Assessment, Measurement and Evaluation within Public Mental Health
Assessment, Measurement and Evaluation within Public Mental Health

Methodological and Psychometric Challenges

Dissertation

to obtain the degree of Doctor at Maastricht University, on the authority of the Rector Magnificus, Prof. dr. Rianne M. Letschert in accordance with the decision of the Board of Deans, to be defended in public

on Monday 7 May 2018 at 14.00 hours

by

Hubertus (Huub) Maria Hoofs
Promotores
Prof. dr. IJ. Kant
Prof. dr. ir. M. W. J. Jansen

Co-promotor
dr. N. W. H. Jansen

Reading committee
Prof. dr. F. R. H. Zijlstra (Chair)
dr. M. A. van Bokhoven
Prof. dr. P. A. E. G. Delespaul
Prof. dr. E. L. Hamaker, Utrecht University
Prof. dr. ir. H. C. W. de Vet, VU University Medical Center
# Contents

1 Introduction 5

2 The context dependency of the self-report version of the Strength and Difficulties Questionnaire (SDQ): A cross-sectional study between two administration settings 29

3 Structural validity for the Need for Recovery Scale and the Checklist Individual Strength and its equivalence/invariance within different subgroups of workers 55

4 Monitoring of need for recovery and prolonged fatigue within the working population: Evaluation of reliability and agreement over time 71

5 The trait and occasion components of fatigue and their association with sickness absence 93

6 Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA 109

7 Discussion 149

Summary 171

Samenvatting 181

Valorisation addendum 191

Dankwoord 197

CV 201
1 Introduction
Public Mental Health

Public health refers to “the science and art of preventing disease, prolonging life and promoting health through organised efforts and informed choices of society, organizations, public and private, communities and individuals” (World Health Organization, 2001). Within public health, health is considered as “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity”. Many subdivisions of public health exist, such as the distinction between work and health (e.g. occupational health professionals) and society and health (e.g. preventive child health care). Both these domains are a combination of several disciplines such as epidemiology and bio-statistics, which shows the multifaceted nature of public health.

An important aspect of public health is mental health. Public mental health can therefore be defined as “a public health or population health science approach to mental health and the mental health variations exhibited by populations” (Davies, 2014). Public mental health can be considered as consisting out of three overlapping domains: mental health promotion, mental illness prevention, and treatment and rehabilitation (Figure 1). At the core of all these domains of public mental health is the measurement of mental health. Measuring mental health is, however, far from trivial. The challenges associated with the measurement of mental health are sometimes evident but often concealed. They are important nevertheless as conclusions within public mental health are often based on results from measurement instruments. Measurement is also important for the assessment of public mental health. It should be noted that assessment comprises other aspects besides results from measurement instruments (e.g. the combination of instruments) which exhibit their own difficulties. The same goes for evaluation of public mental health, in which results from assessments and measurement instruments should be quantified (Huitt, Hummel, & Kaeck, 2001). Within this thesis the focus is on measurement of public mental health, but implications for assessment and evaluation are also taken into account.

While the challenges associated with measurement within public mental health may seem small and unimportant at first, they can actually have significant consequences. They could therefore be considered the latent defects of public mental health. That is, if public mental health policy or research is based on measurements with unknown defects, societal impact can be large. Misinformed policy could for example lead to misallocation of budget or to the implementation of preventive strategies that are in fact ineffective. This thesis illustrates the potential impact of latent defects of measurement instruments. Before we turn to identifying these latent defects, however, this introduction starts with the operationalisation of mental health. Following this short introduction the current state of public mental health will be addressed. Building on these sections the framework for the measurement of mental health will be examined. This general framework will also illustrate the implication of measurement challenges within the public mental health. A definition of latent defects with respect to the measurement of public mental health will be given in the following section. The remainder of this introduction will cover the statistical and methodological tools used in this thesis. It will conclude with an outline of the subsequent chapters.
Mental Health: Dichotomous versus Continuous

The world health organisation (WHO) states that mental health is an integral part of (general) health and is directly associated with physical health and behaviour. The emerging consensus is therefore that mental health is closely interwoven with physical and social health which plays a crucial role during the entire life span (World Health Organization, 2001). The formal definition of mental health is “a state of well-being in which the individual realises his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community” (World Health Organization, 2001).

Mental health is sometimes portrayed as the absence or presence of a certain condition or illness. As with physical health, however, the appearance and development of mental health is rarely strictly dichotomous, but moves along a continuum. For example, a person either suffers from hypertension or doesn’t, just as a person can either have a depression or not. However, the severity of depressive symptoms, analogue to blood pressure, forms the underlying construct of depression (World Health Organization, 2001). It is along this continuum that a distinction is made between people who suffer from a mental disorder and those that do not (Figure 2). Even when depressive symptoms are not classified as a disorder they can still be a risk factor for ill health. As with blood pressure, falling in the higher ranges carries risks, not the least of which is developing the disorder. To clarify this distinction, the terminology in this thesis will be based around disorders and symptoms. Disorders portray a mental disease which is diagnostically assessed using the International Classification of Diseases (ICD) or the Diagnostic and Statistical Manual of Mental Disorders (DSM). Symptoms refer to the number and severity of difficulties a particular person experiences.

The distinction between disorders and symptoms is important as one of the domains of public mental health is the prevention of mental illness. To adequately fulfil this purpose it is crucial

![Figure 1: Decomposition of public mental health into its core dimensions, from Davies (2014).](image-url)
Introduction

Figure 2: Relation between the continuum of depressive symptoms and the dichotomization of symptoms into the absence or presence of a depressive disorder.

that the development of a mental illness can be detected at an early stage (Koplan et al., 2009). It is, therefore, important to acknowledge the continuum that underlies mental health. This allows preventive measures or interventions for people who show an adverse development but are not yet classified as having a disorder (World Health Organization, 2001). Within occupational health, for example, it would be preferable to use a proactive approach for aiding employees who are still working but have mild to severe depressive symptoms (Figure 2; A) than 'waiting' until employees become unable to work as a consequence of mental health complaints possibly already developed into a (rather severe) disorder (Figure 2; B). An active approach could facilitate timely interventions (e.g. cognitive behavioural therapy) or even prevention.

Relevance of Studying Public Mental Health

Key facts

The importance of mental health within public health is somewhat paradoxical. On the one hand it is still an ill-defined field characterised by its lack of clarity over its boundaries and definitions (Davies, 2014). This is also reflected in ever-lasting discussions about the true rate of mental health disorders (e.g. attention deficit disorder) and the stigma people with a mental illness experience (see Box 1). If someone has a broken leg and calls in sick for work, the response is likely to be different than for someone with an episode of depressive symptoms. On the other hand, however, awareness for mental health is increasing and rapidly growing (World Health Organization, 2013, 2015). This important, growing, role of mental health within general (public) health is also stressed by several facts. Within the Netherlands, for example, the total health costs in 2011 for mental disorders was €19,612 million, covering 21.9% of the total health costs in the Netherlands. It is expected that in the next 20 years mental disorders will remain the largest disease burden in the Netherlands along with cardiovascular diseases (Ministry of Health, Welfare and Sport, 2017).
Within the UK mental illness is the largest single cause of disability (Davies, 2014). It is estimated that around 40% of all disability pensions are based on a mental disorder (OECD, 2012). For young people this percentage even exceeds the 50% mark. Many of those young disabled persons have never held a job. These mental health difficulties can have devastating effects on the individual, as this effect is often persistent. Interestingly it has been found that common mental health disorders reduce employee productivity as much as severe mental illness (OECD, 2012; Schultz & Edington, 2007). This also clearly illustrates the added value of addressing mental health as a continuum, as not only employees with mental disorders are at risk of adverse (health) outcomes. This makes that mental illness costs 4.5% of the gross domestic product of the UK. Again, however, these developments are not in line with the decreasing investments in mental health services, which are decreasing (Davies, 2014).

Besides disability and reduced productivity, in case of employees, a number of studies have reported on the quality of life of individuals with mental disorders, concluding that the negative impact is not only substantial but also sustained (e.g., UK700 Group, 1999). It has been shown that quality of life continues to be poor even after recovery from mental disorders as a result of social factors that include continued stigma and discrimination (e.g., Lehman, Steinwachs, &

---

**Box 1: Addressing the ADD/ADHD ‘Epidemic’**

The Huffington Post [17/12/2013]:

Recent data published in *Journal of the American Academy of Child and Adolescent Psychiatry* found that 6.4 million children between 4 and 17 have been diagnosed with ADHD, and roughly 3.5 million of them are on some form of medication. These statistics, compiled by the Centers for Disease Control and Prevention, reflect a 41 percent increase in ADHD diagnoses in the past decade.

So what’s going on here? Are there really that many more children who have ADD/ADHD, or have we simply broadened our diagnostic guidelines? Is it possible we are over-diagnosing these disorders, as well as over-treating them? These are serious questions. Many children are prescribed powerful psychiatric drugs, such as Adderall and Ritalin, which, like most prescriptions, come with their share of side effects.

This excerpt from the Huffington Post is one of many articles devoted to the increasing numbers of ADHD diagnoses and related drugs use. This growth in literature is not only limited to mainstream media, but also finds its way into the scientific literature. While there are numerous reasons that such an emerging growth reflects reality, with increased awareness and a changing society being the most prominent, there are also good reasons to remain cautious. Mine’s Nystagmus (Culpin & Smith, 1930), for example, a condition found mainly in coal miners saw a rapid increase in prevalence after it was declared a compensable disease (Henderson & Madan, 2014). Symptoms were, however, non-specific and open to various interpretations. Eventually this disease ‘disappeared’. This, nevertheless, does not make the mining industry a healthy business again (Karasek & Theorell, 1992), nor is the the current society beneficial for a fair share of children with a short attention span. Quality measurement of mental health is, therefore, crucial as this can potentially reduce some of the skepticism around mental health problems and improve its legitimacy.
The Co-Investigators of the PORT Project, 1998). The impact on quality of life is not limited to severe mental disorders, such as schizophrenia and depression. Anxiety and panic disorders also have a major effect, in particular with regard to psychological functioning (Olatunji, Cisler, & Tolin, 2007; Burckhardt & Anderson, 2003). Decreased mental health is however not merely a risk factor for a decreased quality of life and further degrading mental health but is also associated with decreased physical health and general well being (Prince et al., 2007; Ferketich, Schwartzbaum, Frid, & Moeschberger, 2000). This relation is, however, by no means unidirectional. Many (adverse) physical health outcomes are, for example, a strong predictor for a deterioration of mental health (e.g., Fallowfield, 2002). Again, this can range from some mild episodes of depressive symptoms to severe mental disorders.

Role of public health with respect to mental health

Promotion of mental health is crucial to maintaining a healthy and productive population. In occupational health, for example, prevention (primary, secondary, and tertiary) is important to halt disability arising from, for example, work-related stress. Treatment and rehabilitation is needed to limit long-term adverse effects of a mental illness. As such, while efforts at different levels (individual, organisational, and governmental) are important to enable employees to remain engaged in sustained labor, it is equally important that those who are (temporarily) disabled receive adequate attention in order to successfully return to work. In this respect it is important to note that mental health related disability numbers nearly doubled in the space of two decades (Eurofound, 2016). This is hardly surprising given the change in work content the last decades, which saw a shift from physically towards psychologically demanding work (Cattrell et al., 2011). Mental health problems are, however, also a risk factor for poor physical health (Knudsen et al., 2010; Stansfeld et al., 1995). This underlines the interdependence between mental and physical health. The workplace, as such, is a perfect illustration for the importance of all three domains of public mental health (Egan et al., 2007).

Within the field of society and health the relation between mental and physical health is also apparent. The preventive youth health care within the Netherlands, for example, not only focuses on the presence of physical problems but also monitors the mental health of children (Brugman, Reijneveld, Verhulst, & Verloove-Vanhorick, 2001). As already indicated, early detection and treatment of mental health problems is more effective within these early stages of a mental health disorder than in the later stages (Moffitt, 1993; Moffitt & Caspi, 2001). To increase the effectiveness of monitoring it is important to select (risk) groups that have an increased likelihood to develop (specific) mental health problems. In this way an optimum can be achieved between the potential benefit of monitoring and the burden it places on people and associated costs.

Measuring a Latent Construct

High quality public mental health starts with the successful measurement of mental health. In contrast to physical health, in which most constructs have an objective definition, mental health is almost always latent. As such, mental health can never be directly measured

---

1This is, of course, not to say that the measurement of, for example, blood pressure is without challenges (Schwartz, Haas, Gerin, & Pickering, 2003). The general point is that blood pressure, as with most physical
nor compared with a golden standard. To measure mental health, manifest indicators are a necessity. For illustration purposes the Rorschach test will be used in this introduction. The Rorschach test (Wood, Lilienfeld, Garb, & Nezworski, 2000) consists of 10 plates which consists of inkblots (e.g. card V shown in Figure 3). A subject is shown the inkblot and through free association and inquiry the subject should describe his or her perception of the different inkblots. These responses are subsequently classified.

Especially the translation, and its difficulty, of a latent construct towards manifest indicators is clearly demonstrated with the Rorschach test. While it is impossible to directly measure a mental state it is possible to measure a perception, motivation, or cognitive process. If it is believed that these manifest outcomes represent the underlying mental health construct (e.g. depression), these manifest outcomes tell us something about this mental health construct. If a subject, for example, sees a tired butterfly in card V this perception of tiredness could pertain information about the mental health status of the subject. As such, these inkblots try to provoke a measurable response associated with the construct of interest. While this test is often classified as pseudo science and its validity is at least questionable (Garb, 1999), it does clearly show the complexity of measuring mental health and its associated challenges: Translating the latent construct into measurable manifest indicators. It should be noted that this does not hold for neuroimaging techniques, which have a different approach (see Box 2).

**Goal of this Thesis**

The latent nature of mental health in combination with the necessity of a continuous framework makes the measurement of public mental health a veritable minefield. By the same token, the development of high-quality measurement instruments is superbly challenging. More precisely, it is a valid validation of this quality that is most challenging. That is, while an instrument could have a high quality, assessing and investigating this quality is a complex task. While any domain in research has to be critical regarding its methodology, the fact that mental health is de facto latent makes it susceptible to a wide range of pitfalls that are not directly obvious. These pitfalls are dubbed ‘latent defects’ within this thesis. The ‘label’ latent defects is used here because measurement issues have many parallels with true latent defects in the law of the sale of property. In real estate, for example, a construction can look valid but on closer inspection is actually faulty. The same mechanism can be found in measurement instruments that were (or are) believed to be reliable and valid. Once adverse effects are found, however, it becomes a difficult task to revert their implementation in society. The same goes for actual latent defects which are often time- and money-consuming to restore.

While the characterisation as latent defects may seem somewhat of an exaggeration, it should be noted that the consequences of wrongful measurement or evaluation of measurements within public mental health can be far-reaching. Within public mental health these implications would not directly lie in short-term consequences but mostly in the long-term (and sometimes

---

2Some proponents, however, even claim it is capable to detect cancer (Graves, Mead, & Pearson, 1986)
indirect) consequences. For example, mental health of most people is assessed during their childhood by means of a questionnaire (Brugman et al., 2001; Reijneveld, Vogels, Hoekstra, & Crone, 2006; Reijneveld, Brugman, Verhulst, & Verloove-Vanhorick, 2004). Whilst high (or low) scores on these questionnaires are by no means the sole determinant of further actions, they can hinder informed decision making when latent defects of measurement instruments alter their validity, both on a individual and societal level (Vostanis, 2006; Ministry of Health, Welfare and Sport & Statistics Netherlands, 2014; Engelse & Zwijgers, 2013). For individuals it could, for example, result in an incorrect referral to specialised care (or the failure to do so if needed). On a societal level invalid measurement of mental health could hinder public mental health promotion and prevention as the relevant statistics are incorrect. The main goal of this thesis is therefore:

- To identify potential latent defects of measurement instruments and evaluate their impact within the public mental health

To achieve this goal, potential latent defects of measurement instruments will be analysed for a range of different measurement instruments applied within different settings. The remainder of this introduction consists of four sections. First, a more thorough definition of latent defects is provided. Second, the statistical tools and methods which are mostly used to assess the quality of a measurement instrument within the public mental health are introduced (α-
psychometrics). The third section will illustrate the shortcomings of these tools with respect to the detection of latent defects and introduces $\beta$-psychometrics which is more fit for this purpose. The fourth section will provide an outline of this thesis.

**Latent Defects**

Latent defects can in the context of this thesis be defined as: “actors that can result in the violation of (important) quality criteria of an instrument measuring public mental health, without it being (directly) apparent.” This definition contains several aspects.

### Box 2: Using neuroimaging to measure mental health

While almost all instruments target manifest indicators to measure the underlying (latent) mental health constructs, neuroimaging techniques somewhat ‘mirror’ this route. These techniques (e.g. functional magnetic resonance imaging) measure specific parts of the brains that are thought to be associated with specific mental health (disorders). Neuroimaging, therefore, measures the ‘precursors’ of mental health (disorders), while other instruments (e.g. questionnaires) measure their ‘outcome’. These ‘precursors’ are known as biomarkers that are defined as “characteristics that are objectively measured and evaluated as an indicator of normal processes, pathogenic processes, or pharmacological responses to therapeutic intervention” (Biomarkers Definitions Working Group, 2001). Striking within this definition is the notion of an indicator which shows a clear parallel with the indicators of a questionnaire. Neuroimaging is, however, not exempted from potential pitfalls regarding validity and reliability. While reliability is high with the same scanner, for example, consistency across different scanners is insufficiently assessed (Kruggel, Turner, & Muftuler, 2010). As most mental health disorders are, furthermore, a combination of different biological processes, each pathology has an unique pattern of biomarkers. That is, while the clinical prototype can be the same the underlying neuroimaging phenotype can differ between subjects (Linden, 2012). While this could also be seen as a differentiation of the disease, classification is still mostly based on DSM V criteria. While these criteria are often multifaceted, differentiation is not always available—certainly not on the basis of biomarkers. These biomarkers are, nevertheless, important for both a ‘proof of concept’ and ‘proof of mechanism’ (Soares, 2010). While neuroimaging, or even genetic imaging, is a promising avenue for future research regarding mental health, its current application in public health is still limited. First of all, the cost of neuroimaging is enormous compared with the costs of other instruments. Second, neuroimaging is akin to using sledgehammers to crack nuts. Within public health symptoms are mostly mild, while most applications of neuroimaging focus on severe mental health disorders such as neurodegenerative disorders (e.g. Alzheimer’s and Parkinson’s disease). Third, neuroimaging studies are still relatively small scaled while public health encompasses whole populations. Fourth, and maybe most importantly, heterogeneity is still too large across individuals to use neuroimaging as a diagnostic tool. Even between disease groups positive findings are limited, as most successful studies have healthy groups as comparison (National Institute of Mental Health, 2010).
**Actors.** The source of a latent defect can be highly variable. Of course the instrument itself should have the most prominent role in examining the potential existence of a latent defect within an instrument. That is, the phrasing or provided answer categories could result in invalid results. While the instrument is a prominent actor, this thesis also examines whether other actors can also influence the working of an instrument (e.g. sample or setting). While ‘actors’ are not true persons, the term is selected to emphasise the active nature of the factors that underly the latent defects.

**Quality criteria.** While different sorts of subdivisions are possible for the categorisation of quality criteria the current thesis uses those as outlined by Terwee et al. (2007). These eight criteria are outlined in Table 1. The corresponding article of Terwee et al. (2007) defines some clear measures that can be used to formally assess these criteria. The most simple example of a quality criterion would be the reliability of results. If this criterion is violated results would be unreliable. In the light of latent defects this would imply that due to a specific actor the results of an instrument are not consistent. A double denial within an item of a questionnaire, for example, could result in unreliable outcomes regarding this item—and therefore the questionnaire as a whole. Violation of a quality criterion can, however, also be more ‘concealed’. Scores of an instrument can, for example, be perfectly reliable within specific settings while comparing scores across these settings is invalid, violating the agreement of the instrument across settings. As such, the relation between the actors (source) and quality criteria (outcome) is a many-to-many relation. This many-to-many relation also explicates that there can be multiple actors at the same time that may result in multiple violations of quality criteria at the same time.

**Instrument.** While the current thesis solely focuses on questionnaires, the range of instruments is of course very wide. Instruments include, amongst others, interviews, observations, and biometrics. While the difficulty to find latent defects within these various sorts of instruments greatly differs, this does not incline that this has any relation with the magnitude or severity of latent defects.\(^3\) Of course instruments are not limited to measurement but also assessment of public mental health can be subject to latent defects. While more indirectly, evaluation of public mental health could also be effected by latent defects. That is, if measurement and assessment on which evaluation is based are subject to latent defects it will also hinder a proper and valid evaluation of the results.

**Public mental health.** Although the scope of the current thesis is Public Mental Health, other fields within the (public) health could of course also be subject to the influence of latent defects if instruments are used. The focus on public mental health makes the conceptualisation and assessment of latent defects more profound.

**Apparent.** If all the foregoing aspects are true for an instrument it would be invalid to apply it in practice. However, if this is directly apparent it would be spurious to apply it anyway. For a latent defect to be latent, however, the violation of a quality criterion (inclined by the actor) should, at least to some extent, be concealed. This shows the importance of an active

\(^3\)On the contrary, one could argue that instruments for which latent defects are hard to detect are likely to be prone to latent defects as this would hint on the lack of some formal and objective tools to assess the quality of the instrument.
Table 1: Quality criteria for measurement and assessment properties of public mental health measurement instruments (Terwee et al., 2007)

<table>
<thead>
<tr>
<th>Properties</th>
<th>Content validity</th>
<th>Construct validity</th>
<th>Floor and ceiling effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal consistency</td>
<td>Reproducibility (agreement &amp; reliability)</td>
<td>Responsiveness</td>
<td>Interpretability</td>
</tr>
<tr>
<td>Criterion validity</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

modelling approach. The concealed nature also implies that the underlying problems often ‘slip in’ the application of the instrument, which in turn makes it harder to properly deal with them. As such the list as provided by Terwee et al. (2007) should be used as a baseline for psychometric quality of an instrument, which should be substantiated by an active modelling approach critically assessing important quality criteria for the research question at hand.

It should be noted that this definition does not include any statement regarding the latent nature of the construct being measured with the instrument. This has two reasons. First, public mental health is always latent, as outlined earlier in this chapter. The inclusion of public mental health in the definition therefore implies that it is directed at latent constructs. Second, and even more important, is the fact that instruments measuring manifest outcomes can also be prone to latent defects. As such public mental health could be replaced with health in general to make the definition more widely applicable. As already indicated the different actors and sources of misuse is a many-to-many relationship. As such this thesis includes, by no means, an exhaustive list of all potential latent defects. Especially the sources of actors could be considered inexhaustible. This thesis includes, however, a variation of different latent defects as the chapters vary in the actors and quality criteria that are potentially violated. It furthermore includes different populations, measurement instruments, and study methodologies. The latent defects are, furthermore, based on important (measurement) issues within the respective fields in which the instruments are often applied. As such these chapters also have a strong stand-alone basis as they have important implications for these various fields.

Before addressing the analytical tools that can potentially identify latent defects, ‘contemporary’ analytical tools that assess the quality criteria of Table 1 are addressed. This outline is a necessity to show the strengths of these classic tools but also the associated limitations in identifying latent defects within the measurement of public mental health.

α-Psychometrics

Within this section some contemporary tools will be reviewed which assess some of the quality criteria of Table 1. These statistical tools and methods are within the field of ‘α-psychometrics’. The goal of α-psychometrics is to (Wicherts, 2007): “gain understanding in the hypothetical construct ... and to study the degree to which [this construct] can explain a host of psychological and societal phenomena.” One of the most used reliability measures is the internal consistency of a test (Figure 4), such as the Chronbach’s α (Bland & Altman, 1997). Internal consistency is a measure based on the relationship between the indicators of an instrument (Sijtsma, 2008). As it is hypothesised that all observations, at least to some extent, measure the same construct it is expected that all observations show some form of
Figure 4: Illustration of the importance of reliability and validity for a measurement instrument. This also shows that both aspects are not necessarily correlated with each other.

interdependence. That is, considering the example of the Rorschach test, if a subject sees a sad butterfly in the first inkblot it is not to be expected that he or she sees a happy clown in the next. If such consistency exists it is likely that an instrument is also consistent over time and across subjects. If a subject, for instance, is measured on two different occasions with the Rorschach test and this subject has the same level of depression on these two measurements it is to be expected that the response on the different inkblots should be, at least, comparable.

Convergent and divergent validity (Figure 4), investigate the relation between the instrument with a construct it should (or should not) be associated (Campbell & Fiske, 1959; Watson et al., 1995). Again, with respect to the assessment of physical health such forms of validity are less difficult to assess. If, for instance, a screening instrument is developed to facilitate the detection of colorectal cancer (Burt, 2010), it is possible to compare the results of the screening instrument with samples of tumor tissue or a colonoscopy (Ng et al., 2009). With mental health such clear golden standards do simply not exist. A clinical interview by a trained psychiatrist is still no guarantee of the absence or presence of a mental health outcome (Wittchen, 1994; Collins & Huynh, 2014). Even if a hypothesised relation is found within a study, alternative explanations can be found. If the Rorschach test, for example, measures anxiety instead of depression it could still be predictive of suicide intentions while it does not adhere to the underlying theoretical framework—in which it is the depression that should predict the suicide intentions.

Crucial within the assessment of the quality of an instrument is the structure of a measurement instrument. The structure of an instrument corresponds which it proposed theoretical
framework (Mokkink et al., 2010). To assess this, exploratory factor analyses (EFA) are often used (e.g., Underwood & Teresi, 2002). This is a statistical method used to uncover the underlying structure (latent constructs) of a set of indicators. Each indicator has a certain relation (i.e. place on the target) with each factor that is retrieved from the EFA (Thompson, 2004). Through inspection of these relations it is possible to retrieve the meaning of the extracted latent constructs. Consider, for example, the Hospital Anxiety and Depression Scale (HADS) which is a questionnaire intended to measure depression and anxiety (Bjelland, Dahl, Haug, & Neckelmann, 2002). Questionnaires include items that are thought to reflect (at least partly) these underlying constructs. As such if one wants to measure depression the question is not: “Do you have a depression”, but “I have lost interest in my appearance” or “I feel as if I am slowed down”. Each of these items is thought to be a manifest of the latent construct depression. That is, it is to be expected that if someone is depressed it is also likely that that person has lost interest in her or his appearance. The HADS consists of 14 items, with 7 items covering on depression and the other 7 anxiety. Using an EFA it is expected that the 7 depression items cluster together on one factor and have only small relations with the other (anxiety) factor. For the anxiety items the opposite is expected. In combination with the concept of reliability it is expected that for higher levels of depression the likelihood that the response on all indicators is more inclined to show the manifestation of depression. It is shown that for the HADS most of these assumptions hold (Bjelland et al., 2002), showing that the HADS does seem to measure two constructs independently.

**β-Psychometrics**

**Confirmatory factor analysis**

Exploration of the quality criteria of an instrument through aforementioned means, including EFA and internal consistency, falls within the field of α-psychometrics. Such α-psychometrics is a prerequisite to assess the quality of an instrument but is certainly not conclusive as it fails to thoroughly grasp the underlying structure of an instrument. For a more complete understanding of an instrument it is necessary to employ β-psychometrics. β-psychometrics is concerned with understanding the relation between test or item scores and the latent variable(s) supposed to underlie these scores (Wicherts, 2007). The field of β-psychometrics employs statistical models to understand this complex relation more fully (Borsboom, 2006a). One of the contrasts between both psychometrics is their a-priori and a post-hoc nature. Reconsider, for example, the HADS example. Within α-psychometrics each item is allowed to load on each factor (Figure 5A). While some informative constraints can be examined, the extracted number of factors is furthermore based on several statistics extracted from the data at hand. Within the field of β-psychometrics a theory based model is directly tested (Figure 5B). So for the HADS it is expected that the indicators measuring anxiety will all load on the same factor, hence anxiety ($F_α$), while the depression indicators will load on a different factor thought to represent depression ($F_δ$). As such, all underlying relations between the anxiety and depression items have to be explained through the overarching latent factors (i.e.

---

4While out of the scope of the current thesis, within measurement theory there is some debate if the relation between an item and a latent construct should always be this way around. That is, it could be also suggested that (for some construct especially) a set of items results in a specific construct, and is not a result of this construct (Fleuren, de Grip, Jansen, Kant, & Zijlstra, 2016).
Figure 5: Difference between exploratory factor analysis (EFA; left) and confirmatory factor analysis (CFA; right) for the HADS.

The latent construct can, as such, be regarded as the target in Figure 6 on which bullets (indicators) are fired with a specific rifle (instrument). With the depression ‘rifle’ (triangles) it is expected that all bullets land somewhat in the same location, and the same goes for the anxiety rifle (squares). For the sake of illustration, let’s further state that higher levels of impact indicate a higher observed level on the indicator (i.e. more depressed based on that indicator). This illustration shows the difference in the post-hoc and a-priori approach of both methods. Within EFA (and internal consistency measures), first both rifles are shot and afterwards the target is placed in such a way it fits best with the pattern of the bullets. Within CFA, however, the targets are placed first and than the bullets are fired at the target. As such, one should have an idea where the bullets should land and what the overlap (or absence of overlap) between the two targets is. Using this approach one can validate the instrument against the theoretical framework (Millsap, 2011). It could be, for example, that the second item of the HADS is believed to be a manifest of anxiety. Within EFA, however, it could
turn out that this indicator shows a much higher overlap with the depression indicators, and is placed within this construct (Figure 6). As such, the meaning of an indicator is (at least partly) attributed to the pattern it shows with other indicators. Within the CFA method, however, this indicator is expected to be somewhere in the anxiety target. If it is highly off-target this is incorporated in the model and penalises the model, as this indicator has apparently a deviation. While post-hoc solutions with such a divergent indicator would still be possible, its initial evaluation takes place at an a-priori basis (Markland, 2007; Millsap, 2007). Such an a-priori basis is based on an active modelling approach which requires a thorough theoretical consideration of the measurement instrument at hand and an understanding of methodological and psychometrical concepts in general. These two aspects will be discussed more in-depth within the discussion of this thesis.

![Figure 6: Illustration of the difference between exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Within EFA the targets are placed, after the shots, in such a way that the targets align with the shots. Following this alignment evaluation of the shots is provided against these targets. Within CFA the targets are placed first and the shots are evaluated against these targets. The squares are indicators classified as measuring anxiety and triangles as measuring depression. The triangle in the anxiety target within the 2nd step is classified as measuring depression because it was implied by the (confirmatory) model.](image-url)
Testing assumptions

The subtle but important difference between $\alpha$-psychometrics and $\beta$-psychometrics enables the justification of, often implicit, but significant assumptions. It can be investigated, for example, if different groups respond in the same manner to an instrument (Borsboom, 2006b). This is important, as an instrument should not be influenced by any additional variables or latent factors (Vandenber, & Lance, 2000; Millsap, 2011). A score on an instrument should, as such, only change if the underlying construct changes (Widaman & Reise, 1997; Meredith, 1993). Findings that mental health prevalence differs according to age groups, should reflect true differences between these age groups and not be an artefact of a different functioning of the measurement instruments between these age groups. If a female and a male subject, who have the exact same level of depression, for example look at the the same inkblot of the Rorschach test, it is to be expected that they show the same outcome. If, however, the female subject sees (again) the sad butterfly but the male subject sees a Velociraptor this makes comparisons between males and females, based on this test, highly invalid. It could still be a valid instrument to measure depression within the male and female population (Vandenber, 2002). Pooling the results disregarding gender, however, would be spurious (Steinkamp & Baumgartner, 1998). The same goes for the HADS in which only the level of depression and anxiety should influence the outcome of the instrument. Figure 7 shows how this can be assessed within $\beta$-psychometrics using CFA. It clearly shows that, in this illustration, the HADS does not measure anxiety and depression invariantly between male and female subjects. It would, therefore, be erroneous to state that the instruments (i.e. rifles) work the same for both genders. Following this, it would also be false to draw conclusions between groups based on these instruments. Within EFA such analyses are not possible as the targets are placed relative to the indicators which hinders direct comparisons of the targets (i.e. constructs).

To make a valid assessment of the stability of an instrument it is, however, also important that its validity is not undermined due to within-group variations over time. As such an instrument could work differently if it is repeatedly measured, jeopardising the longitudinal utility of an instrument. This is a clear example of an implicit assumption that is often made; if an instrument measures something at time $t$ if will also measure this at time $t+1$. Only if this assumption holds it would be worthwhile to investigate the aforementioned stability of an instrument over time (Widaman, Ferrer, & Conger, 2010). This stability, in which the score at $t$ predicts the score at $t+1$, gives insight in the working of the instrument and therefore the measured construct. For some constructs it would be expected that this stability is high (e.g. personality traits) while for others a more moderate stability is theorised and expected (e.g. mood).

To even further disentangle the dynamics of a construct over time, the trait and state component can be separated. By doing so, it increases the understanding of the development of a construct over time. Within the occupational health, for example, this could indicate for example that depression is a stable construct due to the large trait component it entails. If so, this can have implications for interventions and monitoring using on this construct. Hence, it could be that it is not worthwhile to measure this construct with short intervals as it is likely that this results in the same scores. To assess this, an instrument can be disentangled over time into a component that is stable over time (a trait) and a component that varies over time (the occasion) (Prenoveau et al., 2011; Prenoveau, 2016). Often an instrument is
‘contaminated’ over time with the trait of a person inflating the stability of the instrument (Cole, Ciesla, & Steiger, 2007). If this trait is separated from the occasion the ‘actual’ relationship of a construct over time can be estimated. For occupation health, for example, psychological fatigue is likely to be a combination of a more-or-less trait of a person and their current state of affairs. It is important to consider which of these aspects are of interest for the (research) question at hand. To correctly characterise a person for the long-term the trait may be more suitable while during an intervention it may be more suitable to investigate the deviation from this trait.

It is, however, not only possible to formally test whether an instrument differs between groups or over time. CFA has many additional modelling possibilities that provide insight in the structural validity of an instrument. It can be examined, for example, whether there is a second-order factor. Such a factor would explain the anxiety and the depression factor in the HADS example, placing a hypothetical big target behind the two construct specific targets (Figure 8). Other structures could also be investigated including, but not limited to, the role of positive and negative wording on the factor structure. As questions that use the same wording tend to cluster regardless of their content, it is important to investigate this effect and, if present, adjust for this effect. This could be illustrated by placing all positively formulated (e.g. “I look forward with enjoyment to things”) indicators in the top half of the target and all negatively formulated indicators (e.g. “I feel as if I am slowed down”) at the bottom half (Figure 8).
These examples show that \(\beta\)-psychometrics should go hand-in-hand with an active modelling approach which is based on valid hypothesis. These hypotheses should be the basis for the models assessed and for the interpretation of the results. The present thesis will, therefore, investigate a broad range of measurement issue within public mental health. These issues will be addressed with the framework of \(\beta\)-psychometrics as outlined in the previous section. While only a limited number of issues will be addressed, these issues are exemplary and address furthermore important issues within their respective field of application.

**Relevance and potential implications**

A potential straightforward implication of the results of this thesis will be the evaluation of several instruments within the public mental health. Besides the theoretical underpinning of the importance of \(\beta\)-psychometrics it will also be applied. Through this approach \(\beta\)-psychometrics will be illustrated but the results also generate direct output for clinical, scientific, and societal relevance. Analyses can, for example, indicate that specific instruments show measurement variance between groups which could place serious limitations on future, but also prior, research using this instruments within these different groups. This could lead to the re-evaluation, re-designing, or different application of the instrument. This thesis will furthermore result in an increased awareness regarding the issues that could be associated with measurement in the public mental health. As such, even without the application of \(\beta\)-psychometrics, researchers and public mental health professionals should be aware of their explicit and implicit assumptions regarding instruments measuring of mental health.

The added value of \(\beta\)-psychometrics can, however, also be that it tells something about the measured construct. That is, as the measurement instrument assesses the underlying construct, results could therefore also be related to the underlying mechanism of the construct. Some constructs, for example, could even be expected to differ between different subgroups. In this case measurement variance is maybe not as much a measurement artefact but more a
logical result from the underlying construct. Studies regarding the stability of an instrument, for example, will not only cover the instrument at hand but also the underlying construct. Some constructs are likely to be stable over time whilst others are expected to show high levels of variation over time.

Outline of this Thesis

As the goal of this thesis is to identify potential latent defects of measurement instruments and evaluate their impact within the public mental health, some specific potential latent defects are addressed. These will be analysed with an active modelling approach using β-psychometrics. While these results are first and foremost applicable to the measurement instrument and setting which was assessed, but can also be extrapolated to other instruments and settings within the public mental health.

The current thesis starts with the potential latent defect of different settings in which an instrument is assessed. Different settings could potentially result in invariance between settings. In chapter 2 it is examined if the setting influences assessment of psychosocial development within adolescents using a the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1999). Another latent defect that is addressed in this chapter is the influence of the wording of items. As already introduced in the previous section, this could alter the underlying structure of a measurement instrument.

In chapter 3 measurement invariance is examined for two instruments, measuring need for recovery and prolonged fatigue within employees, across three different domains: demographic factors, personal factors, and work-related. Different groups in which a measurement instrument is assessed can, as such, be a latent defect. As already addressed, if the measurement varies between groups, it is difficult – if not impossible – to evaluate whether a difference is the result of a true difference or merely an artefact of measurement.

Chapter 4 examines the assumption of measurement stability over time for the same two instruments that are used in chapter 3. The 'latent defect' time can potentially alter the working of a measurement instrument. Multiple assessments can, for example, result in ‘fatigue’ with respect to the questionnaire. This chapter also shows how this analysis of a latent defect using β-psychometric is a necessity to validly assess the quality of a measurement instrument using α-psychometrics.

