest reliabilities ranged approximately from adequate to excellent based on the test review criteria from the European Federation of Psychology Associations (Evers et al., 2013), supporting the use of the questionnaire as a dispositional measure of well-being.

Furthermore, the study provided criterion evidence of validity. Each scale was more strongly related to measures of similar constructs than measures of different constructs. For example, the new happiness scale was more strongly related to previous happiness scales than optimism scales. As hypothesized, the scales were also negatively related to measures of ill-being. Furthermore, a general well-being score was strongly related to other well-being questionnaires and exhibited correlations with measures of the big five personality traits, social desirability and dark triad traits, which were consistent with its hypothesized relationships. Finally, all well-being constructs were rated as important in the current sample, which further supports the notion that well-being states have a high subjective value for individuals.

Taken together, the results provide substantial evidence in support of the dimensionality, reliability and validity of the scale scores in the present sample. The findings of the studies, as well as their implications and limitations, are further discussed in the next chapter.

Chapter 7: General discussion

7.1. Findings

The aim of this thesis was to develop a well-being questionnaire to measure specific well-being constructs as well as well-being as a whole. Several measures of well-being have been developed in recent years. In an effort to measure a relatively comprehensive list of well-being constructs, the thesis identified commonalities among six well-known conceptualizations and their respective measures. Based on these commonalities, fourteen constructs were identified: happiness, vitality, calmness, optimism, involvement, self-awareness, self-acceptance, self-worth, competence, development, purpose, significance, self-congruence and connection. These were defined conceptually and, in study 1, an item pool was developed to measure each construct. In total, 147 items were subjected to content validation by six academics with expertise on the subject. Based on their quantitative and qualitative feedback, 129 items were selected, which clearly and unambiguously tapped into their intended construct.

In study 2, these selected items were administered to a large North American sample ($N = 507$), and the data were factor analyzed. Of the 129 items, 65 were retained, as they adequately fit the hypothesized factor structure. Specifically, each item adequately measured its specific factor (e.g. calmness). The fourteen scales adequately fit a correlated factors model and exhibited residual correlations close to 0. These results suggest that the questionnaire produces scores that approximate a simple structure, which would likely replicate across studies. The fourteen factors also showed evidence of discriminant validity, by explaining a higher amount of variance within each factor (AVE) than the maximum variance shared across factors (MVS). Even though the self-acceptance factor shared slightly more variance with the self-worth factor (MVS) than it explained in its own items (AVE), the two were maintained separate, as they fit a two-factor model but not a one-factor model. Additionally, all scales exhibited HTMT values below .85, indicating that each scale explained more variance in its own items compared to the variance shared with other scales. Additionally, each scale exhibited a good balance between adequate reliability ($\omega_h > .81$) and content coverage, which was assessed during content validation. Two hierarchical models were then used to test whether the scores could be explained by a single general factor of well-being. Both a higher-order and a

bifactor model fit the data well. Most importantly, in both models, the general well-being score was highly reliable ($\omega_h > .86$).

Study 3 attempted to replicate the results in a new North American sample ($N = 989$). Indeed, results of factor analyses were consistent with study 2. Specifically, each scale fit a one-factor model and all fourteen scales fit a correlated factors model. All factors, including self-acceptance, showed strong evidence of discriminant validity. Adequate reliability was found in the fourteen scale scores ($\omega_h > .82$) as well as the general well-being scores ($\omega_h > .86$), and hierarchical models fit the data adequately. Additionally, the higher-order model was used to test measurement invariance of both the lower- and higher-order factors across gender and age, which was supported. Longitudinal invariance and test-retest reliability were tested in a subset of the sample ($n = 431$) that completed the questionnaire again five weeks after the original administration. Although the follow-up sample did not include all the initial participants, its size was adequate for these analyses, which require more than 100 (for test-retest reliability) or 400 (for CFA) respondents (e.g. Kline, 2000; Jackson, 2003). Several scales exhibited a substantial trait component as evidenced by a mean latent factor correlation of .74. This indicates that, on average, 55% of each scale's variance is stable over time. The awareness scale had the lowest retest correlation ($r = .58$), indicating that the scale has a stronger state component than other WB constructs. On the other hand, a general well-being score had the highest retest correlation ($r = .88$), indicating that the majority of the variance in general well-being is stable over a 5-week period. Indeed, further research is needed to examine temporal stability over longer time spans (e.g. several months or years). Most importantly, the hierarchical structure of well-being was invariant across the 5-week period, which supports the use of the scale in short-term well-being interventions. However, it should be pointed that the follow-up sample was a subset of the original sample, and included older (on average by 2.79 years) and more female participants. Although statistically significant, the difference is small and unlikely to have influenced the results in a substantial way. Taken together, the findings from studies 2 and 3 support the hypothesized factor structure of the questionnaire.

Study 3 also assessed the relationships between the new questionnaire and external criteria. First, the relationship between each scale in the SGWB (e.g. happiness) and two measures of the same or similar constructs was examined. Each

scale in the SGWB was more strongly related to measures of the same or similar constructs (e.g. moderate arousal positive affect) than measures of distinct constructs (e.g. purpose in life). In most cases, Steiger's test showed a significant difference between the hypothesized convergent relationships (e.g. happiness and moderate arousal positive affect) and the hypothesized divergent relationships (e.g. happiness and purpose in life), thus providing strong criterion evidence of validity. Only in one case moderate criterion evidence was found. The development scale in the SGWB was correlated with two criterion measures: the intentional behavior subscale of the Personal Growth Initiative Scale – II (Robitscheck et al., 2012) and the love of learning subscale of the VIA-IS (Peterson & Seligman, 2004). These measures did not exactly tap into the same construct as the SGWB development scale. The intentional behavior of the PGIS-II subscale focuses on behaviors related to personal development; the VIA-IS subscale focuses on learning activities and preferences (e.g. reading, visiting museums); on the other hand, the SGWB development scale focuses on the subjective experience of personal growth and improvement. The PGIS-II subscale was most strongly related to the SGWB development scale ($r = .63$), but this correlation was not significantly higher than its second highest correlation with competence ($r = .59$). Additionally, the VIA-IS subscale's highest correlation was with the SGWB competence scale ($r = .41$) rather than development ($r = .34$). These correlation patterns may be due to the fact that the three measures focus on related but slightly different constructs. Nevertheless, this correlation analysis indicated strong convergent and discriminant evidence for thirteen of the fourteen scales of GWB, and moderate convergent and discriminant evidence for the development scale.

Criterion evidence of validity was also found when examining the relationships between the scales of GWB and measures of psychological health problems. As hypothesized, all scales exhibited negative correlations with anxiety, depression, fatigue, sleep disturbance and the ability to participate in social roles. Only the correlation between self-awareness and sleep disturbance was non-significant ($r = -.07, p = .112$). Furthermore, self-awareness seemed to be more poorly correlated to psychological ill-being compared to other well-being constructs. Nevertheless, the consistent negative relationships between the SGWB and psychopathology scales provide further criterion evidence of validity for the SGWB.

A general well-being score from the SGWB was strongly correlated with previous measures of well-being (r range = .61 – .86). The most relevant correlation was between a general SGWB score and the general score of flourishing: a general well-being score derived from previous measures (Keyes, 2002). The two scores correlated at .86. Even though this relationship is slightly lower than those found in, for instance, the general factor of intelligence derived from different test batteries (Johnson, Bouchard, Krueger, McGue & Gottesman, 2004; Johnson, te Nijenhuis & Bouchard, 2008), it nevertheless indicates that the SGWB taps into the same general construct as previous measures of well-being (Brown, 2015).

In comparison, its correlations with the big five personality traits, a general dark triad score, and social desirability were lower and consistent with the hypothesized correlation pattern. In other words, correlations between the general well-being score of the SGWB and comparison measures followed the hypothesized correlation pattern, thus providing further criterion evidence supporting the use of the new measure.

Finally, each construct measured in the SGWB was assigned a high subjective value rating. In other words, individuals in the current sample viewed each subjective experience included in the SGWB as important in their life. This was consistent with the notion that well-being constructs are intrinsically valuable (e.g. Alexandrova, 2012). Previous studies had focused on the subjective value of happiness (e.g. Diener, 2000). The present findings further contribute to this literature by providing evidence about the subjective value of several other well-being constructs.

Taken together, the findings have several implications for the conceptualization and measurement of well-being, which are discussed next.

7.2. Implications

7.2.1. A new measure of well-being. The main implication of the studies is the finding that the new questionnaire produced scores that exhibited a factor structure and correlations consistent with the hypotheses. In other words, the questionnaire is supported by consistent evidence of dimensionality, reliability and validity. Not only it measures many constructs included in a variety of previous measures of well-being, but it does so while overcoming some of the dimensionality and discriminant validity issues found in previous measures of PWB (e.g. Abbott et al., 2006; Chen et al., 2013; Burns & Machin, 2009; Springer & Hauser, 2006), eudaimonic well-being

(Schutte et al., 2013), or Seligman's flourishing (Kern et al., 2015). In these cases, the original factor structure proposed by different authors was not corroborated, leading to a higher or lower number of factors being supported. On the other hand, the evidence suggests that the constructs measured by the new measure are more distinguishable from each other than the constructs presented in some previous measures. Furthermore, the new measure provides better reliability than the alternative single-item measures (Diener et al., 2010; Huppert & So, 2013) as well as short measures of PWB (Ryff, 1989). Based on current evidence, the use of the SGWB can be deemed adequate in adult North American samples drawn from the general population.

7.2.2. From a psychometric view, is there a general factor of well-being? In studies 2 and 3, both higher-order and bifactor models fit the data adequately. This indicates that these models adequately reproduced the interrelationships among items. Most importantly, in both models, the general factor saturation was high. In other words, the general factor in the higher-order model accounted for the majority of the covariance in the lower-order factors, and the general factor in the bifactor model accounted for the majority of the covariance in the items. These results are consistent with previous findings supporting a general factor of well-being using a variety of positive feeling and positive functioning measures, as outlined in Chapter 2 (e.g. Disabato et al., 2016; Jovanović, 2015; Longo, Coyne et al., 2016).

The present thesis and these previous studies indicate that when higher-order or bifactor models are tested using different well-being measures, they consistently identify a strong general factor of well-being. Therefore, if we accept the premise that these measures tap into well-being reasonably well, then we could conclude that well-being exhibits a hierarchical structure with a primarily unidimensional general factor.

The strength of the general factor of well-being (i.e. $\omega_h > .86$) can be compared to the strength of the general factor measuring other constructs. For instance, Revelle and Wilt (2013) examined the ω_h values for the general factor in several personality and mental ability scales. Across data sets, the general factor of personality did not adequately account for the variability in the data ($\omega_h = .37$), while the general factor of mental ability showed greater factor saturation ($\omega_h = .74$). From these findings, Revelle and Wilt concluded that mental ability exhibited a stronger general factor than personality. In comparison, it appears that the general factor of well-being

exhibits a saturation that is much higher than the general factor of personality, and slightly higher than the general factor of intelligence.

The fact that a single general factor adequately reproduces the interrelationships in well-being data is important because “we should pick the simplest explanation of those that fit the facts” (Kline, 2000, p. 137). If a single general factor fits the data adequately, it may not be necessary to specify two or more general factors. Given two models of adequate and comparable fit, it is often appropriate to retain the simpler and more parsimonious of the two, even though its fit to data is often slightly lower (e.g. Bollen, 1989; Kim & Mueller, 1978). Choosing more complex models is not needed because, after controlling for a single higher-order factor, the omega reliability analysis showed that additional factors were unreliable. In short, from a psychometric perspective, the present studies are consistent with previous findings indicating that well-being experiences are explained by a common general factor.

7.2.3. Which hierarchical model is more suited to study SGWB scores? In both studies 2 and 3, the bifactor model fit the data better than the higher-order model. However, the bifactor model includes more parameter estimates than the higher-order model and, as the number of parameter estimates increases, model fit tends to improve (e.g. Kline, 2016). In fact, simulation studies found that the bifactor model tends to fit the data better, even when these have a higher-order factor structure, indicating the presence of an “inherent bias favouring the fit of the bi-factor model” (Murray & Johnson, 2013, p. 407).

When comparing two models, a ΔCFI greater than .010 is often interpreted as a practically significant difference in fit between the two (Vandenberg & Lance, 2000). However, based on simulation results, a greater difference in fit is warranted to choose the bifactor model over the higher-order model (Murray & Johnson, 2013). In study 3, the higher-order and bifactor models had a CFI of .959 and .966, respectively, with a difference (ΔCFI) of .007. In other words, the difference in model fit may not be great enough to justify favoring the less parsimonious bifactor model.

Considerations beyond model fit are also required to model data from the SGWB. When it comes to scoring and modeling the general factor, it makes little difference if we use a higher-order or bifactor model, as the general factor scores in either model tend to be extremely highly correlated (e.g. .96) (see Murray &

Johnson, 2013). The general factor in the bifactor model may be slightly biased when a different number of items is used in different group factors, leading to a general factor score that over-represents some constructs over others. For example, in the SGWB the competence construct may be over-represented (6 items) relative to the vitality construct (4 items). On the other hand, the general factor in the higher-order model would not be influenced by such variations because it is derived by the lower-order factors (Murray & Johnson, 2013). That said, the extremely high correlation between the two suggests that this difference may not have practically significant consequences.

This is consistent with findings indicating that, when the general factor in a bifactor model explains the majority of the variance in the items, a single average score will produce relatively unbiased estimates, even if the questionnaire exhibits some multidimensionality. Specifically, based on recent simulations, high ECV and PUC values, like the ones found in studies 2 and 3, indicated that if all items were simply averaged in a single score (*g*), this would have minimal bias (e.g. below 3%), despite the lack of fit of the one-factor model (Bonifay et al., 2015). That said, considering that the higher order model did not fit much worse than the bifactor model, all specific factor loadings in the bifactor model remained significant, and each specific factor includes a different number of items, the preferred approach to producing a general well-being score would still be first to compute mean scores for each scale and then average these mean scale scores into a *g* score.

While the choice of hierarchical model may have a small influence on general factor inferences, it may produce significantly different estimates in the lower-order and group or specific factors. The bifactor model and the higher-order model allow the testing of different hypotheses. In the bifactor model, the influence of the general factor on the items is direct and specific factors explain the covariance in the items after controlling for the influence of the general factor. However, in the higher-order model, the influence of the general factor on the items is indirect (through the lower-order factors) and the lower-order factors attempt to explain all the covariance in the items. Therefore, the bifactor model may be favored when assessing the relationships between specific factors and external criteria, after controlling for the general factor (Murray & Johnson, 2013). On the other hand, when assessing the relationship between the general factor and specific factors, the higher-order model may be favored because, in bifactor models, factor correlations are often assumed to be 0

(Holzinger & Swineford, 1937) and estimating these factor correlations may lead to non-convergence.

In short, the present studies show compelling evidence supporting the use of a higher-order model to study scores from the SGWB. Nevertheless, the bifactor model can also be appropriate in future studies testing hypotheses that are not consistent with the higher-order model.

7.3. Limitations and future directions

7.3.1. Subjective choice of models and constructs. The main limitation of the studies is a degree of subjectivity in the choice of the models reviewed and the choice of lower-order factors. The final list of constructs measured in the questionnaire was influenced by the operationalizations reviewed prior to scale development. This limitation is somewhat reduced by the fact that the chosen operationalizations were repeatedly presented in major reviews of the field (e.g. Kashdan et al., 2008; Keyes, 2015), and by choosing lower-order factors that were repeatedly presented in these previous operationalizations. Because of this strong overlap among previous operationalizations, the reader will likely notice that including or omitting any one previous operationalization would not substantially alter the final choice of lower-order factors.

