# Aspects of validity in stress research

## Measurement properties and the application of self-reported stress questionnaires

Emina Hadžibajramović

Health Metrics, Department of Public Health and Community Medicine

Institute of Medicine

Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2015

Click here to enter text.

Aspects of validity in stress research

emina.hadzibajramovic@vgregion.se

ISBN 978-91-628-9587-7 (print)
Printed in Gothenburg, Sweden 2015
Printed by Ineko AB

To my parents

Azemina and Smail Hadžibajramović

# Aspects of validity in stress research

## Measurement properties and the application of self-reported stress questionnaires

Emina Hadžibajramović

Health Metrics, Department of Public Health and Community Medicine
Institute of Medicine
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

## ABSTRACT

Aim: To increase knowledge about validity evaluation and interpretability of a multi-item self-report questionnaire used in occupational health and stress research, and to investigate longitudinal associations between the psychosocial work environment and symptoms of burnout.

Method: The data come from a four-wave cohort study of public health care workers from the Region Västra Götaland. Rasch analysis was used for evaluation of measurement properties. A criterion based approach (CBA) was developed, and along with the median proposed for global scores in the Stress-Energy Questionnaire (SEQ). The CBA was applied for the SEQ-Leisure Time (SEQ-LT) and for the measurements of demands, decision authority, effort and reward. Longitudinal associations were analysed using mixed-effects regression models with random intercept.

Results: Good psychometric properties were found for the SEQ and SEQ-LT. The CBA was recommended for the SEQ. The CBA was applied to the SEQ and SEQ-LT, demands, decision authority, effort and reward. Investigated workplace factors were associated with increased symptoms of burnout.

Conclusion: The SEQ and SEQ-LT provide valid and useful tools for assessing work-related and non-work-related affective stress responses respectively. Rasch analysis is proposed for the evaluation of measurement properties. Increased awareness of the construction of global scores is needed. The CBA can be used for identification of the risk groups for adverse health effects, as defined by the theoretical foundations of the questionnaires, provided good measurement properties defined by the Rasch model. Longitudinal associations were found between demands, decision authority, effort and reward) and the symptoms of burnout.

**Keywords**: Affective stress response, Validity, Rasch analysis, Global scores

**ISBN:** 978-91-628-9587-7 (print)

**ISBN:** 978-91-628-9588-4 (PDF)

# SAMMANFATTNING PÅ SVENSKA

Bakgrund: Långvarig stressexponering kan leda till allvarliga hälsokonsekvenser, som till exempel utbrändhet. Individens upplevelse och tolkningen av stressexponeringen spelar också roll för stressreaktioner och eventuella hälsokonsekvenser. Stress-Energi formuläret (SEQ) är ett svenskt instrument och används ofta för skattningar av sinnesstämning i arbetet. Det är också viktigt att ta hänsyn till privatlivet för att få en bild av den totala stressbelastningen.

Syfte: Att validera SEQ och öka kunskapen om valideringsutvärdering av frågeformulär inom stressforskning; samt att undersöka longitudinella samband mellan psykosocial arbetsmiljö och symtom av utbrändhet.

Metod: Data kommer från en longitudinell kohortstudie av anställda inom Västra Götalandsregionen och Försäkringskassan. Rasch-analys användes för utvärdering av mätegenskaper. Kriteriebaserad metod (CBA) för beräkning av skalpoängen föreslogs och tillämpades på SEQ. Sinnesstämning utanför arbetet mättes med SEQ under fritiden (SEQ-LT).

Resultat: Goda mätegenskaper bekräftades för stress- och energiskalor för både SEQ och SEQ-LT. Därmed kunde en metrisk skala på intervallnivå konstrueras, och rekommenderas för användning istället för medelvärden. CBA användes på SEQ för identifiering av riskgrupper med höga och låga stress- och energinivåer. CBA har också tillämpats på SEQ-LT för att bestämma brytpunkter som indikerar höga och låga stress- och energinivåer på en metrisk skala, samt för identifiering av riskgrupper på skalor som mäter psykosociala arbetsfaktorer: krav, påverkansmöjlighet, ansträngning och belöning. Longitudinella samband mellan dessa psykosociala arbetsfaktorer och symptom av utbrändhet bekräftades.