In line with the previous chapter it could be that longitudinal measurements, artificially, result in the same scores. Often an instrument is ‘contaminated’ over time with the trait of a person inflating the stability of the instrument (Cole et al., 2007). As already indicated, if this trait is separated from the occasion, the ‘actual’ relationship of a construct over time can be estimated as already outlined in a previous section. To correctly characterise a person for the long-term the trait may be more suitable while during an intervention it may be more suitable to investigate the deviation from this trait. To illustrate this ‘latent defect’ of longitudinal measurement and the evolution of its results, chapter 5 disentangles fatigue within employees into a trait component and a occasion component over time. It is furthermore investigated how this disentanglement relates to the association between fatigue and sickness absence. As such, this relation between the state of fatigue and sickness absence is also disentangled into a trait and occasion component.
All these chapters take groups into account that consists out of two or three subgroups. While this serves the main purpose of this thesis, groups often include a multitude of subgroups. If, for example, subgroups are based on job title the number of unique subgroups can easily surpass 50 within a sample. Within such samples it can however be difficult to find the ‘edge of the roll of tape’ if the model indicates variance between subgroups. A Bayesian framework can be useful in these situations in which a myriad of subgroups is analysed. Assessment of model quality within this framework has received, however, little attention especially within large samples. In chapter 6 a newly defined fit measure is introduced to assess model quality within such large sample model. This fit measure should pave the way for future use of the Bayesian framework for the application of $\beta$-psychometrics.

In these chapters attention will not only be given to the identification of potential latent defects, but also to the quantification of its potential impact. As such, it can be illustrated what the importance is to use $\beta$-psychometrics combined with an active modelling approach in order to validly assess the quality of measurement instruments within the public mental health. Without such an understanding the foundation of public mental health will be (or remains) fragile. As with latent defects within real estate, problems may not seem evident at first but once introduced in the building it can be very persistent and problematic to eliminate them.

References


validity using health-related data. *Annals of Behavioral Medicine, 24*(1), 22–33. doi:10.1207/s15324796abm2401_04

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139–158. doi:10.1177/1094428102005002001


The context dependency of the self-report version of the Strength and Difficulties Questionnaire (SDQ): A cross-sectional study between two administration settings

Hoofs, H., Jansen, N. W. H., Mohren, D. C. L., Jansen, M. W. J., & Kant, IJ.

PLoS ONE, 2015, 10(4), e0120930.
doi:10.1371/journal.pone.0120930
Abstract

Background

The Strength and Difficulties Questionnaire (SDQ) is a screening instrument for psychosocial problems in children and adolescents, which is applied in ‘individual’ and ‘collective’ settings. Assessment in the individual setting is confidential for clinical applications, such as preventive child healthcare, while assessment in the collective setting is anonymous and applied in (epidemiological) research. Due to administration differences between the settings it remains unclear whether results and conclusions actually can be used interchangeably. This study therefore aims to investigate whether the SDQ is invariant across settings.

Methods

Two independent samples were retrieved (mean age = 14.07 years), one from an individual setting ($N = 6,594$) and one from a collective setting ($N = 4,613$). The SDQ was administered in the second year of secondary school in both settings. Samples come from the same socio-geographic population in the Netherlands.

Results

Confirmatory factor analysis showed that the SDQ was measurement invariant/equivalent across settings and gender. On average, children in the individual setting scored lower on total difficulties (mean difference = 2.05) and the psychosocial problems subscales compared to those in the collective setting. This was also reflected in the cut-off points for caseness, defined by the 90th percentiles, which were lower in the individual setting. Using cut-off points from the collective in the individual setting therefore resulted in a small number of cases, 2 to 3%, while $\sim 10\%$ is expected.

Conclusion

The SDQ has the same connotation across the individual and collective setting. The observed structural differences regarding the mean scores, however, undermine the validity of the cross-use of absolute SDQ-scores between these settings. Applying cut-off scores from the collective setting in the individual setting could, therefore, result in invalid conclusions and potential misuse of the instrument. To correctly apply cut-off scores these should be retrieved from the applied setting.
Introduction

Prevalence rates of childhood psychosocial and behavioural problems are high, and if untreated potentially persistent into later life (Moffitt, 1993; Reijneveld, Vogels, Hoekstra, & Crone, 2006). Their consequences, both individual and societal, are among other aspects associated with an increased risk of long-term work disability and psychiatric disorders in young adults (World Health Organization, 2001; Hofstra, 2001). Early detection of children who are at risk for developing psychopathology is therefore crucial, as this enables early social medical counselling which enhances prognosis (Durlak & Wells, 1998; Licence, 2004). To detect these children, valid and reliable screening instruments are crucial.

The Strength and Difficulties Questionnaire (SDQ) is a questionnaire developed to screen for psychopathology in children and adolescents (Goodman, 1997). The widespread use of the SDQ as a screening instrument is attributed to several characteristics, such as its user-friendliness, briefness (25 items), inclusion of positive attribute items, and the availability for different informants (e.g. teacher, parent, youth; Stone, Otten, Engels, Vermulst, & Janssens, 2010; van Roy, Veenstra, & Clench-Aas, 2008; Niclasen et al., 2012; Goodman, Ford, Simmons, Gatward, & Meltzer, 2000).

Due to these characteristics the SDQ is frequently used in child mental health research (Vostanis, 2006). Its application can roughly be divided into two settings, the ‘individual’ and the ‘collective’ setting. The individual setting can be seen as the clinical application of the SDQ (e.g. preventive child healthcare). In such a setting the instrument facilitates the decision making, by a healthcare professional, regarding an individual. In this setting the SDQ is especially suitable for early detection purposes (Vogels, Crone, Hoekstra, & Reijneveld, 2009; van Widenfelt, Goedhart, Treffers, & Goodman, 2003; Goodman, 2001). In contrast, the collective setting is characterised by the aggregation of information across a population. In such a setting the SDQ is used to characterise the overall well-being of populations, for health policy-making and planning, or to compare these populations across countries (van Roy, Groholt, Heyerdahl, & Clench-Aas, 2006; Becker et al., 2006; Ravens-Sieberer, Erhart, Gosch, & Wille, 2008). The collective setting also covers the research regarding the relation of the SDQ with other health-related outcomes or its psychometric properties (van Roy et al., 2008; Giannakopoulos et al., 2009; Sanne, Torsheim, Heiervang, & Stormark, 2009; Hill & Hughes, 2007; Muris, Meesters, & van den Berg, 2003; Calam, Gregg, & Goodman, 2005).

Based on the research of the foregoing decade it seems evident that the SDQ is an outstanding questionnaire for use in both individual and collective settings. However, the interchangeability of scores (e.g. cut-off points) and conclusions between these settings has never been studied so far. If this is hampered it could violate the cross-use of scores and results between settings. The goal of this study is therefore to investigate whether the SDQ is invariant across administration settings. The most prominent aspect that differs between the settings, and therefore possibly violating a valid comparison, is its level of anonymity. Naturally the assessment in an individual setting is not anonymous, as it is otherwise impossible to relate the score to the individual. The term confidential, instead of non-anonymously, is preferred to stress that results are only available for the health professional involved (Nation, 1997). Nevertheless an indicator is needed for the health professional to relate the questionnaire to the individual (e.g. national service number or handing it in in-person). In the collective setting, however, assessment can be completely anonymously (Ong & Weiss, 2000).
Anonymity is often believed to enhance the trustworthiness of answers, as the pressure for social desirable responses is diminished (Lelkes, Krosnick, Marx, Judd, & Park, 2012). This social desirability bias (SDB) refers to the act of misrepresentation of an individual in order to be perceived more positively by others. This may lead to exaggeration of positive behaviour and understating of negative behaviour (Krumpal, 2013). The effect of SDB, however, does not only increase when the condition of assessment becomes less private, but also when questions are more sensitive (Ong & Weiss, 2000; Brown & Vanable, 2009). The influence of gender should also be taken into account, as the effect of anonymous assessment on SDB varies between males and females depending on the topic of the questionnaire (Krumpal, 2013). Studies focusing on adolescents are, however, limited and show diverse results which are modest at best (Durant, Carey, & Schroder, 2002; van de Looij-Jansen, Goldschmeding, & de Wilde, 2006). In all these studies regarding SDB the distinction between the confidential and anonymous group lies mainly in whether the questionnaires are identifiable. In individual settings, questionnaires are however not merely identifiable, they will in fact be identified as elevated scores can elicit possible implications such as a follow-up or referral. Assuming that these possible consequences are known (and unwanted), this could further enhance the SDB in the individual setting in order to remain ‘off the radar’.

Valid conclusions regarding structural differences of the SDQ between settings, due to SDB, are only possible when there is measurement equivalence/invariance (ME/I). ME/I states that an instrument measures the same construct across different populations or settings (Schmitt & Kuljanin, 2008). That is, individuals who have the same standing on the measured construct should receive the same score irrespective of aspects unrelated to this score, leaving aside general inaccuracy (Millsap, 2011; Vandenberg & Lance, 2000). If ME/I does not hold, it is not possible to unambiguously interpret whether structural differences are true differences or merely artefacts of measurement bias (Horn & McArdle, 1992). Or as Vandenberg and Lance (2000) state, in such a case “a group mean comparison may be tantamount to comparing apples and spark plugs (p. 9).” In contrast to ME/I, structural differences reflect variation on the overall construct such as differing mean scores (Schmitt & Kuljanin, 2008). For groups with the same standing on a construct, structural differences should be absent (Meredith, 1993).

For the valid cross-use of the SDQ between collective and individual settings, at least ME/I have to hold. Otherwise the connotation of the SDQ would differ between settings. Extrapolation of conclusions between settings will therefore also be hampered, prohibiting the use of the SDQ in a specific setting based on findings from other settings. When ME/I does hold, but structural mean differences exist, the picture becomes more nuanced. Although in that case the SDQ has the same connotation across settings, these differences violate comparisons regarding absolute scores between collective and individual settings.

Cut-off scores, along with mean scores, might additionally be susceptible for structural differences. In the early publications of the SDQ (Goodman, 1997, 2001; Goodman, Ford, & Meltzer, 2002), using a mixture of settings, it was shown that children falling within the highest 10% of SDQ scores had increased odds of any DSM-IV diagnoses or being a psychiatric case. Subsequently, in collective setting research, this 10% criterion was used to create cut-off scores to identify the high risk group for various nation-wide samples (van Roy et al., 2006; Bourdon, Goodman, Rae, Simpson, & Koretz, 2005; Smedje, Broman, Hetta, & von Knorring, 1999). As it is shown that the discriminatory power of these cut-off scores is satisfactory, one of the purposes of cut-off scores is to facilitate clinical decision making in individual settings
The cut-off scores used in individual settings are often adopted from collective settings as normative data usually stems from these settings (Koskelainen, Sourander, & Kaljonen, 2000). As cut-off scores are based on the same underlying distribution as mean scores they would also suffer from structural differences between settings. Therefore the legitimacy of transferring cut-off scores between settings is questionable when structural differences are present. As this distortion can have massive implications for actual practice, since the interpretation of 'a case' would be biased, the legitimacy of this procedure will receive special attention in this paper. The main question of this article therefore is:

- Is the self-report version of the SDQ invariant whether it is administered in an individual or collective setting?

This main question can be divided into the following sub-questions:

- Does ME/I hold across settings?
- Are there structural differences across settings?
- What is the influence of gender on ME/I and structural differences between and within settings?
- How does possible invariance between settings influence the validity of the cross-use of cut-off points?

**Methods**

**Data collection**

Data from an individual and a collective sample were included to analyse the impact these different settings may have on the SDQ. Data for both settings was collected in the southern region of Limburg, the Netherlands. In contrast to other regions in the Netherlands, the area is characterised by: Its variation of rural and urban areas; a shrinking and ageing population; a relatively low social economic status (SES); and, a small proportion of people with a non-western ethnic background (3%) and relative high proportion with a, non-Dutch, western ethnic background (16%) (Lamberts et al., 2011). Data was collected by the Public Health Service of South-Limburg. Their services include, among others, child immunisation, preventive child healthcare (PCH) programs, and providing municipalities with epidemiological data for policy purposes.

**Individual setting**

For the individual setting, data from the PCH program is used. All children who live in the Netherlands are automatically included in this program. The PCH is regulated through national regulation and resources, while the financing takes place by local governments without any parental costs (Kuo et al., 2006).

Information from the screening assessment during the second year of secondary school is used in the current study. This is the final of 14 screening assessments spread across different stages in the child’s development, with more frequent screening assessments early in life. During the
screening (20 to 30 minutes) a specially trained youth health care physician assesses the psychosocial and physical health of the child, but also pays attention to the socio emotional development of the child or other issues related to the well-being of the child. If needed the youth health care physician can advise a possible follow-up or referral to secondary or tertiary care. During the screening it is optional for a parent or guardian to accompany the child. The procedure has an opt-out approach, for which the responsibility lies with the parent(s) or guardians(s) of the child and the child themselves. Only information from screenings in regular secondary schools was used, excluding special needs education schools. There were no additional inclusion or exclusion criteria.

Approximately two weeks prior to the screening a questionnaire was sent to the child. This paper-and-pencil self-administered questionnaire (SAQ), used to facilitate the screening, consisted of questions regarding medical history, psychosocial development (using the SDQ), and substance use. The questionnaire was filled in at home and handed in during the screening assessment, in which the youth health care physician reviewed the questions and computed the total difficulties score for the SDQ. Eventually all questionnaires were scanned and processed to obtain a database for each scholastic year. Screening took place at, and during, school. For logistical reasons children were invited and scheduled school-wise. The screenings took place during the whole school year (median = March). For the current study the data was anonymised. Data from the two most recent years of data gathering were included. This resulted in the inclusion of the school years 2010/11 and 2011/12. Circa 90% of the eligible children were present during this screening (Lamberts et al., 2011). The sample included 6,594 available questionnaires. The mean age of the children was 14.26 years ($SD = 0.62$ years), with 51.30% of the sample being female.

**Collective setting**

For the collective setting information from the National Youth Monitor was used. This monitor was used to inform policymakers and researchers about the situation of the youth in the Netherlands covering several domains, such as health and welfare (Statistics Netherlands, 2009). The SDQ was included as indicator of psychosocial well-being of the youth. For the present study questionnaires from 2009 were used.

Administration took place in regular classes in the second year of secondary school. The intended mode of assessment was a computerised self-administered questionnaire (CSAQ). Due to technical issues this was not always possible, which led to a minority of cases which were given a SAQ (17%). Both modes were included in the sample. Administration took place in the last quarter of 2009 for all children (median = November). The procedure has an opt-in approach, in which the parent(s) or guardians(s) of the child had to give informed consent. Only information from regular secondary schools was used, excluding special needs education schools. There were no additional inclusion or exclusion criteria.

The response rate for the collective setting was approximately 70%. This resulted in 4,613 available questionnaires. The mean age of the children was 13.80 years ($SD = 0.57$ years), with 50.01% of the sample being female.
Ethical statement

Research in this article involves participants who were part of the usual governmental (preventive) healthcare program attending regular secondary schools. Two methods were used to collect the data. The first method involved the retrospective patient files (individual setting). Information was anonymised and de-identified by the Public Health Service of South-Limburg before it was received and reviewed by the authors. Retrospective research/research with patient files does not fall under the scope of the Human Subjects Act (Wet medisch-wetenschappelijk onderzoek met mensen; WMO) as the research subject is not physically involved in the research. The second method involved the use of anonymous questionnaire data (collective setting), with informed consent given by the caregiver of the child. This does not require WMO approval since does not utilize any invasive techniques, substance administration or psychological manipulations. Since both methods of data collection do not require any WMO approval by the Dutch law, the approval of an ethical commission is not needed and therefore the CCMO was not contacted for a waiver or approval (CCMO, 2014). None of the authors played any part or had any interaction with the human participants in the two methods of data collection. Both strategies were furthermore registered at the Personal Data Protection Act (Wet bescherming persoonsgegevens, Wbp) (CCMO, 2014).

Measures

SDQ. The SDQ comprises 25 items spread equally across 5 subscales; emotional symptoms, conduct problems, hyperactivity/inattention, peer problems, and pro-social behaviour (Goodman, Lamping, & Ploubid, 2010; Goodman, 1999, 1997). Each item gives a statement regarding negative or positive behaviour (e.g. “I worry a lot”) which can be scored on a three point Likert scale (“not true”, “somewhat true”, or “certainly true”). Five items were positively worded and were therefore reverse coded. Besides a score on each subscale, ranging from 0 to 10, a total difficulties score is also computed, ranging from 0 to 40, by summing the scores from the emotional symptoms, conduct problems, hyperactivity/inattention and peer problems subscales. Higher scores indicated elevated problems on the total difficulties score and all subscales, except for the pro-social behaviour subscale where higher scores indicated higher levels of pro-social behaviour. Further information regarding the computation of the SDQ can be found on www.sdqinfo.org. For the current study the Dutch self-report version was used (van Widenfelt et al., 2003).

Demographic characteristics. Age, gender, and home situation were retrieved from the questionnaires administered in the individual and collective setting. Social economic status (SES) and domestic area were derived from postal codes which were present in the questionnaires in both settings. SES was computed using a national database which provided a SES score for each postal code area (ranging from -7.3 to 3.2). These scores were based on a combination of several aspects such as education, income and employment status. As these scores represent averages they should not be seen as individual indicators, as they merely serve descriptive purposes (Knol, 2012). Domestic area was defined using the classification of Statistics Netherlands. Areas with more than 1,000 addresses per square kilometre, were classified as urban while areas with a lower density were classified as rural (Statistics Netherlands, 2009). Ethnic background was also defined in accordance with the classification system of Statistics Netherlands, in which both parents had to be born in the Netherlands to be classified...
as Dutch. If both the mother and father were born in a foreign country, the country of birth of the mother determined the nature of the ethnic background (western vs. non-western). If only one of the two parents was born in a foreign country this parent determined the nature of the ethnic background (Statistics Netherlands, 2014).

**Statistical analyses**

Cronbach’s $\alpha$ and greatest lower bound ($glb$) were calculated as reliability coefficients. Although $glb$ seems most appropriate, Cronbach’s $\alpha$ was also reported to enhance comparability with previous studies. Since the $glb$ was always larger than the Cronbach’s $\alpha$ of a scale, they were reported as a range (Sijtsma, 2008).

Confirmatory Factor Analysis (CFA) was used to test whether the SDQ was invariant across settings. The original hypothesised 5-factor model proposed by Goodman (2001) was used including correlations between all factors. As all items were rated on a 3-point Likert scale they were handled as categorical data. The weighted least square means and variance adjusted (WLSMV) estimator was selected. This estimator has proven to be preferred over maximum likelihood for a broad range of sample sizes when handling ordered, especially asymmetrical, categorical data with few categories (Beauducel & Herzberg, 2006; Flora & Curran, 2004; Rhemtulla, Brousseau-Liard, & Savalei, 2012). To evaluate model fit the Comparative Fit Index ($CFI$), Tucker-Lewis Index ($TLI$), and Root Mean Square Error of Approximation ($RMSEA$) were used as approximate fit indexes (AFIs). Greater $CFI$ and $TLI$ values and smaller $RMSEA$ values indicate better fit. For the $CFI$ and $TLI$ a cut-off point of .95 is used to indicate good model fit, while values greater than .90 are often used to indicate acceptable model fit. For the $RMSEA$ values smaller than 0.06 indicate good fit (Schreiber, Nora, Stage, Barlow, & King, 2006; Kline, 2011). Given the large sample size in the current study the $\chi^2$ was not interpreted as it was shown to be highly sensitive for sample size, but was reported for sake of completeness (Meade, Johnson, & Braddy, 2008; Cheung & Rensvold, 2002; Kline, 2011).

The first step consisted of testing the overall model fit. Therefore, the hypothesized CFA model was fitted to the combined data of the two settings. This step was performed to assure that, on average, this model reflected the data properly. The second step consisted of ME/I testing, regarding gender, within each setting separately. This step also enabled the free estimation of mean scores between gender. To check whether the model fitted adequately for each gender, models were initially fitted separately for males and females (Meade et al., 2008). Thereafter, four sequential models were conducted to test increasingly restrictive forms of ME/I (Millsap, 2011): (a) Configural (Model 1); all parameters were allowed to be free across groups. (b) Metric (Model 2); factor loadings were constrained to be equal across groups. (c) Scalar (Model 3); factor loadings and thresholds were constrained to be equal across groups. (d) Strict (Model 4); factor loadings, thresholds and unique factor variances (UFVs) were constrained to be 1 in both groups. For identification purposes the means in Model 2, 3, and 4 of the factors in the first group were constrained to be 0, while they were freely estimated in the second group. In Model 2 and 3 the UFVs were furthermore only freely estimated for the second group, while they were constrained to be 1 in the first group. Additionally Model 2 had the first threshold of each item constrained to be equal across groups, as were all thresholds for the items which were used to set scale for each factor (Millsap & Yun-Tein, 2004; Muthen & Asparouhov, 2012).
The third step consisted of ME/I testing between the collective and individual setting. This was done analogous to the second step. Model 1 of the third step therefore included the two final models of the second step which were fitted simultaneously to serve as a test configural ME/I. Model 2 constrained, in addition to the constraints already present within settings between genders, factor loadings to be equal across settings. Model 3 additionally constrained the threshold to be equal across settings, while Model 4 also constrained the UFVs to be equal across settings. In addition to these models, which tested ME/I, two additional models were conducted to investigate structural differences between settings. Model 5, therefore, constrained the variances of the same factor to be equal across all four groups (male-individual, male-collective, female-individual, and female-collective). Model 6 constrained means of the same factor to be equal across settings for the same gender (male-individual = male-collective, and female-individual = female-collective).

As all models in each step were nested, it enabled testing whether models differed from each other. If models do not differ from each other the more parsimonious – more strict form of ME/I – should be preferred (van de Schoot, Lugtig, & Hox, 2012). As \( \chi^2 \), \( \Delta \chi^2 \) was also highly sensitive for sample size and therefore not plausible to interpret for the evaluation of ME/I (Meade et al., 2008; Cheung & Rensvold, 2002). Instead, \( \Delta \)AFIs were used to determine ME/I. However, as \( \Delta \)AFIs have an unknown sampling distribution it is not possible to test whether differences in AFIs are significant (Sass, Schmitt, & Marsh, 2014). Several articles, using different caveats, therefore investigated the performance of \( \Delta \)AFIs to detect ME/I. In the current study two guidelines were used to assess the presence of ME/I, as articles proposed substantially different \( \Delta \)AFIs cut-off points. The conventional criteria by Chen (2007) stated that \( \Delta CFI \leq -0.01 \) and \( \Delta RMSEA \geq 0.01 \) indicate substantial differences between models, therefore not supporting ME/I. Meade et al. (2008) recommended more conservative cut-off points; \( \Delta CFI \leq -0.002 \) and \( \Delta RMSEA \geq 0.007 \). In accordance with these conservative cut-off points Marsh et al. (2010) advocated the use of \( \Delta TLI = 0 \) as an additional cut-off point to evaluate the presence of ME/I (Sass et al., 2014). As \( \Delta \)AFIs can give mixed results, depend on several aspects, should not be used as clear cut-off points, and guidelines differ greatly, ME/I is almost never a clear cut case, prohibiting irrefutable claims (Sass et al., 2014; Marsh, Hau, & Wen, 2004). It seems therefore best to facilitate critical judgement regarding ME/I by presenting the used guidelines, notwithstanding interpretation by the authors for the sake of continuation in the article (Meade et al., 2008).

As there were two modes of assessment (SAQ vs. CSAQ) in the collective setting, it was assessed whether this had any impact on the SDQ scores. A CFA model for the collective setting was therefore conducted including mode of assessment as a covariate for all factors. School level, dichotomised as preparatory vocational secondary education or lower (VMBO or lower) and senior general secondary education or pre-university education (HAVO or VWO), which was known for this setting, correlated strongly with the mode of assessment and was therefore also included as a covariate. The presence of a potential design effect for the collective setting, due to the class-wise assessment, was also assessed. A multilevel model was therefore conducted enabling the estimation of intraclass correlations depicting the level of variance on the class level in relation to the overall variation (Hox, 2010).

To investigate the direction and magnitude of mean differences for the SDQ scores between settings and gender, an ANOVA was performed. Additionally an ANCOVA was also performed (controlling for age and SES). However, only results, from the ANOVA are presented here, but as results only marginally differed, the results from the ANOVA were presented, since:
(1) Adjustment for age and SES increased the effect, although marginally, of the setting on the outcome variables. The more conservative (crude) effect was therefore favoured; (2) Crude SDQ scores are often used to compare different study populations or to define cut-off points and are therefore more valuable to report; (3) Correcting for age could be considered ambiguous as the focus was on the grade of the children (second grade), instead of their age. Percentile scores (80% & 90%) were computed for each setting to give an indication of cut-off scores. Corresponding percentages were furthermore given for these cut-off scores. As a last step the cut-off points from the collective setting were used to categorise the individual sample, in order to investigate the effect of transferring cut-off scores between settings. Analyses were performed using R (version 3.1.1), in combination with Mplus (version 6.11) for the CFA.

Results

Missing data

In accordance with Niclasen et al. (2012) cases with a total of more than one missing value on all SDQ items were excluded (individual setting: $n = 170$, 2.58%; collective setting: $n = 50$, 1.08%). This handling of missing data was furthermore supported by the notion of Kline (2011) that less than 5% of missing data on a single variable is of little concern (0.63–2.31%). The majority of the excluded cases had missings on all items (individual setting: 67.06%, collective setting: 54.00%). It was therefore not appropriate to conduct missing data analysis, as for the majority of the excluded cases background characteristic were also missing. Cases were furthermore deleted when information on gender was not known (individual setting: $n = 2$, 0.03%; collective setting: $n = 3$, 0.07%).

Background characteristics

Table 1 shows the descriptive statistics for the two settings. There were some small differences regarding the ethnic background, home situation, and domestic area between the settings (Table 1). The mean age was higher in the individual setting compared to the collective setting (Cohen’s $d = .76$). The average SES was furthermore somewhat higher in the individual setting compared to the collective setting (Cohen’s $d = .14$).

Reliability

Cronbach’s $\alpha$ and $glb$ were calculated for each subscale and the total difficulties scale. Both the total difficulties scale (Cronbach’s $\alpha$; $glb = .78; .85$) and hyperactivity/inattention subscale (.76; .81) showed high estimates. The pro-social behaviour (.65; .67) and emotional symptoms subscales (.69; .71) showed moderate estimates, while reliability coefficients for the conduct problems (.55; .56) and peer problems (.51; .55) subscale were low. Reliability coefficients were also calculated for each setting, separately for males and females (Table 2). On average, reliability coefficients were equal across setting and gender and along the lines of the overall estimates.
Table 1: Background characteristics

<table>
<thead>
<tr>
<th>Setting</th>
<th>Individual</th>
<th>Collective</th>
<th>Difference Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>6,422</td>
<td>4,560</td>
<td></td>
</tr>
<tr>
<td>Gender (%):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>48.72</td>
<td>49.78</td>
<td>$\chi^2(1) = 1.15, p = .28$</td>
</tr>
<tr>
<td>Female</td>
<td>51.28</td>
<td>50.22</td>
<td></td>
</tr>
<tr>
<td>Age Mean (SD):</td>
<td>14.26 (0.62)</td>
<td>13.81 (0.57)</td>
<td>$t(10978) = 39.36, p &lt; .01$</td>
</tr>
<tr>
<td>Ethnic Background (%):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>84.57</td>
<td>81.51</td>
<td>$\chi^2(2) = 18.80, p &lt; .01$</td>
</tr>
<tr>
<td>Western</td>
<td>8.54</td>
<td>10.64</td>
<td></td>
</tr>
<tr>
<td>Non-Western</td>
<td>6.89</td>
<td>7.86</td>
<td></td>
</tr>
<tr>
<td>SES Mean (SD):</td>
<td>-0.33 (1.13)</td>
<td>-0.49 (1.13)</td>
<td>$t(10953) = 7.40, p &lt; .01$</td>
</tr>
<tr>
<td>Home situation (%):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 biological parents</td>
<td>77.95</td>
<td>75.48</td>
<td>$\chi^2(2) = 66.86, p &lt; .01$</td>
</tr>
<tr>
<td>1 biological parent</td>
<td>20.74</td>
<td>20.84</td>
<td></td>
</tr>
<tr>
<td>Other/unknown</td>
<td>1.31</td>
<td>3.68</td>
<td></td>
</tr>
<tr>
<td>Domestic area (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rural</td>
<td>51.70</td>
<td>47.66</td>
<td>$\chi^2(1) = 13.74, p &lt; .01$</td>
</tr>
<tr>
<td>Urban</td>
<td>48.30</td>
<td>52.34</td>
<td></td>
</tr>
</tbody>
</table>

CFA

CFA were conducted to test for ME/I. The first step consisted of testing the overall model fit for the two settings combined. All factor loadings from the original 5-factor were significant ($p < .05$) and positive, but the model had a mediocre fit, $\chi^2(265) = 8,968.77, p < .001, CFI = .8827, TLI = .8672, RMSEA = 0.0547$. In accordance with van Roy et al. (2008) an extra factor was therefore added comprising the five positively worded items. This resulted in a better fitting model, $\Delta \chi^2(10) = 1,815.93, p < .001, \Delta CFI = -.0513, \Delta TLI = -0.0551, \Delta RMSEA = 0.0129$, with an overall satisfactory fit, $\chi^2(255) = 5,153.24, p < .001, CFI = .9340, TLI = .9223, RMSEA = 0.0418$. All factor loadings were significant ($p < .05$) and in the expected, positive, direction. The positive construal factor was allowed to correlate with all other factors. For the rest of the analysis this adjusted model was used.

In the second step the model was fitted separately for each setting, to test for ME/I regarding gender within settings. For the individual setting the model had a satisfactory fit for females and males. Factor loadings of these models are given in Table 2. The factor loadings for the obedient item on the conduct factor and the good friend item on the positive construal factor were non-significant, for both genders, while all other loadings were significant ($p < .05$) and positive. Model 1 estimated these two models conjointly, which showed that configural ME/I hold as model fit was satisfactory to good (Table 3). Metric ME/I was also supported, as constraining the factor loadings did not surpass any $\Delta$AFIs criteria. Model 3, constraining the thresholds of the same items to be equal, only marginally surpassed the conservative
Table 2: Internal consistency of the SDQ scales and standardized factor loadings of the CFA for the 5-factor model

<table>
<thead>
<tr>
<th>Scale/Factor</th>
<th>Individual setting</th>
<th></th>
<th>Collective setting</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td><strong>Total difficulties</strong></td>
<td>(α; glb)</td>
<td></td>
<td>(α; glb)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.75; .84</td>
<td>.78; .86</td>
<td>.77; .85</td>
<td>.76; .85</td>
</tr>
<tr>
<td><strong>Emotional symptoms</strong></td>
<td>(α; glb)</td>
<td></td>
<td>(α; glb)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.59; .61</td>
<td>.68; .72</td>
<td>.62; .64</td>
<td>.68; .72</td>
</tr>
<tr>
<td>Somatic</td>
<td>.44</td>
<td>.54</td>
<td>.52</td>
<td>.54</td>
</tr>
<tr>
<td>Worries</td>
<td>.62</td>
<td>.67</td>
<td>.59</td>
<td>.60</td>
</tr>
<tr>
<td>Unhappy</td>
<td>.78</td>
<td>.82</td>
<td>.80</td>
<td>.83</td>
</tr>
<tr>
<td>Clingy</td>
<td>.66</td>
<td>.61</td>
<td>.56</td>
<td>.59</td>
</tr>
<tr>
<td></td>
<td>.67</td>
<td>.66</td>
<td>.72</td>
<td>.64</td>
</tr>
<tr>
<td><strong>Conduct problems</strong></td>
<td>(α; glb)</td>
<td></td>
<td>(α; glb)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.50; .53</td>
<td>.49; .50</td>
<td>.57; .59</td>
<td>.53; .54</td>
</tr>
<tr>
<td>Tempers</td>
<td>.65</td>
<td>.67</td>
<td>.69</td>
<td>.66</td>
</tr>
<tr>
<td>Obedient</td>
<td>-.08†</td>
<td>-.03†</td>
<td>.21</td>
<td>-.01†</td>
</tr>
<tr>
<td>Fights</td>
<td>.55</td>
<td>.56</td>
<td>.66</td>
<td>.65</td>
</tr>
<tr>
<td>Lies</td>
<td>.72</td>
<td>.75</td>
<td>.70</td>
<td>.73</td>
</tr>
<tr>
<td>Steals</td>
<td>.54</td>
<td>.49</td>
<td>.66</td>
<td>.53</td>
</tr>
<tr>
<td><strong>Hyperactivity</strong></td>
<td>(α; glb)</td>
<td></td>
<td>(α; glb)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.77; .81</td>
<td>.77; .82</td>
<td>.71; .78</td>
<td>.75; .81</td>
</tr>
<tr>
<td>Restless</td>
<td>.78</td>
<td>.72</td>
<td>.75</td>
<td>.76</td>
</tr>
<tr>
<td>Fidgety</td>
<td>.73</td>
<td>.72</td>
<td>.76</td>
<td>.72</td>
</tr>
<tr>
<td>Distractible</td>
<td>.83</td>
<td>.87</td>
<td>.76</td>
<td>.82</td>
</tr>
<tr>
<td>Think before acting</td>
<td>.31</td>
<td>.27</td>
<td>.51</td>
<td>.36</td>
</tr>
<tr>
<td>Persistent</td>
<td>.47</td>
<td>.61</td>
<td>.54</td>
<td>.52</td>
</tr>
<tr>
<td><strong>Peer problems</strong></td>
<td>(α; glb)</td>
<td></td>
<td>(α; glb)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.51; .55</td>
<td>.52; .57</td>
<td>.51; .56</td>
<td>.46; .50</td>
</tr>
<tr>
<td>Solitary</td>
<td>.59</td>
<td>.56</td>
<td>.51</td>
<td>.50</td>
</tr>
<tr>
<td>Good friend</td>
<td>.40</td>
<td>.36</td>
<td>.47</td>
<td>.20</td>
</tr>
<tr>
<td>Popular</td>
<td>.50</td>
<td>.56</td>
<td>.65</td>
<td>.47</td>
</tr>
<tr>
<td>Bullied</td>
<td>.77</td>
<td>.82</td>
<td>.83</td>
<td>.76</td>
</tr>
<tr>
<td>Adults</td>
<td>.55</td>
<td>.56</td>
<td>.59</td>
<td>.56</td>
</tr>
<tr>
<td><strong>Prosocial behaviour</strong></td>
<td>(α; glb)</td>
<td></td>
<td>(α; glb)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.59; .61</td>
<td>.56; .58</td>
<td>.65; .68</td>
<td>.64; .67</td>
</tr>
<tr>
<td>Considerate</td>
<td>.81</td>
<td>.82</td>
<td>.70</td>
<td>.79</td>
</tr>
<tr>
<td>Shares</td>
<td>.38</td>
<td>.43</td>
<td>.43</td>
<td>.42</td>
</tr>
<tr>
<td>Caring</td>
<td>.57</td>
<td>.57</td>
<td>.65</td>
<td>.63</td>
</tr>
<tr>
<td>Kinds to kids</td>
<td>.61</td>
<td>.63</td>
<td>.67</td>
<td>.69</td>
</tr>
<tr>
<td>Often volunteers to help</td>
<td>.57</td>
<td>.62</td>
<td>.58</td>
<td>.61</td>
</tr>
</tbody>
</table>

Note. First item of each factor was used as indicator; † indicator also loaded on the positive construal factor. † not significant at α of .05.

criteria regarding the difference in CFI and TLI. The ΔRMSEA was however within these conservative limits. As all conventional criteria were furthermore met, there was no evidence for a strong violation of scalar ME/I. Strict ME/I, constraining the UFVs to be equal across groups, was also supported for the individual setting.

For the collective setting separate model testing for males and females resulted in satisfactory fitting models. Factor loadings of these models are given in Table 2. Only the factor loading for the obedient item on the conduct factor for females was non-significant, while all other loadings were significant (p < .05) and positive. Joint estimation of these models supported
Table 3: Model fit and nested model comparisons for multiple-group CFA analyses

<table>
<thead>
<tr>
<th>Model (Constraints)</th>
<th>Model fit indices</th>
<th>Nested model comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$ (df)</td>
<td>CFI</td>
</tr>
<tr>
<td><strong>Individual setting</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male ($n = 3,129$)</td>
<td>1,349.51 (255)</td>
<td>.939</td>
</tr>
<tr>
<td>Female ($n = 3,293$)</td>
<td>1,698.97 (255)</td>
<td>.934</td>
</tr>
<tr>
<td>1. All parameters free</td>
<td>3,046.03 (510)</td>
<td>.936</td>
</tr>
<tr>
<td>2. M.1 + Loadings</td>
<td>3,009.20 (533)</td>
<td>.938</td>
</tr>
<tr>
<td>3. M.2 + Thresholds</td>
<td>3,132.62 (553)</td>
<td>.935</td>
</tr>
<tr>
<td>4. M.3 + UFVs</td>
<td>3,187.98 (578)</td>
<td>.934</td>
</tr>
<tr>
<td><strong>Collective setting</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male ($n = 2,270$)</td>
<td>1,316.38 (255)</td>
<td>.921</td>
</tr>
<tr>
<td>Female ($n = 2,290$)</td>
<td>1,465.97 (255)</td>
<td>.916</td>
</tr>
<tr>
<td>1. All parameters free</td>
<td>2,783.67 (510)</td>
<td>.918</td>
</tr>
<tr>
<td>2. M.1 + Loadings</td>
<td>2,797.13 (533)</td>
<td>.918</td>
</tr>
<tr>
<td>3. M.2 + Thresholds</td>
<td>2,876.39 (553)</td>
<td>.916</td>
</tr>
<tr>
<td>4. M.3 + UFVs</td>
<td>2,968.76 (578)</td>
<td>.914</td>
</tr>
<tr>
<td><strong>Settings combined (Individual + Collective)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. All parameters free</td>
<td>6,157.52 (1156)</td>
<td>.926</td>
</tr>
<tr>
<td>2. M.1 + Loadings</td>
<td>6,341.22 (1179)</td>
<td>.923</td>
</tr>
<tr>
<td>3. M.2 + Thresholds</td>
<td>6,489.06 (1199)</td>
<td>.921</td>
</tr>
<tr>
<td>4. M.3 + UFVs</td>
<td>6,790.19 (1224)</td>
<td>.917</td>
</tr>
<tr>
<td>5. M.4 + F. variances</td>
<td>6,715.55 (1242)</td>
<td>.919</td>
</tr>
<tr>
<td>6. M.5 + F. means</td>
<td>10,348.80 (1254)</td>
<td>.865</td>
</tr>
</tbody>
</table>

*Note. df = degrees of freedom; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; RMSEA = Root Mean Square Error of Approximation; UFVs = unique factor variances. *significant at $\alpha$ of .05.
the presence of configural ME/I, as the model fit of Model 1 was satisfactory. Constraining the factor loadings in Model 2 showed the presence of metric ME/I, as all ΔAFIs were within the conservative criteria. Scalar ME/I also seemed to hold for the collective setting, as constraining the threshold to be equal across gender resulted in an acceptable change of model fit. Only the ΔCFI marginally surpassed the conservative criteria. This was also found when the UFVs were constrained to be 1 in both groups, supporting strict ME/I (Table 3). As the CFA models of both settings showed good fit and supported ME/I, regarding gender, it was justified to proceed to the third step in which ME/I between settings was tested.