Similarly, the process of conceptualizing each construct is usually somewhat subjective (Jaccard & Jacoby, 2010). As is often the case, different researchers are likely to include a different number of constructs to measure well-being (e.g. Ryff, 1989; Waterman et al., 2010). Nevertheless, the SGWB produced an overall well-being score that was strongly correlated to previous measures of well-being, indicating that the questionnaire taps into the same general construct as previous measures. Furthermore, each scale in the SGWB exhibited a pattern of association with previous measures that was consistent with its hypothesized pattern, indicating that the new operationalization was systematically related to previous operationalizations in a way that was consistent with prior hypotheses.

That said, previous conceptualizations and measures were mainly developed in western societies. Consequently, the conceptualization and measure presented here may be influenced by this cultural context. The constructs included in both the SGWB as well as previous measures may reflect specific values and cultural norms. Additionally, the SGWB were validated in North American samples, and specifically

with individuals who had access to computers or portable devices as well as an internet connection. In other words, the samples may not reflect the experiences reported by individuals in worse circumstances. These circumstances were not specifically addressed by the thesis. Similarly, the thesis did not attempt to assess well-being in a representative sample including a more balanced percentage of males and females, or a wider variety of age groups (e.g. children and adolescents). Therefore, further studies are required to assess the validity of the scale in a broader range of socio-economic and cultural contexts (this is further discussed in section 7.3.3).

Finally, while subjective experiences and evaluations seem to be a fundamental component of mental health, based on the psychological conceptualizations and measures reviewed here, it must be acknowledged that well-being, more broadly, may be defined by other constructs, which have not been considered in the thesis. Therefore, while the conceptualization and measure presented here may help provide a fuller picture of mental well-being, it is possible that additional constructs, beyond subjective experiences and evaluations, will need to be further explored in subsequent research. Similarly, the relationship between mental health problems and well-being seems to deserve further attention.

7.3.2. The use of higher- and lower-order factor scores. The SGWB can produce both a general well-being score as well as fourteen specific scale scores. A general well-being score would indeed adequately reflect response variations in the questionnaire. This single score may be intuitively appealing because of its simplicity. However, it may also be beneficial to focus on lower order constructs, as these can be specifically targeted during interventions (Huppert & So, 2013). For example, an individual may exhibit an average well-being score with relatively high scores on most components except purpose and self-acceptance. In this case, targeted interventions focusing on these constructs may be more beneficial than broader interventions aiming to improve one's general well-being. This argument is also in line with personality research on the bandwidth-fidelity dilemma, which found that using specific (facet-level) scales improves measurement precision and prediction than aggregating facets into broader personality traits (Briley & Tucker-Drob, 2012; Cheng, Wang & Ho, 2009; Vasilopoulos, Cucina & Hunter, 2007).

At the same time, individual scale scores are not without limitations. Each lower-order well-being scale is only intended as an imperfect indicator of general

well-being. While these factors are often perceived as valued and are related to health outcomes, they may sometimes lead to undesirable consequences. For example, a study by Fredrickson and Losada (2005) found that experiencing too little, but also too much, positive affect is linked to lower levels of general well-being, measured with Keyes's scale. Similarly, involvement (or flow) has been linked to online game addiction, problematic internet use, and risky behaviors (Chou & Ting, 2003; Schöler & Nakamura, 2013; Thatcher, Wretschko & Fridjhon, 2008). Furthermore, optimism has been linked to a lower likelihood to disengage from gambling even after losses (Gibson & Sanbonmatsu, 2004), and to lower marital satisfaction when one's partner does not meet one's optimistic expectations (McNulty & Karney, 2004). These findings suggest that any well-being dimension need not be experienced all the time, or at its highest level, in order to be beneficial. Instead, each construct is likely to contribute to one's psychological health to some extent and in some contexts.

In short, it may be desirable to study well-being using both a general well-being score and individual scales scores, as both approaches exhibit strengths and limitations. However, further research is needed to explore these strengths and limitations in greater depth.

7.3.3. The structure of well-being across cultures. Based on the sample size and heterogeneity in the present studies, it is expected that the factor structure of well-being identified here will generalize to other adult North American samples. Indeed, samples from the Amazon Mechanical Turk have been shown to be more heterogeneous than student samples (Buhrmester et al., 2011) and to show demographics comparable to large U.S. stratified surveys (Huff & Tingley, 2015). However, the structure of well-being constructs has been shown to vary across cultures or contexts (e.g. Russell, 1980). Therefore, if the questionnaire is adapted to a specific language or culture, some constructs may not fit the current factor structure and may need to be omitted. Alternatively, constructs which adequately measure well-being only in specific contexts or cultures may be added where appropriate. Therefore, if the structure of well-being varies across cultures, the SGWB may need to be modified to account for such variation.

7.3.4. Self-report measures. Criterion validation was carried out with data from external self-report measures. The use of self-report measures was deemed adequate because the constructs under study were *subjective* experiences and evaluations

(Spector, 2013). However, the use of self-report measures has been argued to potentially lead to common method bias (Spector, 2013). That is, two unrelated scales may show a substantial correlation simply because they were measured using the same method (e.g. a self-report questionnaire). However, current results seem to alleviate such concerns. Study 3 examined the relationships between a general well-being scale and external criteria, with which well-being was hypothesized to exhibit small correlations (e.g. $r = .20$). Both general well-being and criterion measures were assessed using the same self-report method. If the data were substantially biased by common method variance, moderate to strong correlations may have been found, for example, between well-being and imagination. However, consistent with previous findings (Donnellan et al., 2006), general well-being was poorly correlated with imagination ($r = .16$). Therefore, current findings do not seem to be unduly influenced by common method bias, caused by the use of self-report measures.

Despite, the belief that common method variance inflates correlations among variables, this has been found not to significantly influence inter-variable relationships (Spector, 2006). Additionally, a simulation study found that even if scores included method variance, this would not substantially influence the results, as its potentially inflating effect is offset by the attenuating effect produced by random error variance in the scores (Lance, Dawson, Birkelbach, & Hoffman, 2010).

In short, current evidence suggests that self-report scales are adequate to measure subjective constructs like the ones studied here. That said, for the purpose of questionnaire validation, future studies could further validate the SGWB by assessing the questionnaire's associations with non-self-report criteria. For example, future studies could investigate the relationships between well-being constructs and biological correlates such as REM sleep duration, daily salivary cortisol and cardiovascular risk (e.g. Ryff, Singer & Love, 2004).

Finally, the construct tapped by each scale could be more or less apparent to survey respondents. Items are direct, when the construct being measured is apparent to the survey respondents (e.g. the happiness item "I am happy"), or indirect, when the construct measured is more difficult to infer. The suitability of a direct or indirect approach to item writing depends on the purpose for which the questionnaire is developed. For selection purposes, or when dealing with sensitive topics, the indirect approach is often warranted. For example, when measuring vindictiveness, the item "I'm vindictive" may produce biased responses due to the social undesirability of the

construct. This item may instead need to be indirectly phrased as follows: “vindictive savagery is the motivation of most revolutionaries” (Kline, 2000, p. 275). Virtually all well-being scales reviewed, and consequently the SGWB use a direct approach, because they do not generally focus on particularly sensitive topics. Beyond the settings outlined thus far, the direct approach may not be suitable in studies aiming to avoid the respondents’ personal opinions about topics such as happiness or vitality. Therefore, it must be acknowledged that the SGWB may not be appropriate as indirect measures.

7.3.5. Cross-sectional design. The present studies adopted mainly a cross-sectional design, with a short follow-up study to establish longitudinal invariance. The cross-sectional design was adequate to provide concurrent evidence of criterion validity. However, longitudinal or experimental designs could be used in future studies to allow for causal inferences while exploring long-term outcomes of well-being over several years.

7.3.6. Short form. Finally, while the SGWB exhibit good psychometric properties, the overall 65-item questionnaire is fairly long. This may preclude its inclusion in large national surveys, where shorter questionnaires are typically selected, such as the 12-item general health questionnaire (GHQ; Goldberg & Williams, 1988), which was included, for instance, in the British Household Panel Survey. Indeed, the long form is preferable to examine the lower- and higher-order factor structure, reliability and criterion validity of the SGWB. Nevertheless, if necessary, future studies could develop a short form of the SGWB by selecting items with the best psychometric properties (e.g. factor loadings, invariance, test-retest reliability) from the current version of the questionnaire. This version would limit the range of feasible analyses on the data, as it would no longer be possible to assess, for instance, lower-order reliability. However, a short form may still be useful by providing a score of general well-being while lowering questionnaire completion times.

7.4. Conclusion

The thesis contributes to the current literature by presenting a psychometrically sound questionnaire measuring general well-being as well as fourteen specific well-being constructs. In three studies, questionnaire scores exhibited adequate content validity ratings, factor structure, internal consistency, test-retest reliability, invariance across age, gender and a 5-week period, and

relationships with external criteria that were consistent with the hypothesized pattern. While some scales exhibited stronger evidence than others at different stages of the validation process, all scales exhibited adequate psychometric properties across the studies. Additionally, the thesis further replicates recent findings supporting the presence of a single strong general factor of well-being (e.g. Disabato et al., 2016; Jovanović, 2015; Longo, Coyne et al., 2016), which allows researchers to model well-being using more parsimonious measurement and structural models.

References

- Abbott, R. A., Ploubidis, G. B., Huppert, F. A., Kuh, D., Wadsworth, M. E., & Croudace, T. J. (2006). Psychometric evaluation and predictive validity of Ryff's psychological well-being items in a UK birth cohort sample of women. *Health and Quality of Life Outcomes*, 4, 76-92.
- Aghababaei, N., & Błachnio, A. (2015). Well-being and the Dark Triad. *Personality and Individual Differences*, 86, 365-368.
- Alexandrova, A. (2012). Well-being as an object of science. *Philosophy of Science*, 79(5), 678-689.
- Allison, P. D. (2009). Missing Data. In R. E. Millsap, & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology*, (pp. 72-89). Thousand Oaks, CA: Sage Publications.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.
- Alfonso, V. C., Allison, D. B., Rader, D. E., & Gorman, B. S. (1996). The extended satisfaction with life scale: Development and psychometric properties. *Social Indicators Research*, 38(3), 275-301.
- Allport, G.W. (1961). *Pattern and growth in personality*. New York: Holt, Rinehart, & Winston.
- Andrews, F. M., & Withey, S. B. (1976). *Social indicators of well-being: America's perception of life quality*. New York, NY: Plenum Press.
- Aristotle. (2009). *Nicomachean ethics* (D. Ross, Trans.). Oxford: Oxford University Press.
- Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, 13(1), 27-45.

- Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42(5), 815-824.
- Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles? *Psychological Science in the Public Interest*, 4(1), 1-44.
- Baumeister, R. F., Smart, L., & Boden, J. M. (1996). Relation of threatened egotism to violence and aggression: The dark side of high self-esteem. *Psychological Review*, 103(1), 5-33.
- Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, 13(2), 186-203.
- Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. New York: Routledge.
- Beaujean, A. A. (2015). John Carroll's Views on Intelligence: Bi-Factor vs. Higher-Order Models. *Journal of Intelligence*, 3(4), 121-136.
- Beck, J. S. (2011). *Cognitive behavior therapy: Basics and beyond*. New York: Guilford Press.
- Bentler, P. M. (2006). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.
- Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, 16(1), 78-117.
- Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20(3), 351-368.
- Blore, J.D., Stokes, M.A., Mellor, D., Firth, L., & Cummins, R.A. (2011). Comparing multiple discrepancies theory to affective models of subjective wellbeing. *Social Indicators Research*, 100(1), 1-16.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

- Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling*, 22(4), 504-516.
- Bostic, T. J., Rubio, D. M., & Hood, M. (2000). A validation of the subjective vitality scale using structural equation modeling. *Social Indicators Research*, 52(3), 313-324.
- Bradburn, N. M. (1969). *The structure of psychological well-being*. Chicago: Aldine.
- Brandstätter, M., Baumann, U., Borasio, G. D., & Fegg, M. J. (2012). Systematic review of meaning in life assessment instruments. *Psycho-Oncology*, 21(10), 1034-1052.
- Briley, D. A., & Tucker-Drob, E. M. (2012). Broad bandwidth or high fidelity? Evidence from the structure of genetic and environmental effects on the facets of the Five Factor Model. *Behavior Genetics*, 42(5), 743-763.
- Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, 18(1), 36-52.
- Brown, T.A. (2015). *Confirmatory factor analysis for applied research*. London: The Guilford Press.
- Brown, T. A., Chorpita, B. F., & Barlow, D. H. (1998). Structural relationships among dimensions of the DSM-IV anxiety and mood disorders and dimensions of negative affect, positive affect, and autonomic arousal. *Journal of Abnormal Psychology*, 107(2), 179-192.
- Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, 84(4), 822-848.
- Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, 36(1), 111-150.
- Buhler, C. (1935). The curve of life as studied in biographies. *Journal of Applied Psychology*, 19, 405-409.

- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1), 3-5.
- Burns, R. A., & Machin, M. A. (2009). Investigating the structural validity of Ryff's psychological well-being scales across two samples. *Social Indicators Research*, 93(2), 359-375.
- Butler, J. & Kern, M.L. (in preparation). The well-being theory questionnaire.
- Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, 13(2), 287-321.
- Campbell, A., Converse, P.E., & Rodgers, W. L. (1976). *The quality of American life*. New York, NY: Russell Sage Foundation.
- Cardaciotto, L., Herbert, J. D., Forman, E. M., Moitra, E., & Farrow, V. (2008). The assessment of present-moment awareness and acceptance the Philadelphia mindfulness scale. *Assessment*, 15(2), 204-223.
- Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6), 2156-2160.
- Cattell, R.B. (1978). *The Scientific Use of Factor Analysis*. New York, Plenum.
- Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., ... & Group, P. C. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, 63(11), 1179-1194.
- Chen, G., Gully, S. M., & Eden, D. (2001). Validation of a new general self-efficacy scale. *Organizational Research Methods*, 4(1), 62-83.
- Chen, F. F., Jing, Y., Hayes, A., & Lee, J. M. (2013). Two concepts or two approaches? A bifactor analysis of psychological and subjective well-being. *Journal of Happiness Studies*, 14(3), 1033-1068.

- Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, 41(2), 189-225.
- Cheng, Y. Y., Wang, W. C., & Ho, Y. H. (2009). Multidimensional Rasch analysis of a psychological test with multiple subtests: A statistical solution for the bandwidth–fidelity dilemma. *Educational and Psychological Measurement*, 69, 369-388.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255.
- Chou, T. J., & Ting, C. C. (2003). The role of flow experience in cyber-game addiction. *CyberPsychology & Behavior*, 6(6), 663-675.
- Christoffersson, A. (1977). Two-step weighted least squares factor analysis of dichotomized variables. *Psychometrika*, 42(3), 433-438.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319.
- Coffey, J. K., Wray-Lake, L., Mashek, D., & Branand, B. (2016). A multi-study examination of well-being theory in college and community samples. *Journal of Happiness Studies*, 17, 187-211.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78(1), 98-104.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349-354.

- Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. London: Harper.
- Csikszentmihalyi, M., & Csikszentmihalyi, I. S. (Eds.). (1988). *Optimal experience: Psychological studies of flow in consciousness*. Cambridge: Cambridge university press.
- Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease*, 175, 526-536.
- Cummins, R. A. (2013). Measuring happiness and subjective well-being. In S. A. David, I. Boniwell, & A. C. Ayers (Eds.), *The Oxford handbook of happiness*, (pp. 185-200). Oxford: Oxford University Press.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1(1), 16-29.
- David, S., Boniwell, I., & Ayers, A. C. (2013). Introduction. In S. A. David, I. Boniwell, & A. C. Ayers (Eds.), *The Oxford handbook of happiness*, (pp. 1-8). Oxford: Oxford University Press.
- Davies, D.S.C. (Ed.). (2013). Annual report of the chief medical officer 2013 - Public mental health priorities: Investing in the evidence. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/413196/CMO_web_doc.pdf
- de Bruin, G. P., & du Plessis, G. A. (2015). Bifactor analysis of the mental health continuum—short form (MHC–SF). *Psychological Reports*, 116(2), 438-446.
- de Jager, M., Coetzee, S., & Visser, D. (2008). Dimensions of Social Well-Being in a Motor Manufacturing Organisation in South Africa. *Journal of Psychology in Africa*, 18(1), 57-64.
- De Vaus, D. (2002). *Surveys in social research*. Crows Nest: Allen & Unwin.

- De Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, 39(4), 695-710.
- Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, 19(2), 109-134.
- Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11(4), 227-268.
- Delle Fave, A., Brdar, I., Freire, T., Vella-Brodrick, D., & Wissing, M. P. (2011). The eudaimonic and hedonic components of happiness: Qualitative and quantitative findings. *Social Indicators Research*, 100(2), 185-207.
- DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, 36(2), 171-181.
- DeVellis, R. F. (2012). *Scale development: Theory and applications*. London: Sage Publications.
- Diener, E. (1984). Subjective well-being. *Psychological Bulletin*, 95, 542-575.
- Diener, E. (2000). Subjective well-being: The science of happiness and a proposal for a national index. *American Psychologist*, 55(1), 34-43.
- Diener, E. (2006). Guidelines for national indicators of subjective well-being and ill-being. *Applied Research in Quality of Life*, 1(2), 151-157.
- Diener, E., Colvin, C. R., Pavot, W. G., & Allman, A. (1991). The psychic costs of intense positive affect. *Journal of Personality and Social Psychology*, 61(3), 492-503.
- Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, 49(1), 71-75.

- Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D. W., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, 97(2), 143-156.
- Diogenes Laërtius. (1972). *Lives of eminent philosophers* (R.D. Hicks, Trans.). London: Harvard University Press.
- Disabato, D.J., Goodman F.R., Kashdan T.B., Short J.L., & Jarden A. (2016). Different types of well-being? A cross-cultural examination of hedonic and eudaimonic well-being. *Psychological Assessment*, 28, 471-482.
- Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment*, 18(2), 192-203.
- Ekkekakis, P. (2013). *The measurement of affect, mood, and emotion: A guide for health-behavioral research*. Cambridge: Cambridge University Press.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8(3), 430-457.
- Engeser, S., & Rheinberg, F. (2008). Flow, performance and moderators of challenge-skill balance. *Motivation and Emotion*, 32(3), 158-172.
- Englund, J. A., Decker, S. L., Woodlief, D. T., & DiStefano, C. (2014). Development and evaluation of an online, multicomponent working memory battery. *Assessment*, 21, 543-561.
- Epicurus. (1994). Letter to Menoeceus (B. Inwood & L.P. Gerson, Trans.). In B. Inwood, & L.P. Gerson (Eds.), *The Epicurus reader: Selected writings and testimonia*, (pp. 28-31). Indianapolis: Hackett.
- Erikson, E. (1959). Identity and the life cycle. *Psychological Issues*, 1, 18-164.
- Evers, A., Hagemeister, C., Høstmælingen, A., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). *EFPA Review Model for the description and evaluation of psychological and educational tests*. Version 4.2.6. Brussels: European Federation of Psychology Associations.

- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory factor analysis*. Oxford: Oxford University Press.
- Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3), 272-299.
- Fastame, M. C., Penna, M. P., & Hitchcott, P. K. (2015). Life satisfaction and social desirability across the late life span: What relationship? *Quality of Life Research*, 24(1), 241-244.
- Feldman, G., Hayes, A., Kumar, S., Greeson, J., & Laurenceau, J. P. (2007). Mindfulness and emotion regulation: The development and initial validation of the Cognitive and Affective Mindfulness Scale-Revised (CAMS-R). *Journal of Psychopathology and Behavioral Assessment*, 29(3), 177-190.
- Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology*, 74(4), 967-984.
- File, T., & Ryan, C. (2014). *Computer and Internet Use in the United States: 2013*. Retrieved May 6, 2016 from [census.gov/content/dam/Census/library/publications/2014/acs/acs-28.pdf](https://www.census.gov/content/dam/Census/library/publications/2014/acs/acs-28.pdf)
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9(4), 466-491.
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, 16(4), 625-641.
- Forgeard, M. J., Jayawickreme, E., Kern, M. L., & Seligman, M. E. (2011). Doing the right thing: Measuring wellbeing for public policy. *International Journal of Wellbeing*, 1(1), 79-106.

- Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, 18(1), 39-50.
- Fredrickson, B. L., & Losada, M. F. (2005). Positive affect and the complex dynamics of human flourishing. *American Psychologist*, 60(7), 678-686.
- Furr, M. (2011). *Scale construction and psychometrics for social and personality psychology*. London, UK: SAGE Publications Ltd.
- Gagné, M. (2003). The role of autonomy support and autonomy orientation in prosocial behavior engagement. *Motivation and Emotion*, 27(3), 199-223.
- Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, 87(3), 564-567.
- Gallagher, M. W., Lopez, S. J., & Preacher, K. J. (2009). The hierarchical structure of well-being. *Journal of Personality*, 77(4), 1025-1050.
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological methods*, 21(1), 93-111.
- Gavrilov-Jerković, V., Jovanović, V., Žuljević, D., & Brdarić, D. (2014). When less is more: A short version of the Personal Optimism Scale and the Self-Efficacy Optimism Scale. *Journal of Happiness Studies*, 15(2), 455-474.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Gibson, B., & Sanbonmatsu, D. M. (2004). Optimism, pessimism, and gambling: The downside of optimism. *Personality and Social Psychology Bulletin*, 30(2), 149-160.
- Goldberg, D. P., & Williams, P. A. (1988). *User's Guide to the GHQ*. Windsor: NFER Nelson.

- Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, 59(2), 93-104.
- Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability what they are and how to use them. *Educational and Psychological Measurement*, 66(6), 930-944.
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60, 549-576.
- Greco, L. A., Baer, R. A., & Smith, G. T. (2011). Assessing mindfulness in children and adolescents: development and validation of the Child and Adolescent Mindfulness Measure (CAMM). *Psychological Assessment*, 23(3), 606-614.
- Green, S. B. (1991). How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research*, 26(3), 499-510.
- Green, S. B., & Yang, Y. (2009). Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74(1), 121-135.
- Hair, J. F., Black, W. C., Babin, B. J, Anderson, R. E. (2010). *Multivariate data analysis: A global perspective*. New York: Pearson.
- Hardesty, D. M., & Bearden, W. O. (2004). The use of expert judges in scale development: Implications for improving face validity of measures of unobservable constructs. *Journal of Business Research*, 57(2), 98-107.
- Harrison, D. A., & McLaughlin, M. E. (1996). Structural properties and psychometric qualities of organizational self-reports: Field tests of connections predicted by cognitive theory. *Journal of Management*, 22(2), 313-338.
- Hashemi, T., Hekmati, I., Vahedi, S., & Babapour, J. (2014) Psychometric properties of short-form social well-being questionnaire. *International Journal of Behavioral Sciences*, 8(1), 11-19.
- Hau, K. T., & Marsh, H. W. (2004). The use of item parcels in structural equation modelling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology*, 57(2), 327-351.

- Health and Safety Executive. (2015). *Work related Stress, Anxiety and Depression Statistics in Great Britain 2015*. Retrieved from <http://www.hse.gov.uk/statistics/causdis/stress/stress.pdf>
- Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, 43(1), 115-135.
- Hogan, M. J., Johnston, H., Broome, B., McMoreland, C., Walsh, J., Smale, B., ... & Groarke, A. M. (2015). Consulting with citizens in the design of wellbeing measures and policies: Lessons from a systems science application. *Social Indicators Research*, 123, 857-877.
- Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, 65(2), 202-226.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70.
- Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, 2(1), 41-54.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1-55.
- Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics*, 2(3), 1-12.
- Huppert, F. A., & So, T. T. (2013). Flourishing across Europe: Application of a new conceptual framework for defining well-being. *Social Indicators Research*, 110(3), 837-861.

- Huta, V. (2013a). Eudaimonia. In S. A. David, I. Boniwell, & A. C. Ayers (Eds.), *The Oxford handbook of happiness*, (pp. 200-213). Oxford: Oxford University Press.
- Huta, V. (2013b). Pursuing eudaimonia versus hedonia: Distinctions, similarities, and relationships. In A. S. Waterman (Ed.), *The best within us: Positive psychology perspectives on eudaimonia*, (pp. 139-158). Washington, DC: American Psychological Association.
- Huta, V., & Ryan, R. M. (2010). Pursuing pleasure or virtue: The differential and overlapping well-being benefits of hedonic and eudaimonic motives. *Journal of Happiness Studies*, 11(6), 735-762.
- Ipeirotis, P. G., Provost, F., & Wang, J. (2010). Quality management on amazon mechanical turk. In R. Chandrasekar, E. Chi, M. Chickering, P. G. Ipeirotis, W. Mason, F. Provost, J. Tam et al. (Eds.), *Proceedings of the ACM SIGKDD workshop on human computation* (pp. 64-67). New York: Association for Computing Machinery.
- Jaccard, J., & Jacoby, J. (2010). *Theory construction and model-building skills: A practical guide for social scientists*. New York: The Guilford Press.
- Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling*, 8(2), 205-223.
- Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N: q hypothesis. *Structural Equation Modeling*, 10(1), 128-141.
- Jackson, S. A., & Eklund, R. C. (2004). *The flow scales manual*. Morgantown, WV: Fitness Information Technology.
- Jahoda, M. (1958). *Current concepts of positive mental health*. New York: Basic Books.
- Jamieson, S. (2004). Likert scales: How to (ab) use them. *Medical Education*, 38(12), 1217-1218.

- Jensen, A. R., & Weng, L. J. (1994). What is a good g? *Intelligence*, 18(3), 231-258.
- Jovanović, V. (2015). Structural validity of the Mental Health Continuum-Short Form: The bifactor model of emotional, social and psychological well-being. *Personality and Individual Differences*, 75, 154-159.
- Johnson, W., 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.
- Johnson, W., te Nijenhuis, J., & Bouchard, T. J. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence*, 36(1), 81-95.
- Johnston, M. M., & Finney, S. J. (2010). Measuring basic needs satisfaction: Evaluating previous research and conducting new psychometric evaluations of the Basic Needs Satisfaction in General Scale. *Contemporary Educational Psychology*, 35(4), 280-296.
- Joiner Jr, T. E., Catanzaro, S. J., & Laurent, J. (1996). Tripartite structure of positive and negative affect, depression, and anxiety in child and adolescent psychiatric inpatients. *Journal of Abnormal Psychology*, 105(3), 401-409.
- Jonason, P. K., Kaufman, S. B., Webster, G. D., & Geher, G. (2013). What lies beneath the Dark Triad Dirty Dozen: Varied relations with the Big Five. *Individual Differences Research*, 11(2), 81-90.
- Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: a concise measure of the dark triad. *Psychological Assessment*, 22(2), 420-432.
- Jung, C. G. (1933). *Modern man in search of a soul*. New York: Harcourt.
- Kaczmarek, L. D., Bujacz, A., & Eid, M. (2015). Comparative Latent State–Trait Analysis of Satisfaction with Life Measures: The Steen Happiness Index and the Satisfaction with Life Scale. *Journal of Happiness Studies*, 16(2), 443-453.
- Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. London, UK: Sage Publications.

- Kaplan, R., & Saccuzzo, D. (2012). *Psychological testing: Principles, applications, and issues*. Belmont, CA: Wadsworth.
- Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience: The day reconstruction method. *Science*, 306(5702), 1776-1780.
- Kashdan, T. B., Biswas-Diener, R., & King, L. A. (2008). Reconsidering happiness: The costs of distinguishing between hedonics and eudaimonia. *The Journal of Positive Psychology*, 3(4), 219-233.
- Kee, Y. H., & Wang, C. J. (2008). Relationships between mindfulness, flow dispositions and mental skills adoption: A cluster analytic approach. *Psychology of Sport and Exercise*, 9(4), 393-411.
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44, 486-507.
- Kern, M. L., Waters, L. E., Adler, A., & White, M. A. (2015). A multidimensional approach to measuring well-being in students: Application of the PERMA framework. *The Journal of Positive Psychology*, 10, 262-271.
- Kesebir, P., & Diener, E. (2008). In pursuit of happiness: Empirical answers to philosophical questions. *Perspectives on Psychological Science*, 3(2), 117-125.
- Keyes, C. L. M. (1998). Social well-being. *Social Psychology Quarterly*, 61(2), 121-140.
- Keyes, C. L. M. (2002). The mental health continuum: From languishing to flourishing in life. *Journal of Health and Social Behavior*, 43(2), 207-222.
- Keyes, C. L. M. (2005). Mental illness and/or mental health? Investigating axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology*, 73(3), 539-548.
- Keyes, C. L. M. (2007). Promoting and protecting mental health as flourishing: A complementary strategy for improving national mental health. *American Psychologist*, 62(2), 95-108.

- Keyes, C. L. M. (2013). Promoting and protecting positive mental health: Early and often throughout the lifespan. In C. L. M. Keyes (Ed.), *Mental well-being: International contributions to the study of positive mental health*, (pp. 3-28). New York: Springer.
- Keyes, C. L. M. (2015). Human Flourishing and Salutogenetics. In *Genetics of Psychological Well-Being: The role of heritability and genetics in positive psychology* (pp. 3-19). Oxford: Oxford University Press.
- Keyes, C. L. M., Dhingra, S. S., & Simoes, E. J. (2010). Change in level of positive mental health as a predictor of future risk of mental illness. *American Journal of Public Health*, 100(12), 2366-2371.
- Keyes, C. L. M., Shmotkin, D., & Ryff, C. D. (2002). Optimizing well-being: the empirical encounter of two traditions. *Journal of Personality and Social Psychology*, 82(6), 1007-1022.
- Khaw, D., & Kern, M. (2014). A Cross-Cultural Comparison of the PERMA Model of Well-being. *Undergraduate Journal of Psychology at Berkeley*, 8, 10-23.
- Kim, J. O., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*. London: Sage.
- King, L. A., Hicks, J. A., Krull, J. L., & Del Gaiso, A. K. (2006). Positive affect and the experience of meaning in life. *Journal of Personality and Social Psychology*, 90(1), 179-196.
- Kline, P. (2000). *The handbook of psychological testing*. New York: Routledge.
- Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York: Guilford Press.
- Kraut, R. (1979). Two conceptions of happiness. *Philosophical Review*, 87, 167–196.
- Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of Board of Scientific Affairs' Advisory Group on the Conduct of Research on the Internet. *American Psychologist*, 59(2), 105-117.

- Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods, 13*(3), 435-455.
- Lang, K. M., & Little, T. D. (in press). Principled Missing Data Treatments. *Prevention Science*. doi: 10.1007/s11121-016-0644-5
- Larsen, R. J., & Diener, E. (1992). Promises and problems with the circumplex model of emotion. In M. S. Clark (Ed.), *Review of personality and social psychology: Emotion* (pp. 25-59). Newbury Park: Sage.
- Lavallee, L. F., Hatch, P. M., Michalos, A. C., & McKinley, T. (2007). Development of the contentment with life assessment scale (CLAS): Using daily life experiences to verify levels of self-reported life satisfaction. *Social Indicators Research, 83*(2), 201-244.
- Layard, R. (2010). Measuring subjective well-being. *Science, 327*(5965), 534-535.
- Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Quality and Quantity, 43*(3), 495-507.
- Lei, P-W., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 164–179). New York, NY: Guilford Press.
- Li, C. H. (2014). *The performance of MLR, USLMV, and WLSMV estimation in structural regression models with ordinal variables* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses.
- Liao, P. S. (2014). More happy or less unhappy? Comparison of the balanced and unbalanced designs for the response scale of general happiness. *Journal of Happiness Studies, 15*(6), 1407-1423.
- Linley, P. A., Maltby, J., Wood, A. M., Osborne, G., & Hurling, R. (2009). Measuring happiness: The higher order factor structure of subjective and psychological well-being measures. *Personality and Individual Differences, 47*(8), 878-884.

- Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, 47(2), 519-528.
- Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198-1202.
- Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, 13(1), 59-72.
- Long, J. S. (1983). *Confirmatory factor analysis: A preface to LISREL*. London, UK: Sage Publications.
- Longo, Y. (2015). The Simple Structure of Positive Affect. *Social Indicators Research*, 124(1), 183-198.
- Longo, Y., Gunz, A., Curtis, G. J., & Farsides, T. (2016). Measuring need satisfaction and frustration in educational and work contexts: the need satisfaction and frustration scale (NSFS). *Journal of Happiness Studies*, 17(1), 295-317.
- Longo, Y., Coyne, I., Joseph, S., & Gustavsson, P. (2016). Support for a general factor of well-being. *Personality and Individual Differences*, 100, 68-72.
- Luszczynska, A., Gutiérrez-Doña, B., & Schwarzer, R. (2005). General self-efficacy in various domains of human functioning: Evidence from five countries. *International Journal of Psychology*, 40(2), 80-89.
- Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: Does happiness lead to success? *Psychological Bulletin*, 131(6), 803-855.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130-149.

- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4(1), 84-99.
- Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, 48, 61-83.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519-530.
- Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior Genetics*, 34(6), 593-610.
- Marsh, H. W. (2007). Application of confirmatory factor analysis and structural equation modeling in sport and exercise psychology. In G. Tenenbaum & R. C. Eklund. (Eds), *Handbook of sport psychology* (3rd Ed.). (pp. 774-798). Hoboken, NJ, US: John Wiley & Sons Inc.
- Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33(2), 181-220.
- Martin, A. J., & Jackson, S. A. (2008). Brief approaches to assessing task absorption and enhanced subjective experience: Examining ‘short’ and ‘core’ flow in diverse performance domains. *Motivation and Emotion*, 32, 141–157.
- Maslow, A. H. (1968). *Toward a psychology of being* (2nd ed.). New York: Van Nostrand.
- McArdle, J. J., & Nesselroade, J. R. (2014). *Longitudinal data analysis using structural equation models*. Washington, DC: American Psychological Association.
- McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, 51(6), 882-888.
- McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain*. New York, NY: Springer.

- McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.
- McDougle, L., Handy, F., Konrath, S., & Walk, M. (2014). Health outcomes and volunteering: The moderating role of religiosity. *Social Indicators Research*, 117(2), 337-351.
- McDowell, I. (2010). Measures of self-perceived well-being. *Journal of Psychosomatic Research*, 69(1), 69-79.
- McNulty, J. K., & Karney, B. R. (2004). Positive expectations in the early years of marriage: Should couples expect the best or brace for the worst? *Journal of Personality and Social Psychology*, 86(5), 729-743.
- Meade, T., & Dowswell, E. (2015). Health-related quality of life in a sample of australian adolescents: Gender and age comparison. *Quality of Life Research*, 24(12), 2933-2938.
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93(3), 568-592.
- Meisenberg, G., & Woodley, M. A. (2015). Gender differences in subjective well-being and their relationships with gender equality. *Journal of Happiness Studies*, 16(6), 1539-1555.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Miller, G. A. (1956). The magic number seven plus or minus two: Some limits on our automatization of cognitive skills. *Psychological Review*, 63, 81-97.
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.
- Morgan, J., & Farsides, T. (2009). Measuring meaning in life. *Journal of Happiness Studies*, 10(2), 197-214.

- Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, 5(2), 159-168.
- Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, 41(5), 407-422.
- Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 205-243). Newbury Park, CA: Sage.
- Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. *Self and Identity*, 2(3), 223-250.
- Neugarten, B. L. (1973). Personality change in late life: A developmental perspective. In C. Eisdorfer & M. P. Lawton (Eds.), *The psychology of adult development and aging*. Washington, D.C.: American Psychological Association.
- Newburger, E. C. (2001, September, U.S. Census Bureau). *Home computers and Internet use in the United States: August 2000*. Retrieved May 6, 2016, from <http://www.census.gov/prod/2001pubs/p23-207.pdf>
- Newman, D. A. (2014). Missing data five practical guidelines. *Organizational Research Methods*, 17(4), 372-411.
- Niemiec, C. P., & Ryan, R. M. (2013). What makes for a life well lived? Autonomy and its relation to full functioning and organismic wellness. In S. A. David, I. Boniwell, & A. C. Ayers (Eds.), *The Oxford handbook of happiness*, (pp. 214-226). Oxford: Oxford University Press.
- Norton, D.L. (1976). *Personal destinies*. Princeton, NJ: Princeton University Press.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Osborne, J. W. (2014). *Best practices in exploratory factor analysis*. CreateSpace.

- Oxford Dictionaries. (2014). English Synonyms. Retrieved from <http://www.oxforddictionaries.com/>
- Paolacci, G., & Chandler, J. (2014). Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3), 184-188.
- Patterson, T. G., & Joseph, S. (2006). Development of a self- report measure of unconditional positive self-regard. *Psychology and Psychotherapy: Theory, Research and Practice*, 79(4), 557-570.
- Pavot, W., & Diener, E. (2008). The satisfaction with life scale and the emerging construct of life satisfaction. *The Journal of Positive Psychology*, 3(2), 137-152.
- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4), 1023-1031.
- Peterson, C., Park, N., & Seligman, M. E. (2005). Orientations to happiness and life satisfaction: The full life versus the empty life. *Journal of Happiness Studies*, 6(1), 25-41.
- Peterson, C., & Seligman, M. E. (2004). *Character strengths and virtues: A handbook and classification*. New York: Oxford University Press.
- Proctor, C., Tweed, R., & Morris, D. (2015) The naturally emerging structure of well-being among young adults: “Big Two” or other framework? *Journal of Happiness Studies*, 16, 257-275.
- R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL <http://www.R-project.org/>.
- Rattray, J., & Jones, M. C. (2007). Essential elements of questionnaire design and development. *Journal of Clinical Nursing*, 16(2), 234-243.
- Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, 21(2), 173-184.

- Reise, S. P., Waller, N. G., & Comrey, A. L. (2000). Factor analysis and scale revision. *Psychological Assessment*, 12(3), 287-297.
- Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment* (pp. 13–40). New York, NY: Routledge.
- Revelle, W. (2015). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, <http://CRAN.R-project.org/package=psych>
- Revelle, W. and Condon, D. M. (in press). Reliability. In Irwing, P., Booth, T., and Hughes, D., (Eds.), *Wiley-Blackwell Handbook of Psychometric Testing*. Wiley-Blackwell.
- Revelle, W., & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of Research in Personality*, 47, 493-504.
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74(1), 145-154.
- Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe - Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 38(1), 119-125.
- Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354-373.
- Rigdon, E. E., & Ferguson Jr, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research*, 28, 491-497.
- Robitschek, C., Ashton, M. W., Spering, C. C., Geiger, N., Byers, D., Schotts, G. C., & Thoen, M. A. (2012). Development and psychometric evaluation of the

- Personal Growth Initiative Scale–II. *Journal of Counseling Psychology*, 59(2), 274-287.
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(2), 137-150.
- Rogers, C. R. (1961). *On becoming a person: A therapist's view of psychotherapy*. London: Constable & Robinson Ltd.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36.
- Rossiter, J. R. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research in Marketing*, 19(4), 305-335.
- Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161-1178.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145-172.
- Russell, J. A., & Feldman Barrett, L. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5), 805-819.
- Rust, J., Golombok, S. (2009). *Modern psychometrics: The science of psychological assessment*. London, UK: Routledge.
- Ryan, R. M., & Deci, E. L. (2001). On happiness and human potentials: A review of research on hedonic and eudaimonic well-being. *Annual Review of Psychology*, 52(1), 141-166.
- Ryan, R. M., & Deci, E. L. (2002). An overview of self-determination theory: An organismic-dialectical perspective. In E. L. Deci & R. M. Ryan

- (Eds.), *Handbook of self-determination research*, (pp. 3-33). Rochester, NY: The University of Rochester Press.
- Ryan, R. M., & Frederick, C. (1997). On energy, personality, and health: Subjective vitality as a dynamic reflection of well-being. *Journal of Personality*, 65(3), 529-565.
- Ryan, R. M., Huta, V., & Deci, E. L. (2008). Living well: A self-determination theory perspective on eudaimonia. *Journal of Happiness Studies*, 9(1), 139-170.
- Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of Personality and Social Psychology*, 57(6), 1069-1081.
- Ryff, C. D., & Keyes, C. L. M. (1995). The structure of psychological well-being revisited. *Journal of Personality and Social Psychology*, 69(4), 719-727.
- Ryff, C. D., & Singer, B. H. (2008). Know thyself and become what you are: A eudaimonic approach to psychological well-being. *Journal of Happiness Studies*, 9(1), 13-39.
- Ryff, C. D., Singer, B. H., & Love, G. D. (2004). Positive health: Connecting well-being with biology. *Philosophical Transactions of the Royal Society of London, B Biological Sciences*, 359, 1383-1394.
- Ryu, E. (2011). Effects of skewness and kurtosis on normal-theory based maximum likelihood test statistic in multilevel structural equation modeling. *Behavior Research Methods*, 43(4), 1066-1074.
- Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: John Wiley & Sons.
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, 66(4), 507-514.
- Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: Extensions of complete data formulae and a Monte Carlo comparison. *Structural Equation Modeling*, 17(2), 241-264.
- Scheier, M. F., & Carver, C. S. (1985). Optimism, coping, and health: Assessment and implications of generalized outcome expectancies. *Health Psychology*, 4(3), 219-247.
- Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, 67(6), 1063-1078.
- Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, 22(1), 53-61.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, 8(4), 350-353.
- Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, 9(4), 367-373.
- Schriesheim, C. A., & Eisenbach, R. J. (1995). An exploratory and confirmatory factor-analytic investigation of item wording effects on the obtained factor structures of survey questionnaire measures. *Journal of Management*, 21(6), 1177-1193.
- Schriesheim, C. A., Solomon, E., & Kopelman, R. E. (1989). Grouped versus randomized format: An investigation of scale convergent and discriminant validity using LISREL confirmatory factor analysis. *Applied Psychological Measurement*, 13(1), 19-32.
- Schuerger, J. M., & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology*, 45(2), 294-302.

- Schulenberg, S. E., Schnetzer, L. W., & Buchanan, E. M. (2011). The purpose in life test-short form: Development and psychometric support. *Journal of Happiness Studies*, 12(5), 861-876.
- Schüler, J., & Nakamura, J. (2013). Does flow experience lead to risk? How and for whom. *Applied Psychology: Health and Well-Being*, 5(3), 311-331.
- Schutte, L., Wissing, M. P., & Khumalo, I. P. (2013). Further validation of the questionnaire for eudaimonic well-being (QEWB). *Psychology of Well-Being: Theory, Research and Practice*, 3(1), 1-22.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464.
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93-105.
- Schwarz, N., & Strack, F. (2000). Reports of subjective well-being: Judgmental processes and their methodological implication. In D. Kahneman, E. Diener, and N. Schwarz (Eds.), *Well-being: Foundations of hedonic psychology* (pp. 61-84). New York: Russell Sage Foundation.
- Seligman, M. E. P. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. New York: Free Press.
- Seligman, M. E. P. (2011). *Flourish: A new understanding of happiness and well-being – and how to achieve them*. London: Nicholas Brealey.
- Seligman, M. E. P., & Csikszentmihalyi, M. (2000). Positive psychology: An introduction. *American Psychologist*, 55(1), 5-14.
- semTools Contributors. (2016). semTools: Useful tools for structural equation modeling. Retrived from <http://cran.r-project.org/web/packages/semTools/index.html>
- Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2012). *Research methods in psychology*. New York: McGraw-Hill.

- Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, 1(2), 213-220.
- Sheldon, K. M. (2013). Individual daimon, universal needs, and subjective well-being: Happiness as the natural consequence of a life well lived. In A. Waterman (Ed.), *The best within us: Positive psychology perspectives on eudaimonic functioning* (pp. 119-137). Washington, D.C.: American Psychological Association.
- Sheldon, K. M., Elliot, A. J., Kim, Y., & Kasser, T. (2001). What is satisfying about satisfying events? Testing 10 candidate psychological needs. *Journal of Personality and Social Psychology*, 80(2), 325-339.
- Sheldon, K. M., & Hilpert, J. C. (2012). The balanced measure of psychological needs (BMPN) scale: An alternative domain general measure of need satisfaction. *Motivation and Emotion*, 36(4), 439-451.
- Sortheix, F. M., & Lönnqvist, J. E. (2014). Personal value priorities and life satisfaction in Europe: The moderating role of socioeconomic development. *Journal of Cross-Cultural Psychology*, 45(2), 282-299.
- Spector, P. E. (2006). Method variance in organizational research truth or urban legend? *Organizational Research Methods*, 9(2), 221-232.
- Spector, P. E. (2013). Survey design and measure development. In T. D. Little (Ed.), *The Oxford Handbook of Quantitative Methods, Volume 1: Foundations*, (pp. 170-188). Oxford: Oxford University Press.
- Springer, K. W., & Hauser, R. M. (2006). An assessment of the construct validity of Ryff's scales of psychological well-being: Method, mode, and measurement effects. *Social Science Research*, 35(4), 1080-1102.
- Steel, P., Schmidt, J., & Shultz, J. (2008). Refining the relationship between personality and subjective well-being. *Psychological Bulletin*, 134(1), 138-161.
- Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2), 245-251.

- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680.
- Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, 18(5), 429-442.
- Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford: Oxford university press.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics*. New York: Pearson.
- Tafarodi, R. W., & Swann, W. B. (2001). Two-dimensional self-esteem: Theory and measurement. *Personality and Individual Differences*, 31(5), 653-673.
- Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, 69(4), 613-625.
- Thatcher, A., Wretschko, G., & Fridjhon, P. (2008). Online flow experiences, problematic Internet use and Internet procrastination. *Computers in Human Behavior*, 24(5), 2236-2254.
- Thayer, R. E. (1986). Activation-deactivation adjective check list: Current overview and structural analysis. *Psychological Reports*, 58(2), 607-614.
- Thurstone, L. L. (1947). *Multiple Factor Analysis. A Development and Expansion of Vectors of the Mind*. Chicago, University of Chicago Press.
- Trower, P., Jones, J., Dryden, W., & Casey, A. (2011). *Cognitive behavioural counselling in action*. Los Angeles: Sage.
- Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology*, 90(2), 288-307.
- Tsouana-McKirahan, V. (2004). *The epistemology of the Cyrenaic school*. Cambridge: Cambridge University Press.