Slutsats: SEQ och SEQ-LT kan användas för skattning av sinnesstämning i arbetet respektive på fritiden. Rasch-analys rekommenderas för validitetsutvärdering av självskattningsinstrument. CBA rekommenderas för identifiering av riskgrupper och för att underlätta tolkningen av skalpoängen. Ökad kunskap behövs om att skalpoäng kan konstrueras på flera olika sätt. Då de ovan nämnda arbetsfaktorerna visade samband med symtom av utbrändhet, är det viktigt att regelbundet mäta samt minimera upplevelser av dålig psykosocial arbetsmiljö. Resultatet kan användas i arbetsmiljöundersökningar för tidig upptäckt av personer som ligger i riskzonen för utveckling av klinisk utbrändhet.

# LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals:

I. Hadžibajramović E., Svensson E., Ahlborg G Jr. Construction of a global score from multi-item questionnaires in epidemiological studies. Working paper series 4/2013, Örebro University, 2013.

II. Hadžibajramović E., Ahlborg G. Jr., Grimby-Ekman A., Lundgren-Nilsson Å., Internal Construct Validity of the Stress-Energy Questionnaire in a working population, a cohort study. BMC Public Health, 2015, 15:180

III. Hadžibajramović E., Ahlborg G. Jr., Håkansson C., Lundgren-Nilsson Å., Grimby-Ekman A., Affective stress responses during leisure time – validity evaluation of a modified version of the Stress-Energy Questionnaire, Scan J Public Health 2015 doi: 10.1177/1403494815601552, Epub ahead of print September 2015. Sage Publications.

IV. Hadžibajramović E., Ahlborg G. Jr., Grimby-Ekman A., A longitudinal study of the impact of psychosocial job stressors on symptoms of burnout, synchronous and delayed effects, *Manuscript*

# ABBREVIATIONS

| | |
|---|---|
| CBA | Criterion based approach |
| CI | Confidence intervals |
| CTT | Classical test theory |
| D | Measure of disordered pairs |
| DCQ | Demand-Control Questionnaire |
| DIF | Differential item functioning |
| ERI | Effort-Reward Imbalance |
| IRT | Item response theory |
| JDC | Job demand-control |
| MTT | Modern test theory |
| PCA | Principal component analysis |
| PSI | Person Separation Index |
| SD | Standard deviation |
| SEQ | Stress-Energy Questionnaire at work |
| SEQ-LT | Stress-Energy Questionnaire during leisure time |
| SMBQ | Shirom-Melamed Burnout Questionnaire |
| WPF | Workplace factors |

# CONTENTS

# 1 INTRODUCTION

Work-related stress is common in many European countries and is a growing occupational health concern. Approximately 25% of workers in Europe experience work-related stress for all or most of their working time, and report that this has a negative impact on their health [1]. Psychosocial stress at work was found to be one of the most important factors behind the increase in sick-leave in recent decades [2, 3]. In terms of sectorial and occupational differences, the prevalence of psychosocial risk factors was greatest among employees in healthcare and social work [1].

The effect of prolonged exposure to stress at work can have serious consequences for health and well-being. One example is burnout, which is a mental condition, described as the result of long-term stressors related to psychosocial conditions at work. In addition to health and well-being, stress is also linked to performance-related outcomes such as absenteeism, presenteeism and work ability. As the burden of stress-related disorders is high and long-lasting, early identification of people at risk is of crucial public health interest. Consequently, it is important that measurement and evaluation of stress is done in a way that is both valid and reliable. Measuring stress exposures, stress responses and stress-related health outcomes is mostly based on self-reported questionnaires. Increased knowledge is needed of the validation process and the validity of questionnaires used to measure stress exposures, stress responses and stress-related health outcomes.

This thesis will focus on the measurement properties of a multi-item self-report questionnaire used in occupational health and stress research - the Stress-Energy Questionnaire (SEQ). An additional focus will be on evaluating longitudinal associations between the psychosocial work environment, affective stress response and symptoms of burnout.

## 1.1 Stress and health

### 1.1.1 Work-related stress

The word *stress* conveys a variety of meanings and there is no common definition of stress in the literature. The fact that *stress* can refer to stress exposure (stressors), stress reactions or responses (strain) as well as consequences in terms of stress-related ill-health, can lead to confusion when using this term.

Work-related or occupational stress refers to different aspects of the organisation, management and work design that can have a negative impact on an employee's health and well-being. Work-related stress has been defined as

a pattern of stress responses/reactions (emotional, cognitive, behavioural and/or physiological) caused by the adverse aspects of work stressors (work content, organisation, environment) and is a state of high levels of arousal, distress and feelings of not coping [4]. The consequences of these reactions could then result in health problems (physical, mental or both) [4, 5].