In the third step the collective setting and individuals setting models were jointly estimated to assess ME/I. Model 1 showed satisfactory fit supporting configural ME/I. Metric ME/I was supported according to the conventional criteria and the conservative ΔRMSEA criteria, this indicated that factor loadings were similar across settings. The test for scalar ME/I resulted in the same conclusion, with also the ΔCFI falling within the conservative criteria. This indicated that thresholds of the SDQ items were equal across the collective and individual setting. Constraining the UFVs to be equal across settings resulted in a change of model fit surpassing the conservative criteria regarding the ΔCFI and ΔTLI. Changes were, however, within the conventional criteria and the conservative ΔRMSEA criteria indicating that there was no strong violation of strict ME/I (Table 3). As metric and scalar ME/I was supported, testing structural differences between settings was legitimate.

To test for structural differences, additional constraints were added to Model 4 of the combined estimation (Table 3). First, variances of the same factors were set to be equal across both genders for both settings. This resulted in a model change falling within the conservative criteria, supporting that there were no differences in the variances of any factor across settings or gender. In the last model, factor means were additionally constrained to be equal for the same gender across settings. This implied that the average scores of the same gender across settings were approximately equal for all SDQ factors. The decrease in model fit implied, however, the presence of structural differences between settings regarding mean scores.

To test for a potential mode effect, SAQ vs. CSAQ, a CFA model for the collective setting, female and male combined, was conducted with mode of assessment and school level as covariates of all factors. This resulted in a good fitting model, $\chi^2(293) = 2,620.90, p < .001, CFI = .9180, TLI = .9020, RMSEA = 0.0420$, with no statistically significant association between mode of assessment and any SDQ-factor ($p < .05$). The presence of a design effect, through classes, in the collective setting was assessed with a multilevel model. This multilevel model showed a good model fit $\chi^2(510) = 4,248.96, p < .001, CFI = .9260, TLI = .9130, RMSEA = 0.0410$. Intraclass correlations for all factors (range = .02; .08) and items (range = .00; .07) were low, indicating that no more than 7.81% of the variance in any factor was due to the class-wise assessment.

**Mean scores**

To investigate the structural mean differences of the subscales, and the total difficulties score, an ANOVA was performed. Table 4 shows the results of this analysis. Setting was significant ($p < .05$) for all scale scores, with a small association for hyperactivity and a more moderate effect for pro-social behaviour, conduct problems and the total difficulties score. Scores from the collective setting were higher than the scores from the individual setting, except for the pro-social behaviour subscale. However as this subscale was an index of positive behaviour,
Table 4: Comparisons for SDQ total difficulties and subscales mean scores for each setting and gender

<table>
<thead>
<tr>
<th>Setting; mean (SD)</th>
<th>Setting</th>
<th>Gender</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Individual</td>
<td>Collective</td>
<td></td>
</tr>
<tr>
<td>Total difficulties score</td>
<td></td>
<td></td>
<td>.04*</td>
</tr>
<tr>
<td>Male</td>
<td>7.31a (4.49)</td>
<td>8.96b (5.12)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>7.81c (4.83)</td>
<td>10.27d (5.20)</td>
<td></td>
</tr>
<tr>
<td>Emotional symptoms</td>
<td></td>
<td></td>
<td>.02*</td>
</tr>
<tr>
<td>Male</td>
<td>1.42e (1.54)</td>
<td>1.66f (1.77)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2.40g (2.05)</td>
<td>3.12h (2.29)</td>
<td></td>
</tr>
<tr>
<td>Conduct problems</td>
<td></td>
<td></td>
<td>.05*</td>
</tr>
<tr>
<td>Male</td>
<td>1.27 (1.26)</td>
<td>1.89 (1.65)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1.09 (1.14)</td>
<td>1.70 (1.44)</td>
<td></td>
</tr>
<tr>
<td>Hyperactivity</td>
<td></td>
<td></td>
<td>.01*</td>
</tr>
<tr>
<td>Male</td>
<td>3.58i (2.46)</td>
<td>3.93j (2.43)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>3.35k (2.38)</td>
<td>4.08l (2.45)</td>
<td></td>
</tr>
<tr>
<td>Peer problems</td>
<td></td>
<td></td>
<td>.02*</td>
</tr>
<tr>
<td>Male</td>
<td>1.04 (1.33)</td>
<td>1.47 (1.60)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.97 (1.30)</td>
<td>1.37 (1.48)</td>
<td></td>
</tr>
<tr>
<td>Pro-social behaviour</td>
<td></td>
<td></td>
<td>.05*</td>
</tr>
<tr>
<td>Male</td>
<td>7.92m (1.62)</td>
<td>7.02n (2.00)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>8.67o (1.34)</td>
<td>7.95p (1.73)</td>
<td></td>
</tr>
</tbody>
</table>

Note. If interaction effect is significant different subscripts indicate significant differences between groups. *significant at α of .05.

the direction of the effect of setting was the same for each subscale and the total difficulties score.

Gender was significantly (p < .05) associated with all scale scores except hyperactivity. Interaction effects were small, but nonetheless significant (p < .05) for emotional symptoms, hyperactivity, pro-social behaviour and total difficulties score. Inspection of Table 4 shows that if an interaction effect was present, females showed larger differences between settings than males.

The ANCOVA included age and SES, which strongly relates with domestic area and home situation, as covariates. The overall effect of age (in years) on the total difficulties score was b = 0.68 (p < .01). The individual setting showed a somewhat smaller effect, b = 0.48 (p < .01), while the effect in the collective setting was higher, b = 1.01 (p < .01). The overall effect of SES was not significant at b = −0.04 (p = .29), which remained when analysed separately for both groups (individual setting, b = 0.08 [p = .08], collective setting, b = −0.09 [p = .84]). Subsequently, an ANCOVA was performed, with age and SES as covariates, using the overall effects as correction. The effect size of the setting only marginally increased while there was no change in the effect size of gender and the interaction term (gender × setting). The effects of the covariates on the total difficulties score were generalisable to all subscales, except the prosocial subscale which showed the inverse pattern. As indicated in the methods section the results of the ANOVA were favoured above those of the ANCOVA.
Cut-off points

To further illustrate the impact of the structural differences between the settings, cut-off points were calculated for the current samples, separately for each setting. As shown in Table 5 all cut-off points were lower in the individual setting as compared to the collective setting. Differences between females and males were in accordance with the differences in mean scores as found in the ANOVA. There were for instance more females with a borderline or abnormal score regarding hyperactivity in the individual setting, while in the collective setting these cases were more prevalent among males.

As in daily practice cut-off scores from collective settings are often applied in individual settings, this process was simulated using the current data. The sample from the individual setting was therefore differentiated in abnormal, border, and normal cases using the cut-off scores from the collective setting (Table 5). This resulted in much lower prevalence rates of abnormal case on all scales, except hyperactivity. On the total difficulties score for instance only 8% of the children were classified as a borderline or abnormal case, instead of the original 18%. If only abnormal cases were selected this percentage even drops to 3% instead of the original 8%.

Discussion

Using an observational approach, this study investigated whether assessment of the self-report version of the SDQ is sensitive to its administration setting. Using two large samples results show that, although the SDQ has the same connotation across both settings, there are significant structural differences between the collective and individual setting.

Structural differences complicate the cross-use of cut-off points between settings. In the individual setting these cut-off points facilitate the screening purpose of the SDQ to differentiate between children with and without an elevated risk for developing psychopathology. Cut-off points in individual settings are often retrieved from validation studies, which are typically conducted in collective settings. There are, however, structural differences between the mean scores of the subscales and total difficulties score of the SDQ while ME/I and equal factor variances hold. These three findings combined imply that the sample distribution of the SDQ has the same form and represents the same construct for both settings, but shifts, as a whole, between the two settings (Figure 1). The cross-use of cut-off scores between settings is therefore problematic. If, for example, the cut-off score from the collective setting, based on its 90th percentile (segment B + C; Figure 1), is applied in the individual setting only 3% of the scores are abnormal (segment C). Only the most extreme scores are, therefore, seen as abnormal instead of the expected ∼10% (segment A + C). Predictive validity studies are based on these ∼10% highest scores (B + C). Children falling within the most extreme scores (segment C) are, however, more at risk for psychopathology compared to children who do not score as high but do score above the individual 90th percentile (segment A), as previous research indicates that an increase on the SDQ across the full range is related with increased odds for psychopathology (Goodman & Goodman, 2009). Using the cut-off score from the collective setting in the individual setting would therefore increase the ratio of children who are correctly characterized as a case (true positives; i.e. increased sensitivity), while the ratio of children who are correctly characterized as a non-case (true negatives) would decrease (i.e.
Table 5: Cut-off scores and percentage of cases for the individual and collective setting based on the 80th and 90th percentile of these settings and the percentage of cases in the individual setting based on the cut-offs of the collective setting.

<table>
<thead>
<tr>
<th></th>
<th>Normal: 0-80 percentile</th>
<th>Borderline: 80-90 percentile</th>
<th>Abnormal: &gt;90 percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cut-off</td>
<td>T%</td>
<td>M%</td>
</tr>
<tr>
<td><strong>Collective setting</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total difficulties score</td>
<td>0-14</td>
<td>83</td>
<td>86</td>
</tr>
<tr>
<td>Emotional symptoms</td>
<td>0-4</td>
<td>83</td>
<td>91</td>
</tr>
<tr>
<td>Conduct problems</td>
<td>0-3</td>
<td>87</td>
<td>86</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>0-6</td>
<td>84</td>
<td>85</td>
</tr>
<tr>
<td>Peer problems</td>
<td>0-3</td>
<td>89</td>
<td>88</td>
</tr>
<tr>
<td>Pro-social behaviour</td>
<td>10-6</td>
<td>85</td>
<td>79</td>
</tr>
<tr>
<td><strong>Individual setting</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total difficulties score</td>
<td>0-11</td>
<td>82</td>
<td>83</td>
</tr>
<tr>
<td>Emotional symptoms</td>
<td>0-3</td>
<td>82</td>
<td>90</td>
</tr>
<tr>
<td>Conduct problems</td>
<td>0-2</td>
<td>88</td>
<td>86</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>0-5</td>
<td>80</td>
<td>79</td>
</tr>
<tr>
<td>Peer problems</td>
<td>0-2</td>
<td>87</td>
<td>87</td>
</tr>
<tr>
<td>Pro-social behaviour</td>
<td>10-7</td>
<td>87</td>
<td>81</td>
</tr>
<tr>
<td><strong>Individual setting using cut-off scores from Collective setting</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total difficulties score</td>
<td>0-14</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>Emotional symptoms</td>
<td>0-4</td>
<td>90</td>
<td>95</td>
</tr>
<tr>
<td>Conduct problems</td>
<td>0-3</td>
<td>95</td>
<td>94</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>0-6</td>
<td>88</td>
<td>86</td>
</tr>
<tr>
<td>Peer problems</td>
<td>0-3</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>Pro-social behaviour</td>
<td>10-6</td>
<td>94</td>
<td>92</td>
</tr>
</tbody>
</table>

*Note.* T = Total; M = Male; F = Female.
decreased specificity). This would bias the interpretation of caseness, as the chance that a child, scoring above the cut-off point, develops psychopathology is higher than expected. For an informed and valid usage of the SDQ in the individual setting the 'actual' 90\textsuperscript{th} percentile cut-off has to be determined. Or, the accuracy in this individual setting has to be reassessed using the collective setting cut-off scores.

For the collective setting these results shows that it is important that attention is paid to the setting of the original assessment. As Mieloo et al. (2012, p. 7) point out in their discussion regarding their use of SDQ scores retrieved from an individual setting: “Thus, generalizing our findings to an anonymous [collective] research setting probably requires caution.” Based on the results of the present article this caution seems warranted, as the individual setting may yield an underestimation of SDQ scores compared to the collective setting. Besides the possible underestimation of SDQ scores, and associated caseness, interaction effects between gender and setting adds another potential hazard. No significant difference, for instance, is present on the hyperactivity subscale in the collective setting whilst there is in the individual setting, with males scoring higher than females. The magnitude, or even direction, of differences
between groups, such as gender, can therefore depend on the assessment setting. While the assessment setting showed structural differences, the present study showed no differences between scores derived from administration through CSAQ or SAQ. Overall, these findings illustrate the importance of taking sources of variation regarding the mode of assessment (e.g. interview versus self-administration) into account when interpreting or comparing the results of studies (Bowling, 2005).

The direction of the structural differences between the settings is as hypothesized based on the literature regarding SDB. For all scores, except the pro-social subscale, mean scores of the individual setting are lower compared to the collective, setting. This enhanced SDB is likely to be a result of the confidential nature of the individual setting, compared to the anonymous nature of the collective setting, in combination with the potential desirability to remain ‘off the radar’ as high scores can have personal consequences (Krumpal, 2013). The impact of SDB was however not limited to the understating of negative behaviour, but also resulted in the exaggeration of positive behaviour. This is shown by the higher scores on the pro-social subscale in the individual setting compared to the collective setting. It is furthermore noteworthy that not only an effect is present for the more sensitive questions, regarding conduct problems, but also for the relative insensitive questions regarding hyperactivity. This shows that it remains difficult to determine which questions are sensitive and which are not, especially for youths (Krumpal, 2013). Previous research for example found that substance use seems relatively unaffected by the level of anonymity, whilst such an effect has been more readily established for the disclosure of sexual and antisocial behaviour (Ong & Weiss, 2000; O’Malley, Johnston, Bachman, & Schulenberg, 2000; Durant et al., 2002). Also in line with previous research is the finding that females show more profound differences on sensitive health indicators between varying levels of anonymity, compared to males (Durant et al., 2002; van de Looij-Jansen et al., 2006).

CFA show that ME/I seems to hold between the settings and within each setting regarding gender. Mean comparisons, between gender and settings, regarding the SDQ scores and interpretation of the direction of the SDB are therefore justified as they are not an artefact of measurement bias. It furthermore validates the extrapolation of conclusions between settings, such as the established relation between the SDQ and psychopathology – the basis for its usage in the individual setting (Goodman, 2001). The establishment of metric ME/I also supports the absence of a difference regarding extreme response styles (ERS), between the settings (Cheung & Rensvold, 2000). Despite the elevated SDB, there is therefore, no excessive reliance on the extreme categories in the individual setting compared to the collective setting.

For both the individual and collective setting this article shows that the SDQ is a reliable instrument to use. As previous research also shows, the total difficulties score should be preferred to the subscales (Stone et al., 2010). This is due to the reliability of some subscales. Especially the conduct problems subscale scores low, the reliability of the peer problems, pro-social, and emotional symptoms subscales is however also below preferred levels. Favouring the total difficulties score above the subscales is in line with the primary purpose of the SDQ as an early screening instrument. That is, although the SDQ can differentiate between different psychopathologies (e.g. ADHD vs. OCD), it is by no means a psychiatric diagnostic instrument (van Widenfelt et al., 2003; Goodman, 2001).

Some strengths and limitations of the present study should be noted. The sample size from both samples was large and the amount of missingness was very small. In contrast to most studies investigating SDB, the current study had furthermore an observational design instead
of an experimental. It is, however, not possible to capture the true nature of the individual setting in an experimental design. Samples from both settings represent, however, the same socio-geographic population. Some, mostly small, differences on the background characteristics, nevertheless, occurred. Controlling for SES and age did, however, only marginally influence the results and, if anything, resulted in a larger effect size for setting. Comparison seems therefore valid, despite the difference in selection procedures (opt-in vs. opt-out), response rates, and background characteristics between the two settings. It could, furthermore, be argued that differences between the two settings are a characteristic of the settings as they illustrate the true nature of such settings. Although the sample size was large and the sample heterogeneous, the presented cut-off points, both for the collective and individual setting, should not be used as cut-off points for any of these two settings since the study was limited to a specific socio-geographic population. These cut-off points however do illustrate the mechanisms that will be present when cut-off points from actual, collective setting, validation studies will be used in the individual setting.

Regarding the CFA models it has to be noted, that since the original 5-factor structure proposed by Goodman (2001) shows a mediocre fit an adaptation of this structure was used for all following models which shows good fit throughout. This adaptation includes a positive construal factor for all positively worded items (Dickey & Blumberg, 2004). The increased fit supports previous research advocating inclusion of this factor (van Roy et al., 2008; Palmieri & Smith, 2007). Alternative factor models can however not be ruled out as they are not tested as this was beyond the focus of this article. There is furthermore no clear consensus regarding the procedure for establishing ME/I, especially regarding categorical variables and the use of $\Delta$AFI for this purpose (Sass et al., 2014; Stark, Chernyshenko, & Drasgow, 2006). By providing and interpreting multiple $\Delta$AFI along two common guidelines, the process is presented in a transparent and insightful manner (Meade et al., 2008; Chen, 2007).

To conclude, this study shows that the SDQ has the same meaning when measuring psychological problems in children in a collective and in an individual setting. Due to the SDB, seemingly inflicted by the confidential nature and possible personal consequences of high scores in the individual setting, there are however structural differences between the settings. These structural differences result in invalid mean comparisons and cross-use of cut-off points between settings and should be taken into account in future studies and should also be considered when interpreting finding from earlier studies. One should, therefore, be aware of the potential pitfalls that are instigated by the structural differences between the settings. For the collective setting this implies that (cross-country) comparisons of absolute scores between samples could be seriously hampered if the setting of assessment differs, while for the individual setting these structural differences result in an increased sensitivity and decreased specificity when cut-off points from the collective setting are used. Although the use of cut-off points on itself will always be debated, it is crucial that the SDQ works in the way it is ‘advertised’ achieving a certain accuracy (Goodman & Goodman, 2009). It is therefore of the utmost importance that if cut-off scores are used they are retrieved from the same setting as they are applied in.
References


3 Structural validity for the Need for Recovery Scale and the Checklist Individual Strength and its equivalence/invariance within different sub-groups of workers

Hoofs, H., Jansen, N. W. H., Jansen, M. W. J., & Kant, IJ.

Submitted
Abstract

Need for recovery (NFR) and prolonged fatigue are important facets in the complex relation between physical and psychological job demands and adverse health outcomes among workers. In the continuing unraveling of the etiology of NFR and prolonged fatigue, instruments with high structural validity are therefore important. It is, furthermore, important that NFR and prolonged fatigue are measured invariantly across subgroups of these grouping variables. The current article, therefore, investigates the factor structure and its equivalence across groups for two instruments, the NFR scale and the Checklist Individual Strength (CIS), measuring the concept NFR and prolonged fatigue respectively. The Maastricht Cohort Study (MCS) was used, capturing 12,140 employees from 45 different companies. Different factor structures were examined. The final structure for both instruments was examined for equivalence across three domains: demographic, personal, and work-related grouping variables. A 1-factor model showed a good fit for the NFR scale. For the CIS a 4-factor structure was most adequate. Subsequent models showed equivalence of measurement for each factor from the three different domains. The results significantly substantiate the practical application of the NFR scale and the CIS to measure NFR and prolonged fatigue.
Introduction

Work is associated with many positive objectives. It facilitates, amongst others, personal growth, a source of income, social contacts, and supports self-esteem (Eurofound, 2012). The number of employees who face high physical and psychological demands at work (e.g. high work load) on a daily basis is, nevertheless, high and still increasing (OECD, 2012). High demands can induce a short-term effect of increased neurophysiological activity which can be accompanied by elevated stress levels (Sluiter, Frings-Dresen, van der Beek, & Meijman, 2001). Elevated job demands can, therefore, influence the mental and physical health of employees both in the short and long-term. In the present study, the concept need for recovery (NFR) will be used to represent the adverse short-term effect of high job demands and insufficient resources to deal with these demands. High job demands can also independently lead to adverse long-term effects, represented by prolonged fatigue in the current study. NFR is, within the normal recuperation cycle, defined as the extent to which the job demands induces a need to recuperate from work which is ideally completed before the next working day (van Veldhoven, 2008). Prolonged fatigue is the presence of fatigue for a more extended period (Bultmann et al., 2000).

Due to the potential burden of NFR and prolonged fatigue for employees it is of strong social significance to study their etiology (Eurofound, 2012; van Veldhoven, 2008). Within the current study the NFR scale (van Veldhoven, 2003) and the Checklist Individual Strength (CIS; Vercoulen et al., 1994) will be used to measure NFR and prolonged fatigue, respectively. These instruments are widely used within occupational health (research) also within an international perspective (de Vries, Michielsen, & van Heck, 2003; Ergin & Yildirim, 2012; Moriguchi, Alem, van Veldhoven, & Coury, 2010; van Veldhoven, 2008). While research using these instruments show promising results regarding their reliability and validity, structural validity of the instruments is rarely explicitly assessed. Furthermore, structural validity should be stable irrespective for specific subgroup, e.g. by age or sex. While this assumption has probably been implicitly presumed in studies using the NFR scale and the CIS it has rarely been tested. The aim of present article is, therefore, to investigate the structural validity, and its equivalence in different subgroups, of the NFR scale and the CIS.

Structural validity assesses the underlying, theoretical structure of an instrument and can enhance the construct validity which encompasses whether a test measures what it claims to measure (de Vet, Terwee, Mokkink, & Knol, 2011). While face validity is an important aspect in this respect, a confirmatory approach assessing the structure of an instrument greatly enhances the applicability of an instrument (Terwee et al., 2007). As the NFR scale was specifically developed to have a unidimensional structure, a 1-factor model seems most suited to describe the underlying structure of the instrument (van Veldhoven, 2008). While previous research already shows strong indications for such a 1-factor structure, to our knowledge its structure has never been tested confirmatively using CFA to date (Jansen, Kant, & van den Brandt, 2002). The CIS consists of four subscales which also can be summed into a total score. This hierarchical component of the CIS can imply different factor structures (Figure 1). Based on the total score a 1-factor structure could be examined (Figure 1a; Bultmann, Kant, Kasl, Beurskens, & van den Brandt, 2002). A 4-factor structure seems also applicable to assess the structural validity of the CIS (Figure 1b; Jansen et al., 2002). The 2-level structure (Figure 1c), which has an overarching latent factor predicting the four subscales, is a hybrid form of these two models as it encompasses both the total score and the four subscales.
The structure of an instrument should, subsequently, function in parallel fashion in different subgroups. Many studies show the potential implication if such an equivalence does not hold (for reviews see, Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). This could result in differences that are either an artefact of this inequivalence and/or a true difference between groups. Besides scientific consequences, a different connotation across groups has also direct practical implications. That is, due to this inequivalence the instrument should be interpreted and applied differently across groups. Regarding NFR, for example, employees who work in shifts could differently interpret specific questions compared to employees with regular working hours. This could result in different scores on the instrument, while their ‘underlying’ true NFR could be equal, which could subsequently impact the conclusions based on these results. These (hampered) conclusions could be on an individual level (e.g. monitoring employees within an occupational health context) but could also be related to policy making (e.g. investigating the relation between NFR and different work schedules). The equivalence of factor structures is often denoted as measurement equivalence/invariance (ME/I). Technically ME/I considers whether or not, under different conditions of observing and studying phenomena, instruments yield measures of the same attributes (Millsap, 2011).

A myriad of relevant subgroups could be inspected regarding the presence of ME/I. Naturally, the selection of groups (grouping variables) for which the equivalence of an instrument should be assessed depends on the research question at hand. This article will therefore focus on some core grouping variables from different domains (demographic, personal, and work-
related) in the field of occupational health. Demographic grouping variables for which ME/I is assessed are: age, gender, and education. Personal grouping variables that are assessed include: perceived general health, and home situation. Grouping variables regarding health are included as they are closely related with NFR and prolonged fatigue (e.g. Jansen et al., 2002; Bultmann et al., 2002). Home situation (i.e. living situation/arrangement) is included as this could alter the perception of items, especially those that refer to the home situation (e.g. When I get home from work, I need to be left in peace for a while). Work-related grouping variables include: working time arrangements, and job demands. Working time arrangements and job demands are included as these grouping variables are associated with increased levels of NFR and prolonged fatigue (e.g. Sluiter, de Croon, Meijman, & Frings-Dresen, 2003; van Veldhoven, 2008). More importantly, however, is the fact that these grouping variables can, to a differing degree, be modified and therefore be used for preventative or treatment purposes.

**Methods**

**Sampling and procedures**

This study is based on data from the ongoing Maastricht Cohort Study (MCS). The MCS was set up in May 1998 and included 12,140 participants from 45 different companies. At this baseline measurement, all included participants were aged between 18 and 65 (Mohren, Jansen, van Amelsvoort, & Kant, 2007). Written informed consent was obtained from all participants. The study was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. Employees who were absent from work, pregnant, or had multiple jobs when they completed the questionnaires were excluded ($n = 10,800; 25.72\%$).

**Measures**

**Need for recovery.** NFR was measured at baseline using a subscale from the Dutch Questionnaire on the Experience and Evaluation of Work (VBBA; van Veldhoven, 2003). The scale comprises 11 dichotomous items (yes/no), representing short-term effects of a day at work. Cronbach’s alpha of the scale was .88 within the present study. Item number 4 was positively formulated while the other 10 items had a negative formulation. The mean score was computed for descriptive purposes ranging from 0 to 100, with a higher score indicating a higher NFR (van Veldhoven, 2003).

**Prolonged fatigue.** Fatigue was measured at baseline with the CIS (Vercoulen et al., 1994) stemming from hospital studies of chronic fatigue syndrome. The questionnaire was subsequently also validated within the working population to measure prolonged fatigue (Beurskens et al., 2000). The CIS includes four subscales: subjective fatigue experience (8 items), concentration (5 items), motivation (4 items), and physical activity level (3 items). Each item includes a statement which is scored on a 7-point Likert scale (ranging from yes, this is true to no, this is not true). In a study involving 351 employees the Cronbach’s $\alpha$ was .96 for the total CIS score and ranged from .88 to .96 for the different subscales (de Vries et al., 2003).
For descriptive purposes the composite CIS total score (ranging from 20 to 140) was obtained by summing the scores of four subscales.

**Demographic grouping variables.** All demographic grouping variables were assessed at the baseline measurement. Age was divided into three subgroups, that is young (< 35 years), middle aged (35-45 years), and older workers (> 45 years). Educational level was recoded into three categories: low, medium, and high.

**Personal grouping variables.** To measure the perceived general health status of the participants, one item from the SF-36 Health Survey (Aaronson et al., 1998) was used. This measure was dichotomised by grouping the first three of the five response categories (Excellent, Very good, or Good) into “good general health” and the last two (Moderate, or Bad) into “poor general health”. To measure home situation three dichotomous (yes/no) items were combined which assessed whether they lived alone, lived with their spouse or partner, and/or had any children at home. If employees responded positive to the latter two items they were categorised as having children at home.

**Work-related grouping variables.** With regard to working hours, employees were asked to indicate their working hours per week. The answer options were recoded into two categories: a full-time job (≥ 36 hours per week) or a part-time job (< 36 hours per week). For work schedule, employees who indicated that they worked in shifts (3, 4, 5, or irregular-shift) and worked frequently during the night (1-item question) were categorized as shift workers. The Dutch version of the self-administered Job Content Questionnaire (JCQ) was used to measure psychological job demands and decision latitude (Karasek et al., 1998). The psychological job demands scale ranges from 12 to 48, and the decision latitude scale from 24 to 96. To construct subgroups (low, medium, and high) the tertiles of each scale were used. Cronbach’s $\alpha$ was .77 for the decision latitude and .74 for the psychological job demands scale in a Dutch sample (Karasek et al., 1998).

**Statistical analysis**

First the descriptives of the sample, including the number of missing values, were examined. Cronbach’s $\alpha$ and greatest lower bound ($glb$) were calculated as reliability coefficients. Although $glb$ seems most appropriate, Cronbach’s $\alpha$ was reported to facilitate comparisons with previous studies. Since the $glb$ was always larger than the Cronbach’s $\alpha$ of a scale, they were reported as a range (Sijtsma, 2008).

Confirmatory Factor Analysis (CFA) was used to test both structural validity and ME/I for the NFR scale and the CIS. As indicated in the Introduction section a 1-factor model of the NFR scale was tested. For the CIS three alternative models were tested (Figure 1). As the items of the NFR scale were dichotomous they were handled as categorical data, using the weighted least square means and variance adjusted (WLSMV) estimator. The CIS was scored on a 7-point Likert scale. Initial analysis indicated that due to skewness of the items the WLSMV should be preferred. The first step consisted of testing the overall model fit. Therefore, the hypothesised CFA models were fitted to the total population. If model fit was unsatisfactory, modification indices were requested to indicate possible improvements for the factor structure. Based on these model building results, a final model was selected for both scales.
The final models for the NFR scale and the CIS were used to test for ME/I, regarding the different subgroups for each grouping variable. Four sequential models were conducted to test increasingly restrictive forms of ME/I (see Millsap, 2011): (a) Configural (Model 1); (b) Metric (Model 2); (c) Scalar (Model 3); (d) Strict (Model 4). Following recommendations of Muthen and Muthen (1998-2012) metric and scalar ME/I were jointly tested.

Respondents were included if they had a valid response on at least one item. To control for the company based sampling procedure the complex estimation type within Mplus was used to correct for potential confounding effects (Muthen & Muthen, 1998-2012). To evaluate model fit the Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA) were used as approximate fit indexes (AFIs). Greater CFI values and smaller RMSEA values indicate better fit. For the CFI a cut-off point of .95 was used to indicate good model fit and values greater than .90 to indicate acceptable model fit. For the RMSEA values smaller than .06 indicate good fit (Hu & Bentler, 1999). In the current study two guidelines were used to assess model difference. The conventional criteria by Chen (2007) stated that $\Delta CFI \leq -.01$ and $\Delta RMSEA \geq .01$ indicate substantial differences between models, therefore not supporting ME/I. Meade, Johnson, and Braddy (2008) recommended more conservative cut-off points; $\Delta CFI \leq -.002$ and $\Delta RMSEA \geq .007$. For the assessment of ME/I parsimonious models – stricter form of ME/I – should be preferred if no difference is indicated based on the $\Delta$AFIs (Millsap, 2011).

**Results**

**Missing data & internal consistency**

The descriptives of the sample are reported in Table 1. The items on the NFR scale and CIS had few missing data (<5 %). For each analysis employees were included if they had at least one valid response on an item. Employees were furthermore excluded for ME/I analysis if they had a missing value on the grouping variable.

Internal consistency was computed for the NFR scale and the CIS. The internal consistency of the NFR scale was good (Cronbach’s $\alpha = .88$, $glb = .90$). Internal consistency for the CIS was also good (Cronbach’s $\alpha = .93$, $glb = .94$).

**Factor structure**

The 1-factor model of the NFR resulted in an acceptable fit, $\chi^2(44) = 1,686.14, p < .001, CFI = .959, RMSEA = 0.059$. Including two error correlations resulted in a good fitting model, $\chi^2(42) = 668.31, p < .001, CFI = .984, RMSEA = 0.037$, with an improved fit over the baseline model, $\Delta \chi^2(2) = 1,307.68, p < .001, \Delta CFI = -.025, \Delta RMSEA = 0.022$. The standardised factor loadings of the final CFA model for the NFR scale are presented in Figure 2.

The initial fit of the 4-factor model of the CIS was mediocre, $\chi^2(164) = 7,831.75, p < .001, CFI = .915, RMSEA = 0.066$. The addition of some error correlations resulted in a model with a good, $\chi^2(158) = 4,392.36, p < .001, CFI = .953, RMSEA = 0.050$, and increased fit, $\Delta \chi^2(6) = 4,614.17, p < .001, \Delta CFI = -.038, \Delta RMSEA = 0.016$. 

Table 1: Descriptive statistics for the study population for the NFR scale, the CIS, and the grouping variables for which ME/I is assessed

<table>
<thead>
<tr>
<th></th>
<th>n (%)</th>
<th>Mean (SD)</th>
<th>Range (min - max)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Need for recovery</strong></td>
<td>37.12 (32.01)</td>
<td>0 - 100</td>
<td></td>
</tr>
<tr>
<td><strong>Prolonged fatigue</strong></td>
<td>57.19 (23.69)</td>
<td>20 - 140</td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 35 years</td>
<td>3,546 (29.23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35-45 years</td>
<td>4,109 (33.87)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 45 years</td>
<td>4,477 (36.90)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>3,263 (26.88)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>8,874 (73.12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>2,369 (20.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>5,349 (45.45)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>4,051 (34.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Perceived health</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Good</td>
<td>10,268 (85.23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poor</td>
<td>1,780 (14.77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Home situation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>1,121 (9.64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Partner</td>
<td>4,022 (34.60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children</td>
<td>6,482 (55.76)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Psychological job demands</strong></td>
<td>33.27 (5.70)</td>
<td>12 - 4</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>4,401 (37.39)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>4,452 (37.82)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>2,919 (24.80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Decision latitude</strong></td>
<td>71.61 (11.43)</td>
<td>24 - 96</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>4,371 (36.46)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>3,872 (32.30)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>3,746 (31.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Work schedule</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day work</td>
<td>8,114 (79.63)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shift work</td>
<td>2,075 (20.37)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Working hours per week</strong></td>
<td>3,174 (26.36)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-time</td>
<td>8,865 (73.64)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full-time</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: For each continuous variable the observed range was equal to the theoretical range.

The two alternative models were also tested against the final 4-factor structure. Using the same error correlations as in the final 4-factor model the 1-factor model showed a poor fit, $\chi^2(164) = 15,547.52, p < .001, CFI = .829, RMSEA = 0.093$. This fit was also significantly worse compared to the 4-factor solution, $\Delta \chi^2(6) = 4,816.85, p < .001, \Delta CFI = .124, \Delta RMSEA = −0.043$, regardless of the criterion used for model comparison. The second order factor structure, including the same error correlations as in the final 4-factor model, showed an acceptable fit, $\chi^2(164) = 15,547.52, p < .001, CFI = .940, RMSEA = 0.056$. The standardised loading of the four subscales on the second order factor were all
significant ($p < .05$) and substantial ($\beta > .75$). Whilst acceptable the fit of the second order factor model deteriorated compared to the 4-factor structure, $\Delta \chi^2(2) = 594.23, p < .001$, $\Delta CFI = .013$, $\Delta RMSEA = -.006$. Both the 4-factor structure and the second order structure seemed, therefore, to be valid. Based on model fit, however, the 4-factor structure should be preferred. The factor loadings of this final model are presented in Figure 2.
Measurement equivalence/invariance

The final factor models of the CIS and the NFR scale were used to assess the ME/I across various grouping variables (see Table 2 for results of model fit comparison). For the demographic grouping variables, age, gender, and education, the model fit of the different subgroups (e.g. male and female) was good for both the NFR scale and the CIS. Configural ME/I was also shown for each demographic factor as model fit of the configural CFA was good for the NFR scale and the CIS. Subsequent, more stringent, nested CFA models showed increased fit compared to the configural model, adhering to both the conventional and conservative cut-off criteria, indicating strict ME/I for both the NFR scale and the CIS for the demographic grouping variables age, gender, and education.

For the personal grouping variables, perceived health, and home situation the CFA models showed good fit for both the NFR scale and the CIS. Models testing indicated that there was strict ME/I for all grouping variables.

The grouping variables psychological job demands, decision latitude, work schedule, and working hours per week (work-related grouping variables) also showed strict ME/I for the NFR scale and the CIS.

Discussion

NFR and prolonged fatigue play a prominent role in the association between high physical and psychological job demands and adverse long-term health outcomes. Reliable and valid measures for these constructs are therefore crucial. The NFR scale and the CIS have already proved to be instruments with good psychometric properties, including high internal consistency and discriminative (Bultmann et al., 2000; van Veldhoven, 2003), convergent (Beurskens et al., 2000; de Vries et al., 2003) and predictive validity (Sluiter et al., 2003). While these facets are imperative for valid measurement instruments, they are not indecisive for the structural validity of the instruments and its equivalence across (multiple) groups. For this purpose confirmative studies are vital as they do not rely on post-hoc models assessment but test a-priori theories. Such studies enhance the theoretical underpinning of an instrument and its comparisons across groups, which is tantamount in the ongoing unraveling of the etiology of NFR and prolonged fatigue. The ME/I of an instrument is often assumed but rarely tested. Goal of the present article was, therefore, to examine the structural validity of the NFR scale and the CIS and to what extent they assessed NFR and prolonged fatigue invariantly across different grouping variables.