- van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: Let's learn from cows in the rain. *PloS ONE*, 8(7), e68967.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.
- Vasilopoulos, N. L., Cucina, J. M., & Hunter, A. E. (2007). Personality and training proficiency: issues of bandwidth - fidelity and curvilinearity. *Journal of Occupational and Organizational Psychology*, 80(1), 109-131.
- Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the likert scale comparing different numbers of options. *Educational and Psychological Measurement*, 72(4), 533-546.
- Waterman, A. S. (1993). Two conceptions of happiness: Contrasts of personal expressiveness (eudaimonia) and hedonic enjoyment. *Journal of Personality and Social Psychology*, 64(4), 678-691.
- Waterman, A. S. (2008). Reconsidering happiness: A eudaimonist's perspective. *The Journal of Positive Psychology*, 3(4), 234-252.
- Waterman, A. S. (2013). Introduction: Considering the nature of a life well lived – intersections of positive psychology and eudaemonist philosophy. In A. S. Waterman (Ed.), *The best within us: Positive psychology perspectives on eudaimonia*, (pp. 3-17). Washington, DC: American Psychological Association.
- Waterman, A. S., Schwartz, S. J., Zamboanga, B. L., Ravert, R. D., Williams, M. K., Agocha, V. B., Kim, S. Y., & Donnellan, M. B. (2010). The Questionnaire for Eudaimonic Well-Being: Psychometric properties, demographic comparisons, and evidence of validity. *The Journal of Positive Psychology*, 5(1), 41-61.

- Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality*, 38(4), 319-350.
- Watson, D., & Clark, L. A. (1992). Affects separable and inseparable: on the hierarchical arrangement of the negative affects. *Journal of Personality and Social Psychology*, 62(3), 489-505.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063-1070.
- Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, 18(3), 320-334.
- Weijters, B., De Beuckelaer, A., & Baumgartner, H. (2014). Discriminant validity where there should be none: Positioning same-scale items in separated blocks of a questionnaire. *Applied Psychological Measurement*, 38(6), 450-463.
- Weinstein, N., Przybylski, A. K., & Ryan, R. M. (2012). The index of autonomous functioning: Development of a scale of human autonomy. *Journal of Research in Personality*, 46(4), 397-413.
- Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, 14(4), 490-495.
- Wong, N., Rindfleisch, A., & Burroughs, J. E. (2003). Do reverse-worded items confound measures in cross-cultural consumer research? The case of the Material Values Scale. *Journal of Consumer Research*, 30(1), 72-91.
- Wood, A. M., Linley, P. A., Maltby, J., Baliousis, M., & Joseph, S. (2008). The authentic personality: A theoretical and empirical conceptualization and the development of the Authenticity Scale. *Journal of Counseling Psychology*, 55(3), 385-399.
- Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception research over the internet: A tutorial review. *PeerJ* 3:e1058 <https://doi.org/10.7717/peerj.1058>.

- Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, 28(3), 186-191.
- Yik, M., Russell, J. A., & Steiger, J. H. (2011). A 12-point circumplex structure of core affect. *Emotion*, 11(4), 705-731.
- Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, 14(3), 435-463.
- Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* (Doctoral dissertation, University of California Los Angeles).
- Yuan, K. H., & Bentler, P. M. (2000). Three likelihood - based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, 30(1), 165-200.
- Yung, A., Cardoso-Leite, P., Dale, G., Bavelier, D., & Green, C. S. (2015). Methods to test visual attention online. *Journal of Visualized Experiments*, 96, e52470.
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70, 123-133.

Appendices

Appendix 1

Conceptual development of the 14 scales of GWB

Components	Huppert & So (2013)	Keyes (2002)	Diener et al. (2010)	Seligman (2011)	Ryan et al. (2008)	Waterman et al. (2010)
Happiness Happiness includes moderate-arousal pleasant feelings, such as feeling happy, satisfied and pleased.	Happiness: feeling “happy, cheerful, contented” (Huppert and So, 2013, p. 857).	Feeling “in good spirits”, “cheerful” and “satisfied” (Keyes et al., 2002, p. 1011).		Feeling “cheerful” and “happy” (Kern et al., 2015, p. 266).		
Vitality Vitality includes high-arousal pleasant feelings, such as feeling energetic and lively.	Vitality: “feeling energetic” (Huppert & So, 2013, p. 856).	Feeling “full of life” (Keyes et al., 2002, p. 1011).		Feeling “energetic”, “active” and “lively” (Kern et al., 2015, p. 266)		
Calmness Calmness includes low-arousal pleasant feelings, like serenity and peacefulness.	Emotional stability: “feeling calm or relaxed, emotionally stable” (Huppert & So, 2013, p. 856).	Feeling “calm and peaceful” (Keyes et al., 2002, p. 1011).		Feeling “calm” (Kern et al., 2015, p. 266)		
Optimism A positive outlook on and expectations about the future.	Optimism: “feeling optimistic or hopeful about the future” (Huppert & So, 2013, p. 857).	Social actualization: being “hopeful about the condition and future of society” (Keyes, 1998, p. 123).	Being “optimistic about the future” (Diener et al., 2010, p. 155)			
Involvement A feeling of absorption in which the individual is completely focused on the task at hand.	Engagement: feeling “interested and involved in a wide range of activities” (Huppert & So, 2013, p. 857).		Being “engaged and interested in [...] daily activities” (Diener et al., 2010, p. 154)	Engagement: “being absorbed by and focused on what [one is] doing” (Forgeard et al., 2011, p. 84; also Seligman, 2011).		Involvement: intense involvement and flow in one’s activities (Waterman et al., 2010).

Components	Huppert & So (2013)	Keyes (2002)	Diener et al. (2010)	Seligman (2011)	Ryan et al. (2008)	Waterman et al. (2010)
Self-awareness Knowing oneself and noticing different aspects of one's experience: thoughts, feelings, and physical sensations.					“Awareness of what is occurring at the present moment” (Ryan et al., 2008, p. 158).	Self-discovery: knowing oneself (Waterman et al., 2010).
Self-acceptance Experiencing different aspects of oneself (e.g. one's past, personality, thoughts, and feelings) in a tolerant, receptive and non-judgmental way.		Self-acceptance: “acknowledg[ing] and accept[ing] multiple aspects of self including good and bad qualities; feel[ing] positive about past life” (Ryff, 1989, p. 1072).			“Open and receptive processing of events” (Ryan et al., 2008, p. 158).	
Self-worth Positive evaluations and feelings about oneself.	Self-esteem: “feelings of worth” (Huppert & So, 2013, p. 857).	Self-acceptance: “Possess[ing] a positive attitude toward the self” (Ryff, 1989, p. 1072).	Perceiving oneself as a “good person” (Diener et al., 2010, p. 155)			
Competence Feeling and perceiving oneself as effective and able to overcome challenges and achieve desired outcomes.	Competence: “Capable of thinking clearly, concentrating, making decisions” (Huppert & So, 2013, p. 856).	Environmental mastery: “a sense of mastery and competence in managing the environment; control[ing] complex array of external activities; mak[ing] effective use of surrounding opportunities; [being] able to choose or create contexts suitable to personal needs and values” (Ryff, 1989, p. 1072).	Being “competent and capable” in one's important activities (Diener et al., 2010, p. 154)	Accomplishment and competence: “achievement, success, or mastery” (Forgeard et al., 2011, p. 87; Seligman, 2011).	Competence: “the sense of efficacy one has with respect to both internal and external environments” (Ryan et al., 2008, p. 153).	

Components	Huppert & So (2013)	Keyes (2002)	Diener et al. (2010)	Seligman (2011)	Ryan et al. (2008)	Waterman et al. (2010)
Development Experiencing continuous growth and improvement. This is not a fixed state that can be attained at a single point in time. Instead, it represents a perceived improvement and expanding of the self that continues over time.		Personal growth: having “a feeling of continued development; see[ing] self as growing and expanding; [being] open to new experiences; [having] sense of realizing [one’s] potential; see[ing] improvement in self and behavior over time; [...] changing in ways that reflect more self-knowledge and effectiveness” (Ryff, 1989, p. 1072).				Perceived development: developing one’s talents and potentials (Waterman et al., 2010).
Purpose Having clear goals, a sense of direction and a larger aim in life.	Meaning: a “sense of purpose” (Huppert & So, 2013, p. 856).	Purpose in life: having “goals in life and a sense of directedness; [...] hold[ing] beliefs that give life purpose; hav[ing] aims and objectives for living” (Ryff, 1989, p. 1072).	Leading a purposeful life (Diener et al., 2010, p. 154)	A sense of purpose (Seligman, 2011; also Forgeard et al., 2011; Butler & Kern, in preparation).		Purpose and meaning in life: having purposes that allow one to use one’s talents (Waterman et al., 2010).
Significance The feeling that what we do is worthwhile, rewarding and valuable.	Feeling that what one does “in life is valuable and worthwhile” (Huppert & So, 2013, p. 843)	Social contribution: the belief that our activities produce something valuable, important and worthwhile for society (Keyes, 1998).	“Actively contribut[ing] to the happiness and well-being of others” (Diener et al., 2010, p. 154)	Meaning and purpose (also above): a sense of significance, and serving something larger than oneself (Seligman, 2011; also Forgeard et al., 2011; Butler & Kern, in preparation)		

Components	Huppert & So (2013)	Keyes (2002)	Diener et al. (2010)	Seligman (2011)	Ryan et al. (2008)	Waterman et al. (2010)
Self-congruence The perception that one's actions are compatible with one's interests, values, and beliefs.		Autonomy: being "self-determining and independent; able to resist social pressures to think and act in certain ways; regulate[ing] behavior from within; evaluat[ing] self by personal standards" (Ryff, 1989, p. 1072).			Autonomy or self-determination: "the experience of regulation by the self" (Ryan et al., 2008, p. 157); "to have activity be concordant with one's integrated sense of self" (Deci & Ryan, 2000, p. 231).	Personal expressiveness: feeling authentic (Waterman, 2013; Waterman et al., 2010).
Connection Feeling of belonging, mutual caring, love, and closeness.	Positive relationships: feeling cared about (Huppert & So, 2013).	Positive relationships: having "warm, satisfying, trusting relationships with others; [being] concerned about the welfare of others; capable of strong empathy, affection, and intimacy; understand[ing] give and take of human relationships." (Ryff, 1989, p. 1072). Social integration: feeling one belongs to a community and society (Keyes, 1998, p. 122).	Having "social relationships [that] are supportive and rewarding" and feeling respected (Diener et al., 2010, p. 154, 155).	Positive relationships: "the belief that one is cared for, loved, esteemed and valued" (Forgeard et al., 2011, p. 86; see also Seligman, 2011).	Relatedness: "feeling connected to and cared about by others" (Ryan et al., 2008, p. 153); "caring for and being cared for by those others, [...] having a sense of belongingness both with other individuals and with one's community" (Ryan & Deci, 2002, p. 7).	

Appendix 2

Past measures consulted during item development (study 1)

Construct	Past measures
Happiness	Positive emotion subscale of the PERMA-Profilier (Butler & Kern, in preparation); Positive experience subscale of the Scale of Positive and Negative Experience (Diener et al., 2010); Pleasant scale (Feldman Barrett & Russell, 1998); Pleasant scale (Larsen & Diener, 1992).
Vitality	Pleasant Activated scale (Feldman Barrett & Russell, 1998); Activated Pleasant Affect scale (Larsen & Diener, 1992); Subjective Vitality Scale (Ryan & Frederick, 1997); Energy subscale of the Activation-Deactivation Adjective Check List (Thayer, 1986).
Calmness	Pleasant Deactivated scale (Feldman Barrett & Russell, 1998); Unactivated Pleasant Affect scale (Larsen & Diener, 1992); Calmness subscale of the Activation-Deactivation Adjective Check List (Thayer, 1986).
Optimism	Life orientation test-revised (LOT-R, Scheier, Carver & Bridges, 1994); Social Actualization subscale of the Scales of Social Well-Being (Keyes, 1998); Personal Optimism scale (Gavrilov-Jerkovic et al., 2014).
Involvement	Engagement subscale of the PERMA-Profilier (Butler & Kern, in preparation); Flow Questionnaire (Csikszentmihalyi & Csikszentmihalyi, 1988); Flow Short Scale (Engeser & Rheinberg, 2008); Flow State Scale 2 (Jackson & Eklund, 2004); Core Flow Scale (Martin & Jackson, 2008).
Self-awareness	Nonreactivity to inner experience and Observing/noticing/attending to sensations/perceptions/thoughts and feelings subscales from the Five Facet Mindfulness Questionnaire (Baer et al., 2006); Mindful Attention Awareness Scale (Brown & Ryan, 2003); Awareness subscale of the Philadelphia Mindfulness Scale (Cardaciotto et al., 2008); Awareness subscale of the Cognitive and Affective Mindfulness Scale-Revised (Feldman et al., 2007).
Self-acceptance	Nonjudging of experience subscale from the Five Facet Mindfulness Questionnaire (Baer et al., 2006); Acceptance subscale of the Philadelphia Mindfulness Scale (Cardaciotto et al., 2008); Acceptance subscale of the Cognitive and Affective Mindfulness Scale-Revised (Feldman et al., 2007); Self-kindness subscale of the Self-Compassion Scale (Neff, 2003); Self-acceptance subscale of the Ryff Scales of Psychological Well-Being (Ryff, 1989).
Self-worth	Self-regard subscale of the Unconditional Positive Self-Regard Scale (Patterson & Joseph, 2006); Rosenberg Self-Esteem Scale (Rosenberg, 1965); Self-acceptance subscale of the Ryff Scales of Psychological Well-Being (Ryff, 1989); Self-liking subscale of the Self-Liking/Self-Competence Scale-Revised Version (Tafarodi & Swann, 2001).

Competence	Accomplishment subscale of the PERMA-Profilier (Butler & Kern, in preparation); New General Self-Efficacy Scale (Chen et al., 2001); Competence subscale of the General Need Satisfaction Scale (Gagné, 2003); Competence subscale of the Need Satisfaction and Frustration Scale (Longo, Gunz et al. 2016); Rosenberg Self-Esteem Scale (Rosenberg, 1965); Environmental mastery subscale of the Ryff Scales of Psychological Well-Being (Ryff, 1989); Competence subscale of the Balanced Measure of Psychological Needs (Sheldon & Hilpert, 2012); Self-competence subscale of the Self-Liking/Self-Competence Scale-Revised Version (Tafarodi & Swann, 2001).
Development	Personal Growth Initiative Scale-II (Robitscheck et al., 2012); Personal growth subscale of the Ryff Scales of Psychological Well-Being (Ryff, 1989).
Purpose	Meaning subscale of the PERMA-Profilier (Butler & Kern, in preparation); Purposeful life subscale of the Meaningful Life Measure (Morgan & Farsides, 2009); Purpose in life subscale of the Ryff Scales of Psychological Well-Being (Ryff, 1989); Purpose in Life Test-Short Form (Schulenberg et al., 2011).
Significance	Meaning subscale of the PERMA-Profilier (Butler & Kern, in preparation); Social contribution subscale of the Scales of Social Well-Being (Keyes, 1998); Valued life subscale of the Meaningful Life Measure (Morgan & Farsides, 2009).
Self-congruence	Autonomy subscale of the General Need Satisfaction Scale (Gagné, 2003); Autonomy subscale of the Need Satisfaction and Frustration Scale (Longo, Gunz et al. 2016); Autonomy subscale of the Ryff Scales of Psychological Well-Being (Ryff, 1989); Autonomy subscale of the Balanced Measure of Psychological Needs (Sheldon & Hilpert, 2012); Authentic living subscale of the Authenticity Scale (Wood et al., 2008).
Connection	Relationship subscale of the PERMA-Profilier (Butler & Kern, in preparation); Relatedness subscale of the General Need Satisfaction Scale (Gagné, 2003); Social integration subscale of the Scales of Social Well-Being (Keyes, 1998); Relatedness subscale of the Need Satisfaction and Frustration Scale (Longo, Gunz et al., 2016); Positive relations with others subscale of the Ryff Scales of Psychological Well-Being (Ryff, 1989); Relatedness subscale of the Balanced Measure of Psychological Needs (Sheldon & Hilpert, 2012).