### 1.1.2 Stress exposure

Occupational or job stressors are events or conditions in the work environment that bring about strain. Simplified occupational stressors can be divided into: physical, psychosocial and management stressors [4]. As regards temporal aspects and duration, some stressors can be the result of discrete events (e.g. an accident of some kind) or a change process (e.g. a reorganisation), while other stressors are measures of more chronic working conditions that are indefinite in duration. Employees in different occupational groups in different sectors and in different cultural settings can be exposed to several different stressors that vary in duration and intensity. In this thesis, the focus is on psychosocial stressors in the public health care sector.

In the field of occupational health research, the focus of many studies has been mainly on work-related stressors and their effects on health. Two predominant job-stress models are the job demand-control (JDC) model [6, 7] and the effort-reward imbalance (ERI) model [8]. The job demand-control model or job-strain model is based on measurements of job demands combined with measurements of control or decision latitude. Job demands are the workload put on the individual. The control dimension refers to the employee's decision authority and skill discretion. The model predicts that job strain is a function of both job demand and control. This implies that demands are not the most important contributors to strain experiences. The amount of strain experience is influenced by the amount of control over demands the worker need to deal with. In other words, control will buffer the impact of demands on the level of strain. The most stressful situation is thus identified by the combination of high job demands and low control. The demand-control model was developed for work environments in which stressors are long-lasting.

The effort-reward imbalance model emphasizes both effort and the reward structure at work. Effort represents workload and obligations. Job reward consists of money, esteem and career opportunities, including job security. The model assumes that lack of reciprocity between costs and gains i.e. high effort and low rewards situations are experienced as stressful, and are a state of emotional distress with a particular propensity towards autonomic arousal and associated strain reactions. The ERI model seems to evoke adverse health by stimulating both psychophysiological and behavioural mechanisms [9]. Similar to the JDC model, the ERI is a measure of chronic working conditions.

In addition to working conditions, it is also important to study non-work-related stressors [10-12]. In a study investigating which stressors were reported to be important for the onset of exhaustion disorder, closely related to burnout, non-work-related stressors were almost as prevalent as work-related stressors [13]. It is well-supported that a work-life balance, i.e. the amount of each everyday activity, as well as the total amount of activities in relation to the available resources, has a relationship to health and well-being [14-16]. The opportunity to recover from the temporary effects of stress exposure, both during and after working hours, is important in order to avoid accumulation of strain [17, 18]. Consequently, non-work-related stressors also need to be considered in studies of work-related stress and health.

### 1.1.3 Affective stress response

An increasing volume of knowledge has been built up over the years about different pathways and the interplay between psychosocial stressors and health. In the field of stress research, a great deal of effort has been devoted to the understanding of psychological, physiological and behavioural mechanisms leading from stress exposure to stress response and the development of stress-related health problems. There are several different theories about the mechanisms behind physiological and psychological responses to stressor overload, potentially resulting in health problems, e.g. allostatic load theory [19], conservation of resources [20] and the cognitive activation theory of stress [21]. Some stress responses or reactions to exposure may occur immediately whilst others may take longer time to develop. Stress responses can be affective, behavioural or biological [22]. In this thesis, one of the focuses is on affective stress response.

The subjective evaluation of the stressfulness of a certain situation is referred to as appraisals or perceptions of stress. Negative emotional response is a reaction to a situation which a person perceives as stressful. As explained by Cohen et al. [22], a psychological model of stress posits that an affective stress response, i.e. a negative emotional response to the stressful situation is a requirement for a physiological stress reaction, which in turn increases the risk of adverse health effects. Negative emotional responses, such as mood changes, anxiety and frustration, are often immediate psychological reactions and are associated with physiological changes in the body [23]. Emotional states can be classified using Russell's circumplex model of affect [24]. This model posits that affective states arise from two fundamental neurophysiological systems, one related to a pleasure-displeasure continuum, the other to arousal or alertness. Negative states include: high arousal strains such as anxiety and irritation, low-arousal strains such as depression and exhaustion, and general negative psychological well-being [25].

Several questionnaires are available to assess the presence and magnitude of various aspects of affective stress response. One example is the Perceived Stress Scale, which is designed to assess whether situations in everyday life are perceived as stressful [26]. Another example is a Swedish questionnaire called the Stress-Energy Questionnaire (SEQ) based on Russell's model of affect [27, 28]. The SEQ is designed to measure affective stress response at work and was used in many Scandinavian studies [29-35]. It is the instrument in focus in this thesis.