The structural validity of the NFR scale confirms to the 1-factor structure which is posited by the original scale. The current study is the first to assess this factor structure using CFA, showing strong support of the hypothesised underlying structure of the scale (van Veldhoven, 2003). For the CIS the 1-factor structure was found to be unsatisfactory. The 4-factor structure and the second order factor structure showed a much better fit to the model, with the 4-factor structure slightly outperforming the second order structure. Selection of one of these two structures should, therefore, also depend on the conceptual framework and theoretical considerations of the specific study at hand. Both structures enhance, however, the theoretical support for the divergent validity of the four different subscales (Vercoulen et al., 1994). Inclusion of several error correlations within both models indicated a strong
dependence between some items. While such findings can be a result of the current sample (Markland, 2007), the similarity of the items and the magnitude of the correlation suggest that this finding is not inflicted by change. This finding implies that internal consistency measures will be an overestimation (Raykov, 2012). Future research should take this into account and could possibly examine whether some items could be omitted or parcelled.
With regards to the structural equivalence across groups, strict ME/I was found for all grouping variables on the NFR scale and the CIS. While van Veldhoven (2003) already found comparative reliability coefficients across gender and age groups, this does not automatically imply equality of measurement across groups. ME/I does imply such equivalence of structural validity across groups. Mean scores, and their differences, and associations found with other constructs can therefore be validly interpreted as the same construct is measured in each subgroup. While it is too simplistic to state that structural validity of NFR and CIS is confirmed or that both scales are ME/I, all bodes well for the application of these instruments within different settings and populations. The current study included the employees of 45 different companies which resulted in a large sample enabling the investigation of ME/I for a myriad of grouping variables. This, furthermore, resulted in a large within- and between variation of work content and demands of the employees (Mohren et al., 2007). This in combination with the robustness of the results makes the findings broadly applicable. A remaining issue which should be examined is whether the NFR scale and the CIS are ME/I over time (e.g. measure the same construct at each time-point), which is crucial in understanding the development of the constructs over time (Vandenberg & Lance, 2000). Cultural and ethnic factors may affect the interpretation of the items of the CIS and NFR scale and as may therefore affect ME/I. However, the impact of cultural and ethnic factors could not be determined in this study and should therefore also be addressed in future research. This is important to further integrate the NFR scale and the CIS within the international research context in which NFR and prolonged fatigue are important concepts to study the relation between high physical and psychological demands and long-term adverse health outcomes (e.g. Ergin & Yildirim, 2012; Moriguchi et al., 2010).

The presence of the structural validity and ME/I also strengthens the practical implementation of these instruments assessing NFR and prolonged fatigue. The strong ME/I shows that occupational health services, for example, do not have to differentiate their assessment due to a biased measurement for the grouping variables assessed. This does not imply, however, that the constructs are equivalent in their predictive validity. Hence, while NFR and prolonged fatigue are, for example, invariantly measured for different age groups, associations with other grouping variables could still differ (Millsap, 2011). Recent studies, for example, show that prolonged fatigue does predict sickness absence for men but not for women (Roelen, van Rhenen, Groothoff, van der Klink, & Bultmann, 2014). These differentiations in associations are, however, only valid through the presence of ME/I.

To conclude, this study shows that within the current population the NFR scale and the CIS show good structural validity which is also equivalent across groups. This is an important prerequisite for etiological research unraveling the complex nature of need for recovery and prolonged fatigue. For group-wise comparisons an equivalent factor structure across groups is especially important as otherwise such comparisons are not feasible. It should be noted that this prerequisite is an important quality criteria for all (mental health) measurement instruments that are used to compare different groups. As such this study strengthens the further implication of the NFR scale and the CIS to be used for comparisons across different (important) grouping variables within the occupational health. That is, since strict ME/I is demonstrated, mean scores, associations with other constructs, and item reliability can be validly compared for the respective grouping variables (e.g. gender). Some specific issues, such as ME/I in cross-country comparisons, warrant further research. The findings in this article, nevertheless, make a strong point for the structural validity of the NFR scale and the
CIS, especially within the light of group-wise comparisons. Future usage of these instruments is, therefore, significantly substantiated by the results of the current study.

References


Sluiter, J. K., de Croon, E. M., Meijman, T. F., & Frings-Dresen, M. H. W. (2003). Need for recovery from work related fatigue and its role in the development and prediction of subjective health complaints. *Occupational and Environmental Medicine, 60*(suppl 1), i62–i70. doi:10.1136/oem.60.suppl_1.i62


Monitoring of need for recovery and prolonged fatigue within the working population: Evaluation of reliability and agreement over time

Hoofs, H., Jansen, N. W. H., Jansen, M. W. J., & Kant, IJ.

doi:10.3233/WOR-172624
Abstract

Purpose

Need for recovery (NFR) and prolonged fatigue are two important concepts for monitoring short- and long-term outcomes of psychological job demands within employees. For effective monitoring it is, however, important to gain insight in the reproducibility of the instruments that are used.

Objective

The objective was to assess reproducibility of the NFR scale and Checklist Individual Strength (CIS), measuring NFR and prolonged fatigue respectively, in the working population.

Methods

Longitudinal data from the Maastricht Cohort Study (MCS) study was used, capturing 12,140 employees from 45 different companies at baseline. A ‘working’ and ‘returning to work’ sample was conceived for different intervals; 4-month, 1-year, and 2-year.

Results

Reliability, assessed with the interclass correlation, was high within employees with a stable work environment for the NFR scale (.78) and CIS (.75). The smallest detectable change, assessing the agreement, was 41.20 for the NFR scale and 31.10 for the CIS.

Conclusions

Reliability was satisfactory for both the NFR scale and CIS. The agreement of both scales to detect changes within employees was, however, less optimal. It is, therefore, suggested that, ideally, both instruments are placed within a broader range of instruments to effectively monitor the outcomes of psychological job demands.
Monitoring NFR and prolonged fatigue

Introduction

Due to demographic, economic and social changes, work has changed substantially in western societies over the last decades. Work has shifted from physical challenging towards work which is more psychological demanding (Eurofound, 2012). Global competition and the 24 hour economy have led to intensification of labor with negative consequences for contractual arrangements and working time arrangements (Eurofound, 2012). Also techniques and working methods are rapidly changing, requiring workers skills to be up to date (Statistics Netherlands, 2015). Moreover, labor market policies nowadays aim at keeping workers in the labor market for as long as possible, and in good health. These developments generate some challenging issues with regard to maintaining productivity of employees in good health. There are, however, growing indications that workers increasingly face negative consequences for their health and well-being and have difficulties to obtain a job in which they can maintain or develop their skills. The Organisation for Economic Cooperation and Development (OECD) has recently published a report on ‘Mental Health and Work’ about the state of affairs in the Netherlands (OECD, 2014). From this report it becomes clear that mental health is an important issue since about 25% of the working population is expected to experience a period with mental health complaints at some point during their working life.

Occupational health surveillance should therefore include mental health outcomes which are a direct (short-term) consequence of the psychosocial work environment. Monitoring these outcomes can play a crucial role in occupational and public health, especially if these health outcomes are reversible. Short-term mental health outcomes can, over time, result in health outcomes which encompass multiple domains in life and are often more persistent. For informative monitoring and research in occupational health, as such, both sorts of outcomes can have an important role. Need for recovery (NFR) is a short-term mental health outcome which captures the need for and extent of recuperation after work. Within the normal recuperation cycle this is ideally completed before the next working day (van Veldhoven, 2008). Prolonged fatigue is the presence of fatigue, including an overwhelming sense of tiredness, for a more extended period (Bultmann et al., 2000; Kalkman, Zwarts, Schillings, van Engelen, & Bleijenberg, 2008). Both NFR and prolonged fatigue are not only undesirable themselves, but are also associated with adverse (health) outcomes such as sickness absence, cardiovascular disease, subjective health complaints, and occupational mobility (de Croon, Sluiter, & Frings-Dresen, 2003; van Amelsvoort, Kant, Bultmann, & Swaen, 2003; de Raeve, Kant, Jansen, Vasse, & van den Brandt, 2009; van Amelsvoort, Kant, Beurskens, Schroer, & Swaen, 2002; Silva-Costa, Griep, Fischer, & Rotenberg, 2012). In combination with other facets, monitoring NFR and prolonged fatigue within employees might therefore be vital for facilitating or enabling preventive strategies targeting at prevention or limiting such adverse outcomes. Effective and informative monitoring of NFR and prolonged fatigue requires, however, information regarding the quality of the instruments for this purpose. The reproducibility of an instrument, encompassing stability of scores over time within stable employees, is of special interest for monitoring purposes and associated research (Dimitrov, Rumrill, Fitzgerald, & Hennessey, 2001; Terwee et al., 2007). For two instruments, measuring NFR and prolonged fatigue, the reproducibility, consisting of the reliability and agreement over time, will therefore be examined in the present study.

Measuring mental health, including NFR and prolonged fatigue, can be achieved with different types of instruments. Bio-markers are, for example, used to measure NFR and prolonged fa-
tigue. While these bio-marker studies are promising, results are mixed (Russell, Koren, Rieder, & van Uum, 2012; Sluiter, Frings-Dresen, van der Beek, & Meijman, 2001). Questionnaires are often used to grasp the mental health construct and its context. Within the current study the NFR scale (van Veldhoven, 2003) and the Checklist Individual Strength (CIS; Bultmann et al., 2000) will be used to measure NFR and prolonged fatigue, respectively. As the main goal of monitoring is to detect differences between and within employees over time it is important that questionnaires measure the same construct irrespective of the timing of measurement (Vandenberg & Lance, 2000). This assumption is known as longitudinal measurement equivalence/invariance (ME/I) and is a prerequisite for valid comparisons of scores of time (Millsap, 2011). To assess a questionnaire’s reproducibility, which is pertinent for all sorts of instruments (e.g. bio-markers or interviews), longitudinal ME/I should therefore be evaluated additionally.

Reliability and agreement within a monitoring context

One of the monitoring goals would be to adequately differentiate between employees with different levels of NFR or prolonged fatigue. This goal is captured in the reliability of an instrument. For continuous measures reliability is expressed using the intraclass correlation (ICC), in which higher values indicate a higher reliability (Shrout & Fleiss, 1979). The ICC indicates the proportion of true change in relation to the measurement error and systematic differences between measurements. Variation in the true score over time is associated with the stability of the environment of the employees (Hopkins, 2000; Lexell & Downham, 2005). Especially changes in an environment closely related to the outcome, such as the psychosocial work environment, might alter this variation. Employees with a stable work environment were, as such, found to show higher reliability on the NFR scale over time compared to employees with a changing work environment (e.g. reorganization de Croon, 2006). Furthermore, for short time intervals the reliability is generally higher, compared to longer intervals (Schuerger, Zarrella, & Hotz, 1989). Reliability is also affected by the presence of measurement error (Choi & Park, 2005; Weir, 2005), which lowers the potential ICC an instrument can achieve. It is, furthermore, assumed that the ICC is unrelated to the true score of an employee (de Vet, Terwee, Mokkink, & Knol, 2011). Hence, reliability should be comparable for employees with (generally) high and low levels of NFR or prolonged fatigue.

Apart from differentiating between employees, monitoring can also be used to detect changes within employees. This goal is captured by the agreement of an instrument over time, comprising the effect of measurement error on consecutive measurements if no change is expected. Its inverse, which change can be considered statistically relevant, is especially interesting for monitoring purposes. Agreement over time is expressed using the smallest detectable change (SDC) within continuous measures (Beckerman et al., 2001; de Vet, Bouter, Bezemer, & Beurskens, 2001). The SDC separates those effects which are ‘important’ (signal) from those which are not (noise; Lexell & Downham, 2005). Expressed in the same unit as the instrument the SDC indicates the minimal change in the outcome, independent of clinical relevance, which can be considered relevant with a certainty of 95% (Beckerman et al., 2001). Resembling the signal-to-noise ratio this range can shed light on the effectiveness of the instrument for mental health surveillance for both clinical practice and research (Hopkins, 2000). For individual monitoring of employees the SDC should be expressed on an individual level (SDCind). For the monitoring of groups or for the application of an instrument within (public
health) research the \( SDC \) is expressed on a group level (\( SDC_{\text{group}} \); de Vet et al., 2001). This \( SDC_{\text{group}} \) can shed light, for example, on the effectiveness of an intervention within a group of employees.

The NFR scale and CIS are frequently implemented in research and practice. For neither of these two instruments, however, longitudinal ME/I has ever been assessed so far. The reliability of the NFR scale was only examined in one earlier study. This study assessed the stability across different levels of stability in the work environment of the employees, and showed high \( ICC \) for the stable group (de Croon, 2006). The study population, however, included hospital nurses and truck drivers, only hampering external validity. To our knowledge, the agreement for the NFR scale has never been examined. For the CIS the reliability and agreement were assessed in multiple studies (Ergin & Yildirim, 2012; Rietberg, van Wegen, & Kwakkel, 2010; Koopman, Brehm, Heerkens, Nollet, & Beelen, 2014). While the reliability and agreement were found to be adequate, these studies were exclusively based on patient populations (e.g. multiple sclerosis). This hampers extrapolation of these results to the working population, also because the reliability and agreement of an instrument depends on the population, and its heterogeneity, in which it is assessed.

It is, therefore, important that the reproducibility of the NFR scale and CIS are assessed within the population where they are applied, that is, generally a heterogeneous working population. Within such a population, the reliability of the NFR scale and CIS should be comparable for different subgroups. This implies that the quality of the instruments to differentiate between employees should not be altered by group membership (e.g. gender). For characteristics that are related with changes in the outcome it is, however, expected that the reliability does differ. Employees with an unstable work environment are, therefore, expected to show lower levels of reliability on both scales. As the NFR is considered to be more domain (work) specific compared to prolonged fatigue (Lexell & Downham, 2005) this difference in reliability is expected to be more pronounced for the NFR scale compared to the CIS. Conversely, different levels of stability within other domains (e.g. private life) are expected to show more differences in the reliability of prolonged fatigue whilst less within NFR.

Not only should the preventive goal of monitoring be substantiated, however, but also its role in the rehabilitation process. A heterogeneous working population also enables such an investigation among employees who return to work after a long period of (sickness) absence. This sheds light on the applicability of the instrument within different contexts (Hoefsmit, Houkes, & Nijhuis, 2014). The present article, therefore, investigates the reproducibility of the NFR scale and CIS as monitoring instruments both within currently ‘active’ employees and employees who return to work after a long period of (sickness) absence. As the reliability of an instrument is associated with the timing of measurement, different intervals will be explored (i.e. 4-month, 1 year, and 2 year). The following three questions will be specifically addressed:

- Does longitudinal ME/I hold for the NFR scale and CIS?
- What is the reliability and agreement over time of the NFR scale and CIS within employees with a stable work environment?
- What is the reliability of the NFR scale and CIS within various subgroups of employees, including various levels of work environment stability?
Methods

Sampling and procedures

This study is based on data from the ongoing Maastricht Cohort Study (MCS), set up in May 1998 including 12,140 participants from 45 different companies. At baseline measurement, all included participants were aged between 18 and 65 (Kant et al., 2003; Mohren, Jansen, van Amelsvoort, & Kant, 2007). The baseline (T0) cohort consists of 8,840 (73%) men and 3,255 (27%) women. All employees who returned the baseline questionnaire received the two short questionnaires T1 in September 1998 (response rate 88%, \(n = 10,592\)) and T2 in January 1999 (85%, \(n = 10,270\)) as well. Employees returning the baseline questionnaire and at least one of the short questionnaires received the extensive questionnaire T3 in May 1999 (80%, \(n = 9,655\)). Employees returning the T3 questionnaire also received the short questionnaires T4 in September 1999 (74%, \(n = 8,956\)) and T5 in January 2000 (72%, \(n = 8,692\)). Employees who returned the questionnaire at T3 and at least one of the consecutive short questionnaires also received the extensive questionnaire T6 in May 2000 (67%, \(n = 8,707\)). All procedures followed were in accordance with the ethical and with the Helsinki Declaration of 1975, as revised in 2000. Informed consent was obtained from all patients for being included in the study.

‘Working’ samples

To control the amount of variation in the job characteristics over time for each employee several in- and exclusion criteria were applied which resulted in three ‘working’ samples. These were defined for the interval between T0-T1 (4-month), T0-T3 (1-year), and T0-T6 (2-year). Employees were only included if they had a response on each questionnaire in the interval assessed (remaining: \(n_{4\text{-month}} = 10,592, n_{1\text{-year}} = 8,673, n_{2\text{-year}} = 6,908\)). Next, employees were excluded if they had a change in employer (\(n_{4\text{-month}} = 10,473, n_{1\text{-year}} = 8,383, n_{2\text{-year}} = 6,457\)) or currently received any form of benefit or had a long-term (> 4 weeks) sickness absence (\(n_{4\text{-month}} = 9,834, n_{1\text{-year}} = 7,241, n_{2\text{-year}} = 5,049\)). Employees were also excluded if they were pregnant, had multiple jobs, or were currently on sick leave, as this could confound the interpretation of the questionnaire. This resulted in a ‘working’ sample of 9,173 employees for the 4-month interval, 6,545 for the 1-year interval, and 4,429 for the 2-year interval.

‘Returning to work’ samples

To examine the NFR scale and CIS within employees who returned to work after a long period of (sickness) absence two ‘returning to work’ samples were defined. As it was not possible to define a 2-year interval for this subsample only a 4-month and 1-year interval were defined. Employees had to report that they received a disability benefit at T2 or were sick for a period of four weeks or longer in between T3 and T2 (remaining: \(n = 689\)). As such, the ‘working’ and ‘returning to work’ samples were mutually exclusive. Further in- and exclusion criteria were the same as for the ‘working’ samples. This resulted in a ‘returning to work’ sample of 257 employees for the 4-month interval and 153 for the 1-year interval, which had T3 as baseline measurement.
Measures

Need for recovery. NFR was measured using a subscale from the Dutch Questionnaire on the Experience and Evaluation of Work (Dutch abbreviation: VBBA) (van Veldhoven, 2003). The scale comprises 11 dichotomous items (yes/no), representing short-term effects of a day at work (e.g.: At the end of a working day I am exhausted.). Cronbach’s $\alpha$ of the scale was .88 within the present study. The mean score was computed by transforming the sum of all items, resulting in a scoring range of 0 to 100, with a higher score indicating a higher NFR. The NFR scale was characterised by a strong floor effect (i.e. approximately 25% of the employees had the minimum score).

Prolonged fatigue. Prolonged fatigue was measured with the CIS (Vercoulen et al., 1994; Beurskens et al., 2000). The questionnaire is also validated within the working population to measure prolonged fatigue (Beurskens et al., 2000; Bultmann et al., 2000). This questionnaire includes four subscales: subjective fatigue experience (8 items), concentration (5), motivation (4), and physical activity level (3). Each item includes a statement (e.g.: Thinking requires effort) applying to the last two weeks which has to be scored on a 7-point Likert scale (yes, this is true to no, this is not true). The composite CIS total score (20-140) was obtained by adding the scores of the four subscales. Within the MCS the Cronbach’s $\alpha$ was .93 (Bultmann et al., 2000).

Demographic, health, and work-related factors. Age was divided into the categories, < 35, 35-44, and $\geq$ 45 years. Educational level was recoded into three categories: low, medium, and high. Long-term illnesses were measured with the use of one dichotomous item (yes/no). For work schedule, employees who indicated that they worked in shifts (3, 4, 5, or irregular-shift) were categorised as shift workers whilst does who did not were indicated as day workers. A change in the private life of an employee was based on two items. The first item indicated if an employee moved houses between waves. The second item considered whether the employee experienced a change in her/his private life within the last 4 months. This item was scored on a 3-point scale (worsened, stable, or improved) and was inventoried at T1-2, T4-6. These two items were recoded into one variable with two categories: no or one or more changes in private life.

Perceived change in work environment. Four self-report items, which were inventoried at T1-2, T4-6, assessed whether the employee perceived a change in her/his work environment. This reflects a subjective change in the work environment. These items referred to workload, working conditions, content of work, and relationship with the supervisor and colleagues. Each item was scored on a 3-point scale indicating whether this aspect improved, worsened or was stable, compared with the situation 4 months ago. The number of aspects on which at least one change was indicated across the waves of interest was summed. This sum was recoded into three categories: no, one, and two or more perceived changes in the work environment.

Reported change in work environment. Three aspects were considered to indicate a self-reported, but, more objective change in the work environment of the employee: function, work schedule, working hours per week. A possible change in function (or workplace) was inventoried in each questionnaire. For work schedule a change was scored if an employee’s
work schedule changed between day and shift work between waves. Working hours were first recoded into part- (≤ 35 hours per week) and full-time (≥ 36 hours per week). A change was scored if working hours changed between part- and full-time between waves. Each change was summed to retrieve a total number of changes in the work environment. This sum was recoded into three categories: no, one, or two or more reported changes in the work environment.

**Statistical analysis**

First the descriptives of the ‘working’ and ‘returning to work’ sample at each time interval, including the number of missing values, were examined.

Confirmatory Factor Analysis (CFA) was used to test longitudinal ME/I. A 1-factor model was used for the NFR scale. For the CIS a 4-factor model, based on the subscales, was used including correlations between the factors. The weighted least square means and variance adjusted (WLSMV) estimator was used for categorical data as the NFR scale items were dichotomous and item responses on the CIS were heavily skewed (Rhemtulla, Brosseau-Liard, & Savalei, 2012). Each model was identified by using the first factor loading as anchor and constraining the latent means to be zero. To incorporate the longitudinal component into the CFA the latter latent factor of the (sub)scale was regressed on the latent factor at T0. Respondents were included if they had a valid response on at least one item (Lei & Wu, 2012). Modification indices were requested to indicate possible improvements for the factor structure (Kline, 2011). The final models for were used to test for ME/I. To evaluate model fit the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA) were used. For the CFI and TLI values above .95 indicate good model fit and values greater than .90 indicate acceptable model fit. For the RMSEA values smaller than .06 indicate good fit (Hu & Bentler, 1999). Significance of the $\chi^2$ was not interpreted as it was shown to be highly sensitive for large sample size (Meade, Johnson, & Braddy, 2008; Cheung & Rensvold, 2002). To test for full ME/I, factor loadings and thresholds of each item were constrained to be equal over time, freeing the the mean(s) of the latent factor(s) on the latter interval (Muthen & Asparouhov, 2012). If the full ME/I model did not differ from the configural model the former should be preferred (van de Schoot, Lugtig, & Hox, 2012). The following guidelines were therefore used to indicate a difference between models Meade et al. (2008), Marsh et al. (2010); $\Delta CFI \leq -.002$, $\Delta RMSEA \geq .007$, $\Delta TLI \leq 0$.

Intraclass correlation coefficient (ICC) and corresponding 95% confidence intervals (95% CI) were used to assess the reliability. A 2-way mixed effects model with an absolute agreement definition was used; ICC(A,1). An ICC > .70 is considered acceptable (Nunnally & Bernstein, 1994). Using the 95% CI it was examined whether ICC were significantly different between subgroups. As the NFR scale had a strong floor effect, a sensitivity analysis was conducted to inspect which effect thid had on the reliability. Employees were, therefore, excluded if they had at least one zero score on the NFR scale for the interval assessed. For comparative purposes, this sensitivity analyses was also conducted for the CIS (excluding CIS < 26).

The $SDC$, to assess the agreement, was defined as $1.96 \times \sqrt{2} \times SEM$, in which SEM is the standard error of measurement including systematic differences. The $SDC_{ind}$ was also expressed as a percentage of the range of the scale. The impact of the $SDC_{ind}$ was also graphically displayed in Bland and Altman plots for the ‘returning to work’ samples. These figures plot the difference between the two intervals against the score at the first measurement.
These plots enable the inspection of the assumption that the agreement is unrelated to the mean value on the instrument. To examine the $SDC_{\text{group}}$ of the instruments, the $SDC_{\text{ind}}$ was divided by $\sqrt{n}$.

**Results**

**Descriptive statistics**

In Table 1 the descriptives for the ‘working’ sample on each interval are given. Descriptives were based on T0, except the variables regarding the change in the work environment which were based on all measurements within the interval assessed. The distribution of most variables seems consistent for the different time intervals. There was, however, a small decrease in average NFR scale and CIS scores if the interval of the sample increased. The percentage of missingness was below 3% for each variable.

To analyse the reliability and agreement within a group of employees with a stable work environment, employees were included if they had no reported changes and no more than one perceived change (if only employees with neither perceived changes nor reported changes had been included, the sample size would have become too small). The number of employees classified as having such a stable work environment was 6,932 at the 4-month interval, 3,521 at the 1-year interval, and 1,373 at the 2-year interval (Table 2).

Descriptives were also provided for the ‘return to work’ sample (Table 2). Gender and education were still based on T0, while information on age, work schedule, and long-term illness were based on T3. The average score on the NFR scale and CIS were also based on T3. The percentage of missing data was below 6% for each variable. While not tested for significance especially the differences between the ‘working’ and ‘returning to work’ sample with respect to NFR and prolonged fatigue are noteworthy

**ME/I**

Models with the shortest time interval were modelled first as their estimation was most complex. Models with longer time intervals were based on these models. For the NFR the initial models had a poor fit. Two pairs of item residuals were therefore correlated, on the basis of the modification indices, also over time. The residual term of each item was, furthermore, correlated over time. Using this model for the NFR scale, ME/I was supported for each interval as model fit increased with the constraints of full ME/I (Table 3).

The initial CIS model had implausible parameters (i.e. not positive definite). Six pairs of item residuals were therefore correlated, on the basis of the modification indices, and the residual terms of each item was therefore correlated over time. This modified model was also used for the 1- and 2-year interval. As shown in Table 3 full ME/I was supported for each interval as model fit increased with the constraints of scalar ME/I. These results showed that both the NFR scale and CIS were ME/I over time. The scores on these instruments can, therefore, be meaningfully compared.
Table 1: Descriptive statistics for the working sample and its subsample including only employees with a stable work environment at different time intervals

<table>
<thead>
<tr>
<th></th>
<th>4-Month</th>
<th>1-Year</th>
<th>2-Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
<td>Mean (SD)</td>
<td>n (%)</td>
</tr>
<tr>
<td><strong>Working sample</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>41.26 (8.79)</td>
<td>41.79 (8.58)</td>
<td>42.10 (8.24)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>2,347 (26)</td>
<td>1,568 (24)</td>
<td>997 (23)</td>
</tr>
<tr>
<td>Male</td>
<td>6,825 (74)</td>
<td>4,977 (76)</td>
<td>3,432 (77)</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1,666 (19)</td>
<td>1,104 (17)</td>
<td>674 (16)</td>
</tr>
<tr>
<td>Medium</td>
<td>4,026 (45)</td>
<td>2,844 (45)</td>
<td>1,904 (44)</td>
</tr>
<tr>
<td>High</td>
<td>3,223 (36)</td>
<td>2,434 (38)</td>
<td>1,761 (41)</td>
</tr>
<tr>
<td>Long-term illness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>6,962 (78)</td>
<td>5,045 (79)</td>
<td>3,485 (80)</td>
</tr>
<tr>
<td>Yes</td>
<td>1,967 (22)</td>
<td>1,345 (21)</td>
<td>851 (20)</td>
</tr>
<tr>
<td>Work schedule</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daywork</td>
<td>7,601 (83)</td>
<td>5,521 (84)</td>
<td>3,834 (87)</td>
</tr>
<tr>
<td>Shiftwork</td>
<td>1,572 (17)</td>
<td>1,024 (16)</td>
<td>595 (13)</td>
</tr>
<tr>
<td>NFR</td>
<td>36.82 (31.85)</td>
<td>35.12 (31.33)</td>
<td>33.31 (30.65)</td>
</tr>
<tr>
<td>CIS</td>
<td>55.85 (22.73)</td>
<td>54.72 (22.36)</td>
<td>53.55 (22.03)</td>
</tr>
<tr>
<td>Perceived change in work environment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>5,221 (57)</td>
<td>2,454 (37)</td>
<td>772 (17)</td>
</tr>
<tr>
<td>1</td>
<td>2,265 (25)</td>
<td>1,815 (28)</td>
<td>972 (22)</td>
</tr>
<tr>
<td>&gt;1</td>
<td>1,687 (18)</td>
<td>2,276 (35)</td>
<td>2,685 (61)</td>
</tr>
<tr>
<td>Reported change in work environment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>8,324 (91)</td>
<td>5,055 (77)</td>
<td>2,896 (65)</td>
</tr>
<tr>
<td>1</td>
<td>794 (9)</td>
<td>1,022 (16)</td>
<td>814 (18)</td>
</tr>
<tr>
<td>&gt;1</td>
<td>55 (1)</td>
<td>468 (7)</td>
<td>719 (16)</td>
</tr>
<tr>
<td>Combination of perceived and reported change</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>5,004 (55)</td>
<td>2,115 (32)</td>
<td>639 (14)</td>
</tr>
<tr>
<td>Medium</td>
<td>4,128 (45)</td>
<td>4,143 (63)</td>
<td>3,217 (73)</td>
</tr>
<tr>
<td>High</td>
<td>41 (0)</td>
<td>287 (4)</td>
<td>573 (13)</td>
</tr>
</tbody>
</table>

Reliability

The reliability (i.e. ICC) for the NFR scale within the employees with a stable work environment in the ‘working’ sample was .78 for the 4 month interval (Table 4). The ICC for the CIS within the 4 month interval was .75. For longer intervals this reliability remained fairly stable.

The reliability was also examined for different subgroups among the employees with a stable work environment (Table 4). Older employees had higher ICC compared to younger employees for both scales. Male employees showed higher ICC on the CIS compared to females. This was also found for the NFR scale, but the finding was only significant for the 1 year interval. A significant difference between ICC on the CIS was found between employees with a stable and unstable private life while this difference was not significant for the NFR scale.
Table 2: Descriptive statistics subsamples including only employees with a stable work environment at different time intervals and the employees who return to work

<table>
<thead>
<tr>
<th></th>
<th>4-Month</th>
<th>1-Year</th>
<th>2-Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n (%)</td>
<td>Mean (SD)</td>
<td>n (%)</td>
</tr>
<tr>
<td><strong>Working sample: stable work environment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>41.63 (8.77)</td>
<td>42.65 (8.43)</td>
<td>43.59 (8.01)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>1,736 (25)</td>
<td>829 (24)</td>
<td>313 (23)</td>
</tr>
<tr>
<td>Male</td>
<td>5,195 (75)</td>
<td>2,692 (76)</td>
<td>1,060 (77)</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1,238 (18)</td>
<td>605 (18)</td>
<td>238 (18)</td>
</tr>
<tr>
<td>Medium</td>
<td>3,067 (45)</td>
<td>1,590 (46)</td>
<td>627 (46)</td>
</tr>
<tr>
<td>High</td>
<td>2,443 (36)</td>
<td>1,247 (36)</td>
<td>485 (36)</td>
</tr>
<tr>
<td><strong>Long-term illness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>5,339 (79)</td>
<td>2,750 (80)</td>
<td>1,082 (81)</td>
</tr>
<tr>
<td>Yes</td>
<td>1,425 (21)</td>
<td>699 (20)</td>
<td>257 (19)</td>
</tr>
<tr>
<td><strong>Work schedule</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daywork</td>
<td>5,814 (84)</td>
<td>3,038 (86)</td>
<td>1,217 (89)</td>
</tr>
<tr>
<td>Shiftwork</td>
<td>1,118 (16)</td>
<td>483 (14)</td>
<td>156 (11)</td>
</tr>
<tr>
<td><strong>NFR</strong></td>
<td>34.49 (31.43)</td>
<td>30.90 (30.31)</td>
<td>25.67 (28.52)</td>
</tr>
<tr>
<td><strong>CIS</strong></td>
<td>54.47 (22.40)</td>
<td>52.03 (21.80)</td>
<td>48.91 (21.07)</td>
</tr>
<tr>
<td><strong>Returning to work sample</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>42.83 (8.52)</td>
<td>42.73 (7.95)</td>
<td></td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>77 (30)</td>
<td>44 (29)</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>180 (70)</td>
<td>109 (71)</td>
<td></td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>83 (33)</td>
<td>40 (26)</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>120 (48)</td>
<td>75 (50)</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>49 (19)</td>
<td>36 (24)</td>
<td></td>
</tr>
<tr>
<td><strong>Long-term illness</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>128 (53)</td>
<td>77 (52)</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>115 (47)</td>
<td>71 (48)</td>
<td></td>
</tr>
<tr>
<td><strong>Work schedule</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Day work</td>
<td>199 (78)</td>
<td>122 (80)</td>
<td></td>
</tr>
<tr>
<td>Shift work</td>
<td>56 (22)</td>
<td>31 (20)</td>
<td></td>
</tr>
<tr>
<td><strong>NFR</strong></td>
<td>49.21 (35.56)</td>
<td>49.28 (35.64)</td>
<td></td>
</tr>
<tr>
<td><strong>CIS</strong></td>
<td>66.18 (24.91)</td>
<td>64.47 (24.97)</td>
<td></td>
</tr>
</tbody>
</table>

Reliability was also compared between the different levels of work environment stability (Table 4). Whilst not significant at each step the ICC showed a clear pattern, with higher ICC for higher levels of stability. For the NFR scale the combination of the reported and perceived changes in the work environment showed the strongest pattern, significant at each step. For the ‘returning to work’ sample the reliability of the NFR scale and CIS were comparable with the reliability of the instruments as found within the ‘working’ sample (Table 4). The relia-
### Table 3: Model fit and nested model comparisons for longitudinal measurement invariance/equivalence for the three different intervals for the NFR scale and the CIS

<table>
<thead>
<tr>
<th>Interval</th>
<th>Need for Recovery (NFR scale)</th>
<th>Prolonged Fatigue (CIS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-Month</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Configural</td>
<td>Configural</td>
<td>Configural</td>
</tr>
<tr>
<td>Scalar</td>
<td>Scalar</td>
<td>Scalar</td>
</tr>
<tr>
<td>Model (Constraints)</td>
<td>Model fit indices</td>
<td>Model fit indices</td>
</tr>
<tr>
<td>χ² (df)</td>
<td>CFI</td>
<td>TLI</td>
</tr>
<tr>
<td>0.000 - 0.006</td>
<td>634.87 (132)</td>
<td>0.976</td>
</tr>
<tr>
<td>0.007 - 0.013</td>
<td>883.54 (132)</td>
<td>0.974</td>
</tr>
<tr>
<td>0.014 - 0.020</td>
<td>751.11 (132)</td>
<td>0.974</td>
</tr>
<tr>
<td>0.021 - 0.027</td>
<td>549.00 (132)</td>
<td>0.974</td>
</tr>
<tr>
<td>0.028 - 0.034</td>
<td>515.27 (132)</td>
<td>0.974</td>
</tr>
<tr>
<td>0.035 - 0.041</td>
<td>12.7629 (824)</td>
<td>0.974</td>
</tr>
<tr>
<td>0.042 - 0.048</td>
<td>3.6955 (824)</td>
<td>0.974</td>
</tr>
<tr>
<td>0.049 - 0.055</td>
<td>12.7629 (824)</td>
<td>0.974</td>
</tr>
<tr>
<td>0.056 - 0.062</td>
<td>3.6955 (824)</td>
<td>0.974</td>
</tr>
<tr>
<td>0.063 - 0.069</td>
<td>12.7629 (824)</td>
<td>0.974</td>
</tr>
</tbody>
</table>

Note: The NFR scale and the CIS model fit indices.
bility of both scales was therefore also positive for employees who return to work after a long period of (sickness) absence.

For the sensitivity analysis, controlling for the floor effect, the $ICC$ were re-examined within the ‘working’ samples for employees with a stable work environment without a floor effect score on the NFR scale ($< 1$) or the CIS ($< 26$; Figure 1a). The $ICC$ dropped .09 for the NFR scale and .03 for the CIS. This pattern, in which the effect of the floor effect on the $ICC$ was stronger for the NFR scale compared to the CIS, was also found for the 1-year and 2-year interval (Table 4). Figure 1a illustrates the strong impact the floor effect has on the reliability, especially noteworthy for the NFR scale in which a large amount of employees scored zero on both occasions.

**Agreement**

For the NFR the $SDC_{ind}$ was 41.20 which equals the percentage it covers of the scale range. The $SDC_{ind}$ for the CIS was 31.10, covering 25.92% of the scale range. Figure 1b shows the Bland and Altman plots for the NFR scale and CIS within the ‘returning to work’ samples. Employees who had a difference score which exceeded the $SDC$, hence had a statistically substantiated change, were annotated in the plot. These plots also show that there is no relation between the mean NFR scale or CIS score and the agreement over time of the instruments. Figure 1c shows the $SDC_{group}$ for an increasing sample size. For the ‘returning to work’ samples, for example, a change of 3.10 on the NFR scale and 2.21 on the CIS can be reliably detected at $\alpha = 0.05$ for the 4-month interval ($n = 257$).

**Discussion**

With the growing burden of the psychosocial working environment, mental health complaints among employees are widespread (OECD, 2012). Occupational health surveillance should therefore include mental health outcomes which are related to the psychosocial work environment to facilitate preventive strategies targeted at prevention or limiting adverse (mental) health outcomes. The quality of two instruments, the NFR scale and CIS, measuring these consequences, NFR and prolonged fatigue respectively, was therefore assessed. Both the NFR scale and CIS have shown high levels of internal consistency (Bultmann et al., 2000; van Veldhoven, 2003). While this is an important quality criterion of an instrument, it does not automatically imply that the instruments can be used for monitoring purposes. The reproducibility of both scales was, therefore, examined in the current study. Reproducibility consists of the reliability and the agreement of an instrument over time which, respectively, indicate the quality of an instrument to differentiate between employees and to detect changes over time, both within individuals and groups (Terwee et al., 2007). Reliability was, furthermore, assessed within a heterogeneous working population. Doing so, this study is the first to give insight in the applicability of the instruments for monitoring purposes within the general working population, including various subgroups and employees who return to work after a long period of (sickness) absence. Examining reliability of the two instruments also for this latter group of employees gave further insight in the applicability of the NFR scale and CIS.