Appendix 3

Item pool submitted for content validation (Study 1)

Instructions: Below you'll find several statements about your experiences. Please indicate how true each statement is regarding the EXPERIENCES IN YOUR LIFE OVERALL. There are no right or wrong answers. Please, choose the answer that best reflects your experience rather than what you think your experience should be.

Happiness

STEM: I feel...

1. Happy
2. Pleased
3. Satisfied
4. Glad
5. Cheerful
6. Content
7. Delighted
8. Good
9. Merry
10. Good humored

Vitality

STEM: I feel...

1. Energetic
2. Lively
3. Vigorous
4. Full of life
5. Full of vitality
6. Alive
7. Full of energy
8. Strong
9. Active
10. Excited
11. Alert

Calmness

STEM: I feel...

1. Calm
2. Peaceful
3. Relaxed
4. Serene
5. Tranquil
6. Even-tempered
7. At ease
8. Comfortable
9. Leisurely
10. A sense of harmony

Optimism

NO STEM

1. I'm optimistic
2. I hope for the best
3. I feel hopeful
4. I expect good things to happen
5. I'm positive about my plans
6. I look on the bright side
7. I believe things will work out in the end
8. I believe things will go well
9. I believe good things will happen to me
10. I believe the future is promising

Involvement

STEM: In my ACTIVITIES (e.g. work/study, free time), I feel...

1. Completely involved
2. Absorbed by what I'm doing
3. Fully engaged
4. Focused
5. Centered
6. Deeply immersed in my activities
7. Engrossed
8. In tune with what I do
9. Gripped by what I'm doing
10. My mind is fixed on what I'm doing

Self-awareness

NO STEM

1. I'm in touch with how I really feel inside
2. I feel like I know myself very well
3. I find it easy to describe my thoughts and feelings
4. I pay attention to what I think and what I feel
5. I am aware of my emotions as I experience them
6. I am aware of changes in my mood
7. I observe what goes on in my mind
8. I am conscious of my thoughts as I think them
9. I notice different sensations in my body
10. I notice the feelings in my body

Self-acceptance

NO STEM

1. I accept myself the way I am
2. I accept most aspects of myself
3. I appreciate most parts of my personality
4. I'm tolerant of my weaknesses and shortcomings
5. I'm tolerant of my flaws
6. I tolerate my experiences whether they're pleasant or unpleasant
7. I accept my thoughts and feelings
8. I accept the mistakes I've made in the past
9. I accept the choices I've made
10. I'm kind and affectionate towards myself
11. I treat myself in a caring way
12. When I make a mistake I'm patient with myself

Self-worth (self-liking)

NO STEM

1. I really like myself
2. I feel great about myself
3. I value myself highly
4. I respect myself a lot
5. I have high regard for myself
6. I admire myself
7. I know I'm a person of worth
8. I think I'm a very good person
9. I believe I have many good qualities
10. I have a very positive attitude toward myself

Competence

NO STEM

1. I am highly effective at what I do
2. I tend to perform very well
3. I am confident in my ability to master new things
4. I can obtain the results I want
5. I am able fulfil my goals
6. I can accomplish what I attempt to do
7. I can successfully complete difficult tasks
8. I can handle difficult situations
9. I'm very good at overcoming challenges
10. I have a great ability to deal with problems

Development**STEM:** I feel I am...

1. Growing
2. Developing
3. Learning new things
4. Expanding my horizons
5. Improving
6. Progressing
7. Advancing
8. Enhancing my skills
9. Stretching my limits
10. Increasing my capabilities

Purpose**STEM:** In my life, I have...

1. Clear aims and goals
2. A mission
3. A clear idea of what I want
4. A purpose
5. A vision for my life
6. A plan of what I want
7. Something to strive for
8. Something I aspire to
9. Something I work towards
10. Something I seek

Significance**STEM:** *What I do* in my life is...

1. Worthwhile
2. Valuable
3. Significant
4. Important
5. Meaningful
6. Rewarding
7. Remarkable
8. Relevant
9. Helpful
10. Precious

Self-congruence

NO STEM

1. What I do is in line with my interests
2. I base my decisions on my true values and interests
3. I decide what to do based on what I really care about
4. What I do is consistent with what I believe I should do
5. I only do something if it fits my values and beliefs
6. I act in accordance with who I am and what I believe
7. My decisions are compatible with who I really am as a person
8. I stand firmly by my beliefs
9. I'm true to myself when I make a decision
10. I feel free to be who I am
11. My actions are congruent with how I see myself
12. I only do something if it feels right

Connection

NO STEM

1. I feel a strong friendship with the people I regularly interact with
2. I feel close and connected with the people around me
3. I feel a deep connection with some people
4. I feel I belong to a social group
5. I feel accepted
6. I feel loved
7. I feel appreciated
8. I really like the people in my life
9. There are people who truly care about me
10. I feel really cared for
11. I really care for the people in my life
12. I receive support and help when I need it

Appendix 4

Content validation (Study 1)³

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Happiness			
1. Happy	3		
2. Pleased	2.5		
3. Satisfied	2.67		
4. Glad	2.67		
5. Cheerful	2.67		
6. Content	2.67		
7. Delighted	2.33		
8. Good	2.17 ⁴	could possibly have another meaning (morally good)	Agreed.
9. Merry	2.83	old-fashioned language?	Yes, this may be true. However, since the rating is high I'll keep it for now and keep this comment in mind during factor analysis.
10. Good humored	2.17	old-fashioned language? General comment: I think it's more about whether one can experience the specific feelings but still not be happy. So I have indicated the ones that for me are closer to the definition of happiness. General comment: yes for hedonic happiness (SWB & positive affect) do you also want to measure absense of negative affect? doesn't cover eudaimonic happiness	Agreed. The assumption of the scale is that if one generally feels these feelings in their life, one is happy. Negative affect overlaps with Ill-being. Eudaimonic aspects come later in the questionnaire.

³ To keep this section concise, only criticisms have been included. Positive comments, such as “the concept is well represented”, were omitted.

⁴ Content validation scores below the cut-off point of 2.30 are highlighted in brown.

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Vitality			
1. Energetic	3		
2. Lively	3		
3. Vigorous	3		
4. Full of life	3		
5. Full of vitality	2.83		
6. Alive	1.83	Alive could be taken too literally!; Alive is too broad.	Agreed.
7. Full of energy	2.83	Instead of full of energy, I would say energetic.	"Energetic" has already included. This item will only be included in the final scale if it outperforms "energetic."
8. Strong	1.5	Strong could be taken too literally!; Strong, Active and Alert are captured by the PANAS, so perhaps they are redundant.	Agreed.
9. Active	2.17	Strong, Active and Alert are captured by the PANAS, so perhaps they are redundant	One reviewer gave this item the lowest rating because he found an overlap with an item in the PANAS. However, such an overlap does not prevent this item from performing well in this scale
10. Excited	1.67	Excited has ambiguous meaning (both positive and negative).	Agreed.
11. Alert	2	Strong, Active and Alert are captured by the PANAS, so perhaps they are redundant.	One reviewer gave this item the lowest rating because he found an overlap with an item in the PANAS. However, such an overlap does not prevent this item from performing well in this scale.
		General comment: do you think you capture all aspects of work engagement (Bakker) or engagement according to Shorim.	The reviewer refers to Bakker's vigor (i.e. a component of Engagement). While vigor is likely to be highly related to vitality, it is not precisely the construct this scale attempts to measure.

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Calmness			
1. Calm	3		
2. Peaceful	3		
3. Relaxed	3		
4. Serene	3		
5. Tranquil	2.67		
6. Even-tempered	2		
7. At ease	2.33		
8. Comfortable	2.67		
9. Leisurely	2.5		
10. A sense of harmony	2.17	A sense of harmony, in my opinion, is more closely related to a broader construct of spirituality General comment: several items here do not make sense to me as a non-native speaker General comment: I've noticed, this being the 3rd construct, that there are no reverse-coded items (yet), e.g., stressed or anxious as the opposite of calmness. I'm not sure if such items will come up, just something to think about in the scale construction.	I will keep that in mind, although the target sample will include mainly native speakers. Furthermore, hopefully the two omitted items are the ones that are the least clear. This choice is justified in the paper/dissertation

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Optimism			
1. I'm optimistic	3		
2. I hope for the best	2.5	2. perhaps use "expect" instead of "hope"; Hope is distinct from optimism. Hope refers to positive expectations in the context of low personal agency.	The term "expect" is already used in item 4; Within academic research that may be so. However, in everyday speech the two are synonyms.
3. I feel hopeful	2		
4. I expect good things to happen	3	I'm positive about my plans is more specific than other items, because it refers to plans.	
5. I'm positive about my plans	2.5		
6. I look on the bright side	3		
7. I believe things will work out in the end	3		
8. I believe things will go well	3		
9. I believe good things will happen to me	2.83		
10. I believe the future is promising	3		
		General comment: In general, I would not use the same term to define a concept. So I would not say I'm optimistic to define optimism. This applied to all questions. The items above vary in terms of whether one is optimistic about themselves or in general. So you need to reconsider how items are phrased.	Most items focus on general optimism. Only 2 items are more specific and focus on "my plans" and "me." Nevertheless, even the other items assume that one is optimistic about one's own future

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Involvement			
1. Completely involved	2.67		
2. Absorbed by what I'm doing	3		
3. Fully engaged	2.83		
4. Focused	2.33		
5. Centered	2.17	5 possibly confusing needs elaboration	Agreed. This will be omitted, also because its rating is low.
6. Deeply immersed in my activities	2.83	6. the stem contains the word "activities" and so does the item, and this doesn't work well grammatically: "In my activities, I feel deeply immersed in my activities". you could stop at immersed.	Agreed: the phrase "in my activities" will be omitted from this item.
7. Engrossed	2.83		
8. In tune with what I do	2.5		
9. Gripped by what I'm doing	2.67		
10. My mind is fixed on what I'm doing	2.67	General comment: this aspect is probably multi-dimensional. hard to see that these capture the same dimension. Our own work consistently distinguish between at least 3 of these, that is 2, 3 and 4	I will keep this in mind during factor analysis.

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Self-awareness			
1. I'm in touch with how I really feel inside	2.67		
2. I feel like I know myself very well	3		
3. I find it easy to describe my thoughts and feelings	1.83	Item 3 is different because it is more closely related to the construct of alexithymia or maybe emotional intelligence	It has also been conceptualized as part of mindfulness. However, since this item has such a low rating it will be omitted.
4. I pay attention to what I think and what I feel	2.67		One reviewer later points out that it's generally preferable to omit "and." Indeed, that is a good point in most cases. Therefore, this item will be revised and the last phrase ("and what I feel") will be omitted.
5. I am aware of my emotions as I experience them	3		
6. I am aware of changes in my mood	2.67		
7. I observe what goes on in my mind	2.67		
8. I am conscious of my thoughts as I think them	2.83		
9. I notice different sensations in my body	2.83		
10. I notice the feelings in my body	2.5		

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Self-acceptance			
1. I accept myself the way I am	3		
2. I accept most aspects of myself	3		
3. I appreciate most parts of my personality	2.67		
4. I'm tolerant of my weaknesses and shortcomings	2.67	do not use "and"	I can omit the term "shortcomings" and include it in item 9
5. I'm tolerant of my flaws	2.67		
6. I tolerate my experiences whether they're pleasant or unpleasant	2.33	6 seems more about accepting one's experiences (coping?) than one's self.; maybe 6 is different - about external events?	that is one aspect of self-acceptance in the mindfulness literature and is expected to load on the same factor
7. I accept my thoughts and feelings	2.33	do not use "and"	while that's a good point, separating this item may result cumbersome – (see Saris & Gallhofer, 2014). If it performs poorly, it will be replaced
8. I accept the mistakes I've made in the past	2.33	Items 8 refers to past experiences and is more closely related to self-forgiveness.	self-forgiveness and self-kindness are part of this conceptualization
9. I accept the choices I've made	2.33	Item 9 is, in my opinion, more specific than other items (refers to personal choices).	This item could be revised into, "I accept my shortcomings" - see above notes to item 4
10. I'm kind and affectionate towards myself	2.17	do not use "and"	
11. I treat myself in a caring way	2.5		
12. When I make a mistake I'm patient with myself	2.5		
		General comment: It would be good to have the term "non-judgemental" included in an item. General comment: I don't think this concept is captured well. One can accept themselves but actively trying to change certain things about themselves. So they may not be tolerant of certain aspects or accept them.	This term is very infrequent in everyday speech see https://books.google.com/ngrams/ This comment assumes that - and is only valid if - tolerance and acceptance are not the same. Previous factor analyses show that they load strongly on the same factor, so for now I will keep these terms

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Self-worth			
1. I really like myself	2.67		
2. I feel great about myself	2.67		
3. I value myself highly	2.83		
4. I respect myself a lot	2.83		
5. I have high regard for myself	2.83		
6. I admire myself	2.33		
7. I know I'm a person of worth	3		
8. I think I'm a very good person	2.33	8 perhaps too relative	Agreed. If this item performs poorly in factor analyses it will be omitted.
9. I believe I have many good qualities	2.67		
10. I have a very positive attitude toward myself	3		
general note: this dimension is a little problematic.			The two most problematic items have been omitted

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Competence			
1. I am highly effective at what I do	2.67		
2. I tend to perform very well	2.67		
3. I am confident in my ability to master new things	2.67		
4. I can obtain the results I want	2.67		
5. I am able fulfil my goals	2.5		
6. I can accomplish what I attempt to do	2.67		
7. I can successfully complete difficult tasks	2.67		
8. I can handle difficult situations	2.67		
9. I'm very good at overcoming challenges	2.67		
10. I have a great ability to deal with problems	2.67		
		General note: I think al the above items are too absolute. And also I don't think that the definition is reflected well in the items that refer to new problems for example. General note: From my perspeticve these items capture self-efficacy	The items aim to reflect Deci & Ryan's definition of competence - specifically, a felt sense of confidence in one's abilities. Indeed, that is by design. Self-efficacy and competence are very closely linked, and the latter focuses more on one's perceptions rather than actual skills acquired.
		General note: I have a problem with that these items i projected onto the future	Perhaps this was only this reviewer's interpretation, since all the verbs are in the present tense and may be answered while thinking about past actions.
		General note: I think that several constructs are captured by these items: competence, self-efficacy and resilience. Maybe it would be difficult to make a distinction between them.	I believe so, these facets tap into Deci and Ryan's definition of competence and have been shown to load strongly on one factor (e.g. see Longo, Gunz et al., 2016).

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Development			
1. Growing	2.67	1 may be taken too literally	Agreed. This will be removed, as it is vague and may be interpreted as physical growth.
2. Developing	2.67		
3. Learning new things	3	perhaps include an item that better represents learning from experience rather than learning new things (#3)	This item had high ratings, so I would not change it. Perhaps, the item "I learn from experience" could also be included, but that would include a redundant root: "learn."
4. Expanding my horizons	3		
5. Improving	3		
6. Progressing	3		
7. Advancing	3		
8. Enhancing my skills	3		
9. Stretching my limits	2.67		
10. Increasing my capabilities	3		

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Purpose			
1. Clear aims and goals	2.83	do not use "and"	In this case, aims and goals are so closely linked that separating them would not make much sense (Saris & Gallhofer, 2014).
2. A mission	2.67		
3. A clear idea of what I want	3		
4. A purpose	2.67		
5. A vision for my life	3		
6. A plan of what I want	2.67		
7. Something to strive for	3		
8. Something I aspire to	3		
9. Something I work towards	3		
10. Something I seek	2.5	10 perhaps ambiguous - search vs. presence of meaning in life; 10. this could be answered affirmatively by people who've lost a possession but don't necessarily have purpose. Alternative: "In my life, I have... a clear sense of purpose"	Agreed, this item will be omitted. The term "purpose" is already included

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Significance			
1. Worthwhile	3		
2. Valuable	3		
3. Significant	2.67		
4. Important	3		
5. Meaningful	3		
6. Rewarding	2.5		
7. Remarkable	2.33	7 could have more than one meaning	Agreed. This item will be omitted.
8. Relevant	2.67		
9. Helpful	2.33		
10. Precious	2.17		

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Self-congruence			
1. What I do is in line with my interests	2.67		
2. I base my decisions on my true values and interests	2.83	and;	To avoid using the word “and”, which may create a double-barreled item, the words “and interests” will be omitted.
3. I decide what to do based on what I really care about	3		
4. What I do is consistent with what I believe I should do	3		
5. I only do something if it fits my values and beliefs	2.67	and;	The words “values and” will be omitted.
6. I act in accordance with who I am and what I believe	2.83		
7. My decisions are compatible with who I really am as a person	3		
8. I stand firmly by my beliefs	2.83		
9. I’m true to myself when I make a decision	2.83		
10. I feel free to be who I am	2.5		
11. My actions are congruent with how I see myself	3		
12. I only do something if it feels right	2.17	Avoid being absolute in the above items. Also self-congruence does not necessarily mean the something is right	Agreed. This item will be omitted.

Scale - Items	Mean rating	Reviewer feedback	Notes from the scale developer
Connection			
1. I feel a strong friendship with the people I regularly interact with	2.67		
2. I feel close and connected with the people around me	2.83		
3. I feel a deep connection with some people	2.83		
4. I feel I belong to a social group	2.67		
5. I feel accepted	2.67		
6. I feel loved	2.33		
7. I feel appreciated	2.67		
8. I really like the people in my life	2.33		
9. There are people who truly care about me	2.33		
10. I feel really cared for	2.33		
11. I really care for the people in my life	2.5		
12. I receive support and help when I need it	2.33		
		General comments: I think some of the items are overstretching the definition or are absolute	Agreed. Some items (e.g. 12) may tap into constructs like social support. I will keep this in mind during factor analyses.

Appendix 5

Item selection in individual CFAs (Study 2)

Scale and Item	Order ⁵	Correlated error (CE) ⁶	Notes	Con ⁷
Happiness				
1 Happy				1
2 Pleased				
3 Satisfied	E3	CE with item 2	in the past this item had some overlap with low-arousal PA	
4 Glad				
5 Cheerful				1
6 Content	E1	CE with item 3	strong conceptual overlap with low-arousal PA	
7 Delighted	E4	CE with 5	it has lower content validity than 5 and overlaps more with high-arousal PA	2
8 Merry	E2	CE with 7	most confusing item; infrequent use of the term	7

⁵ Items were excluded from the scale iteratively. This column describes the order of exclusion from the scale.

⁶ The highest modification indices showed that this item had a correlated error with another item on the same scale.

⁷ How many participants found this item confusing / ambiguous.

Vitality				
1 Energetic		CE with 6		2
2 Lively		CE with 4		1
3 Vigorous				2
4 Full of life	E3	CE with 2	more confusing than "lively"	4
5 Full of vitality	E2	CE with 1, 3 and 7	Redundant wording; although item 7 had the highest MI, it was less redundant than this item. Also, this item had many CE with other items.	3
6 Full of energy	E1	CE with 1	Redundant wording and More complex than "energetic"	2
7 Active		CE with 1	This item is very simple and unambiguous, so we initially refrained from omitting it. Furthermore, the CE later disappears	
8 Alert	E4	CE with 1 and 7	Alertness may not be pleasant; had the lowest loading on the factor.	

Calmness					
1 Calm					1
2 Peaceful					
3 Relaxed					
4 Serene	E2	CE with 2 and 3	serene is more infrequent than relaxed or peaceful		4
5 Tranquil	E1	CE with 4	The term is used more rarely than item 4		4
6 At ease					1
7 Comfortable	E3	CE with 2	may not refer to affect		
8 Leisurely	E4	CE with 1 and 3	lowest loading; more confusing and infrequent than 1 or 3		3
Optimism					
1 I'm optimistic					
2 I hope for the best					1
3 I expect good things to happen					
4 I'm positive about my plans					
5 I look on the bright side	E1	CE with 1	Longer item than "I'm optimistic"		
6 I believe things will work out in the end					
7 I believe things will go well	E3	CE with 3	Redundant wording with 3, 6 and 9		
8 I believe good things will happen to me	E2	CE with 3	Longer item than 3; more redundant		
9 I believe the future is promising	E4		Redundant wording with 6; also it's vague		

Involvement

1 Completely involved					
2 Absorbed by what I'm doing					
3 Fully engaged					
4 Focused	E4	CE with 2	According to one of the content reviewers this may be different from absorption.		
5 Deeply immersed	E1	CE with 6	less frequent than 6		
6 Engrossed	E2	CE with 8	both 6 and 8 seem confusing, and eventually they are both omitted	2	
7 In tune with what I do					
8 Gripped by what I'm doing	E3	CE with 4	Confusing and longer than 4		1
9 My mind is fixed on what I'm doing	E5	CE with 1	Longest item; more complex than 1		

Awareness

1 I'm in touch with how I really feel inside					
2 I feel like I know myself very well	E2	CE with 1	Taps more into self-knowledge than self-awareness		
3 I pay attention to what I think					
4 I am aware of my emotions as I experience them	E3	CE with 5	longer / more complex than 5		
5 I am aware of changes in my mood	E4	CE with 8	more redundant than 8		
6 I observe what goes on in my mind	E5	CE with 8	was confusing; also item 8 is the only one tapping into physical sensations, while items 3 and 7 already tap into the same facet as item 6	2	
7 I am conscious of my thoughts as I think them					
8 I notice different sensations in my body					
9 I notice the feelings in my body	E1	CE with 8	Item 8 taps better into physical sensations	1	

Acceptance

1 I accept myself the way I am	E5	CE with 2	Item 2 is more realistic than item 1	
2 I accept most aspects of myself				
3 I appreciate most parts of my personality	E6	CE with 4	appreciation taps into self-worth	
4 I'm tolerant of my weaknesses				
5 I'm tolerant of my flaws	E1	CE with 4	"flaw" is a more infrequent term than "weakness"	
6 I tolerate my experiences whether they're pleasant or unpleasant	E7		Very long item	1
7 I accept my thoughts and feelings				
8 I accept the mistakes I've made in the past				
9 I accept my shortcomings	E2	CE with 4	"shortcoming" is a more infrequent term than "weakness"	
10 I treat myself in a caring way	E4	CE with 1	more of a behavior than an experience	
11 When I make a mistake I'm patient with myself	E3	CE with 4	Long item	

Self-Worth

1 I really like myself

2 I feel great about myself

3 I value myself highly

E3 CE with 4

slightly more abstract than 4

4 I respect myself a lot

5 I have high regard for myself

6 I know I'm a person of worth

--

This item could also fit into self-acceptance

7 I believe I have many good qualities

E1 CE with 6

More complex than 6

8 I have a very positive attitude toward
myself

E2 CE with 2

Longer item than 2, was written as a back-up item

Competence				
1 I am highly effective at what I do				
2 I tend to perform very well				
3 I am confident in my ability to master new things				
4 I can obtain the results I want	E2	CE with 5	worsens fit (BIC) more than 5	
5 I am able fulfil my goals	E4	CE with 9	this item is more redundant while without 9 the scale becomes more narrow	
6 I can accomplish what I attempt to do				
7 I can successfully complete difficult tasks				
8 I can handle difficult situations	E3	CE with 9	It's more vague than 9	
9 I'm very good at overcoming challenges				
10 I have a great ability to deal with problems	E1	CE with 8	Longer than 8	
Development				
1 Developing				
2 Learning new things	E3	CE with 4	This item is more specific than 4	
3 Expanding my horizons				
4 Improving				2
5 Progressing	E1	CE with 6	advancing is a more frequently used term	2
6 Advancing				1
7 Enhancing my skills				
8 Stretching my limits	E2	CE with 3	stretching one's limits may be a sign of stress	1
9 Increasing my capabilities	E4	CE with 6	more complex than 6	

Purpose				
1 Clear aims and goals	E2	CE with 6	it's generally best to avoid "and." this item may be misinterpreted because of this ambiguity.	
2 A mission				
3 A clear idea of what I want				
4 A purpose				
5 A vision for my life				
6 A plan of what I want	E4	CE with 3	3 and 6 have a similar wording, but plan is more specific than the rest of the items	
7 Something to strive for	E1	CE with 9	7,8,9 have similar wording. Eventually both 7 and 9 are omitted due to poor fit.	
8 Something I aspire to				
9 Something I work towards	E3	CE with 8	this was a back-up item	
Significance				
1 Worthwhile				
2 Valuable	E2	CE with 7	may tap into monetary return	
3 Significant	E1	CE with 4	slightly more vague than 4	1
4 Important				
5 Meaningful				
6 Rewarding				
7 Relevant				1
8 Helpful	E3		may tap into altruism; was initially included as a back-up item	

Congruence

1 What I do is in line with my interests	E2	CE with 3	it's more abstract than 3
2 I base my decisions on my true values			
3 I decide what to do based on what I really care about			
4 What I do is consistent with what I believe I should do			
5 I only do something if it fits my beliefs	E1	CE with 8	5 and 8 are very similarly worded, but 5 taps more into behavior
6 I act in accordance with who I am and what I believe	E6	CE with 11	4,6,8 contain root "belief", it seems more vague than 11
7 My decisions are compatible with who I really am as a person	E3	CE with 4	it's more abstract and more complex than 4
8 I stand firmly by my beliefs	E5	CE with 11	4,6,8 contain the word "belief", 8 is more vague than 11
9 I'm true to myself when I make a decision	-		potential redundancies: 9 and 11 contain the word "myself"; also 9 and 3 contain the root "decision"
10 I feel free to be who I am	E4	CE with 2	it's vague
11 My actions are congruent with how I see myself			very face valid

Connection

1 I feel a strong friendship with the people I regularly interact with			some Ces but not substantial
2 I feel close and connected with the people around me			
3 I feel a deep connection with some people	E6	CE with 9	Connection is already expressed in item 2
4 I feel I belong to a social group	E4	CE with 9	Has several Ces; 9 is a more central aspect of the construct
5 I feel accepted	E5	CE with 7	the root of "accepted" may create errors and cross-loadings with self-acceptance items
6 I feel loved	E3	CE with 9	Has several Ces
7 I feel appreciated			
8 I really like the people in my life			
9 There are people who truly care about me			
10 I feel really cared for	E1	CE with 6	has more complex wording than 6
11 I really care for the people in my life	E2	CE with 8	has the highest skewness
12 I receive support and help when I need it	E7	CE with 9	May tap into social support; this was a back-up item

Appendix 6

Final scales

Instructions: Below you'll find several statements about your experiences. Please indicate how true each statement is regarding the EXPERIENCES IN YOUR LIFE OVERALL. There are no right or wrong answers. Please, choose the answer that best reflects your experience rather than what you think your experience should be.

Happiness

I feel...

- Happy
- Pleased
- Glad
- Cheerful

Vitality

I feel...

- Energetic
- Lively
- Vigorous
- Active

Calmness

I feel...

- Calm
- Peaceful
- Relaxed
- At ease

Optimism

- I'm optimistic
- I hope for the best
- I expect good things to happen
- I'm positive about my plans
- I believe things will work out in the end

Involvement

In my ACTIVITIES (e.g. work/study, free time), I feel...

- Completely involved
- Absorbed by what I'm doing
- Fully engaged
- In tune with what I do

Self-awareness

- I'm in touch with how I really feel inside
- I pay attention to what I think
- I am conscious of my thoughts as I think them
- I notice different sensations in my body

Self-acceptance

- I accept most aspects of myself
- I'm tolerant of my weaknesses
- I accept my thoughts and feelings
- I accept the mistakes I've made in the past

Self-worth

- I really like myself
- I feel great about myself
- I respect myself a lot
- I have high regard for myself
- I know I'm a person of worth

Competence

- I am highly effective at what I do
- I tend to perform very well
- I am confident in my ability to master new things
- I can accomplish what I attempt to do
- I can successfully complete difficult tasks
- I'm very good at overcoming challenges

Development

I feel I am...

- Developing
- Expanding my horizons
- Improving
- Advancing
- Enhancing my skills

Purpose

In my life, I have...

- A mission
- A clear idea of what I want
- A purpose
- A vision for my life
- Something I aspire to

Significance

WHAT I DO in my life is...