The reliability was, first of all, examined among employees with a stable work environment. Both scales showed high levels of reliability, which indicate low levels of measurement error over
### Table 4: Intraclass Correlations and their 95% Confidence Intervals of the CIS and NFR Scale the Different Samples and Subgroups at Each Time Interval

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Total</th>
<th>Working Sample: Stable Work Environment</th>
<th>Age</th>
<th>Gender</th>
<th>Education</th>
<th>Long-term Illness</th>
<th>Work Schedule</th>
<th>Change in Private Life</th>
<th>Floor Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-Month</td>
<td>.78 (.76; .79)</td>
<td>.76 (.74; .77)</td>
<td>.74 (.71; .76)</td>
<td>.75 (.74; .76)</td>
<td>.71 (.70; .73)</td>
<td>.72 (.69; .74)</td>
<td>.78 (.76; .79)</td>
<td>.74 (.72; .75)</td>
<td>.68 (.66; .70)</td>
</tr>
<tr>
<td>1-Year</td>
<td>.77 (.74; .79)</td>
<td>.74 (.70; .77)</td>
<td>.73 (.69; .76)</td>
<td>.76 (.74; .77)</td>
<td>.72 (.69; .73)</td>
<td>.74 (.72; .75)</td>
<td>.74 (.71; .77)</td>
<td>.74 (.71; .77)</td>
<td>.68 (.66; .70)</td>
</tr>
<tr>
<td>2-Year</td>
<td>.76 (.74; .78)</td>
<td>.73 (.69; .76)</td>
<td>.71 (.69; .73)</td>
<td>.76 (.74; .77)</td>
<td>.72 (.69; .74)</td>
<td>.74 (.72; .75)</td>
<td>.74 (.71; .77)</td>
<td>.74 (.71; .77)</td>
<td>.68 (.66; .70)</td>
</tr>
<tr>
<td>4-Month</td>
<td>.74 (.71; .77)</td>
<td>.73 (.69; .76)</td>
<td>.71 (.67; .74)</td>
<td>.75 (.73; .77)</td>
<td>.71 (.67; .75)</td>
<td>.72 (.67; .76)</td>
<td>.72 (.63; .79)</td>
<td>.72 (.63; .79)</td>
<td>.67 (.63; .70)</td>
</tr>
<tr>
<td>1-Year</td>
<td>.73 (.70; .76)</td>
<td>.72 (.68; .74)</td>
<td>.71 (.67; .74)</td>
<td>.75 (.73; .77)</td>
<td>.71 (.67; .75)</td>
<td>.72 (.67; .76)</td>
<td>.72 (.63; .79)</td>
<td>.72 (.63; .79)</td>
<td>.67 (.63; .70)</td>
</tr>
<tr>
<td>2-Year</td>
<td>.72 (.69; .75)</td>
<td>.70 (.65; .73)</td>
<td>.69 (.65; .72)</td>
<td>.74 (.72; .76)</td>
<td>.70 (.66; .72)</td>
<td>.71 (.67; .75)</td>
<td>.71 (.62; .78)</td>
<td>.71 (.62; .78)</td>
<td>.66 (.62; .70)</td>
</tr>
<tr>
<td>4-Month</td>
<td>.71 (.68; .73)</td>
<td>.68 (.63; .69)</td>
<td>.67 (.63; .69)</td>
<td>.73 (.71; .75)</td>
<td>.69 (.65; .71)</td>
<td>.70 (.65; .72)</td>
<td>.70 (.60; .73)</td>
<td>.70 (.60; .73)</td>
<td>.64 (.57; .71)</td>
</tr>
<tr>
<td>1-Year</td>
<td>.70 (.67; .72)</td>
<td>.66 (.61; .67)</td>
<td>.65 (.61; .66)</td>
<td>.72 (.70; .74)</td>
<td>.68 (.64; .70)</td>
<td>.69 (.64; .70)</td>
<td>.69 (.59; .69)</td>
<td>.69 (.59; .69)</td>
<td>.64 (.57; .71)</td>
</tr>
<tr>
<td>2-Year</td>
<td>.69 (.66; .71)</td>
<td>.65 (.60; .66)</td>
<td>.64 (.60; .65)</td>
<td>.71 (.69; .73)</td>
<td>.67 (.63; .69)</td>
<td>.68 (.63; .69)</td>
<td>.68 (.58; .68)</td>
<td>.68 (.58; .68)</td>
<td>.63 (.56; .69)</td>
</tr>
</tbody>
</table>

Note: Per interval and grouping variable different subscripts indicate significant differences between groups, if no subscripts are given all groups are equal.
Table 4 continued

<table>
<thead>
<tr>
<th>Perceived change in work environment</th>
<th>4-Month</th>
<th>1-Year</th>
<th>2-Year</th>
<th>4-Month</th>
<th>1-Year</th>
<th>2-Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.78 (.76; .79)</td>
<td>.76 (.74; .78)</td>
<td>.76 (.73; .79)</td>
<td>.75 (.74; .77)</td>
<td>.72 (.71; .74)</td>
<td>.77 (.74; .80)</td>
</tr>
<tr>
<td>1</td>
<td>.75 (.73; .77)</td>
<td>.73 (.71; .75)</td>
<td>.70 (.66; .73)</td>
<td>.74 (.72; .76)</td>
<td>.69 (.67; .72)</td>
<td>.67 (.64; .71)</td>
</tr>
<tr>
<td>2 or more</td>
<td>.67 (.64; .70)</td>
<td>.68 (.66; .70)</td>
<td>.62 (.59; .65)</td>
<td>.68 (.65; .70)</td>
<td>.67 (.65; .69)</td>
<td>.62 (.59; .64)</td>
</tr>
<tr>
<td>Reported change in work environment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.77 (.76; .78)</td>
<td>.75 (.74; .77)</td>
<td>.71 (.68; .74)</td>
<td>.74 (.73; .75)</td>
<td>.71 (.70; .72)</td>
<td>.69 (.66; .71)</td>
</tr>
<tr>
<td>1</td>
<td>.66 (.61; .70)</td>
<td>.67 (.63; .70)</td>
<td>.60 (.55; .65)</td>
<td>.71 (.67; .75)</td>
<td>.69 (.66; .72)</td>
<td>.59 (.54; .64)</td>
</tr>
<tr>
<td>2 or more</td>
<td>.48 (.23; .67)</td>
<td>.63 (.57; .68)</td>
<td>.60 (.55; .64)</td>
<td>.53 (.29; .70)</td>
<td>.64 (.58; .69)</td>
<td>.65 (.60; .69)</td>
</tr>
<tr>
<td>Combination of perceived and reported change</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>.78 (.76; .80)</td>
<td>.77 (.75; .78)</td>
<td>.77 (.73; .80)</td>
<td>.75 (.74; .77)</td>
<td>.72 (.70; .74)</td>
<td>.77 (.74; .80)</td>
</tr>
<tr>
<td>Medium</td>
<td>.72 (.70; .73)</td>
<td>.72 (.70; .73)</td>
<td>.66 (.63; .69)</td>
<td>.72 (.70; .73)</td>
<td>.69 (.68; .71)</td>
<td>.64 (.62; .67)</td>
</tr>
<tr>
<td>High</td>
<td>.46 (.16; .68)</td>
<td>.58 (.49; .65)</td>
<td>.57 (.51; .62)</td>
<td>.53 (.24; .72)</td>
<td>.59 (.51; .66)</td>
<td>.62 (.56; .67)</td>
</tr>
<tr>
<td>Total</td>
<td>.75 (.69; .80)</td>
<td>.78 (.71; .83)</td>
<td>.77 (.71; .81)</td>
<td>.75 (.67; .81)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Per interval and grouping variable different subscripts indicate significant differences between groups, if no subscripts are given all groups are equal.
Figure 1: Panel a shows the scatterplot with associated histograms for the 4-month interval for the CIS and NFR scale, in which the floor effect is grayed out. The main reliability analysis are performed on the total range, the sensitivity analysis excluded the grayed out area. Panel b shows the Bland and Altman plots of the CIS and NFR scale for the ‘returning to work’ sample for the 4-month and 1-year interval. Panel c shows the $SDC_{group}$ of the CIS and NFR scale for increasing sample size in which the x-axis is on a logarithmic scale.
time. Sensitivity analysis, controlling for floor effects however showed that the reliability of the NFR scale substantially decreased if only scores above zero were considered. The assumption that the reliability of the instrument is unrelated with the standing on the underlying construct was, therefore, violated. This shows that while the NFR scale is capable to differentiate between employees with high and low levels of NFR among employees with a stable work environment, further differentiation within smaller contrasts is cumbersome. For the CIS the floor effect on the reliability was limited, which substantiates differentiation across the whole range of the CIS. The CIS can be, therefore, regarded as a good instrument for differentiating employees on prolonged fatigue for monitoring purposes.

Reliability was high for almost all subgroups, with some noteworthy differences. On average, for example, older employees showed higher levels of reliability compared to younger employees. This finding could indicate that older employees have more stable levels of NFR and prolonged fatigue. Female employees were found to have lower reliability scores compared to male employees, especially for the CIS. It was, furthermore, found that the reliability of the CIS, but not for the NFR scale, was significantly lower for employees with a change in their private life compared to employees without. As NFR concerns a work domain specific construct, changes in other domains will have less influence on its reliability over time. Subsequently, changes in the work environment were found to be associated with a decreased reliability of the NFR scale. That is, employees who had an unstable work environment had lower ICC compared to employees with a stable work environment. This effect was present both for reported and perceived changes, but became especially apparent when those two sources were combined. For the CIS this effect was also present but less pronounced.

To substantiate the reproducibility of the NFR scale and CIS over time longitudinal ME/I was assessed (Vandenberg & Lance, 2000). CFA showed that both factor structures were equivalent/invariant over time for the different time intervals. Comparing scale scores over time, including the assessment of reliability and agreement, is therefore substantiated (Millsap, 2011). It should be noted that the CFA also showed the presence of item residuals correlations. This indicates that items have a dependency over time which is not explained by the association of the factors of time. While such a finding is not exceptional, it does show that correlation and reliability of the construct over time is augmented by this factor (Cole, Ciesla, & Steiger, 2007).

Implications for practice and research

While reliability was substantial for most subgroups of workers, as well as within employees who return to work after a long period of absence, monitoring NFR and prolonged fatigue within subgroups with lower levels of reliability should be exercised more cautiously. That is, it is more likely that differentiation between employees within these groups is erroneous. This information should, first of all, be taken into account if the instrument is implemented and interpreted in practice. Caution should be exercised, for example if the scale is applied within an environment in which changes in the work environment are expected (e.g. reorganisation), as this may decrease the reliability. A potential solution would be the use of a Likert scale instead of a dichotomous response category, as this could potentially also capture those nuances at the lower levels of the NFR continuum (Moriguchi, Alem, van Veldhoven, & Coury, 2010). Second, these findings could also instigate additional measures, such as shortening the time interval between assessments or including additional measures. Hence, while the reliability
of both instruments is satisfactory, especially for the CIS, their strength particularly lies when incorporated in a broader monitoring test battery assessing different aspects associated with the mental health of employees. Doing so would not only increase the reliability of the monitoring but also increase the amount and depth of information obtained in comparison with the assessment of only one scale.

To assess the ability of the scales to detect change over time within employees, the $SDC$ was evaluated within employees with a stable work environment. The $SDC$ of the CIS and especially the NFR scale were high. This does, however, not imply that the NFR scale would be unfit for monitoring purposes. The NFR scale is most suitable to identify employees who show a change from low levels of NFR towards more elevated levels of NFR. This is not only in line with the sensitivity analysis, but also with the conceptualisation of NFR (van Veldhoven, 2003): “…gradually the concept loses its salience to more serious long term …” Nevertheless, the $SDC$ remains large. To objectively assess its relevance it should be compared with a change that is found to be clinically relevant for the scale (de Vet et al., 2011). While absent for the NFR scale and CIS it seems implausible that such a clinically relevant change would surpass the $SDC$ found in the current article. Future studies should, therefore, examine the clinically relevant change and address it in relation to the $SDC$ of the scales to provide information on the responsiveness and interpretation of the instruments. If the purpose of monitoring is to detect change within groups of employees, both scales show to be good instruments. These instruments are therefore suitable to apply within, for example, public health and associated research. The effect of policy changes can be established, for example, on a group level.

**Strengths and limitations**

The main strength of the current study is the large sample within a heterogeneous group of employees. To inspect the reliability and agreement within a group of employees in which no change was expected, a selection of this sample was made. This selection was based on the notion that if employees have a stable psychosocial work environment, levels of NFR and prolonged fatigue were expected to be stable over time. While no direct measures of the psychosocial work environment were available both reported but also perceived changes were assessed. Employees within a stable working environment had no reported changes and no more than one perceived (subjective) change. As such, employees had no change in, employer, function, workplace, or working hours and only a small change was allowed in the perceived work environment. While this does not necessarily guarantee that the psychosocial working environment is stable, it is expected that no significant differentiation in the psychosocial working environment occurs, especially as perceived changes are also limited. Inclusion of both reported and perceived changes is, therefore, an important strength of the current study. This was stressed by the finding that results were most pronounced if these two aspects were analysed in conjunction. The 4-month measurement interval over a prolonged period, enabled the examination of the reliability and agreement within a context which closely mirrored the actual implementation of the instruments. This interval is also likely to reflect timing of monitoring in practice, as short intervals are impractical for monitoring employees. That is, the number of employees being monitored is often voluminous and changes do not often occur – which is also substantiated by the high reliability for the 2-year interval. A short period (i.e. 2 weeks) between assessments would have been useful, however, for descriptive purposes.
Within such a short period a change in the work environment is, in general, unlikely. Short intervals as these are, however, difficult to reconcile with the etiological purpose of the cohort (Kant et al., 2003). It should, further be noted that the heterogeneity of the cohort also results in a higher SDC (de Vet et al., 2011).

Conclusion

Since each aspect of the psycho social work environment may have a different time course of cause and effect, monitoring associated mental health outcomes such as NFR and prolonged fatigue is of crucial importance within occupational health surveillance and associated research (OECD, 2012; Grooten et al., 2016; Eurofound, 2012). The goal of monitoring can be considered to be threefold: differentiation between employees (i.e. reliability), detecting changes within an employee (i.e. agreement over time), and detecting changes within a group of employees (i.e. agreement over time). The present study shows that both the NFR scale and CIS have somewhat limited capabilities to detect changes within individual employees, while detecting changes on a group level and differentiating between employees with low and high levels of NFR and prolonged fatigue would be well justified. To outweigh these limitations both instruments seem especially valuable as part of a more extensive test battery, including measures of other mental health outcomes related to the psychosocial work environment and measures of this psychosocial work environment itself. As such a reliable and comprehensive monitoring can be established. Justification of the scales for monitoring purposes does not only hold for the working population but also for employees who return to work after a long period of (sickness) absence. It should be noted that the reliability of NFR scale is hampered by its floor effect. While future research should further substantiate the application of the NFR scale and CIS as monitoring instrument, especially with regards to the sensitivity of the instrument, the current article clearly gives an indication of the potential of both instruments for monitoring purposes.

References


van Amelsvoort, L. G. P. M., Kant, I. J., Bultmann, U., & Swaen, G. M. H. (2003). Need for recovery after work and the subsequent risk of cardiovascular disease in a working population. Occupational and Environmental Medicine, 60(90001), i83–i87. doi:10.1136/oem.60.suppl_1.i83


The trait and occasion components of fatigue and their association with sickness absence

Hoofs, H., Jansen, N. W. H., Jansen, M. W. J., & Kant, IJ.

doi:10.1016/j.jpsychores.2017.09.007
Abstract

Objective

Fatigue is an important aspect within various health domains, including work. To correctly understand and treat high levels of (prolonged) fatigue it is important to breakdown the state of fatigue into a time-varying (occasion) and -invarying (trait) component. This also applies for its relation with other adverse health outcomes such as sickness absence.

Methods

Longitudinal data from the Maastricht Cohort Study (MCS) study was used, which assessed fatigue across 4-month intervals using the Checklist Individual Strength (CIS).

Results

It was found that the occasion components explains 27.60% (95%-CI [25.80%; 29.40%]) of the variance of fatigue and the trait component 71.00% (95%-CI [69.00%; 72.90%]). The trait component was, furthermore, found to be a significant predictor of sickness absence.

Conclusion

Fatigue has a considerable time-invariant component. As this component is also related with other adverse health outcomes, preventive measures and treatment should takes this breakdown into account.
Introduction

Fatigue is a common condition within the general population and is an important health problem (Bultmann, Kant, Kasl, Beurskens, & van den Brandt, 2002; Pawlikowska et al., 1994; Ranjith, 2005). Fatigue is, however, not a discrete disorder, but ranges from mild, frequent complaints to severe and prolonged characteristics. This multifaceted prolonged fatigue is found prevalent in the general population and can be a disabling condition being associated with various adverse health outcomes (Akerstedt, Axelsson, Lekander, Orsini, & Kecklund, 2014; Nijrolder, van der Horst, & van der Windt, 2008). Also within the working population, fatigue plays an important role as it is found to be related to various negative health outcomes, such as sickness absence and subjective health complaints (Bultmann, Nielsen, Madsen, Burr, & Rugulies, 2012; Janssen, Kant, Swaen, Janssen, & Schroer, 2003; de Raeve, Kant, Jansen, Vasse, & van den Brandt, 2009). This high impact, in combination with an estimated 2-week prevalence rate of fatigue of around 38%, makes it an important concept within occupational health (Ricci, Chee, Lorandeau, & Berger, 2007). While its importance seems irrefutable, the dynamics of fatigue are, however, poorly understood. That is, on the one hand, fatigue could be seen as a rather static construct, based on high correlations over time, with its state found to be rather robust (Akerstedt et al., 2014; Nijrolder et al., 2008). This is also in line with findings which show that treatment of prolonged fatigue is difficult (Leone, Huibers, Knottnerus, & Kant, 2008). On the other hand, however, varying prevalence rates per sector, trade, and company and associated dynamic prognostic factors seem to suggest a more time-varying concept (Kant et al., 2003). Based on these findings, it is likely that, on the continuum of fatigue employees show variation between occasions (time varying), but also to posit a certain trait-like (time-invariant) construct. Understanding these dynamics of fatigue is crucial for an enhanced understanding of the concept, its potential treatment and prevention.

To grasp the dynamics of fatigue it is important to disentangle its state. The state of fatigue is the observed level of fatigue of an employee at a specific moment in time. This observed state is, however, a combination of a time-invariant component and a component that varies over time. If measured longitudinally the observed level of fatigue can be disentangled into this time-invariant and time-varying component using a trait state occasion (TSO) model. The time-invariant component is the trait of fatigue which is stable over time. The trait component can, therefore, be regarded as the ‘baseline’ of an individual. A high trait implies, in general, high levels of fatigue. The trait component as such is a static characteristic of an individual. To fully account for the observed state, however, the trait should be combined with the occasion-specific level of fatigue. The occasion component can be thought of as the result of variations in the (working) environment of an employee. This trait-occasion breakdown illustrates that, for individual employees, the same score may have different compositions. That is, an employee with an observed level of fatigue of ‘15’ can have a high trait fatigue (i.e. ‘12’) and low occasion fatigue (i.e. ‘3’) or have a low trait fatigue (i.e. ‘4’) and a high occasion fatigue (i.e. ‘11’). It should be noted, however, that in this simple example the former employee will have a higher state of fatigue as her or his trait level is higher. In other words, the scores of this employee will fluctuate around this baseline as a function of the current occasion component.

The disentanglement of the state of fatigue reveals whether it is strongly trait-like, entails a strong occasion-specific variance, or somewhere in between. For example, variance in depres-
The trait and occasion components of fatigue are explained for 46% by a trait component and for 48% by an occasion-specific component (Prenoveau et al., 2011). The remaining 6% is explained by the auto-regressive (regression between consecutive occasions) component which connects successive occasions. Depending on the spacing between intervals, rapid successive occasions are more alike than more distant successive occasions (Cole, Martin, & Steiger, 2005). Disentangling fatigue into a trait and occasion component can enhance further understanding of the concept of fatigue, which in turn can enhance successful development and implementation of preventive measures (Hertzog & Nesselroade, 2003; Ployhart & Vandenberg, 2010). A strong time-varying component would, for instance, merit a different approach for effective treatment compared to a concept in which the trait component is most prominent. Moreover, employees with a different composition of occasion-specific and trait fatigue are likely to be receptive for different sorts of treatments and interventions. Occasion-specific fatigue could be associated with time-varying aspects such as a (temporary) high job strain. Potential interventions should, therefore, target the occasional aspects. As a high trait of fatigue is more static over time, more thorough interventions would be preferred (e.g. cognitive behavioural therapy).

Besides an enhanced understanding of fatigue and its potential treatment, disentanglement of fatigue into an occasion and trait component can also shed light on its relation with other health outcomes (Prenoveau et al., 2011). A significant health outcome is sickness absence which is a source of marked distress for the employee and has high direct and indirect costs for employers, employee, and society (Eurofound, 2016). Sickness absence is, as such, a pressing problem urging for preventive measures. Studies focus, therefore, on the observed state of fatigue and its association with sickness absence (e.g. Bultmann et al., 2012; Janssen et al., 2003). This association between fatigue (i.e. its state) and sickness absence can, however, also be disentangled into the association between sickness absence with the (a) occasion and the (b) trait component of fatigue using the TSO model. This disentanglement may greatly clarify the dynamics that form the association between fatigue and sickness absence (Cole & Maxwell, 2009; LaGrange et al., 2011). Furthermore, it gives insight into the role of the different components in the development of sickness absence and their potential as preventive measure (Cole, 2006). That is, different sorts of preventive measures are likely to be more or less successful depending on whether the occasion or the trait component is the dominant factor in the association between fatigue and sickness absence.

To gain a deeper understanding of the relation between fatigue and sickness absence it is important to distinguish different forms of absence. In the literature short-term sickness absence and a high absence frequency are found to be partly related to attitude (Kristensen, 1991). Long term sickness absence is suggested to be related to poor health and inability to perform work tasks (Feeney, North, Head, Canner, & Marmot, 1998). In this context, long-term sickness absence is often referred to as a primarily involuntary and necessary absence measure, while short term absence and absence frequency can be seen as primarily voluntary absence measures. This difference shows how important it is to differentiate the underlying mechanisms of both forms of sickness absence. Notice that, the relation between the occasion and trait component of these different forms of sickness absence could differ as well. In conclusion, both the breakdown of fatigue into an occasion and trait component and the breakdown of the association between fatigue and short- and long-term sickness absence across these components are crucial for a deeper understanding on how a reduction of fatigue may lead to the prevention and reduction of sickness absence. The goal of the current study
The trait and occasion components of fatigue is, therefore, twofold:

1. Breakdown the observed state of fatigue, which is longitudinally measured with an interval of 4 months, into a trait and occasion component.

2. Breakdown the association of fatigue and short- and long-term sickness absence among these trait and occasion component.

Methods

Sampling and procedures

This study is based on data from the ongoing Maastricht Cohort Study (MCS), using its first 10 waves (T0-T9). The MCS was set up in May 1998 including 12,140 participants from 45 different companies. At baseline measurement, all included participants were aged between 18 and 65 (Kant et al., 2003; Mohren, Jansen, van Amelsvoort, & Kant, 2007). Written informed consent was obtained from all participants. The study was of a strict observational nature and was conducted in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki. Between wave T8 and T9, which was a 16-month interval, company absence records were also available to measure sickness absence. These records were available for 13 of the 45 companies, covering 6,603 employees (21.53% female) at baseline (T0). It was, furthermore, necessary that employees did not switch companies as otherwise incomplete information was present regarding sickness absence during this period. Employees were therefore excluded if they were non-respondent on any of the 9 waves. Spacing between the waves was 4 months, except between T8 and T9 for which the interval was 16 months.

All employees who returned the baseline questionnaire received the two short questionnaires T1 (response rate [rr] 87.61%, n = 5,785) and T2 (rr = 79.48%, n = 5,248) as well. Employees returning the baseline questionnaire and at least one of the short questionnaires received the extensive questionnaire T3 (rr = 71.83%, n = 4,743). Employees returning the T3 questionnaire also received the short questionnaires T4 (rr = 67.01%, n = 4,425) and T5 (rr = 63.32%, n = 4,181). Employees who returned the questionnaire at T3 and at least one of the consecutive short questionnaires also received the extensive questionnaire T6 (rr = 57.50%, n = 3,797). Again, employees returning the T6 questionnaire also received the short questionnaires T7 (rr = 54.93%, n = 3,627) and T8 (rr = 52.76%, n = 3,484). The T9 questionnaire was the last wave included in the current study (rr = 39.59%, n = 2,614). For a detailed description of the MCS study population see Kant et al. (2003).

From these 2,614 employees only the 2,316 employees were included who did not indicate a change of employer. Employees were additionally excluded if they were pregnant or had multiple jobs. Pregnancy could highly influence the fatigue measurement and for employees with multiple jobs not all sickness absence information was present. This resulted in a final study population of 2,168 employees.

Measures

Prolonged fatigue. Fatigue was measured with the Checklist Individual Strength (Beurskens et al., 2000; Vercoulen et al., 1994). This questionnaire includes four subscales of which only the subjective fatigue experience (8 items) was used in the present study. Each item includes
The trait and occasion components of fatigue

Figure 1: Trait-state-occasion (TSo) model showing a compromised version of the model. The mean structure is excluded for reasons of clarity. The different segments of the model (measurement, TSo, method, and regression) are also indicated. The grey part is included if the association with a different construct is inspected, in which the dotted arrow represents the classic representation of the relation between the observed (state) of fatigue and sickness absence without the disentanglement into an occasion and trait component.
a statement (e.g.: “Thinking requires effort”) applying to the last two weeks which has to be scored on a 7-point Likert scale (yes, this is true to no, this is not true). Within the MCS the Cronbach’s $\alpha$ was .93 (Bultmann et al., 2000). Due to model complexity issues it was not possible to construct a model incorporating all 20 items and four latent factors at once. For the main analysis, therefore, the subjective fatigue subscale, with its corresponding 8 items, was used to illustrate the dynamics of an important aspect of fatigue also in relation to sickness absence.

**Sickness absence.** The use of objective absence data was expected to reduce self-report bias. Sickness absence data were expressed in calendar days. Partly resumption of work or working on a therapeutically basis counted as full sickness absence. Three dichotomous operationalisations of sickness absence were used: (a) Spell (or streak) duration, 21 consecutive sickness days or more vs. less than 21 consecutive sickness days. Employees who reported sick 21 days in a row were coded as having a sickness absence spell. The 21st day of this spell was subtracted from 15 January 2001 to compute the number of days until this event. (b) Total duration, a total of 42 sickness days. Sickness days were not required to be consecutive within this operationalisation. The 42nd sickness day was subtracted from 15 January 2001 to compute the number of days until this event. (c) Frequency, 3 (or more) different spells of sickness absence. Regardless of the length of these spells employees who had three different spells were coded as having a sickness absence frequency. The start of the third spell was subtracted from 15 January 2001 to compute the number of days until this event.

**Confounders.** The following factors were assessed at the baseline measurement: Age, gender, educational level, and long-term illness. Educational level was recoded into three categories: low, medium, and high. The presence of a long-term illness was assessed by self-report (yes/no) of 19 frequent medical condition (groups), such as cancer, mental problems, and multiple sclerosis. Smoking status at T8 was included as confounder using a dichotomous question: “Do you smoke daily? (yes/no)”. T8 was selected for smoking status at it is less static compared to the other confounders, diminishing its effect if smoking status at T0 was regressed on sickness absence between T9 and T8.

**Statistical analysis**

The TSO model as specified by Cole et al. (2005) was used (Figure 1). To our knowledge, this is the only suited model to disentangle a latent factor (state) into an occasion-specific factor and a trait factor (Prenoveau, 2016). It furthermore included an auto-regressive (AR) component for the occasion factors. Each occasion factor was, therefore, regressed on the previous occasion factor. In addition, a method factor was included for each item, except the first item (M-1 approach), as suggested by Prenoveau (2016). Within this TSO model the state score would, as such, be the observed score for an employee at the corresponding wave. This is, subsequently, disentangled into a trait component, which is the shared variance of all states over all waves (i.e. baseline), and an occasion component, which is the occasion or unique variance not predicted by the trait component. For TSO models it is important that the factor structure is measurement equivalent/invariant (ME/I) over time (Cole, 2012). For the CIS this ME/I was already demonstrated (unpublished results). Full ME/I was, therefore, imposed on the baseline TSO models of both scales (i.e. $\lambda_{10} = \lambda_{11} = \ldots = \lambda_{18}$).
For the first goal of the analysis, breakdown of the observed state of fatigue into a trait and occasion component, several assumptions were tested for each TSO model through the inclusion of several constraints (Cole, 2012; Prenoveau, 2016). Model 1A tested for homogeneity of the auto-regressive (AR) pathways and occasion variance ($\beta_{0-1} = \beta_{1-j} = \ldots = \beta_{j-s}; \sigma^2_{O1} = \sigma^2_{O1} = \ldots = \sigma^2_{O8}$). Model 1B examines equivalence of the trait component over time ($\lambda_{T0} = \lambda_{T1} = \ldots = \lambda_{T8}$). Model 1C tested for mean differences in the occasion component over time ($E_{T0} = E_{T1} = \ldots = E_{T8}$). Model 2A tested whether the auto-regressive component had a significant contribution to the model ($\beta_{0-1} = \beta_{1-j} = \ldots = \beta_{j-s} = 0$). Model 2B tested whether the trait component had a significant contribution to the model ($\lambda_{T0} = \lambda_{T1} = \ldots = \lambda_{T8} = 0$). It was not possible to test a model without occasion factor residuals as this factor also includes the error term of the state factor. Based on these different models, a final TSO model was selected for the subjective fatigue scale. To further illustrate the TSO model, this final model was applied to three different age groups to detect any difference in the breakdown of fatigue.

To evaluate model fit the Comparative Fit Index ($CFI$), Tucker-Lewis Index ($TLI$), and Root Mean Square Error of Approximation ($RMSEA$) were used. For the $CFI$ and $TLI$, values above .95 indicate good model fit and values greater than .90 indicate acceptable model fit. For the $RMSEA$ values smaller than 0.06 indicate good fit (Hu & Bentler, 1999). Significance of the $\chi^2$ was not interpreted as it was shown to be highly sensitive for large sample size (Meade, Johnson, & Braddy, 2008; Cheung & Rensvold, 2002).

It was tested if the hierarchical models differed from each other. If models do not differ from each other the more parsimonious should be preferred (van de Schoot, Lugtig, & Hox, 2012). As $\chi^2$, $\Delta \chi^2$ was also highly sensitive for sample size (Meade et al., 2008; Cheung & Rensvold, 2002). In the current study the following guidelines were used to indicate a difference between models Meade et al. (2008), Marsh et al. (2010); $\Delta CFI \leq -0.002$, $\Delta RMSEA \geq 0.007$, $\Delta TLI \leq 0$.

For the second goal of the study, disentangling the association of fatigue and sickness absence among the trait and occasion component, a survival analysis was performed. Since individual record linkage with the company registers on sickness absence was established, survival analyses were used as this makes the best use of the presence of this detailed information. Both the occasion component at T8 and the trait component were used in this analysis to predict sickness absence in the different operationalisations (spell, frequency, and duration). First a crude model was estimated followed by a model including several confounders. Estimates were given in the form of a standardised hazard ratio (HR) with corresponding 95% confidence intervals (CI). For the survival analyses employees were excluded if they were sick during the first week after the questionnaires at wave T8 were sent out, or if they indicated that they were on sick-leave at the time they filled out the questionnaire (excluding prevalent cases). An additional model to estimate the HR of the state and sickness absence will also be provided to show the ‘classical’ approach relating the observed state of fatigue with sickness absence (i.e. the dotted line in Figure 1).
The trait and occasion components of fatigue

<table>
<thead>
<tr>
<th>Model</th>
<th>Model fit indices</th>
<th>Nested model comparisons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$ (df)</td>
<td>CFI</td>
</tr>
<tr>
<td>Baseline</td>
<td>4,433.65 (2549)*</td>
<td>.982</td>
</tr>
<tr>
<td>1A</td>
<td>4,504.09 (2564)*</td>
<td>.982</td>
</tr>
<tr>
<td>1B</td>
<td>4,470.98 (2557)*</td>
<td>.982</td>
</tr>
<tr>
<td>1C</td>
<td>4,535.04 (2557)*</td>
<td>.981</td>
</tr>
<tr>
<td>2A</td>
<td>4,756.24 (2557)*</td>
<td>.979</td>
</tr>
<tr>
<td>2B</td>
<td>6,215.78 (2558)*</td>
<td>.966</td>
</tr>
<tr>
<td>Final</td>
<td>4,632.28 (2580)*</td>
<td>.981</td>
</tr>
</tbody>
</table>

Results

Descriptives

Mean age at T8 was 43.81 years ($SD = 7.29$ years) and 19.05% of the employees were female. For educational level there were three categories, low (19.94%), medium (47.77%), and high (32.29%). 29.00% of the employees reported at least one long-term illness. 22.10% of the employees indicated that they were a current smoker at T8.

Breakdown of the observed state of fatigue

The baseline model, based on all measures from T0 to T8, showed a good fit (Table 1). Model 1A through 1C did not show a substantial decrease, compared to the baseline model, in model fit. Homogeneity of the auto-regressive pathways, occasion variance, occasion means, and trait stability was therefore substantiated. Model 2A, but especially model 2B, showed a decrease in model fit compared to the baseline model. All variance components had therefore a significant contribution in the final TSO model. The AR component was, however, small at 1.50% (95%-CI [1.00%; 2.00%]) which was also reflected in the relatively minor decrease in model fit in model 2A compared to the baseline model. The occasion and trait variance were more substantial at 27.60% (95%-CI [25.80%; 29.40%]) and 71.00% (95%-CI [69.00%; 72.90%]) respectively. The method factor, additionally, explained between 7.56; 16.24% of the variance in the different indicators.

To illustrate the descriptive purpose of the TSO model. This final model was stratified for three age groups. All parameters were constrained between these three groups except the TSO part of the model. This model showed that with increasing age the occasion variance decreases and the trait variance increases (Figure 2). A Wald test showed, however, that this trend was not significant ($p = .22$).

These analyses were also repeated for the presence of a long-term illness. These results show that the trait component of fatigue is more stable for employees with at least one long-term illness compared to employees without a long-term illness ($p < .01$). It was, furthermore, found that this trait component of fatigue was significantly ($p < .01$) higher (0.65 on a standardised scale) within employees with at least one long-term illness compared to employees without...
a long-term illness. The trait component explained, as such, more variance and was higher within employees with a long-term illness compared to employees without a long-term illness (Figure 3).

**Disentangling the association of fatigue and sickness absence**

For the survival analysis 265 employees (12.22%) were excluded since they were a prevalent case (sick-leave) at wave T8. The survival rates of each operationalization of sickness absence are shown in Figure 4. This survival rate was subsequently regressed on the trait and occasion component at T8 of final TSO model of subjective fatigue. Table 2 shows the hazard ratios of each component (per standard deviation) for the different operationalisations of sickness absence. Regardless of the operationalisation of sickness absence, the relation between the trait component of fatigue and sickness absence was statistically significant. This in contrast to the occasion component where no statistically significant associations with sickness absence
The trait and occasion components of fatigue

**Figure 4:** (1 year) Survival plots for each of the operationalisations of sickness absence. Prevalent cases were excluded for these analysis.

**Table 2:** Crude and adjusted hazard ratios for different operationalisations of sickness absence on the occasion and trait component of fatigue, crude and adjusted hazard ratio for sickness absence on the state component of fatigue is given for illustrative purposes

<table>
<thead>
<tr>
<th></th>
<th>Spell (21 days)</th>
<th>Duration (41 days)</th>
<th>Frequency (3 times)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR^a</td>
<td>HR^b</td>
<td>HR^a</td>
</tr>
<tr>
<td><strong>Occasion</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.30 (0.93; 1.68)</td>
<td>1.40 (0.98; 1.82)</td>
<td>1.31 (0.91; 1.72)</td>
</tr>
<tr>
<td><strong>Trait</strong></td>
<td>1.37 (1.23; 1.51)</td>
<td>1.30 (1.15; 1.44)</td>
<td>1.41 (1.26; 1.56)</td>
</tr>
<tr>
<td><strong>State</strong></td>
<td>1.26 (1.17; 1.35)</td>
<td>1.23 (1.14; 1.33)</td>
<td>1.28 (1.19; 1.38)</td>
</tr>
</tbody>
</table>

^aCrude standardised (per increase in standard deviation) hazard ratio (HR) and 95% CI; ^bAdjusted standardised (per increase in standard deviation) hazard ratio (HR) and 95% CI, controlling for gender, age, education, long-term illness and smoking status (T8).

were observed. This pattern did not change when controlling for confounders, although the hazard ratio of the trait component within the frequency operationalisation was no longer significant (Table 2). It is, therefore, likely that the (significant) associations between the state of fatigue and sickness absence (Table 2), is caused by the trait component of fatigue.