- Worthwhile
- Important
- Meaningful
- Rewarding
- Relevant

Congruence

- I base my decisions on my true values
- I decide what to do based on what I really care about
- What I do is consistent with what I believe I should do
- I'm true to myself when I make a decision
- My actions are congruent with how I see myself

Connection

- I feel a strong friendship with the people I regularly interact with
- I feel close and connected with the people around me
- I feel appreciated
- I really like the people in my life
- There are people who truly care about me

Appendix 7**Item loadings (Study 2)**

Item	Loading	Item	Loading
Happiness 1	0.94	Worth 4	0.92
Happiness 2	0.88	Worth 5	0.82
Happiness 3	0.89	Competence 1	0.88
Happiness 4	0.89	Competence 2	0.88
Vitality 1	0.92	Competence 3	0.83
Vitality 2	0.90	Competence 4	0.87
Vitality 3	0.85	Competence 5	0.88
Vitality 4	0.85	Competence 6	0.82
Calmness 1	0.90	Develop 1	0.86
Calmness 2	0.89	Develop 2	0.81
Calmness 3	0.91	Develop 3	0.91
Calmness 4	0.91	Develop 4	0.89
Optimism 1	0.90	Develop 5	0.88
Optimism 2	0.80	Purpose 1	0.87
Optimism 3	0.90	Purpose 2	0.84
Optimism 4	0.88	Purpose 3	0.88
Optimism 5	0.90	Purpose 4	0.90
Involve 1	0.92	Purpose 5	0.86
Involve 2	0.82	Significance 1	0.90
Involve 3	0.88	Significance 2	0.91
Involve 4	0.80	Significance 3	0.93
Aware 1	0.80	Significance 4	0.87
Aware 2	0.77	Significance 5	0.89
Aware 3	0.80	Congruence 1	0.84
Aware 4	0.61	Congruence 2	0.76
Accept 1	0.88	Congruence 3	0.80
Accept 2	0.78	Congruence 4	0.84
Accept 3	0.76	Congruence 5	0.82
Accept 4	0.76	Connect 1	0.84
Worth 1	0.92	Connect 2	0.90
Worth 2	0.93	Connect 3	0.83
Worth 3	0.92	Connect 4	0.83
		Connect 5	0.81

Appendix 8

Higher-order reliability analysis (Study 2)

Schmid Leiman Factor loadings greater than 0.20

	g	F1	F2	F3	h2	u2	p2
Happiness	0.75		0.55		0.87	0.13	0.65
Vitality	0.72		0.43		0.70	0.30	0.73
Calmness	0.69		0.35		0.61	0.39	0.77
Optimism	0.81		0.28		0.77	0.23	0.86
Involvement	0.65	0.29			0.51	0.49	0.83
Awareness	0.57	0.28			0.43	0.57	0.75
Acceptance	0.79			0.49	0.86	0.14	0.72
Self-worth	0.84			0.28	0.81	0.19	0.86
Competence	0.76	0.29			0.67	0.33	0.86
Development	0.70	0.34			0.62	0.38	0.79
Purpose	0.72	0.38			0.67	0.33	0.77
Significance	0.78	0.31			0.72	0.28	0.84
Congruence	0.71	0.29			0.62	0.38	0.82
Connection	0.63		0.25		0.48	0.52	0.84
ω_h	0.86						
ω_{hs}		0.15	0.19	0.17			

Note: Only loadings greater than 0.20 are shown; h2 = communalities; u2 = uniquenesses; p2 = percentage of variance attributable to the general factor; ω_h = Omega hierarchical; ω_{hs} = omega hierarchical subscale (i.e. reliable subscale variance after controlling for the general factor).

Appendix 9

Survey information (Study 3)

Construct	Questionnaire and source	Items	Rating	<i>α</i>
Section 1				
	Scales of general well-being:		5-point rating scale from “not at all true” to “very true”	
	Happiness	4		0.92
	Vitality	4		0.91
	Calmness	4		0.92
	Optimism	5		0.86
	Involvement	4		0.89
	Self-awareness	4		0.82
	Self-acceptance	4		0.84
	Self-worth	5		0.91
	Competence	6		0.88
	Development	5		0.91
	Purpose	5		0.88
	Significance	5		0.90
	Self-congruence	5		0.85
	Connection	5		0.86
Section 2a				
Happiness	Pleasant affect scale (Feldman Barrett & Russell, 1998)	3	5-point rating scale from “not at all” to “extremely”	0.91
	Happiness scale (Huppert & So, 2013)	1	11 points from “Extremely unhappy” to “Extremely happy”	-
Vitality	Subjective vitality scale (Ryan & Frederick, 1997; Bostic et al., 2000)	6	7 points from “not at all true” to “very true”	0.87
	energy / general activation scale (Thayer, 1986)	5	5 points from “not at all” to “extremely”	0.91
Calmness	Pleasant deactivated (Feldman Barrett & Russell, 1998)	5	5 points from “not at all” to “extremely”	0.91
	Calmness / general deactivation scale (Thayer, 1986)	5	5 points from “not at all” to “extremely”	0.79
Optimism	Life orientation test-revised (LOT-R, Scheier, Carver & Bridges, 1994)	6	5 points from “strongly disagree” to “strongly agree”	0.81
	optimism scale (Huppert & So, 2013)	1	5 points from “strongly disagree” to “strongly agree”	-
Involvement	Core flow scale (Martin & Jackson, 2008)	10	5 points from “strongly disagree” to “strongly agree”	0.87
	Total concentration subscale of the Dispositional Flow Scale 2 (DFS2; Jackson & Eklund, 2004)	4	5 points from “never” to “always”	0.89
Self-awareness	Awareness subscale of the Cognitive and Affective Mindfulness Scale-Revised (Feldman et al., 2007)	3	4 points from “rarely/not at all” to “almost always”	0.84
	Awareness subscale of the Philadelphia Mindfulness Scale (Cardaciotto et al., 2008)	10	5 points from “never” to “very often”	0.78
Self-acceptance	Acceptance subscale of the Cognitive and Affective Mindfulness Scale-Revised (Feldman et al., 2007)	3	4 points from “rarely/not at all” to “almost always”	0.74

	Self-kindness subscale of the self-compassion scale (Neff, 2003)	5	5 points from “almost never” to “almost always”	0.84
Self-worth	Rosenberg self-esteem scale (Rosenberg, 1965)	10	4 points from “strongly disagree” to “strongly agree”	0.80
	Self-esteem scale (Sheldon et al., 2001)	3	5 points from “not at all” to “very much”	0.88
Competence	New general self-efficacy scale (Chen et al. 2001)	8	5 points from “strongly disagree” to “strongly agree”	0.85
	Competence satisfaction subscale of the balanced measure of psychological needs (Sheldon & Hilpert, 2012)	3	5 points from “no agreement” to “much agreement”	0.89
Development	Intentional behavior subscale of the Personal Growth Initiative Scale – II (Robitscheck et al., 2012)	4	6 points from “disagree strongly” to “agree strongly”	0.89
	Love of learning subscale of the VIA-IS (Peterson & Seligman, 2004)	10	5 points from “not at all like me” to “very much like me”	0.69
Purpose	Purpose in life test-short form (Schulenberg et al., 2011)	4	5-point semantic differential scale	0.87
	Purposeful life subscale of the meaningful life measure (Morgan & Farsides, 2009)	4	Both 7-point semantic differential scale and Likert scale from “strongly disagree” to “strongly agree”	0.87
Significance	Adapted ⁸ valued life subscale of the meaningful life measure (Morgan & Farsides, 2009)	4	7 points from “strongly disagree” to “strongly agree”	0.87
	Meaning in life scale (Huppert & So, 2013)	1	5 points from “disagree strongly” to “agree strongly”	-
Self-congruence	Authorship subscale of the index of autonomous functioning (Weinstein et al., 2012)	5	5 points from “not at all true” to “completely true”	0.85
	Autonomy satisfaction subscale of the balanced measure of psychological needs (Sheldon & Hilpert, 2012)	3	5 points from “no agreement” to “much agreement”	0.77
Connection	Relatedness satisfaction subscale of the balanced measure of psychological needs (Sheldon & Hilpert, 2012)	3	5 points from “no agreement” to “much agreement”	0.89
	Relatedness satisfaction subscale of the Need Satisfaction and Frustration Scale (Longo, Gunz et al., 2016)	3	7 points from “strongly disagree” to “strongly agree”	0.89
Social desirability	Short form (MC-A; Reynolds, 1982) of the Marlowe-Crowne Social Desirability Scale (Crowne & Marlow, 1960),	11	2 points: “true” and “false”	0.63
Personality	Mini-IPIP measure of the big five personality factors (Donnellan et al., 2006):		5 points from “very inaccurate” to “very accurate”	
	Extraversion	4		0.81
	Agreeableness	4		0.71
	Conscientiousness	4		0.71
	Neuroticism	4		0.71
	Imagination	4		0.68

⁸ Adapted to one’s activities rather than one’s life.

Section 2b				
SWB	Satisfaction with life scale (Diener, Emmons, Larsen & Griffin, 1985)	5	7 points from “strongly disagree” to “strongly agree”	0.86
	Positive and negative affect scale (PANAS; Watson et al., 1988)	20	5 points from “very slightly or not at all” to “extremely”	
	Positive affect	10		0.86
	Negative affect	10		0.84
PWB	Ryff’s scale of psychological well-being (Ryff, 1989) used in the MIDUS II:		7 points from “strongly disagree” to “strongly agree”	
	Autonomy	7		0.75
	Environmental Mastery	7		0.78
	Personal Growth	7		0.75
	Positive Relations	7		0.73
	Purpose in Life	7		0.68
	Self-Acceptance	7		0.81
Social WB	Keyes’ social well-being scale (Keyes, 1998):		7 points from “strongly disagree” to “strongly agree”	
	Social Integration	3		0.89
	Social Acceptance	3		0.71
	Social Contribution	3		0.80
	Social Actualization	3		0.80
	Social Coherence	3		0.72
Ill-being	Patient-reported outcomes measurement system (PROMIS-29 v2) (Cella et al., 2010):			
	Physical Function	4	5 points from “without any difficulty” to “unable to do”	0.85
	Anxiety	4	5 points from “never” to “always”	0.88
	Depression	4	5 points from “never” to “always”	0.90
	Fatigue	4	5 points from “Not at all” to “Very much”	0.90
	Sleep Disturbance	4	5 points from “very poor” to “very good” (item 1) and 5 points from “Not at all” to “Very much” (items 2-4)	0.77
	Ability to Participate in Social Roles and Activities	4	5 points from “never” to “always”	0.89
	Pain Interference	4	5 points from “Not at all” to “Very much”	0.93
	Pain Intensity	1	11 points from “no pain” to “Worst imaginable pain”	-
Construct value	Value / desirability scale adapted from Diener (2000)	14	7 points from “No importance whatsoever” to “Extraordinarily important and valuable”	N/A
Follow-up				
Dark triad	The dirty dozen questionnaire measuring narcissism, psychopathy and Machiavellianism (Jonason & Webster, 2010).	12	9 points from “strongly disagree” to “strongly agree”	0.73

Appendix 10**Item loadings (Study 3)**

Item	Loading	Item	Loading
Happiness 1	0.94	Worth 4	0.93
Happiness 2	0.92	Worth 5	0.87
Happiness 3	0.93		
Happiness 4	0.91	Competence 1	0.84
		Competence 2	0.86
Vitality 1	0.93	Competence 3	0.85
Vitality 2	0.92	Competence 4	0.83
Vitality 3	0.88	Competence 5	0.85
Vitality 4	0.87	Competence 6	0.84
Calmness 1	0.94	Develop 1	0.89
Calmness 2	0.91	Develop 2	0.86
Calmness 3	0.93	Develop 3	0.91
Calmness 4	0.93	Develop 4	0.91
		Develop 5	0.89
Optimism 1	0.91		
Optimism 2	0.75	Purpose 1	0.88
Optimism 3	0.89	Purpose 2	0.84
Optimism 4	0.90	Purpose 3	0.87
Optimism 5	0.89	Purpose 4	0.89
		Purpose 5	0.84
Involvement 1	0.92		
Involvement 2	0.83	Significance 1	0.92
Involvement 3	0.92	Significance 2	0.90
Involvement 4	0.85	Significance 3	0.93
		Significance 4	0.85
Awareness 1	0.81	Significance 5	0.89
Awareness 2	0.82		
Awareness 3	0.84	Congruence 1	0.82
Awareness 4	0.59	Congruence 2	0.73
		Congruence 3	0.79
Acceptance 1	0.88	Congruence 4	0.80
Acceptance 2	0.75	Congruence 5	0.78
Acceptance 3	0.82		
Acceptance 4	0.74	Connection 1	0.84
		Connection 2	0.89
Worth 1	0.94	Connection 3	0.81
Worth 2	0.92	Connection 4	0.82
Worth 3	0.93	Connection 5	0.75

Appendix 11

Higher-order reliability analysis (Study 3)

Schmid Leiman Factor loadings greater than 0.20

	g	F1	F2	F3	h2	u2	p2
Happiness	0.80	0.47			0.87	0.13	0.74
Vitality	0.74	0.29			0.66	0.34	0.84
Calmness	0.69	0.43			0.67	0.33	0.71
Optimism	0.81	0.27	0.21		0.77	0.23	0.85
Involvement	0.68			0.28	0.54	0.46	0.84
Awareness	0.55			0.44	0.51	0.49	0.58
Acceptance	0.73			0.41	0.75	0.25	0.72
Self-worth	0.80	0.32			0.75	0.25	0.84
Competence	0.72			0.21	0.59	0.41	0.88
Development	0.71		0.34		0.62	0.38	0.81
Purpose	0.77		0.45		0.79	0.21	0.74
Significance	0.78		0.29		0.70	0.30	0.86
Congruence	0.73			0.34	0.65	0.35	0.82
Connection	0.67	0.26			0.53	0.47	0.85
ω_h	0.86						
ω_{hs}		0.16	0.16	0.17			

Note: Only loadings greater than 0.20 are shown; h2 = communalities; u2 = uniquenesses; p2 = percentage of variance attributable to the general factor; ω_h = omega hierarchical; ω_{hs} = omega hierarchical subscale (i.e. reliable subscale variance after controlling for the general factor).

Appendix 12

Longitudinal invariance – Time 1 and 2 model fit

Time 1

Scale	YB χ^2	df	p	CFI	SRMR	RMSEA
Happiness	1.02	2	.602	1.00	.003	.000 (.000-.055)
Vitality	0.81	2	.667	1.00	.004	.000 (.000-.041)
Calmness	0.39	2	.824	1.00	.002	.000 (.000-.019)
Optimism	3.60	5	.608	1.00	.006	.000 (.000-.044)
Involvement	0.92	2	.631	1.00	.003	.000 (.000-.053)
Awareness	3.80	2	.150	.996	.011	.046 (.000-.107)
Acceptance	7.74	2	.021	.989	.016	.082 (.039-.130)
Self-worth	6.32	5	.277	.999	.006	.025 (.000-.061)
Competence	17.46	9	.042	.993	.014	.047 (.020-.072)
Development	4.58	5	.469	1.00	.006	.000 (.000-.054)
Purpose	19.97	5	.001	.981	.018	.083 (.057-.111)
Significance	13.50	5	.019	.991	.013	.063 (.037-.090)
Congruence	11.90	5	.036	.987	.022	.057 (.027-.086)
Connection	17.89	5	.003	.986	.019	.077 (.047-.111)
14 correlated factors	2927.22	1924	.000	.958	.041	.035 (.032-.037)
Higher-order	3366.65	2001	.000	.943	.059	.040 (.038-.042)

n = 431

Time 2

Scale	YB χ^2	df	p	CFI	SRMR	RMSEA (90% CI)
Happiness	1.39	2	.499	1.00	.003	.000 (.000-.067)
Vitality	2.23	2	.328	1.00	.004	.016 (.000-.081)
Calmness	0.73	2	.693	1.00	.002	.000 (.000-.042)
Optimism	16.52	5	.006	.988	.013	.073 (.045-.103)
Involvement	2.72	2	.256	.999	.007	.029 (.000-.084)
Self-awareness	5.78	2	.056	.991	.017	.066 (.015-.121)
Self-acceptance	7.98	2	.019	.987	.018	.083 (.042-.130)
Self-worth	10.85	5	.054	.995	.009	.052 (.020-.083)
Competence	11.69	9	.231	.998	.011	.026 (.000-.056)
Development	6.34	5	.274	.999	.009	.025 (.000-.058)
Purpose	16.65	5	.005	.986	.017	.074 (.047-.102)
Significance	10.35	5	.066	.995	.009	.050 (.017-.081)
Congruence	9.43	5	.093	.994	.015	.045 (.000-.081)
Connection	7.82	5	.166	.998	.010	.036 (.000-.078)
14 correlated factors	2981.15	1924	.000	.956	.038	.036 (.033-.038)
Higher-order	3326.74	2001	.000	.945	.052	.039 (.037-.041)

n = 431