**Discussion**

Understanding the dynamics of fatigue among employees is an important prerequisite for intervention and its application within preventive measures. Whilst the results need further replication, this study does warrant awareness for the differentiation between the occasion component and the trait component of fatigue.
Using a 4-month time interval it was found within the current 2-year time interval that the biggest part of fatigue was explained by its trait component (71%). The occasion component was, nevertheless, still substantial (28%). This is in contrast to the small auto-regressive component (2%) found in the TSO model. The observed state of fatigue, as such, is almost exclusively explained by a time-invariant component combined with an occasion-specific variance. The role of the previous occasion-specific variance on the subsequent measurement appeared to be trivial. As the trait component explained 71% of the observed state variance, fatigue might be more static than often assumed. Although this trait component could still be modifiable over an extended period, it does indicate a certain consistency among employees regarding their level of fatigue. As interventions could target the time-varying occasion component or the trait component it is important to take this breakdown into account. Whilst targeting the time-varying component may be more feasible it does account for only 28% of the observed state variance. Targeting the trait component might therefore be more sustainable, but probably requires a different intervention. The current study shows that this trait component of fatigue is significantly related to sickness absence in a 1-year period whilst the occasion component fatigue was not. This is consistent for the different operationalisations of sickness absence, with only the frequency operationalisation showing no significant association if confounders were included. The long-term (i.e. spell and total duration) sickness absence was, as such, predicted by the trait of fatigue also if the hazard ratio was adjusted for several confounders (i.e. age, gender, education, long-term illness, and smoking behaviour). This also corresponds with notion that long term sickness absence is a primarily involuntary and necessary absence measure (Feeney et al., 1998).

The absence of a significant association between the occasion of fatigue and sickness absence over time could be related to the small auto-regressive component between occasions. As the occasion-specific variance hardly influenced the occasion on the subsequent measurement, other longitudinal effects may also be less likely. Occasion-specific variance, as such, could be cross-sectionally associated with sickness absence but less so for longitudinal associations. It should be noted that, however, while there was a difference in significance, the effect size of both hazard ratios was comparable and, for some conditions, the effect of the occasion component even surpassed that of the trait component. The confidence intervals of the trait component were, however, smaller compared to those of the occasion component. This is in line with the finding that the trait component encompasses more variance compared to the occasion component. As with the treatment of fatigue itself, this finding provides important insight regarding the development and implementation of preventive measures targeting fatigue in order to prevent sickness absence. Bultmann et al. (2012), for example, point out that: “In view of the high social and economic cost implications of sickness absence, the early detection and treatment of fatigue . . . should be high on the stakeholders’ agenda”. Nevertheless, the current study points out that it is especially the time-invariant (trait) component of fatigue that predicts future sickness absence. Instead of altering fatigue, coping with fatigue could be therefore a potential viable mechanism to reduce fatigue-related sickness absence. This would take the significant role of the time-invariant component of fatigue into account.

Some strengths and limitations of the current study should be addressed. Only a limited number of confounders was selected for the survival analyses. These confounders, except smoking, were rather static. Including more dynamic factors, such as work-related factors (job characteristics and demands) and personal factors (i.e., alcohol use and smoking behaviour over
The trait and occasion components of fatigue

time), would be informative also within the general TSO model. It would, however, draw a heavy burden on the theoretical and computational component of the model. Placing such dynamic factor within the current TSO model is not straightforward, and would potentially diverge from the main goal of the current study – disentanglement of fatigue. This study does not intend to detect nor draw any conclusion regarding the source of fatigue. It intended to disentangle the state of fatigue into a trait and an occasion component, ‘regardless’ of the source of the state of fatigue of an employee. Two examples were, however, provided to illustrate potential difference in the trait and occasion breakdown between groups. As 29% of the employees had a least one long-term illness at baseline, and previous research (also within this cohort) has shown that the presence of a long-term illness strongly correlates with fatigue (Bultmann et al., 2002), employees with and without a long-term illness were compared. This illustration did not only show that the trait component was more profound within employees with a long-term illness, this trait component was on average also higher compared to employees without a long-term illness. Further research is needed, however, to examine this relation more closely (e.g. considering the specific type/severity of the illness).

A second example showed that whilst a pattern did emerge for different age groups, with an increase in the proportion of the occasion component for younger employees, it was not found to be significant. The inclusion of the (static) confounders within the survival analysis did, furthermore, alter the hazard ratios of the occasion and trait component into different directions. The hazard ratios for the occasion component increased in contrast to the decreasing hazard ratios of the (static) trait component. The nature of the confounder, therefore, seems to affect its adjustive role.

An important aspect within TSO models is the spacing between observations. As more wide intervals could have increased the occasion-specific variance, or vice-versa, it is important to take the 4-month interval of the 9 measurements into account. Longer time intervals could, potentially, alter the results, revealing a different breakdown of the state of fatigue. The same goes for a longer span of the time-window taken into account. As the time between intervals and the total time covered have an impact on the results, generalisation of the results to a longer period is not directly possible. It does give, nevertheless, an indication of the nature of fatigue for longer periods over time as a strong trait like construct. The timing between intervals plays also a role for the (16 month) interval for which sickness absence was assessed. A longer period interval could, potentially, introduce more ‘room’ for the effect of the (long-term) trait component. This study includes objective information for sickness absence based on company record. Different operationalisations of this sickness absence were furthermore assessed in the survival analysis to substantiate these analyses.

An important strength of the TSO model is that method variance can also be taken into account. This method component explains possible shared variance of the same item over time (Eid, Lischetzke, Nussbeck, & Trierweiler, 2003). It was found that method variance explained between 8-16% of the variance of the different indicators. Method variance is not only informative, it also results in a more ‘pure’ measurement component of the model as method bias is filtered out. It should be noted, however, that, whilst this limits the effect of method bias, it could still influence the results found in this study as the trait component could also be regarded (at least partly) as an overall method factor. This could inflate the variance that is attributed to this trait component. However, the result that the trait of the CIS predicts sickness absence, which was measured using company records, increases the validity of the components within the TSO model – as it would be counter intuitive to reason
that a common method factor could predict future sickness absence. Moreover, the current analysis is probably subject to selection bias. This could influence both the findings of the disentanglement of the state of fatigue and the findings regarding the prediction of sickness absence using the survival analysis. That is, for the prediction of future sickness absence, the effect of the occasion component could be confounded by the selective drop-out within the cohort. Whilst survival analysis is, under a normal statistical model, rather robust against this issue, the current hybrid confirmatory factor analysis (CFA) model, combined with the survival analysis, prohibited the inclusion of all employees even if they dropped out of the study. Due to model complexity, it was not possible to include all four subscales of the CIS. The short version of the CIS, consisting of the 8-item subjective fatigue subscale, correlates highly, however, with the overall score of the CIS. This short version of the CIS has shown good reliability and validity (Hewlett, Dures, & Almeida, 2011). Although the CIS has been shown to be a reliable and valid (single) subjective measure for fatigue supporting its implementation, it would be interesting to apply the TSO framework to different questionnaires assessing fatigue.

While structural equation models can be found to be complex at first, they ‘force’ a clear theoretical model of a construct. This theoretical model can reveal potentially important and often overlooked dynamics of a construct. In this context, the TSO model does justice to the complex and dynamic nature of most theories within psychosomatic research at the same time it makes them testable. This article clearly shows the theoretical benefit of the TSO model, when applied to fatigue. It noticeably distinguishes between the different dynamic components of fatigue. Often these components are only implicitly assumed to play a certain role during research. Clearly defining and assessing the components show that (even) a construct as fatigue is highly time-invariant. Even more so, it is found that it is this time-invariant component that predicts future sickness absence of employees. These findings should direct future research regarding fatigue and its treatment, pointing to the importance of the trait component of fatigue.

References


6 Evaluating model fit in Bayesian confirmatory factor analysis with large samples: Simulation study introducing the BRMSEA

Hoofs, H., Van de Schoot, R., Jansen, N. W. H., & Kant, IJ.

Educational and Psychological Measurement, 2017, Advanced online publication.
doi:10.1177/0013164417709314
Discussion
Discussion

Background

The goal of this thesis was to identify potential latent defects of measurement instruments and evaluate their impact within the public mental health. Therefore, some specific potential latent defects of measurement instruments were addressed using $\beta$-psychometrics. Latent defect were in the context of this thesis defined as: “Actors that can result in the violation of (important) quality criteria of an instrument measuring public mental health, without it being (directly) apparent”. Before the results are discussed a short overview of the findings in previous chapters will be provided.

In chapter 2 it was found that the setting influences assessment of psychosocial development in adolescents using the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1999). This violates direct cross-use of absolute scores, such as cut-off points, across settings. Within chapter 3 it was found that different factors (i.e. demographic, personal, and work-related) did not alter the measurement structure of the need for recovery (NFR) scale and the Checklist Individual Strength (CIS) within employees. Chapter 4 showed that these instruments are also invariant over time, both within a 4-month and a 1-year interval. The disentanglement of the state of fatigue in an occasion and a trait component in chapter 5 showed that the trait component was a significant predictor for different operationalisations of sickness absence. Chapter 6 shows that the Bayesian root mean square error of approximation ($RMSEA$) is a valid fit index within Bayesian confirmatory factor analysis (CFA). This fit measure should pave the way for future use of the Bayesian framework for the application of $\beta$-psychometrics.

This discussion will address these results in more detail and will explicate how these findings illustrate how to deal more successfully with latent defects within the measurement of public mental health. Three aspects are crucial for this, that is, identification of latent defects, awareness for latent defects, and strategies to prevent and deal with latent defects. Following this, some conceptual considerations regarding $\beta$-psychometrics will be addressed. The general strengths and limitations of this thesis along with its implications and implementation will be subsequently discussed. The last section will be the overall conclusion of this discussion and the thesis as such.

Identification of latent defects

For successful identification of latent defects two aspects are conditional, an active modelling approach and corresponding statistical tools. Active modelling starts with the critical assessment of the underlying hypotheses (quality criteria) of a measurement instrument. A very common assumption, for example, is that a measurement instrument functions irrespective of the group (or time window) that is assessed (i.e. construct validity). This is a common assumption as almost every single study that uses a measurement instrument includes some sort of a comparison. Often, however, this assumption goes unnoticed or is at least not addressed. This leaves the results somewhat vulnerable as no critical assessment is made regarding an important underlying hypothesis that should hold for the results to be valid. Furthermore, in those cases that it is addressed its discussion is often still somewhat sparse – or even illegitimate. That is, justification for a comparison between groups is often based on the reliability of an instrument within these groups individually. This does, however, state nothing about the construct validity of an instrument. Furthermore it does not consider the comparability of this construct validity across groups. It could be, for example, that an instrument functions...
Discussion

perfectly well in different groups while still measuring something (completely) different. To
assess this quality criterion of an instrument, active modelling calls for a clear hypothesis.
Without such a hypothesis it is deemed impossible to properly assess the underlying qual-
ity criteria of a measurement instrument and therefore its overall quality. Active modelling,
furthermore, provides a clear direction for the (statistical) assessment of the hypothesis.

The active modelling approach is clearly visible in chapter 3 of this thesis. In this chapter it
was stated that for a valid comparison across employees on the NFR scale, this scale should
be invariant across employees. The construct validity of the NFR scale should, as such, be
comparable across these employees. This hypothesis makes clear that not the reliability or
construct validity of the NFR scale in a sample is assessed but the comparability of its score
across groups (of employees).\(^1\) This is illustrative for the assessment of a potential latent
defect as it addresses underlying issues that warrant the assessment of the overall quality
criteria of a measurement instrument. As the active modelling approach elicits the clear
proposition of a hypothesis underlying the valid application of a measurement instrument it
makes way for a theoretical consideration of such an hypothesis.

In chapter 3, for example, different mechanisms could influence the functioning of the NFR
scale in shift versus day workers and therefore its underlying construct validity. First, through
the difference in work schedule the NFR scale could be entangled differently between day and
shift workers with other constructs. Employees who work in night-shifts are, for example,
at risk for a distortion of their circadian rhythm which is amongst other aspects related to
severe sleep deprivation (Boivin & Boudreau, 2014). The NFR scale could, as such, be
strongly intertwined with sleep deprivation in shift workers, while this is less prominent in day
workers. Shift workers, therefore, may perform differently on a test compared to day workers
as this nuisance factor (i.e. sleep deprivation) differentiates the measurement of the construct
(Marquie, Tucker, Folkard, Gentil, & Ansiau, 2015). Second, differences could occur through
a difference between groups in their perception of an item, for example: “By the end of
the working day, I feel really worn out”. Feeling worn out could be interpreted differently
by shift compared to day workers. That is, shift workers may have become accustomed to
their strenuous work (Akerstedt, Fredlund, Gillberg, & Jansson, 2002), possibly resulting in
an altered threshold of their perception of being “worn out”. Third, applicability of an item
could be divergent for the different groups, for example: “When I get home from work, I need
to be left in peace for a while.” Such questions regarding unwinding after work could have
da different connotation for shift workers, who often come home during the night, compared
to day workers, who come home during daytime (Shields, 2002). For the latter group the
family is probably still awake when the employee returns from work, while for the former group
they are probably already asleep. Such a theoretical discussion is conditional for the active
modelling approach to identify latent defects. The statistical tools used in this thesis are
exactly that, tools. As such these tools will not find latent defects which are not targeted.
Furthermore, these statistical tools will often not explicate the exact actors, or the underlying
mechanism, that result in the potential misuse of an instrument.

Chapter 2 has somewhat the same approach as chapter 3 but focuses on a different actor
(i.e. settings). As such, both chapter 2 and 3 focus on the same potential quality criteria
(i.e. construct validity) of an instrument that is potentially violated but focus on different

\(^1\)It is important to note that without a good reliability comparison across groups would also become
hampered. Nevertheless, an instrument could function invariantly across groups while it is unreliable – and
therefore is unreliable within each group.
Chapter 2 considered the SDQ (Goodman, 1999, 2001). The SDQ is a screening instrument for psychosocial problems in children and adolescents, and it is applied in individual and collective settings. For comparison of scores, such as cut-off points and population averages, it is however important that the instrument is measured invariantly between such settings (Vandenberg & Lance, 2000; Millsap, 2011). Due to this explicit hypothesis it is made clear that the SDQ should function invariantly across settings. Figure 1 shows the results of this analysis in a graphical form. The indicators in the middle of each target are the ‘anchor’ indicators which are used as reference for the other indicators of each subscale (Kline, 2011).

Figure 1 shows that although the SDQ showed measurement invariance between the two settings, scores were higher within the collective setting (illustrated with the larger impact of the indicators). As the samples were from the same population, without any notable differences, such a difference is likely an artefact of the setting in which the SDQ is applied. It could be stated that within the null hypothesis, implying no expected differences between the two settings, these indicators should have the same impact. Testing this hypothesis in combination with equal mean scores between the settings however resulted in a rejection of the model. As such, there seems to be no specific indicator responsible for the difference between the settings (Meade & Lautenschlager, 2004). While this conclusion was eventually derived using a ‘simple’ analysis of variance (ANOVA), this ANOVA has only credibility because the scores are comparable. If the indicators had a completely different lay-out on the targets for the two settings such a comparison and therefore conclusion would be impossible (Schmitt & Kuljanin, 2008). β-pyrometric as such provides an identification of the limitations and possibilities of an instrument within the public mental health. This example also illustrates that no definite answers can be given on the exact source or mechanism that leads to the misuse. Whilst the actor is evident in this example (i.e. settings) the mechanism through which it acts can only be hypothesised. In chapter 2, for example, the differential effect of social desirability bias between the settings was hypothesised to affect the difference in measurement between the two settings. Other forms of bias, however, cannot be ruled out.

The illustration from chapter 2 also shows the identification of a somewhat more technical latent defect, the presence of positive and negative formulated indicators. Figure 1 shows these indicators as squares (e.g. “I think before I do things”). Within the SDQ it is found that these indicators have the ‘tendency’ to go to the same direction (Niclasen, Skovgaard, Andersen, Somhovd, & Obel, 2013). These indicators have an interdependency which is unexpected as they have different underlying latent factors. As indicated in the introduction of this thesis latent factors should, ideally, explain all underlying variance of the indicators. If other patterns emerge that are unrelated to the ‘true’ underlying latent factors, this could distort reliability measures such as the Chronbach’s alpha (i.e. quality criteria). That is, inclusion of these indicators is likely to artificially increase the reliability of an instrument (Niclasen et al., 2013). This is deemed artificial as it is expected that indicators are related to each other because of the underlying construct they address and not as a result of artefacts of the indicator. CFA enables the active modelling of this dependence between these indicators (Kline, 2011). As such, the relation between the indicators based on this positive construal factor is ‘filtered out’. The resulting latent factors of the subscales are more ‘pure’ which enhances the interpretability of these factors and their underlying relations (covariance). Inclusion of residual correlations

2A term that is often used for such an ‘additional’ factor is a nuisance factor (Millsap, 2011). Examples of nuisance factors are numerous such as gender, education or attitude.
could also tackle this problem, which is especially helpful if only one pair of indicators has an increased dependency (Muthen & Asparouhov, 2012). Such error correlations should, however, also be based on a theoretical underpinning.

A measurement instrument should as such strive towards an inclusive selection of indicators that grasp the whole depth of a latent construct. While this is likely to decrease the reliability of an instrument somewhat, it will greatly enhance the information provided by an instrument increasing its potential as construct within public mental health. If an instrument is highly monotonous it is, furthermore, possible that it lacks variation. With a lack of variation the likelihood for a relation with other constructs also decreases (de Vet, Terwee, Mokkink, & Knol, 2011). This is in line with the importance of the predictive validity of an instrument as underpinning of its overall usefulness. In chapter 5 it was illustrated, for example, that the trait component of fatigue significantly predicts future sickness absence. Such a prediction gives meaning to the latent construct, apart from its hypothesised meaning based on the underlying indicators. Another issue with monotonous instruments is furthermore that it is also less attractive to fill out the questionnaire. This example, furthermore, also shows that latent defects are ‘subjective’ to the research question at hand. If a positive construal factor (as within chapter 2) does not alter the results within a setting, it does not violate any quality criteria. The positive construal factor was modelled to rule out the possibility that it had a different effect within the different settings. As such, latent defects are always related to the context in which the instrument is applied. Different contexts, applications, and goals of the measurement instrument will also alter the likelihood and importance of latent defects.
While these threats are sometimes seen as the Achilles heel of questionnaire research they can be considered a strength at the same time. That is, the use of multiple indicators enable more stringent tests of specific hypotheses that should hold for a measurement instrument to be valid within a given context. If single indicators are used to measure a latent construct many issues cannot be tackled nor identified. Take, for example, blood pressure measurement. As outlined in the introduction this is often seen as a rather reliable measurement of a manifest outcome. Nevertheless, imagine the situation in which the blood pressure of 100 males and 100 females is measured. Once this is conducted it is not possible to afterwards check for a potential violation of ME/I. Of course this could be validated against a gold standard, but this requires the presence of such a gold standard within the research data. It is, therefore, difficult to test such assumptions post-hoc. For single indicator measurement (e.g. a single quality of life question) the problem is even more eminent as no check of whatsoever is available to test potential variance of the instrument between populations, settings, time, or cultures. This is, however, by no means an argument for the dismissal of critical assessment of the measurement quality of an instrument before its implementation. Identification of latent defects in a measurement instrument after implementation makes it, however, possible to implement an instrument within (slightly) different population without the need to re-assess its quality every time. As such β-psychometrics can function as formal check of many assumptions that are often believed to hold but are rarely checked (Borsboom, 2006). This thesis shows that many of these assumptions can still be checked afterwards. It should be noted, however, that such checks can of course only include variables that are already present. If education, for example, is thought to be an actor within a specific latent defect this can only be taken into account if this variable is also assessed. Such post-hoc checks increase, nevertheless, the internal validity of the overall results of study. It also provides insights in the potential generalisation (external validity) of the findings (e.g., Milojev, Osborne, Greaves, Barlow, & Sibley, 2013). This should, however, not be considered as an alternative or replacement of a thorough design phase of an instrument as these checks can also be implemented in this phase which would strengthen its implementation.

The scrutiny of testing for latent defects should always be tailor-made and depend on many different aspects. One of the most important aspects to take into account is the amount of previous research and the context in which this research was conducted. If previous research, that has substantive quality, shows positive features for the instrument at hand, formal checks could be sparse. This, however, also depends on the comparability between the population and setting of this previous research and of the current population in which the instrument is applied. The Checklist Individual Strength (CIS; Beurskens et al., 2000; Vercoulen et al., 1994) originates, for example, from application within patient populations. As such it seems fruitful to check some core quality issues if the instrument is applied within a cohort of employees. Differences may emerge as a result of lower levels of fatigue within employees compared to patients. A difference in perception of fatigue between employees and patients could also result in a different functioning of the instrument (i.e. construct validity). The necessity of identification of potential latent defects is furthermore related to the research question at hand. If, for example, a research question includes differences between groups it is advisable to check quality issues that are strongly related with such a comparison. As such, researchers and public mental health professionals should be aware which potential latent defects are lurking to be able to critically assess them. Without such awareness latent defects remain latent as they can only be identified through an ‘active’ search. The next section will cover the importance of this awareness and some key aspects of it.
Discussion 155

Awareness for latent defects

As indicated in the previous section awareness is crucial for the detection of latent defects. As already illustrated, latent defects can appear small but can have far reaching consequences. It should furthermore be noted that even if a latent defect is identified, it can still impact the validity of the results of an instrument. As instruments are, by no means, only used within a scientific context awareness should not only be present among public mental health researchers but also within public mental health professionals in general (Vostanis, 2006).

While it seems evident that it is impossible for researchers to test each and every underlying hypothesis of an instrument, concept, or other methodical aspect, this thesis strives to achieve at least two things. At a minimum researchers should be aware that there are many potential latent defects out there ready to inflict heavy damage on underlying quality criteria, hindering valid conclusions. Researchers should therefore be sufficiently equipped to critically assess these and other assumptions underlying the measurement of public mental health before jumping to strict conclusions. While measurement issues, in general, do not result in splashy findings being published in high impact journals, awareness of them can prevent publication of incorrect splashy findings in high impact journals. If there is any doubt this thesis shows that there are many sophisticated methods to test for the presence of such assumptions building trust in ones conclusion.

An important caveat for researchers is, for example, that many classic psychometric statistics do cover reliability of an instrument but do not address absolute correspondence of the instrument across different groups or contexts (Sijtsma, 2008). The true information a reliability estimate does provide is the amount of consistency of an instrument (DiStefano & Hess, 2005). Within this respect this thesis was also limited as there was a lack of minimal clinical relevant difference for the instruments examined (Terwee et al., 2007; de Vet et al., 2011). Such a difference describes, from a clinical perspective, which difference is deemed relevant (e.g. responsiveness). Most instruments seem, however, to mainly focus on cut-off points (Goodman & Goodman, 2009). Surpassing such a cut-off point implies a certain level on the underlying construct of the instrument. Leaving aside the discussion of the arbitrariness of such cut-off points, it leaves little for the longitudinal use of an instrument. That is, if a respondent scores above a cut-off point how much should the score drop to indicate a positive development? Adopting the rationale that scoring below the cut-off point would indicate a positive development seems a little short-sighted as a respondent’s score could go back and forth just around the cut-off point. A minimal clinical relevant difference is a quantified difference that clarifies which difference should be considered a positive development (Beckerman et al., 2001; Lexell & Downham, 2005). Researchers should, therefore, be aware of such implementation issues and compel themselves to quantify such a difference. Awareness is therefore important with respect to the potential domains in which an instrument is applied. Otherwise researchers should make it clear that an instrument is not designed for a specific domain and application is inadvisable if applied in such a domain.

Awareness for latent defects is furthermore important as it is shown that some pitfalls are impossible to circumvent. It seems, for example, undo-able to completely erase social desirability bias (Krumpal, 2013). This bias is hypothesised to be the underlying factor of the difference found between the individual and collective setting in chapter 2 (see also Figure 1). In this chapter some solutions for this specific situation are addressed. For an informed and valid usage of the SDQ, for example, in each setting the ‘actual’ cut-off scores have
to be determined separately for each setting. Or, the accuracy (e.g. positive and negative predictive value) in each setting has to be reassessed in each setting using the cut-off scores from a specific setting. The underlying message, however, is that if a researcher is aware of a potential latent defect, (preventive) actions can be taken. Such actions do not always imply that the latent defect is eliminated, but such actions can at least make sure that the effect of the latent defect is minimised. While it is best to prevent latent defects, it is still better to be aware than to be inconsiderate.

With respect to public mental health professionals this section will only address awareness with regard to the individual application of measurement instruments. Policy makers, for example, often use results of measurement instruments on an aggregate level. This purpose of assessment is however closely related to the issues already addressed for researchers. Indirect application of measurement instruments in the public mental health can take place within many different settings, such as general practitioners or social workers (Davies, 2014). Measurement instruments are used to facilitate the (decision making) process of the professional in such settings. As such not only the score of the instrument is of interest but also the rationale of the answer provided. A measurement instrument is therefore part of the process and by no means the defining aspect (Goodman, Ford, Simmons, Gatward, & Meltzer, 2000; Jaspers et al., 2010; Glascoe, 2000).

Two caveats are especially important for the individual application of a measurement instrument. First of all it is important to keep in mind the ‘naming fallacy’. This implies that an instrument is believed to represent a certain concept because it is ought (named) to measure this. For instance the HADS includes a depression scale and an anxiety scale. Such a naming convention can imply a ground truth while this is not necessarily present. Leaving aside the notion that such a ground truth is by definition impossible within (latent) mental health constructs, a different label for the scale could also result in a different connotation. It is, therefore, important that mental health professionals are aware on which grounds a questionnaire is attributed specific characteristics. Second, if a questionnaire is developed to measure a construct within a specific population it is not necessarily valid in a different population. If a questionnaire is, for example, designed to measure fatigue within a patient population, the underlying quality criteria, such as construct validity and floor effects (e.g. chapter 4), do not automatically translate to a working population. Fatigue could have a different manifestation within such a population calling for a different approach. Professionals should, therefore, be aware that instruments may be difficult to transpose between settings and population.

As options for post-hoc analysis are limited in the individual context (setting), awareness for latent defects is especially important. That is, if female employees continuously show higher rates of fatigue compared to males a researcher should at least conduct a ME/I analysis to check whether this is not due to measurement issues. This option is difficult to apply in clinical practice and it is therefore crucial that public mental health professionals are aware of the potential violations of quality criteria that could explain (maybe partly) an observed difference between females and males. An important example can be found in chapter 4 & 5. These chapter showed that the CIS is fairly stable over time. For public mental health professionals

---

3Of course this dichotomisation between aggregate and individual application of a measurement instrument is somewhat unjust. It could even be considered a persistent myth within public mental health. If an instrument is justly applied within an individual setting, it can be perfectly used for the collective setting. This requires, however, thoughtful planning as two branches of public mental health purposes have to merged into one process.
such information can be crucial not only with respect to the application of the instrument but also for the attitude in general to fatigue. Professionals, for example, should take into account that scores on the CIS are for a large extent time-invariant which could justify treatments that are not directly targeted at decreasing fatigue but at coping strategies to deal with higher levels of fatigue. Such a focus is furthermore substantiated by the finding that it is especially this trait component of fatigue that explains future adverse health outcomes. Being aware of such issues with respect to measurement instruments will greatly increase the evidence-based practice within public mental health. This example furthermore shows that a deeper understanding of a measurement instrument (e.g. the CIS) is also closely intertwined with a deeper understanding of the underlying concept in general (e.g. fatigue). The implications of an increased awareness and implementation to achieve this increased awareness will be discussed later.

Preventive and treatment strategies for latent defects

If latent defects are successfully identified and researchers and professionals are aware of them, preventive strategies can be considered and implemented. Such strategies can be implemented at different levels depending on the issue at hand. Some issues, for example, are bound to a specific context and are difficult to circumvent whilst others can be successfully prevented by tailoring the instrument at hand. Put more simply, treatment of an instrument implies limiting the effects of inherit flaws of an instrument and preventive strategies imply some sort of (re)development of the instrument.

Prevention

A clear example of a preventive strategy can be found in the results of chapter 4. In this chapter it was shown that it was difficult to discriminate on the lower end of the NFR spectrum using the NFR scale van Veldhoven (2003). Such a problem could be prevented using a Likert scale instead of a dichotomous scale (e.g., Moriguchi, Alem, van Veldhoven, & Coury, 2010). Using a Likert scale will likely enable differentiation on the lower end of the spectrum. Besides this increased variation it is also likely to ease the interpretation of questions as black-or-white representations are often difficult for respondents.\footnote{A small anecdote on this matter entails a specific questionnaire of a respondent within the Maastricht Cohort Study (MCS). In this questionnaire the respondent drew a little box between the yes and the no box and filled this ‘custom-made’ box. This simple example does show that respondents cannot always represent their feelings towards a specific indicator with the options provided. Although standardised implementation of a questionnaire is of the upmost importance it is also important to not ‘scare’ or discourage the respondents of a questionnaire with either the questions or the associated response options.}

The original rationale of the NFR scale was that dichotomisation of the response categories greatly simplified the application within the individual setting. Public mental health professionals can simply sum the items and compare this sum score with some reference score. Such an addition is more complex if items have a Likert scale or have other rules to compute the score (e.g. the SDQ). While user friendly, the NFR scale is found to have its limitation for individual monitoring purposes (chapter 4). The NFR scale appears most suitable to identify employees who show a change from low levels of NFR towards more elevated levels of NFR. Further differentiation, especially within employees with high levels of NFR, is cumbersome. Especially for monitoring purposes this would greatly hinder the valid evaluation of the results.
of this measurement instrument. In a monitoring context evaluation of progress could only be valid if an instrument is susceptible to assess such progress. As indicated in the previous section, researchers should be aware of the different domains in which an instrument can be applied. Thorough development and validation of a questionnaire along a multitude of quality issues is therefore very important (Terwee et al., 2007).

Another important issue that is important to address is the difference between indicators regarding their threshold. A threshold of an indicator can be regarded as the ‘difficulty’ of an indicator. That is, how high should a respondent score on the underlying (latent) construct to positively score on this indicator. Using indicators which have threshold across a broad range are likely to better capture the full spectrum of a construct. Again, if all items have almost the same threshold the instrument is very likely to ‘perfectly’ distinguish between high and low scoring in the population but nothing more. Adding additional questions could result in, for example, a low, medium, and high scoring group. The SDQ for example has some pretty ‘vigorous’ items regarding delinquent behaviour for conduct problems, but combines this with some more ‘easy’ items regarding whether a child is obedient or not. As such it is possible to distinguish the children without problems at all, children with some problems, and children with severe issues. If all five items would have the same vigorousness a dichotomisation would appear in the population which would be reliable but would not reflect the, likely continuous, underlying latent construct.

**Treatment**

While prevention regarding latent defects should always be favoured, treatment is often still possible when a latent defect is identified. In chapter 5, for example, it was identified that employees show a strong trait-like characteristic of fatigue over time. If undesirable, this could be (partly) prevented by clearly clarifying the short-term (2-week) scope of the items. It seems, nevertheless, that this trait component is a recurring aspect within questionnaires or even public mental health instruments in general (Cole, Martin, & Steiger, 2005). It could even be stated that this is related to the underlying construct, which could have an underlying trait-component embedded in its manifestation (Prenoveau, 2016). Often, however, researchers are interested in the component of a public mental health construct that changes over time. Such components are more likely to be modifiable over time, making them more suitable for preventive public mental health strategies. Studying these occasion specific (time-variant) components was made possible by the inclusion of an extensive CFA model, which models these trait and occasion components (Prenoveau et al., 2011). This shows that identification and treatment of latent defects within public mental health are often parallelised. That is, the (statistical) model that detects/identifies the latent defect is often also capable to prevent, at least partly, its influence on the results of the study. In fact, such a model even gives credit to the potential underlying dynamics of fatigue giving input for the further theoretical development of its construct (Cole, 2012). As such these potential threats for the validity of mental health research in the public health can be identified and modelled in such a way that their impact can be quantified and minimised (e.g., Cole & Preacher, 2014; Kline, 2011; Muthen & Asparouhov, 2012).
Conceptual Considerations of $\beta$-Psychometrics

The previous sections considered three important challenges regarding the measurement of public mental health in the presence of potential latent defects. As the chapters in this thesis form the basis of these sections, and this discussion as a whole, it is important to consider the methodology used in these chapters. The current section will, therefore, address some (specific) conceptual considerations of $\beta$-psychometrics which is the core of these chapters. First, the Bayesian framework will be introduced and what it could add in contrast to the frequentist framework. Second, the CFA methodology used in this thesis will be compared to item response theory (IRT) analysis. Third, the use of anchor items and the effect of method bias will be discussed.

Bayesian versus frequentist framework

All empirical chapters (2-5) in this thesis used a frequentist framework. An appealing alternative for $\beta$-psychometrics is the Bayesian framework (Lynch, 2007). It most appealing feature is the use of prior information (Kaplan & Depaoli, 2012; van de Schoot, Kaplan, et al., 2013). Such prior information specifies the expected value of specific model parameters (with a certain amount of uncertainty). Doing so, makes it possible to express the amount of certainty that two parameters are equal. Going back to the example of the introduction of this thesis it could be expected that the HADS does have a slight difference for certain questions between female and male respondents. Within the frequentist framework there are (roughly) two options: (i) constrain two (or more) parameters to be equal (square item in the left panel Figure 2), or (ii) freely estimate two (or more) parameters (square item in the centre panel Figure 2). A parameter (i.e. indicator) is, therefore, either equal or different between groups. Within the frequentist framework this is often tackled by assessing potential partial invariance. This states that for a selection of the items of an instrument ME/I does not hold. Using different kind of strategies these items are selected and freely estimated. This results in the estimation of a model in which some items have the same weight across groups whilst others are independent across groups (Meredith, 1993; Millsap, 2011; Vandenberg, 2002).

While partial invariance tackles some of the dichotomy that is present within ME/I assessment in the frequentist framework, it can become impractical if a multitude of different groups are assessed or if the instrument at hand has a large amount of indicators. Within such situations it is an almost impossible task to determine the starting point for assessing partial invariance, if the test for full invariance fails (Millsap, 2007). Using the example of the introduction, it is clear that some rifles on some targets are not behaving as expected but it is not clear which particular rifles. Even more so, it is also not clear to what extent these rifles behave out of fashion. While there are some data-driven ‘work-arounds’ within the frequentist framework, the Bayesian framework has a more subtle and informative approach. 

---

5 These latter items, that are freely estimated across groups, have no impact on the mean difference between groups with respect of their associated latent factor(s). For example, for a 4-item questionnaire (with one underlying latent factor) item 1, 2, and 3 are constrained to be equal across groups and item 4 is freely estimated. Only (mean) differences with respect to item 1, 2, and 3 will be reflected in the overall mean difference of the associated latent factor. Differences with respect to item 4 are 'absorbed' into the specific parameters associated to this item 4 and are therefore not reflected in the over latent factor.
Partial invariance (Free)

Full invariance (Constrained)

Approximate invariance

Females

Males

Figure 2: Illustration of approximate invariance within the Bayesian framework as opposed to (partial) invariance within the frequentist framework. The first set of targets shows that the squared indicator shows a small difference between males and females, which indicates partial invariance. The second set of targets shows full invariance as both squared indicators are ‘forced’ (constrained) to be equal between males and females, although this seems somewhat unjust. The last set of targets shows approximate measurement invariance as the squared indicators are given some ‘wiggle room’ within the dashed circle. This wiggle room regulates the difference between the squared indicators between males and females. The greyed out squares is the ‘original’ (freely) estimated position of the indicator.

(van de Schoot, Kluytmans, et al., 2013). This approach does place a restriction on the difference between groups for each parameter. This restriction is small enough to make sure that items are still comparable, but large enough to allow for some ‘wiggle-room’ (right panel Figure 2; van de Schoot, Kluytmans, et al., 2013). This method has two advantages: (i) it does not place strict constraints on parameters but mean scores are still comparable, and (ii) it provides information on the location and amount of potential sources of variances between items and groups (van de Schoot, Kluytmans, et al., 2013). This approximate ME/I has great benefits for the public mental health, as instruments are often applied to multiple groups (or even countries) in which deviations between groups are likely. For example, employees can differ on their job title. As there are many different jobs, it would be difficult to assess ME/I. Approximate ME/I could place a boundary on the allowed variation between job titles to have a solid model in which scores are still comparable across groups. Another situation in which approximate ME/I can be applied is the comparison between many different countries (e.g., Cieciuch, Davidov, Schmidt, Algesheimer, & Schwartz, 2014). Such comparisons are frequent within public mental health (e.g., World Health Organization, 2015), and approximate ME/I could provide a valid tool for reliable and more informative comparisons between countries. As this technique is, however, relatively new it has some caveats. One of these caveats is that it
is difficult to assess the model fit of Bayesian CFA models with large samples (Asparouhov & Muthen, 2010). It is therefore that within chapter 6 a new fit index is developed and validated to assess model fit within Bayesian CFA with large samples. This chapter shows that this new fit index, the Bayesian Root Mean Square Error of Approximation ($BRMSEA$), is a viable fit index to assess model fit within such models. Using the $BRMSEA$ can support the application of, for example, approximate ME/I within public mental health.\(^6\) While all bodes well for the $BRMSEA$ based on the results of the simulation study – and its accompanying empirical illustration – future empirical studies have to show what the exact added value of the $BRMSEA$ is. This chapter 6 could play an important role in the increased understanding of the quality of measurement instruments in the public mental health.

**Item response theory (IRT)**

All chapters in this thesis use CFA. As all studies use categorical indicators, however, the conceptual framework of IRT could also have been applied (Embretson & Reise, 2000). While it is more than a semantic discussion, it should first of all be noted that the underlying principles are highly similar (Glockner-Rist & Hoijtink, 2003). IRT is, however, more focused on individual testing, evaluation of single items, and dynamic testing. It is, therefore, that CFA was preferred over IRT as it is more common within population based research. Population based research is often more focused on the underlying construct than the items at hand. Nevertheless, studying differential item functioning is highly similar to studying ME/I (Kim & Yoon, 2011; Stark, Chernyshenko, & Drasgow, 2006). Some of the reasoning of chapter 4 draws, therefore, strongly on the IRT paradigm of different reliability of an instrument across its range. That is, IRT posits as one of its core assumptions that an instrument does not have a fixed measurement error (as opposed to classical test theory). As such an instrument can be found to be very precise for low ranges on an instrument but become less stable for scores on the higher end on the continuum (Embretson & Reise, 2000). If a cut-off point is furthermore applied, it is desirable that the precision of the instrument is high around this point (see for example, de Cock, Emons, Nefs, Pop, & Pouwer, 2011). Chapter 4 shows however that around this cut-off point the NFR scale is difficult to use for subsequent measurement within an individual setting as it is difficult to have a score that can be reliably assessed as a significant change (see also, Blanchin et al., 2010). It was also shown that the reliability is highly influenced by the floor effect of the instrument. If this floor effect was excluded the reliability decreased. This implies that when people do ‘move’ around the scale it loses reliability while this is exactly the situation in which reliability is important.

**Anchor items & method bias**

As Figures 1 & 2 in this discussion show, each target has an indicator exactly in the middle. This indicator represents the anchor item within CFA (Kline, 2011). This anchor item is manually selected (Kline, 2011). Without such an anchor item the latent construct has no scale (or metric).\(^7\) Stated differently as the latent factor is not directly measured the target

\(^6\)Within chapter 6 the issues regarding the use of fit indices in general within CFA is also addressed. This line of reasoning is therefore not pointed out in this general discussion.

\(^7\)This example ignores the method in which the variance of the latent factor is fixed. This method has however, in the end the same potential issues as using an anchor item.
could be placed anywhere. The anchor item indicates therefore the middle of the target. As a result the other indicators are placed relative to this anchor item. Whilst within a single-group CFA it does not really matter which indicator is selected as anchor item, it is common practice to select the indicator which is thought to represent the latent factor best and is measured most reliable. The obedient indicator within Figure 1, for example, would be a poor anchor item. The problem is, however, that this becomes only visible after modeling the CFA. Critical assessment should therefore go hand-in-hand with empirical testing.

In the analyses of ME/I selecting an anchor item also has another consequence. This consequence entails that the anchor items are constrained to be equal across groups on a-priori basis. This a-priori basis is however often lacking as this is exactly the assumption that is tested using CFA. In the extreme situation, for example, that only the anchor item differs between groups whilst all other indicators (items) are equal, ME/I analysis would indicate full measurement variance. As all other indicators are assessed relative to the anchor item these remaining indicators are ‘drawn into’ measurement variance. This seems somewhat unjust, however, as only one indicator was truly different between groups. Strategies are therefore developed that eliminate the sole reliance on a single indicator as anchor item (e.g., Meade & Wright, 2012). Another difficult situation is if each indicator has the same difference between groups. So, for example, males shoot each bullet (indicator) 5 cm to the right compared to females. In such a situation it would be clear that there is a difference between males and females. For CFA, however, the latent factors (targets) should first be aligned. As such, regardless of the selected bullet (indicators), each bullet of the males moves 5 cm to the left. This would result seemingly (and unjust) in a perfectly comparable instrument between males and females. Such a situation is difficult to identify. In the example of Figure 1 this was argued to be present. All indicators of the collective setting have a higher impact compared to the individual setting. This issue was only tackled with the assumption that the mean scores of the same subscale should be equal across groups. Applying this assumption resulted in a deteriorated model indicating a differential impact of the indicators between groups. Such an assumption is however not always implied and was only applicable because both samples were from the same population.

While predictive validity is one of the most important checks for a valid instrument it should be noted that it can also be misguiding. Aspects such as common method bias could alter potential relations. Using objective data, or filtering out the method component can prevent such issues (Byrne & Stewart, 2006). This again illustrates the necessity of modelling assumptions before it is possible to tackle them. It should be noted, however, that β-psychometrics does (often) not indicate the ‘source’ of a violation. The section regarding identification of latent defects illustrated, for example, some of the potential mechanisms that were hypothesised to alter the working of the NFR scale between employees who work in shifts compared to employees with day-work. It was found, however, that there were no significant differences regarding the NFR scale between the two groups of employees. It could be hypothesised that employees know what is ‘meant’ with the items resulting in some sort of ‘alignment’ and therefore ME/I. This shows that β-psychometrics is not a silver-bullet to all measurement issues within public mental health but provides clear indications if important assumptions are violated (Borsboom, 2006). If so, the three aspects (identification, awareness, treatment and preventive strategies) already addressed in this discussion can be taken into account to deal appropriately with the potential treat of such a violated assumption (i.e. latent defect).

To illustrate, again, the diverse nature of β-psychometrics the last conceptual issue that is
addressed is the wording used within instruments. While such a small issue can be considered futile, particularly in comparison with cultural differences between instrument, it can have significant implications for an instruments. As shown in chapter 2, and already pointed out in this discussion, it was important to take the formulation of indicators into account. The limitation that is brought up by this mechanism of differential wording within an instrument is that it can negatively influence the results of its reliability and other aspects of β-psychometrics. This is, however, not the only aspect that should be taken into account. The SDQ, for example, has for clear reasons included positively worded items (Goodman & Scott, 1999). As this instrument is often used with children and adolescents it is considered a positive property that it not only includes negative items. On the long run this seems to be a very good virtue of an instrument, as it does decrease the likelihood of aversion of respondents towards the instrument – especially if instruments are combined into a test battery it seems worthwhile to also include positive items. Designing instruments, as such, requires more than only achieving high scores on the reliability and validity scores. Brevity, ease of scoring, clarity, and ‘tone’ are examples of issues that also should be taken into account when designing an instrument which can be successfully implemented within public mental health.

**General Strengths and Limitations**

This section will discuss the strengths and limitations specific to this thesis. First the measurement instruments that were taken into account will addressed. Second the population in which these measurement instruments were applied will be discussed. These two aspects are specifically discussed as this is a recurring theme across the chapters.

**Measurement instruments**

To illustrate the variety of measurement instruments in the public mental health instruments from the work and health and from the society and health domain are included. These measurement instruments are subsequently assessed with various methods using β-psychometrics. The measurement instruments that were used have already shown their added value within the public mental health. This makes it fruitful to assess their quality criteria more critically. The instruments that were assessed in this thesis were, furthermore, applied within various contexts, settings, and countries. As such, the results found in this thesis also have strong merit as stand-alone results.

It should be noted that the absolute number of measurement instruments that were assessed is somewhat limited. Without this selection, however, it would be impossible to rightly address the theoretical foundations of the measurement instruments and their underlying (latent) concepts. The β-psychometrics that were applied are, nevertheless, also applicable to comparable measurement instruments. That is, the issues addressed in chapter 2 regarding the SDQ could for example have also been examined for the Child Behaviour Checklist (CBCL; Achenbach & Edelbrock, 1983). It would furthermore also be interesting to examine different translations of an instrument. The general lessons learned within this thesis are, however, applicable to all measurement instruments within the public mental health. The direction or magnitude of the results depends, however, on the exact measurement instrument at hand. Especially measurement instruments that cover distinctly different concepts could show different effects to
the latent defect assessed in this thesis. This shows that direct generalisation of the findings in this thesis towards other measurement instruments is not straightforward.

Another aspect that is somewhat underexposed in this thesis, but is also related to the nature of public mental health, is the sole inclusion of questionnaires. It could be argued that for other sorts of measurement instruments (e.g. observations or interviews) other actors are at stake that could potentially violate important quality criteria of a measurement instrument. Again, this would show the context specificity of latent defects within the public mental health. Questionnaires, nevertheless, are the most common measurement instruments within public mental health and were therefore the core of this current thesis.

**Populations**

As latent defects in measurement instruments are context-based it was important that the current thesis included heterogeneous and various study populations. This enabled examining a rich variation of different research questions which also elicits specific potential latent defects. The difference between settings, for example, that is addressed in chapter 2 in this thesis is a good example of this specificity. The Maastricht Cohort Study (MCS) that was used to reflect a population from the work and health domain has furthermore some profound strengths. First of all it includes a large amount of employees with heterogeneous background characteristics. This enabled the assessment of various quality criteria that could be violated due to latent defects. The longitudinal nature of the MCS, with low levels of loss-to-follow up, furthermore paved the way for the analysis in chapter 4 and 5 that showed the potential effect of latent defects within longitudinal analysis. For chapter 6 a simulation study was designed in which specific sample characteristics were modelled. This resulted in a successful validation of the BRMSEA. Such simulation studies show that specific aspects of a measurement instrument can have profound effects on its results. This chapter, for example, shows in correspondence with previous literature that lower levels of internal consistency (i.e. factor loading’s) within a measurement instrument make it more difficult to assess its construct validity (Kline, 2011).

A specific shortcoming of the current thesis which would warrant further investigation are cultural differences for the instruments assessed that potentially influence the working of these instruments. It is, for example, often found that there is measurement variance between different translations of an instrument (Schmitt & Kuljanin, 2008). This variance is, however, not always a result of different languages used in the instrument but can also be a result of cultural differences that are present between the countries in which the instrument is assessed. This issue is important as there is a growing body of literature that aims to compare the public mental health across countries or even across the entire world (World Health Organization, 2015). Such a comparison is meaningless if the functioning of the instruments varies across countries. This is also another example that points out the relation between the measurement theory of an instrument and the theoretical framework of the underlying construct that is assessed with the instrument. Such a theoretical framework should make clear that a construct is expected to function invariantly across cultures. If such a claim is missing, or not substantiated it would be futile to expect that an instrument would measure the same construct across different countries. If it is expected that there are important differences between cultures, it could even be considered a validation if an instrument functions differently across cultures. As with the variation of measurement instruments, the findings with respect
to the studied populations are not directly applicable to other populations. Again, however, it is expected that for populations that are closely related to those studied in this thesis findings could be comparable. This could also hold for distinctly different populations, but with increasing differences this extrapolation will become more uncertain increasing the need for additional research.

**Implications and Implementation**

The main implication of this thesis is that it shows that latent defects can greatly hamper the validity of a measurement instrument and therefore the conclusion that are based on its results. The current thesis, however, also shows that it is possible to identify these latent defects using an active modelling approach within a $\beta$-psychometric context. Besides this identification this discussion also showed the possibilities to prevent and treat latent defects within measurement instrument within the public mental health. Most chapters show however positive results for the measurement instruments assessed. Some actors (i.e. setting) were, nevertheless, found to violate the quality criteria of a measurement instrument. As these violations of these quality criteria could have severe consequences it is crucial that the awareness for latent defects is increased.

As indicated in this discussion awareness is key to successfully deal with latent defects within measurement instrument in the public mental health. While a myriad of stakeholders within the public mental health could be identified as having a crucial role in this increased awareness, this discussion will focus on two of them: public mental health professionals and developers of measurement instruments. These two groups are somewhat analogous of the buyer and seller of a property (which may contain latent defects). The seller (developer) should make the buyer (mental health professional) beware of potential latent defects if known to the seller (or reasonably should be). This does, however, not release the buyer from his or her general duty to inspect their purchase before taking possession.

To successfully inspect this ‘possession’ (i.e. measurement instrument) the education of public mental health professionals should include ($\beta$-)psychometrics. In most educational programs the curriculum seem to jump from reliability and simple validity issues directly towards inferential statistics (e.g. regression). While it could be considered too complex for first year students, post-graduate courses should be a perfect place to introduce the importance of a critical assessment of quality criteria of measurement instruments. Not only will this increase the statistical knowledge of public mental health professionals it will also increase their general conceptual awareness regarding public mental health. Since the assessment of latent defects requires a profound understanding of the (research) context, it provokes a theoretical understanding of public mental health. That is, $\beta$-psychometrics is not simply computing a reliability estimate for a given scale, it requires an active modelling approach in which a deeper understanding of measurement issues and public mental health in general is needed. Public mental health professionals, as such, will increase their critical appraisal of this topic which will make them more aware of potential latent defects if they apply or select a measurement instrument. The Critical Appraisal of a Topic (CAT) methodology applied within the medical curriculum of the Maastricht University would be a perfect vehicle for this purpose (de Brouwer, Kant, Smits, & Voogd, 2014). Within the CAT methodology students will bring a case from the internship which they translate into a clinical question. This
clinical question is, subsequently, answered with a critical literature search and application of the selected literature. Doing so students are actively involved with the topic and also learn a much needed skill in being critical (towards measurement instruments) and being able to substantiate their critical thinking with appropriate literature. This approach, as such, perfectly aligns with the active modelling approach to assess latent defects as it requires a deep(er) understanding of the research question at hand.

A second important aspect that could greatly increase the awareness for latent defects in measurement instrument within the public mental health is the research regarding measurement instruments, especially its presentation (i.e. publication). While it is impossible to capture all possible caveats of a measurement instrument if it is introduced, it should at least be addressed that such caveats (i.e. latent defects) do exist – also depending on the context in which it is assessed. As such, the first step should be to include such statements when an instrument is introduced. Such statements should be specific and address aspects that are currently not touched upon. Editors of scientific journals should therefore be critical towards manuscripts that introduce (new) measurement instruments, whether such issues are addressed. It could be helpful to use a checklist, as for example proposed by Mokkink et al. (2010), to evaluate the quality of study in relation to the instruments uses. The STROBE guideline, used to evaluate cohort studies, for example is a highly transparent tool to increase the quality of the assessment of a study (Elm et al., 2007). A likewise checklist or guideline could be useful to assess the potential threat of ‘latent defects’ in measurement instruments. Even better, however, would be that during the introduction of a measurement instrument, latent defects are already addressed as much as possible. Especially those quality criteria that are likely to be important for the context in which the instrument is applied, should be critically assessed. Public mental health professionals that design and implement a measurement instrument should therefore be critical when they assess the quality of a measurement instrument. Again, assessment of quality criteria should be theory driven with an active modelling approach. Only with this scientific attitude it is possible to assess the quality of an instrument which result in a higher standard for measurement instruments within the public mental health. This would even be more effective if the aforementioned education is also successful in increasing the awareness of the (end-)users of the public mental health measurement instruments, as these users will also have higher standards. In an ideal situation, therefore, these two aspects will go hand-in-hand resulting in an overall increased awareness for (potential) latent defects within measurement instruments in the public mental health.

Conclusion

This thesis demonstrates that with the use of β-psychometrics ‘latent defects’ can be successfully identified. The analyses in the foregoing chapters show that latent defects can have a significant impact on the quality of an instrument. Awareness within the field of (public) mental health for latent defects should therefore be increased. This awareness is especially important as the consequences of latent defects are often not directly noticeable or perceived as problematic. Its detection is however important as measurement instruments are often the starting point within many aspects of public mental health, being it in research, clinical practice, or policy making. If latent defects prevail in this stage they will be lurking throughout the whole process hindering valid judgment and decision making.
This thesis shows that both the developers and users of a measurement instrument should be critical towards the application of an instrument and the underlying literature that validates its potential implementation. As an active modelling approach is a necessity within β-psychometrics it is furthermore important that a clear theoretical framework is present to assess the quality of measurement instruments. Only with such a conceptual understanding of an instrument can it be validly evaluated against the potential presence of latent defects.

References


Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139–158. doi:10.1177/1094428102005002001


Summary
Public mental health can be defined as “a public health or population health science approach to mental health and the mental health variations exhibited by populations” (Davies, 2014). Public mental health can be considered as consisting out of three overlapping domains: Mental health promotion, mental illness prevention, and treatment and rehabilitation (Figure 3). At the core of all these domains of public mental health is the measurement and assessment of mental health. Measuring mental health is, however, not a walk in the park. The challenges associated with the measurement of mental health are sometimes evident but are often concealed. They are important nevertheless as conclusions within public mental health, both for individuals but also for policies on different levels, are often (partly) based on the measurement of mental health. At first these challenges may appear small and unimportant, but they can have far-reaching consequences and should therefore be considered the ‘latent defects’ of public mental health. That is, if the public mental health (research) is based on measurements with an unknown blind spot this can have far-reaching implications. The main goal of this thesis was therefore:

- To identify potential latent defects and evaluate their impact within the public mental health

In this summary the rationale, main methods and results of this thesis will be examined and shortly reviewed.

In chapter 1 the conceptualisation of public mental health and its measurement are first introduced. This introduction focusses on four focal points: (1) Mental health moves along a continuum as opposed to a dichotomous representation (healthy/sick). Within a dichotomous representation, or other forms of categorisation, the underlying concept is still a continuum. (2) This underlying continuum is a latent construct. That is, public mental health has no directly observable (manifest) representation. As such public mental health should be inferred from indicators. Measurement instruments use such indicators to grasp the underlying construct. Assessing the quality of such instruments is not straightforward. Quality of a mea-

**Figure 3:** Decomposition of public mental health into its core dimensions, from Davies (2014).
surement instrument has so far mostly been tested on an exploratory and post-hoc basis. This approach has limitations which leaves the application of an instrument vulnerable for possible latent defects. (3) Latent defects are “actors that can result in the violation of (important) quality criteria of an instrument measuring public mental health, without it being (directly) apparent.” The different aspects of this definition were further discussed in the introduction of the thesis. (4) β-Psychometrics are used to assess the quality of a public mental health instrument with an a-priori theoretical framework. Using this framework it is possible to use confirmative statistical testing. With this approach it is, therefore, possible to test specific hypotheses that could rule out (or confirm) the existence of a potential latent defect.

Chapter 2 includes the first empirical study of the thesis. Within this chapter it is examined if the administration of an instrument in different settings could alter the influence of the social desirability bias. Such a difference, in turn, could potentially invalidate the interchangeability of results between the settings. The ‘latent defect’ addressed, in this chapter was the potential misuse of scores (of an instrument) between the different settings. For this chapter the Strengths and Difficulties Questionnaire (SDQ) was used. This is a screening instrument for psychosocial problems in children and adolescents. The compared administration settings were the individual and the collective setting.

The individual setting can be viewed as the clinical application of an instrument (e.g. preventive child healthcare). In this setting an instrument is especially suitable for screening purposes (Vogels, Crone, Hoekstra, & Reijneveld, 2009; van Widenfelt, Goedhart, Treffers, & Goodman, 2003; Goodman, 2001). For this setting data from the preventive child health (PCH) care was used. 6,594 available questionnaires, administered during the screening assessment in the second year of secondary school, were included. In contrast, within the collective setting the information of an instrument is aggregated across a population (van Roy, Groholt, Heyerdahl, & Clench-Aas, 2006; Becker et al., 2006; Ravens-Sieberer, Erhart, Gosch, & Wille, 2008). For this setting information from the National Youth Monitor was used (Ministry of Health, Welfare and Sport & Statistics Netherlands, 2014). This monitor includes the SDQ as an indicator of psychosocial well-being. This resulted in 4,613 available questionnaires. As both samples were retrieved from the same population it was hypothesised that any significant difference in scores between the samples would be a result of the differences in administration between the settings.

Confirmatory factor analysis (CFA) showed that there was no difference in the underlying construct between the two samples. Children in the individual setting, however, scored lower on total difficulties and the psychosocial problems subscales compared to those in the collective setting. To further ‘quantify’ these results the validity of the cross-use between the settings was examined. The cross-use examined was the implementation of cut-off points within the individual setting which are derived from the collective settings. This procedure is comparable with the current practice of implementing cut-off points of the SDQ within the individual setting. Using the 90th percentiles of the collective setting in the individual setting therefore resulted in a small number of cases, 2 to 3%, while ~10% was expected – if there would be no impact of the setting.

These results shows that the SDQ has the same connotation across the individual and collective setting. Nevertheless, the difference in mean scores shows that there is a difference in the ‘trigger-happiness’ within children to tick certain boxes between the setting. This observed structural difference undermines the validity of the cross-use of absolute SDQ-scores between these settings. Applying cut-off scores from the collective setting in the individual setting
could, therefore, result in invalid conclusions and potential misuse of the instrument. To correctly apply cut-off scores these should be retrieved from the applied setting.

The following three chapters (3-5) focus on mental public health in the working population. In these three chapters the same population was used. It is employees who participate in the Maastricht Cohort Study (MCS). It should be noted, however, that the different chapters use (slightly) different subsamples within and between chapters. The MCS was set up in May 1998 and included 12,140 employees from 45 different companies. At this baseline measurement, all included employees were aged between 18 and 65 (Kant et al., 2003; Mohren, Jansen, van Amelsvoort, & Kant, 2007). The baseline (T0) cohort consists of 8,840 (73%) men and 3,255 (27%) women from a working population. For more details regarding the MCS see Mohren et al. (2007). These chapters, as such, focus on the work & health domain of public mental health.

Besides a common sample, these chapters also shared common ground in the public mental health outcomes that were assessed: need for recovery (NFR) and (prolonged) fatigue. NFR is an outcome which captures the need for and extent of recuperation after work. The normal recuperation cycle is ideally completed after the working day or at least before the next working day (van Veldhoven, 2008). Prolonged fatigue is the presence of fatigue, including an overwhelming sense of tiredness, for a more extended period (Bultmann et al., 2000; Kalkman, Zwarts, Schillings, van Engelen, & Bleijenberg, 2008). Both aspects are important public mental health outcomes vital for facilitating or enabling preventive strategies targeting at prevention or limiting further adverse (health) outcomes (e.g., de Croon, Sluiter, & Frings-Dresen, 2003; van Amelsvoort, Kant, Bultmann, & Swaen, 2003; de Raeye, Kant, Jansen, Vanse, & van den Brandt, 2009; van Amelsvoort, Kant, Beurskens, Schroer, & Swaen, 2002; Silva-Costa, Griep, Fischer, & Rotenberg, 2012). NFR was measured with the NFR scale (van Veldhoven, 2003) and prolonged fatigue with the Checklist Individual Strength (Vercoulen et al., 1994; Beurskens et al., 2000). The NFR scale and the CIS have already showed to be instruments with good psychometric properties (e.g., van Veldhoven, 2003; Bultmann et al., 2000; Beurskens et al., 2000; Sluiter, Frings-Dresen, van der Beek, & Meijman, 2001; van Amelsvoort et al., 2003). While these studies show that both measures seem valid measurement instruments many underlying assumptions (latent defect) are not readily addressed.

The first chapter of this ‘trilogy’ is chapter 3. This chapter examined if the NFR scale and the CIS are measured invariantly across subgroups. Absence of measurement invariance is a typical latent defect as this makes comparisons between groups invalid. Absence of measurement invariance between groups is, however, difficult to observe with classical psychometric tests (Vandenberg & Lance, 2000; Millsap, 2011). Using CFA it was examined if the factor structure was comparable between different subgroups. These subgroups were based on demographic factors (e.g. gender), personal factors (e.g. perceived health), and work-related factors (e.g. psychological job demands). To do so first the hypothesised underlying factor structure was assessed. While many studies examined the factor structure post-hoc, this was the first study to use CFA to assess the factor structure with an a-priori model. With the inclusion of some small adaptations, the hypothesised 1-factor structure (see van Veldhoven, 2008) of the NFR scale showed good fit. For the CIS the hypothesised 4-factor structure (see Janssen, Kant, Swaen, Janssen, & Schroer, 2003) showed good fit, also with the inclusion of some small adaptations. Comparing these baseline models within the different factors (i.e. demographic, personal, and work-related) showed that there was no indication of measurement variance
between any of the subgroups of any of these factors. Chapter 3, as such, shows that the factor structure hypothesised in the literature can be confirmed on an a-priori basis and that this factor structure is comparable across a myriad of subgroups that are often the focus within public mental health.

The following chapter, chapter 4, extended these results to examine whether the NFR scale and CIS were also invariant over time. This is an important assumption within instruments which often goes untested – hence a perfect illustration of a 'latent defect’. Such stability of measurement over time is crucial for monitoring purposes of both instruments. If variance over time in measurement exists, scores cannot be meaningfully compared, hence hindering a valid monitoring. The findings of this chapter show, however, that the CIS and NFR scale are measured invariantly over time. This opens the door for monitoring and the assessment of psychometric properties important for this purpose. The smallest detectable change for both instruments indicates which change, over time, can be deemed statistically significant within an individual. These scores showed that especially within the NFR scale it was difficult to differentiate within an employee who has elevated NFR scores. Reliability was, however, very good for both scales. It was shown, however, that this was (partly) due to the floor effect within the scores of the NFR scale (also over time). It is, therefore, suggested that, ideally, both instruments are placed within a broader range of instruments to effectively monitor the impact of the psychosocial work environment.

The last chapter of this trilogy, chapter 5, delves deeper into this over time stability of the CIS. To correctly understand and grasp this stability it is important to breakdown the state of fatigue into a time-varying (occasion) and -invariant (trait) component. This also applies for its relation with other adverse health outcomes such as sickness absence. Using the trait-state-occasion (TSO) model (Prenoveau et al., 2011; Prenoveau, 2016) this chapter showed that the trait component of fatigue explained 71% of the state of fatigue. The occasion specific component in contrast only accounted for 28%. The remaining 1% was explained by the previous occasion component. These results show that fatigue, or at least the CIS, has a high trait component. An underlying assumption of fatigue seems, however, to be that it is quite reactive to changes in the environment of a person. This study shows that within a population of employees this assumption does not hold. If not correctly identified, such an assumption could lead to a misjudgement of the results and related conclusions would be hampered (Hamaker, Kuiper, & Grasman, 2015).

Besides this decomposition to study the state of fatigue, as measured by the CIS this chapter also examined the relation between the decomposed components and sickness absence. Again, this relation – often believed to be based on changes in fatigue – could be associated with sickness absence. The significance as predictor for future sickness absence of both the occasion and the trait component of fatigue was therefore examined. It was found that the trait component, uniquely, predicted future sickness absence within a (2-year) time interval within the MCS. This was found for all three different (dichotomous) operationalisations of sickness absence that were taken into account: Spell (consecutive sick days; cut-off ≥ 21 days), duration (total number of sick days; cut-off ≥ 42 days), and frequency (number of spells; cut-off ≥ 3 times). As such it can be concluded that fatigue has a considerable time-invariant component. As this component is also related with other adverse health outcomes, preventive measures and treatment should take this breakdown into account.

Chapter 6 includes a simulation study that was designed to validate a newly proposed informative fit index within Bayesian CFA. While all previous chapters used a frequentist framework,
Bayesian CFA models can be advantageous within specific situations over frequentist CFA models. If a multitude of subgroups (e.g. >40) are for example examined within a frequentist framework, it is very difficult to ‘detect’ sources of variance between the subgroups. Bayesian CFA has well suited methods to deal with such issues, also within large samples (van de Schoot, Kaplan, et al., 2013). Within large sample Bayesian CFA models, however, a valid fit index that provides an objective assessment of model fit is lacking. Based on the root mean square error of approximation (RMSEA) from the frequentist framework the Bayesian RMSEA (BRMSEA) was therefore proposed. This study shows that this BRMSEA performs very well within large samples successfully accepting models with no or a minor misspecification and successfully rejecting models with a major source of misspecification. The added value of the BRMSEA was demonstrated within an empirical illustration using the MCS. This illustration examined the factor structure of a subscale of the Job Content Questionnaire (Karasek, 1985). It was shown that the classical fit index becomes uninformative within large samples whilst the BRMSEA successfully indicated that the proposed factor structure of the Bayesian CFA model resulted in a good fitting model. This could, for example, open the way for examining structural equivalence of this subscale between a multitude of subgroups (e.g. job-title) using the approximate invariance method within Bayesian CFA (van de Schoot, Kluytmans, et al., 2013). In conclusion it can be stated that the BRMSEA is well suited to evaluate model fit in large sample BSEM models by taking sample size and model complexity into account.

Within chapter 7, the general discussion, three important aspects of β-psychometrics are highlighted to deal properly with ‘latent defects’ regarding the measurement of public mental health. The first aspect concerns the identification of latent defects. Within this section it is illustrated that an active modelling approach is an absolute necessity to successfully identify latent defects within public mental health. The second aspect covers awareness of latent defects. Without this awareness latent defects will always prevail and identification will be a continuous up-hill battle. The discussion covers specific aspects that could increase this awareness (e.g., education and critical review of scholarly work). The third, and final aspect, covers the treatment and preventive strategies regarding potential latent defects. Besides these aspects some conceptual considerations of β-psychometrics in general and some specific for this thesis were addressed. These considerations showed that other alternatives could provide a different perspective but would not greatly alter the results and conclusion of this thesis. It was furthermore stressed that a variety of instruments and populations were included. While all chapters have their profound merit within this thesis and individually, some limitations were also addressed. The development within the field of β-psychometrics regarding the implementation of Bayesian statistics was also considered. Fit indices within this field are however of limited value if large samples – which are typical within public mental health – are used. The newly designed BRMSEA of chapter 6 was therefore discussed in combination with the potential implication of this new fit index.

This thesis shows, as such, that ‘latent defects’ can be successfully identified using β-psychometrics. Using a range of advanced tools this thesis showed that awareness for latent defects is crucial, as measurement is present in almost all aspects within public mental health being it in research, clinical practice, or policy making. As β-psychometrics is based on a-priori model, a critical assessment of the underlying assumptions of an instrument – and its concepts ought to measure – is tantamount. If such an understanding is present ‘latent defects’ can be potentially identified as shown in this thesis. If latent defects prevail, however, they will
be lurking throughout the whole public mental health hindering valid judgment and decision making.

References


Janssen, N., Kant, IJ., Swaen, G. M. H., Janssen, P. P. M., & Schroer, C. A. P. (2003). Fatigue as a predictor of sickness absence: Results from the Maastricht Cohort Study on fatigue at work. *Occupational and Environmental Medicine, 60*, 71i–76. doi:10.1136/oem.60.suppl_1.i71


van Amelsvoort, L. G. P. M., Kant, IJ., Bultmann, U., & Swaen, G. M. H. (2003). Need for recovery after work and the subsequent risk of cardiovascular disease in a working population. Occupational and Environmental Medicine, 60(90001), i83–i87. doi:10.1136/oem.60.suppl_1.i83


Samenvatting
Publieke mentale gezondheid kan worden gedefinieerd als “een benadering op basis van de publieke gezondheidszorg van de mentale gezondheid en variatie binnen de populatie van de mentale gezondheid (Davies, 2014).” Publieke mentale gezondheid bestaat uit drie overlappende domeinen: Aandacht voor mentale gezondheid, preventie van psychische aandoeningen, en behandeling en revalidatie. Startpunt van al deze drie domeinen omvat meestal het meten en beoordelen van mentale gezondheid. Meten van mentale gezondheid is echter niet eenvoudig omdat aandoeningen zich meestal niet in aan- of afwezigheid manifesteren maar in een continuüm van licht tot matig of ernstig. De uitslag van de meting is dus niet dichtoom maar bestrijkt een breed spectrum. De gevolgen van de metingsuitslag zijn verstrekkend en resulteren op individueel niveau in een behandelingstraject en op populatieniveau in een bepaalde beleidskeuze. Het correct meten van mentale gezondheid is daarom van groot belang. De uitdaging betreft vooral ook het tijdig opsporen van potentiële verborgen gebreken. Als (onderzoek naar) publieke mentale gezondheid (mede) gebaseerd wordt op meetinstrumenten met een onbekende blinde vlek, kan dit een negatieve impact hebben. Het hoofddoel van dit proefschrift was daarom:

- Het identificeren van potentiële verborgen gebreken en de impact hiervan evalueren op de publieke mentale gezondheid

In deze samenvatting worden de motivering, methodes, en resultaten zoals gebruikt in de verschillende hoofdstukken kort uiteengezet en besproken.

In hoofdstuk 1 is het concept publieke mentale gezondheid en het meten ervan voor het eerst geïntroduceerd. De introductie richt zich op vier belangrijke punten: (1) Mentale gezondheid is een continuüm in plaats van een dichotome representatie (ziek/gesond). Zelfs als een dichotome representatie wordt gebruikt, of een andere vorm van categorisatie, is het onderliggende construct nog steeds continu. (2) Dit onderliggende continuüm is latent van aard. Dit wil zeggen dat publieke mentale gezondheid geen direct meetbare (manifeste) fenomenen kent. Daarom moet publieke mentale gezondheid worden gemeten aan de hand van indicatoren. Meetinstrumenten gebruiken dergelijke indicatoren om het onderliggende construct te omvatten. Het bepalen van de kwaliteit van dergelijke instrumenten is niet eenvoudig. Tot dusver wordt de kwaliteit van meetinstrumenten voornamelijk getest door middel van verkennende en ad-hoc analyses. Een dergelijke aanpak kent verschillende beperkingen wat er voor zorgt dat de meetinstrumenten kwetsbaar zijn voor verborgen gebreken. (3) Verborgen gebreken zijn “actoren die resulteren in de schending van (belangrijke) kwaliteitscriteria van een meetinstrument binnen publieke mentale gezondheid, zonder dat dit (direct) duidelijk is.” De verschillende aspecten in deze definitie worden verder bediscussieerd in de introductie van dit proefschrift. (4) β-psychometrie kan worden toegepast om de kwaliteit van instrumenten die publieke mentale gezondheid meten te evalueren door middel van een a-priori theoretisch kader. Dit theoretisch kader maakt het mogelijk om confirmatief te testen. Hierdoor is het mogelijk om specifieke hypotheses te testen op de aan- of afwezigheid van verborgen gebreken.

Hoofdstuk 2 vormt de eerste empirische studie van het proefschrift. In dit hoofdstuk wordt onderzocht of het afnemen van een meetinstrument in verschillende settings, gevolgen heeft voor het effect van sociale wenselijkheid waardoor de uitkomst verandert. Dit zou er voor kunnen zorgen dat de vergelijking tussen deze verschillende settings niet valide is. Het ‘verborgen gebrek’ dat hier wordt onderzocht was dus het potentieel verkerde gebruik van scores (van een meetinstrument) tussen settings. In dit hoofdstuk is gebruik gemaakt van de Nederlandse versie van de ‘Strengths and Difficulties Questionnaire’ (SDQ; Goodman, 1997). Dit is een
screeningsinstrument voor psychosociale problemen bij kinderen en adolescenten. Hierbij is de afname in een individuele setting vergeleken met de afname in een collectieve setting.

De individuele setting kan worden gezien als de één-op-één toepassing van een instrument (bijv. van de jeugdgezondheidszorg). In deze setting is een instrument vooral geschikt voor screeningsdoeleinden (Vogels, Crone, Hoekstra & Reijneveld, 2009; van Widenfelt, Goedhart, Treffers & Goodman, 2003; Goodman, 2001). Voor beide settings is gebruik gemaakt van data afkomstig van de jeugdgezondheidszorg. Voor de individuele setting zijn 6.594 vragenlijsten geanalyseerd die afgenomen zijn tijdens het periodiek gezondheidsonderzoek tijdens het tweede jaar van de middelbare school. In de collectieve setting wordt de informatie van een meetinstrument juist geaggregeerd binnen een populatie (van Roy, Groholt, Heyerdahl & Clench-Aas, 2006; Becker e.a., 2006; Ravens-Sieberer, Erhart, Gosch & Wille, 2008). Voor de collectieve setting is gebruik gemaakt van de informatie uit de jeugdmonitor die anoniem afgenomen wordt in klas 2 van het voortgezet onderwijs (Ministerie van Volksgezondheid, Welzijn en Sport & Centraal Bureau voor de Statistiek, 2014). Deze monitor bevat onder andere de SDQ als een indicator van psychosociaal welbevinden. Dit resulteert in de analyse van 4.613 vragenlijsten. Aangezien beide steekproeven uit dezelfde populatie voortkomen, is het aannemelijk dat eventuele significante verschillen een gevolg zijn van verschillen in de setting van de afname.

Confirmatieve factor analyses (CFA) laten zien dat er geen significante verschillen zijn in het onderliggende construct (psychosociale problemen) vergeleken tussen de twee steekproeven. Kinderen in de individuele setting, scoren desalniettemin lager wat betreft totaal probleemgedrag en de onderliggende subschalen in vergelijking met de kinderen in de collectieve setting. Om deze resultaten verder te kwantificeren is er gekeken naar de vergelijkbaarheid van de scores in de verschillende settings. Dit is onderzocht door het toepassen van afkappunten, gebaseerd op de collectieve setting, binnen de individuele setting. Deze procedure is vergelijkbaar met staand beleid aangaande het gebruik van afkappunten van de SDQ binnen de jeugdgezondheidszorg. Hierbij resulteert het gebruik van het 90e percentiel uit de collectieve setting binnen de individuele setting in slechts een klein aantal gevallen met een afwijkende score (2 tot 3%). Dit terwijl 10% zou worden verwacht – als de setting geen impact zou hebben op de afname.

Deze resultaten laten zien dat de SDQ dezelfde connotatie heeft in de individuele en collectieve setting. Het verschil in de gemiddelde scores laat echter zien dat er een verschil zit in de bereidwilligheid om een bepaald vakje aan te kruisen bij afname in de verschillende settings. Dit geobserveerde structurele verschil ondermijn de validiteit van onderling gebruik van scores tussen settings. De toepassing van de afkappunten, afkomstig uit de collectieve setting, binnen de individuele setting resulteert in conclusies die niet valide zijn. Om deze afkappunten correct toe te passen zullen ze daarom ook afzonderlijk binnen iedere setting moeten worden vastgesteld.

De hoofdstukken 3-5 richten zich op de mentale gezondheid van werkenden. De drie studies maakten allen gebruik van dezelfde studiepopulatie, namelijk de deelnemers aan de Maastrichtse Cohort Studie (MCS). Binnen de verschillende hoofdstukken wordt er echter wel gebruik gemaakt van verschillende selecties binnen deze populatie. De MCS is opgezet in mei 1998 en betrof toen 12.140 werknemers van 45 verschillende bedrijven. Gedurende deze baseline meting (T0) waren alle werknemers tussen 18 en 65 jaar oud, waarvan 8.840 man (73%) en 3.255 vrouw (Kant e.a., 2003; Mohren, Jansen, van Amelsvoort & Kant, 2007). Voor meer details aangaande de MCS zie Mohren e.a. (2007).
Naast dezelfde steekproef delen deze hoofdstukken ook de uitkomstmaten die bestudeerd zijn: herstelbehoeften en (langdurige) vermoeidheid. Herstelbehoeften is een uitkomstmaat welke zich vooral richt op de behoefte en duur van herstel na een werkdag. De normale herstelcyclus is idealiter afgerond na het werk maar in ieder geval voor het begin van de volgende werkdag (van Veldhoven, 2008). Langdurige vermoeidheid betreft de aanwezigheid van vermoeidheid, inclusief een overweldigend gevoel van vermoeidheid, voor een langere periode (Bultmann e.a., 2000; Kalkman, Zwarts, Schillings, van Engelen & Bleijenberg, 2008). Beide aspecten zijn belangrijke uitkomstmaten binnen de publieke mentale gezondheid. Daarnaast kunnen deze aspecten een belangrijke rol spelen bij preventieve strategiën die gericht zijn op het voorkomen of beperken van verdere negatieve (gezondheids-) uitkomsten (bijv., de Croon, Sluiter & Frings-Dresen, 2003; van Amelsvoort, Kant, Bultmann & Swaen, 2003; de Ravee, Kant, Jansen, Vasse & van den Brandt, 2009; van Amelsvoort, Kant, Beurskens, Schroer & Swaen, 2002; Silva-Costa, Griep, Fischer & Rotenberg, 2012). Herstelbehoeften wordt gemeten met de herstelbehoefteschaal (van Veldhoven, 2003) en langdurige vermoeidheid met de Checklist Individuele Spankracht (Vercoulen e.a., 1994; Beurskens e.a., 2000). De herstelbehoefteschaal en de CIS zijn beide gevalideerde meetinstrumenten met goede psychometrische eigenschappen (zie bijv., van Veldhoven, 2003; Bultmann e.a., 2000; Beurskens e.a., 2000; Sluiter, Frings-Dresen, van der Beek & Meijman, 2001; van Amelsvoort e.a., 2003). Ook al laten verschillende studies zien dat het valide meetinstrumenten betreft, verschillende onderliggende assumpties – verborgen gebreken – zijn (nog) niet adequaat onderzocht.

In hoofdstuk 3, het eerste deel van deze trilogie, wordt onderzocht of de herstelbehoefteschaal en de CIS al dan niet variabel functioneren in verschillende subgroepen. Variabele functioneren van een instrument tussen groepen is een typisch verborgen gebrek aangezien dit resulteert in een niet valide vergelijking tussen groepen. Afwezigheid van meetinvariantie tussen groepen is echter moeilijk vast te stellen middels klassieke psychometrische analyses (Vandenberg & Lance, 2000; Millsap, 2011). Met CFA is daarom onderzocht of de factor structuur vergelijkbaar is tussen verschillende subgroepen. Deze subgroepen zijn gebaseerd op demografische factoren (bijv. geslacht), persoonlijke factoren (bijv. ervaren gezondheid), en werk gerelateerde factoren (bijv. soort arbeidspatroon). Om dit te analyseren is eerst de onderliggende factor structuur geëvalueerd. Ook al hebben vorige studies deze factor structuur al op een post-hoc basis bekeken, dit is de eerste studie die CFA gebruikt om de factor structuur te evalueren op basis van een a-priori model. Met een aantal kleine aanpassingen laat de gehypothetiseerde 1-factor structuur van de herstelbehoefteschaal een goed passend model zien (van Veldhoven, 2008). Ook voor de CIS laat de gehypothetiseerde 4-factor structuur met een aantal kleine aanpassingen een goed passend model zien (Janssen, Kant, Swaen, Janssen & Schroer, 2003). Alle modellen worden vervolgens vergeleken met deze baseline-modellen. Bij deze vergelijking is er geen aanwijzing voor meetvariantie tussen groepen. Hoofdstuk 3, toont daarmee aan dat op een a-priori basis de gehypothetiseerde factor structuur kan worden bevestigd. Daarnaast toont dit hoofdstuk aan dat deze factor structuur vergelijkbaar is voor een groot aantal subgroepen die vaak worden vergeleken binnen de publieke mentale gezondheid.

In het hierop volgende deel, hoofdstuk 4, wordt dieper ingegaan op deze resultaten door te onderzoeken of er ook sprake is van meetinvariantie over tijd. Dit is een belangrijke assumptie bij meetinstrumenten die vaak niet getest wordt – hierdoor is dit een perfecte illustratie van een ‘verborgen gebrek’. Dergelijke invariantie van metingen over tijd is cruciaal voor het adequaat monitoren van bijvoorbeeld herstelbehoefte en langdurige vermoeidheid. Als meet-
variantie namelijk aanwezig is dan kunnen scores niet inhoudelijk met elkaar worden vergeleken waardoor kwalitatieve monitoring onmogelijk is. Hoofdstuk 4 laat zien dat zowel de CIS als de herstelbehoefteschaal geen meetvariantie hebben over tijd. Hierdoor is het dus mogelijk om deze belangrijke concepten te monitoren en evalueren. Daarnaast wordt er ook gekeken naar het kleinste verschil in score dat kan worden onderscheiden van meetfouten. Deze score kan dus worden gezien als een significant verschil binnen een persoon over tijd. Deze analyse laat zien dat het vooral bij de herstelbehoefteschaal moeilijk is om te differentiëren tussen werknemers met een verhoogde herstelbehoeftte. De betrouwbaarheid van beide meetinstrumenten blijkt erg goed. Dit komt echter deels door het vloer effect dat optreedt bij de herstelbe- hoefteschaal. De aanbeveling is daarom ook dat idealiter dergelijke meetinstrumenten worden geïncorporeerd binnen een bredere testbatterij om zodoende de effecten van de psychosociale werkomgeving betrouwbaar te meten.

In het laatste deel van deze ‘trilogie’, hoofdstuk 5, wordt dieper ingegaan op de longitudinale stabiliteit van de CIS. Om dit goed te kunnen duiden is het van belang om de geobserveerde toestand van vermoeidheid te splitsen in een component die varieert over tijd en een component die stabiel is over tijd. Daarnaast is er ook gekeken naar de relatie met andere negatieve gezondheidsuitkomsten zoals ziekteverzuim. Door gebruik te maken van het Trait-State-Occasion (TSO) model (Prenoveau e.a., 2011; Prenoveau, 2016) in dit hoofdstuk wordt aangetoond dat de tijd-invariante component van vermoeidheid 71% van de geobserveerde vermoeidheid verklaart. De tijd-variante component daarentegen verklaart slechts 28%. De overgebleven 1% wordt verklaard door het vorige meetmoment. Deze resultaten laten zien dat vermoeidheid, of toch in ieder geval de CIS, een hoge mate van stabiliteit over tijd kent. Een onderliggende aanname bij vermoeidheid lijkt echter te zijn dat het redelijk reactief is op veranderingen in de omgeving van een persoon. Dit hoofdstuk toont echter aan dat binnen een steekproef van werknemers deze assumptie niet wordt onderschreven. Als deze assumptie echter zonder kritische toetsing als startpunt wordt gebruikt, kan dit leiden tot resultaten en conclusies die niet valide zijn (Hamaker, Kuiper & Grasman, 2015).

Naast deze decompositie om de geobserveerde toestand van vermoeidheid te bestuderen, gemeten met de CIS, wordt er in dit hoofdstuk ook de relatie tussen deze verschillende componenten en ziekteverzuim geanalyseerd. Er is daarom gekeken naar de tijd-invariante en tijd-variante component in relatie met ziekteverzuim. Uit de analyses blijkt dat alleen de tijd-invariante component ziekteverzuim voorspelt binnen een 2-jaar durend tijdsinterval. Dit is zo voor drie verschillende dichotome operationalisaties van ziekteverzuim: een langdurige verzuimperiode (oopenvolgende ziektedagen, ≥ 21 dagen), lange totale duur (totaal aantal ziektedagen, ≥ 42 dagen) en een hoge frequentie (aantal periodes; ≥ drie periodes). De tijd-invariante component van vermoeidheid heeft dus de grootste voorspellende waarde met betrekking tot negatieve gezondheidseffecten. Aangezien ook al geconcludeerd is dat deze component een groot deel van de variantie van vermoeidheid verklaart, is het dus belangrijk dat preventieve maatregelen en behandelingen deze decompositie in ogenschouw nemen.

Hoofdstuk 6 beschrijft een simulatie studie die was opgezet om een nieuw ontworpen passings-index binnen Bayesiaanse CFA te valideren. Daar waar alle vorige hoofdstukken een frequentistisch kader gebruikten, kan Bayesiaanse CFA verschillende voordelen bieden in bepaalde situaties in vergelijking met frequentistische modellen. Wanneer er bijvoorbeeld een veelvoud aan subgroepen (bijv. > 40) is, kan het binnen een frequentistisch kader erg moeilijk zijn om de bron van eventuele verschillen te ‘ontdekken’. Bayesiaanse CFA heeft goede methoden om met dergelijke problemen om te gaan, ook als de steekproef groot is (van de Schoot, Kaplan
Binnen dergelijke Bayesiaanse CFA-modellen, met een grote steekproef, ontbreekt het echter aan een bruikbare passings-index die een objectieve beoordeling biedt voor de kwaliteit van een model. Op basis van de root mean square error of approximation (RMSEA) uit het frequentistische kader is daarom de Bayesiaanse RMSEA (BRMSEA) voorgesteld. Dit hoofdstuk toont aan dat deze BRMSEA zeer goed presteert in grote steekproeven en succesvol modellen accepteert met geen of een kleine misspecificatie en modellen afkeurt met significante misspecificatie(s). De toegevoegde waarde van de BRMSEA wordt verder aangetoond in een empirische illustratie met behulp van de MCS. Deze illustratie onderzoekt de factorstructuur van een subschaal van de Job Content Questionnaire (Karasek, 1985). Hier is aangetoond dat de klassieke passings-index (de ppp) niet-informatief is binnen grote steekproeven, terwijl de BRMSEA succesvol detecteert dat de voorgestelde factorstructuur een goed passend model is. De BRMSEA maakt het dus mogelijk om de aanwezigheid van meetinvariantie te onderzoeken bij meetinstrumenten tussen een groot aantal subgroepen (van de Schoot, Kluytmans e.a., 2013). Concluderend kan worden vastgesteld dat de BRMSEA goed geschikt is om in BSEM modellen met een grote steekproef de kwaliteit van een model te evalueren door rekening te houden met deze steekproefomvang en de complexiteit van het (factor-)model.

In hoofdstuk 7, de algemene discussie, worden drie belangrijke aspecten benadrukt om goed om te gaan met ‘verborgen gebreken’ in relatie tot het meten van de publieke mentale gezondheid. Het eerste aspect betreft de identificatie/opsoring van verborgen gebreken. Hierbij is geïlustreerd dat een actieve manier van modelleren een absolute noodzaak is om succesvol verborgen gebreken binnen de publieke mentale gezondheid te identificeren. Het tweede aspect betreft het vergroten van het bewustzijn voor dergelijke verborgen gebreken. Zonder dit bewustzijn zullen verborgen gebreken altijd de overhand hebben. De discussie gaat ook in op specifieke methoden om dit bewustzijn te vergroten (bijv. onderwijs en kritische beoordeling van wetenschappelijk werk). Het derde en laatste aspect betreft de aanpak en preventie van potentiële verborgen gebreken aangaande het meten van publieke mentale gezondheid. Naast deze aspecten worden enkele conceptuele eigenschappen van β-psychometrie behandeld. Hierbij wordt zowel gekeken naar meer generieke eigenschappen als eigenschappen die specifiek voor dit proefschrift van belang zijn. Hierbij wordt aangetoond dat andere methodes een alternatief perspectief kunnen bieden op de conclusies en bevindingen van dit proefschrift. Verder wordt benadrukt dat een verscheidenheid aan instrumenten en populaties is bestudeerd. Ook komen enkele beperkingen van de verschillende hoofdstukken aan bod. De ontwikkeling op het gebied van β-psychometrie met betrekking tot de implementatie van Bayesiaanse statistiek is ook aangestipt. Momenteel zijn passings-indexen hierbinnen echter nog niet erg bruikbaar bij grote steekproeven – welke typisch zijn voor de publieke mentale gezondheid. De BRMSEA in hoofdstuk 6 kan hier verandering in brengen. Dit proefschrift toont aan dat ‘verborgen gebreken’ met succes kunnen worden geïdentificeerd door middel van β-psychometrie. Met behulp van een variëteit aan geavanceerde analyses maken de verschillende studies die zijn beschreven in dit proefschrift duidelijk dat een toeneemend bewustzijn voor verborgen gebreken cruciaal is. Metingen liggen immers aan de basis van bijna alle aspecten binnen de publieke mentale gezondheid, betreffende onderzoek, klinische praktijk en beleidsvorming. Aangezien β-psychometrie gebaseerd is op a-priori modellen is het van groot belang dat de onderliggende aannames van een meetinstrument worden geëxpliceerd alvorens het mogelijk is de kwaliteit van een meetinstrument te beoordelen. Als een dergelijk begripsniveau met betrekking tot een meetinstrument aanwezig is, kunnen ‘ver-
borgen gebreken’ worden geïdentificeerd. Dit is belangrijk omdat verborgen gebreken anders een negatieve uitwerking hebben binnen de gehele publieke mentale gezondheid, waardoor een valide oordeel en besluitvorming worden belemmerd.

**Literatuur**


Janssen, N., Kant, IJ., Swaen, G. M. H., Janssen, P. P. M. & Schroer, C. A. P. (2003). Fatigue as a predictor of sickness absence: Results from the Maastricht Cohort Study on fatigue at work. *Occupational and Environmental Medicine, 60*, 71i–76. doi:10.1136/oem.60.suppl_1.i71


Samenvatting

The Maastricht Cohort Study. *Occupational and Environmental Medicine*, 60, 32i–39. doi:10.1136/oem.60.suppl_1.i32


Valorisation addendum
Introduction

As the findings have various implications for future public mental health practice and policy, valorisation of these results is important. In this chapter an outline is given regarding the valorisation potential of the research findings presented in the thesis. Valorisation during the project (e.g. sharing of results) will also be addressed.

The goal of this thesis was to identify potential latent defects of measurement instruments and evaluate their impact within the public mental health. Therefore, some specific potential latent defects of measurement instruments were addressed using $\beta$-psychometrics.

As addressed in the discussion identification of latent defects and increasing the awareness in public mental health professionals is of the utmost important to increase and maintain a high level of quality in research and practice. This thesis, furthermore, shows that using the right preventive strategies measurement of public mental health can achieve this high level of quality. With adequate preventive measures the measurement, assessment, and evaluation of public mental health can be increased.

This high quality measurement of public mental health is important as conclusions, both for individuals but also for policies on different levels, are often (partly) based on this measurement. The challenges of measuring public mental health may appear small and unimportant at first glance but, they may have far-reaching consequences. That is, if latent defects prevail they will be lurking throughout the whole process hindering valid judgment and adequate decision making. As such, hampered measurement can impede the care for vulnerable children within the preventive youth health care, but also nation-wide policy that addresses preventive programs (e.g. to decrease physical and psychological job demands of employees within specific economic sectors).

Valorisation of findings during the PhD trajectory

Results of the current thesis were shared with both researchers and professionals within the public mental health. Especially the results of a pilot study (Hoofs, Jansen, Mohren, Reijis, et al., 2015) were shared at multiple occasions. The subject of this pilot study was the usability of the electronic health records of the preventive youth health care for research purposes. It was within this same preventive youth health care that the results were shared with doctors, managers, and IT-specialists. All these groups were involved with the electronic health records and were, as such, in a position to increase their quality. With interactive presentations the results of this pilot study provided insight in the data quality of the health records. As this quality was highly diverse these results also illustrated the importance of guidelines to facilitate and assure uniform reporting. Shortcomings in the electronic health records were, furthermore, not only hampering research but also the application within the individual setting. To show the benefits of a higher data-quality potential use-cases were also explored within these presentations. Education is, as indicated in the discussion of this thesis, crucial for the understanding of the importance of the quality of the measurement within public mental health. These results were, therefore, not only shared with current professionals.

\[8\] It should be furthermore noted that some variables within the electronic health records are relevant and important for the Dutch Health Care Inspectorate
within the preventive youth health care but also with doctors in training. The importance of a critical assessment of measurement implications within public mental was therefore addressed by showing the implications of the pilot study. This will not only result in a higher data quality regarding the electronic health records, it can also facilitate their evidence based practice with respect to measurement (instruments).

Valorisation of the results took also place with respect to the findings of chapter 2 investigating the Strengths and Difficulties Questionnaire (SDQ). One of the findings of this chapter was that cut-off points of the SDQ cannot be used directly between different settings. Future application of these cut-off points should therefore be based on thresholds that are based on the setting at hand. If cut-off points from a setting are, nevertheless, applied within a different setting these cut-off points should first be re-evaluated within this different setting. To increase the likelihood that these findings are actually implemented within the public mental health and increase the valorisation of this thesis as such, two specific methods were used. First, the findings were not only published within an international (English) journal but also within a national journal specifically directed at (health) professionals (i.e. TSG). Large parts of the manuscripts were therefore rewritten in order to increase accessibility and readability for a wider public. This was further strengthened by the second method, illustrating the effect of cross-use of cut-off points on actual data. Using the cut-off points based on data of one setting (i.e. collective) on the data from a different setting (i.e. individual) showed that the observed percentage of children that scored above the cut-off point was much lower than expected, further underscoring the relevance of these findings.

Future applications of the generated knowledge

As this thesis shows that latent defects can have a large impact on the quality of measurement instruments within the public mental health, the awareness for latent defects should be increased. Target groups for which this awareness should be raised are considered in a separate section of this valorisation chapter. Nevertheless, this thesis clearly illustrates that for future application of measurement instruments their quality should be critically assessed (both for newly developed and existing instruments). This thesis furthermore shows that this assessment is best served by using statistical methods within the β-psychometrics. As argued in the discussion this assessment of the impact of latent defects is also possible on a post-hoc basis. While a-priori testing should be favoured, β-psychometrics makes it possible to successfully identify latent defects if the measurement instrument is already used. This makes it possible to continuously test (critical) quality issues of a measurement instrument. This is furthermore advantageous as latent defects can strongly differ between assessments.

While the identification of latent defects within existing measurement instruments is crucial, preventing latent defects within measurement instruments is of course preferable. Developers of new measurement instruments should therefore critically assess if latent defects could be (or become) present within their measurement instrument. This does not only increase the quality of measurement instruments, it often also increases the understanding of an instrument during the development phase as β-psychometrics requires an active modelling approach. This approach is only possible if a conceptual framework of the instrument is present.

A specific (future) application of the generated knowledge within this thesis is the newly proposed and validated fit index within Bayesian confirmatory factor analysis (CFA), the Bayesian
Root Mean Square Error of Approximation (BRMSEA). The BRMSEA will increase the accessibility and validity of advanced statistical tools within Bayesian CFA. As the BRMSEA makes it possible to assess model fit within large samples Bayesian CFA, public mental health may be more easily served by the tools this methodology includes. That is, research within the public mental health is often characterised by its large sample size. While analysing large datasets has many benefits, assessing the fit of the statistical model can be challenging. The introduction of the BRMSEA is motivated by this challenge and chapter 6 shows that the BRMSEA is up to this task. Again, chapter 6 also includes a section with a clear (empirical) illustration to increase its valorisation. As such the BRMSEA is not only validated within a simulation study, but its added value is also shown within an illustration based on empirical data. Another future application of the BRMSEA is its potential implementation within Mplus (Muthen & Muthen, 1998-2012), one of the most used software package when it comes to (Bayesian) CFA, and Blavaan (Merkle & Rosseel, 2016), an open source software library for R. This will greatly increase the ease with which researchers can use the BRMSEA for the benefits of their research. With the adaptation of the BRMSEA within a software package that is used by so many researchers within the field of public mental health the impact of the generated knowledge within this thesis is greatly enhanced. Especially as researchers do no longer have to ‘actively’ search for the results of this study as these findings are now ‘brought’ to them through this software package.

As already indicated at the start of this addendum and in the conclusion of this thesis, the main goal of this thesis is to promote identification of latent defects and to increase the awareness within public mental health professionals for these defects. This thesis has all the ingredients to facilitate and achieve this increased awareness for the measurement of public mental health. More precisely, to increase the awareness for the valid assessment of measurement of public mental health. That is, as already indicated in the introduction of this thesis, the validation of measurement instruments within the public mental health is a difficult challenge. This thesis shows, however, that with this increased awareness, the right tools, modelling strategy, and conceptual framework it is possible to successfully identify potential latent defects. With β-psychometrics it is, to some extent, even possible to adjust for these defects without the need to completely re-design and -implement measurement instruments. If these three steps will set foot in the public mental health, it will increase the quality of results and conclusions based on measurement instruments. This will positively influence the public mental health on all different levels and dimensions (World Health Organization, 2001).

**Novelty in the topic of the PhD thesis and novelty in the findings**

When it comes to measurement in the public mental health it sometimes looks a bit sidelined. Measurement has to be shorter, faster, and more appealing but at the same time also more reliable, valid, and dynamic. While it is always a good approach to strive for improvement regarding measurement, its quality sometimes looks somewhat neglected. This can be costly, as it is difficult to change an instrument during research while a population could be potentially more easily increased (if money permits). The main novelty of this PhD is, therefore, that it allows for a moment of contemplation regarding the issues that could potentially violate the quality of measurement within public mental health. Not only does this thesis offers such an
insight with advanced techniques it also shows and suggest potential measures to overcome or at least limit the effects of these issues.

While most techniques applied are not a novelty as such, their application within the current framework of public mental health should be considered a novelty. As such this thesis brings together different paradigms and their corresponding techniques. The trait-state-occasion (TSO) model, for example, is often used in the field of developmental psychology (Cole, 2012). Applying it within the context of measurement within public mental health opened the possibility to analyse the stability of a measurement instrument over a longer period of time. The inclusion, furthermore, of the survival analysis within this TSO-model is, to the best of our knowledge, a first. As such this thesis shows the flexibility of structural equation modelling, and the new insights it can offer for a specific topic.

Another novel finding of this thesis is that the cross-use of cut-off points between different settings is not directly possible. As already addressed, this effect was also quantified by illustrating its effect on empirical data. Especially this quantification of this effect can be considered a novelty, as the effect of social desirability bias (SDB) is addressed in the literature (Krumpal, 2013). The majority of this research regarding SDB, however, focuses on its theoretical underpinning and mostly uses controlled settings to study its effect. As discussed in chapter 2, capturing the actual impact of the different setting within the public mental health (e.g. individual and collective) is almost impossible. Using real-world datasets from these settings makes it therefore possible to not only study its actual effect, but also to directly quantify this effect.

A technical novelty that was introduced in this thesis is the BRMSEA. The added value of the BRMSEA is already addressed in the previous section. It is, nevertheless, noteworthy that the implementation of this finding was met with some hefty criticism during the review process. While there is nothing wrong with strong critiques on novel findings, it is important that the scientific community is open to different approaches. The criticism was, furthermore, in stark contrast to the positive response that the manuscript has attracted. This seems to show that there is sometimes a difference between the gatekeepers of a (specific) methodology and its ‘actual’ users. While not necessarily invalid, it would sometimes be good if this former group would take this latter group into account when assessing the added value of a concept or methodology.

**Target groups**

As already indicated in the discussion chapter of this thesis, many different target groups could be considered for the findings of this thesis. The two groups that were highlighted were, the users of the measurement instruments and their developers. The meetings with the PYHC professionals were a first step towards the goal of informing and creating awareness within this former group. For the latter group the findings in this thesis can be used to quantify specific assumptions that are violated, increasing awareness in researchers that develop measurement instruments. This thesis, as such, can create an increased awareness in both groups as it identified some potential latent defects and showed their impact. The thesis, furthermore, also showed (preventive) measures to these latent defects that these group can apply.

Through the means of eduction of public mental health professionals this increased awareness should become reality. In most educational programs the curriculum seem to jump from
reliability and simple validity issues directly towards inferential statistics (e.g. regression). As indicated in the discussion this introducing $\beta$-psychometrics at an early stage to public mental health professionals will not only be beneficial for their statistical knowledge, but will also enhance their general conceptual awareness regarding public mental health. As $\beta$-psychometrics is much more than only statistical analysis it will increase the critical appraisal of public mental health professionals regarding the measurement, assessment and evaluation of public mental health.

**Conclusion**

The findings of this thesis have a wide range of valorisation targets. From professionals within the preventive youth health care to academics within the field of applied Bayesian statistics. Valorisation was furthermore actively pursued, most noteworthy by publishing in a national journal specifically directed at (health) professionals. I believe that this ‘translation’ of results of academic research is key towards a successful valorisation. It furthermore results in valuable input for further avenues of academic research. As such, the process of doing research becomes increasingly interactive. This will result in a more profound and faster iterative process from which the results will furthermore have wider support.

The findings of this thesis should have a continuing impact on the field of public mental health. It shows that many hidden and latent assumptions within this complex field of mental health can be made manifest. Doing so shows the strengths and weaknesses of currently used (measurement) instruments. While this thesis contains some specific examples, its rationale applies to the complete field of public mental health, and maybe even also within a broader context.

**References**


Dankwoord
Het dankwoord, eindelijk. Voor mij het laatste om te schrijven, voor jullie waarschijnlijk het eerste om te lezen... Voor mij waren de afgelopen vijf jaar vooral een kennismaking met heel veel verschillende leuke mensen en het prachtige Maastricht. Vooral deze eerste groep wil ik in dit hoofdstuk nog even bedanken voor deze periode!

**IJmert**, een aantal jaar geleden zijn we samen met een project gestart waarvan we beide niet echt wisten wat er uit zou komen. Al hadden we toen nooit kunnen bedenken dat de hoofdstukken zo zouden worden ingevuld als ze nu zijn, jouw flexibele geest heeft er voor gezorgd dat iedere wending natuurlijk aanvoelde. Daarnaast vond ik het fijn (en misschien zelfs wel een must) dat mijn promotor een gezonde blik heeft tegenover werken in de wetenschap. Wat me hier vooral altijd in bij blijven is de opmerking dat we ook leuk en interessant zouden vinden als we onderzoek zouden doen naar Keltische talen (o.i.d.). Dit verfrissende standpunt heeft er voor gezorgd dat we ons werk altijd serieus hebben genomen maar wel met een gezonde portie relativering. Louis van Gaal haalt altijd het totale mens principe aan, volgens mij doe jij dat al jaren. Altijd interesse in alles wat ik doe en deed, van de gezondheidszorg van Ama tot het wel een wee van Meeko. Maar ook over belangrijke zaken zoals auto's, en jouw geweldig neusje voor occasions (bijv. Passat), konden we het soms uren hebben. Als ik na vijven thuis was, dan wist Carlijn bijna altijd dat ik niet langer had doorgewerkt maar gewoon met jouw had zitten kletsen. Al stond er op jouw deur helaas ook wel eens 'niet storen', voor mijn gevoel was die altijd open.

**Nicole**, jij blijft de uitzondering die de regel bevestigt dat iedereen vervangbaar is. Ik heb zo onteigelijk veel respect voor jouw continue vasthoudendheid en oog voor detail. Ik zou allang gestopt zijn met het verbeteren van mijn eigen spelfouten (where/were, trough/through, etc.). Daarnaast hield je mij (en misschien IJmert ook wel een beetje) altijd op het spoor als het ging om het maken/halen van een planning. Ook jij had daarnaast een geweldige antenne voor allerhande zaken om het werk heen. Ik vond ook altijd fijn dat je mij de ruimte gaf in mijn soms een beetje onorthodoxe ideeën, zinsconstructies en aanpak. Hierdoor heb ik nooit het gevoel gehad dat het een dertiende punt in een dozijn promotietraject was maar echt eentje met zijn eigen smaak en invulling. Daarnaast heb ik ook geweldig veel geleerd en plezier gehad met het begeleiden van enkele studenten uit de master Work, Health, and Career. Yoy, Maren, Frederika, Manon, Lea, we hebben ze uiteindelijk allemaal geholpen om een mooi eindproduct op te leveren. Als klap op de vuurpijl hebben we met zijn allen de masterthesis van Yoy weten om te zetten in een gepubliceerde artikel!

**Maria**, bedankt voor begeleiding en het vasthouden van de rode draad gedurende het hele traject. Ook toen de focus van de GGD wegtrok heb je je altijd met hart en ziel ingezet voor het boekje zoals het er nu ligt. Daarnaast ben je altijd genereus en oprecht in je feedback wat zorgde voor een fijne samenwerking. Ik hoop dat we ook de komende tijd nog met elkaar kunnen optrekken zodat ik van je kan blijven leren.

**Jos**, jij staat terecht in heel wat dankwoordjes apart genoemd. Ik wil daar geen uitzondering op zijn. Een uitzondering is misschien wel dat ik je vooral wil bedanken voor de rondjes in de pauze, en de fijne gesprekjes in jullie altijd warme kamer. Ook de paar avondjes MVV zullen me bij blijven, met de laatste tegen de Graafschap als hoogtepunt dankzij Beppie Kraft.

**Danielle** en **Rianne** jullie zijn in het begin erg betrokken geweest bij de invulling van het proefschrift en vooral de eerste artikelen. Danielle bedankt voor je feedback en introductie binnen de GGD. Rianne bedankt voor het meedenken en -lezen bij het artikel over het digitale dossier in de JGZ. Jouw input hierbij was zeer waardevol en dit was vooral in het begin erg
prettig om gevoel te krijgen met het reilen en zeilen binnen de GGD.

**Rens**, gelukkig hebben we altijd contact gehouden na mijn master. De laatste jaren hebben we volgens wij allebei heel wat gaatjes in onze agenda’s gedicht met het bekijken van de mogelijkheden en uitwerking van de BRMSEA. Hier moesten heel wat heilige huisjes voor over hoop, maar het heeft uiteindelijk wel geleid tot een prachtige publicatie. Ik vond deze samenwerking erg prettig en het zegt ook genoeg dat deze altijd soepel verliep ondanks het feit dat we elkaar nauwelijks zagen. Ik hoop dat de komende jaren de BRMSEA goed tot zijn recht gaat komen, maar met jou als aanjager moet dat goed komen.

**Fleur**, nooit van kamer gewisseld en nooit van kamergenoot gewisseld. Ik denk dat we schrikken als we zouden uitreken hoeveel uur we samen op die soms snikhete kamer hebben gezeten. Ook jij hebt regelmatig, spelfout op spelfout uit mijn stukken gevist, bedankt daarvoor. Je hebt ooit beloofd dat als je BRMSEA zou googelen en je niet langer krijgt “bedoelde u RMSEA”, dat je me dan zou trakteren op een diner. Ik zou zeggen ... (ik weet het trouwens echt niet want het varieert nogal).

**Dave**, volgens mij ben jij het bewijs dat ieder Nederlands bedrijf minstens één Vlaming op de werkvloer moet hebben. Is het niet voor zijn Bourgondische inslag dan wel voor zijn discours. Jij zorgde hiervoor altijd voor een frisse blik een inzichten, was nooit te beroerd om te zeggen als iets een goed idee was of als het juist trok op niks. **Bram** als vervanger van Fleur had je natuurlijk een moeilijk taak, maar toch hebben we het de laatste maanden erg gezellig gehad en alle pizza’s van **Bancale 61** (aanrader) doorgesproken. **Carin**, heerlijk verstrooid maar scherp als het moest. Ik heb onze gesprekjes over van alles en nog wat altijd erg gewaardeerd en ik hoop dat je op je nieuwe werk lekker op je plek bent.

**Colinda**, ik denk dat ik niks verkeerd zeg dat wij in het begin eigenlijk niet een echte klik hadden. Het lot hielp echter een beetje mee en even min of meer hebben we denk ik heel veel steun, op verschillende vlakken, aan elkaar gehad. Af en toe gezellig lunchen in het ziekenhuis, en schetsen hoe onze carrière er verder uit zou zien en hoe hard ‘ze’ ons eigenlijk nodig hebben. Bedankt nog voor het regelen van het afscheid met de geweldige ‘Huib gaat boeven vangen’ gebakjes. Ik hoop dat we elkaar nog vaak spreken en niet alleen over kleine en grote R-issues.

**Linda**, toen het de laatste jaren een beetje leegliep wat PhD-studenten betreft hebben we toch nog heel veel gezellige momenten gehad op de afdeling. Hoogtepunten was natuurlijk de pubquiz die we kansloos verloren. En ik wil hier toch nog even jullie geweldig dagje-uit benoemen waarin ik voor het eerst kennis heb gemaakt met het fenomeen kajakken zonder water, alhoewel ik denk dat iedereen het heel vernieuwend vond denk ik dat de Koestal na afloop toch beter in de smaak viel.

Daarnaast wil ik natuurlijk alle mensen op de afdeling van **Epi** bedanken. Yvonne, Petra, en Marielle, jullie waren altijd bereid om iets uit te zoeken of uit te helpen. Conny, Nicole, en Jolanda altijd gezellig en toch een beetje de wortels van onze afdeling. Dat ‘onze’ zeg ik trouwens nog steeds, ook een jaar later. Ludo, ondanks dat we nooit formeel hebben samen gewerkt heb ik toch heel veel van je geleerd en regelmatig 2km samen naar huis gefietst. Matty, ik hoop dat je als ‘nieuwe’ vakgroepvoorzitter de veranderingen kan doorvoeren die je voor ogen hebt. Ik denk dat dat met jouw vrolijke en open karakter zeker moet lukken. Nu stop ik met namen anders wordt het echt een lange opsomming en die komt hieronder al.

**FBDA**: Mattijs, Edwin, Sander, Ranieri, Erwin, Ankie, Nivea, Bart, Femke, Hannah, Alle, Rolf, Xandra, Wauter, Daniel wat zijn we toch een geweldig team! Ik heb van iedereen al
zoveel geleerd, en ik ben van plan om dat nog heel lang te blijven doen. Iedere keer als ik een appje de lucht in slingerde dat mijn promotie weer een stapje dichter bij was werd dat altijd met veel enthousiasme ontvangen. Daarnaast ben ik jullie minder dankbaar voor al die dingen die ik de afgelopen tijd heb geleerd die mijn promotie een heel stuk makkelijker hadden gemaakt. Vooral Mattijs wijs ik hierbij aan als schuldige al zal hij waarschijnlijk voor al R als schuldige aanwijzen. Daarnaast ben ik natuurlijk speciale dank verschuldigd aan Edwin die zich al op dag drie op wierp als mijn mentor, je moet het maar durven. Daniëlle had als een van haar stellingen in haar proefschrift: "Het belangrijkste bij het combineren van werk en privé, is te zorgen dat je in ieder geval fluitend naar huis gaat. Fluitend naar het werk gaan, is mooi meegenomen", ik doe beide!

Janneke & Barbara of o-nee paranimfen zijn jullie tegenwoordig van mij. Zoals jullie weten de ‘closest thing’ dat jullie ooit zullen komen om mijn getuige te zijn, dus ik hoop dat het een beetje beviel. Voor ons is iedere dag een feestje behalve als er een goede reden voor is dat dit even niet zo zou zijn. Ik denk dat we genoeg van deze redenen meegemaakt hebben om van een promotie(-traject) gewoon één groot feest te maken. Natuurlijk kunnen Martijn en Gijs ook niet ontbreken in dit stukje. Bedankt dat jullie mijn zussen gek blijven laten zijn. Lilly, ik weet niet of ik jouw echt ergens voor kan bedanken, maar ik ben wel heel blij dat je er bent. Familie Boot, hier ligt dus het bewijs dat ik in al die jaren wel degelijk iets heb gedaan. Bij gebrek aan broers ben ik toch altijd erg blij, als er in een weekend een WIDH, Bili-cup, of pubquiz op het programma staat!

Papa & Mama, dit wordt jullie eerste “diploma uitreiking” sinds de middelbare school. Al heb ik nooit het gevoel gehad dat ik me moest bewijzen voor jullie, hopelijk is dit boekje iets van een proeve van bekwaamheid als ik dat mis heb gehad. Bedankt daarnaast voor het vertrouwen dat ze bij Hotak’68 de scherpe kantjes er wel af zouden krijgen, volgens mij is dat redelijk gelukt! Een dankwoord naar deze prachtige vereniging is dan ook zeker op zijn plaats. Het is fijn dat er een plek is waar je altijd je zelf kan zijn en op ieder moment kan komen binnenvallen. Het is fijn dat je binding hebt met een hele fijne groep mensen die los staat van alle andere dingen in het leven. Hier wil ik toch nog even Roland apart benoemen omdat er denk ik weinig mensen zijn met wie ik zo op een lijn zit als met jou, ondanks de drukte wogen we gelukkig nog steeds tijd te maken voor elkaar. Ook tijdens dit promotietraject heb ik daar altijd erg van genoten!

Carlijn, ons hondje staat achterop dit proefschrift (N50°48'38" O5°41'49"), wat willen we nog meer.
Huub Hoofs was born on April 10th, 1989 in Bergen op Zoom, the Netherlands. After completing secondary school at the Mollerlyceum in Bergen op Zoom in 2007, he studied Health Sciences at Utrecht University in Utrecht, the Netherlands. Huub obtained his BSc. degree in Health Sciences three years later. In 2010, Huub started a researchmaster in Development and Socialisation of Childhood and Adolescence. He obtained his MSc. degree in 2012. In his thesis he analysed the psychometric properties and convergent and predictive validity of an executive function test battery for two-year-olds (Mulder, Hoofs, Verhagen, van der Veen, & Leseman, 2014).

After graduating he started as PhD student at the Department of Epidemiology of the Faculty of Health, Medicine and Life Sciences, Maastricht University in Maastricht, the Netherlands. His research project concerned the assessment, measurement and evaluation within public mental health. This resulted in several national and international publications which form the basis of this PhD thesis. During these years he also obtained his University Teaching Qualification (Dutch: BKO). Huub has taught various courses on the subject of methods and statistics in the curriculum of the health sciences.

Currently Huub works as a data scientist at the Nederlands Forensisch Instituut, The Hague.

Publications