### 1.1.4 Stress-related mental health problems

Psychosocial stressors at work in the form of high workloads, high demands, organisational changes and harassments have been recognized as the most important factors behind the increase in sick leave throughout the EU in recent decades [3]. There is robust evidence for associations between psychosocial risk factors and stress-related disorders [36], depressive disorders [37], common mental disorders [38] and burnout [39].

Burnout is a mental condition that has been described as the result of long-term stressors related to psychosocial conditions at work [40, 41]. The burden of mental and somatic symptoms due to burnout is high, often leads to long-term sick leave and has a high public health impact [39, 42]. Similar to many other conditions, several factors can act in concert to cause burnout. Moreover, the process is mediated by the subjective perception of the environment. Whether unfavourable working conditions are perceived as stressful, is subject to individual variation. The meaning and feelings that workers ascribe to the experiences of the situation is also important to measure. It has been proposed, for example, that an unfavourable work situation according to the JDC, may not lead to negative health consequences if the situation is not perceived as stressful by the worker [30]. On the other hand, according to Siegrist, a negative affect associated with the ERI may not always be consciously appraised, since it is a chronically recurrent everyday experience [8]. One focus of this thesis is to investigate longitudinal associations between psychosocial occupational stressors and burnout, even when affective stress response is not perceived.

## 1.2 Measurement

Measurement is a fundamental activity in both clinical work and scientific research. We observe people, objects, events, behaviours and mechanisms and try to make sense of these observations, i.e. we measure things of interest and try to quantify them. As opposed to, for example, clinical trials, where many clinical variables can be measured directly using various measuring

instruments (blood pressure, height, weight etc.), stressors, stress responses and stress outcomes are not directly observable and are hypothetical in characters.

Measurement has been defined as a set of rules for assigning numbers to objects in a meaningful way to represent quantities of attributes [43]. The most commonly known are laws of physics. For example, the rules for measuring of quantitative attributes such as height and weight, are well defined. Rules that uniquely characterise the object's attribute, such as length in metres, have been developed and consensus regarding the standardisation of units has been reached and is now taken for granted. As opposed to physical measurement, stress is a latent construct and indicates a state of elevated activation of bodily adaptive systems with coordinated manifestations at the affective, cognitive and behavioural levels. Latent variables are often referred to as *constructs or latent traits*. Their manifestations are measured by means of *indicator or manifest variables*, which are postulated to be proxies for constructs that are not directly observable. The measurement is thus not identical to the construct being measured. If it is of interest to draw conclusions about the construct, one must take into account the nature of the correspondence between the construct and the measurements.

Operationalization and measurement of latent constructs rely on theories. Based on the theoretical understanding of the world, we know that these phenomena exist and that they influence behaviour, but the phenomena per se are intangible. Stress reaction can be taken as an example. Although there is some empirical understanding about how this reaction is manifested, researchers need to agree on a variable that represents the degree of stress reaction meaningfully. Consequently, theoretical knowledge about the phenomena of interest is crucial for developing a measurement instrument. In this context measurement means estimation of the latent construct. Measuring devices in that case are often multi-item self-reported questionnaires.

The definition of meaningful rules for the measurement of the latent constructs i.e. qualitative variables such as *stress*, varies a lot, depending on the field of application, the paradigm and the measurement theory [43-49]. There are two main measurement paradigms, classical test theory (CTT) and modern test theory (MTT) or item response theory (IRT), which will be described later.

### 1.2.1 Multi-item questionnaires

As mentioned above, in order to measure phenomena that cannot be assessed directly, multi-item questionnaires are commonly used. Various terms are used for measuring instruments for subjectively reported latent variables: *questionnaire, rating scale, inventory, self-reported scale* etc. Irrespective of what they are called, these instruments are essential and are some of the most

valuable tools for data collection in epidemiological studies in general, and in occupational stress research in particular. In this thesis the word *questionnaire* will be used to describe the self-reported variables consisting of multiple items (questions) each answered on a rating scale with several ordered categories.

Items in a multi-item questionnaire are chosen in such a way that they capture the underlying latent construct. Defining which items should be included in a certain questionnaire is a matter of theoretical knowledge and empirical evidence [50]. The latent variable is considered to be what causes the item response. The strength, the magnitude or the quantity of the latent variable is thus presumed to cause an item or a set of items to take on a certain value, assuming that participants respond to items rationally and consistently.

### 1.2.2 Ordinal data

There are different levels or scales of measurement, and the numbers or symbols that constitute the measurement have different properties. Scales are commonly classified as nominal, ordinal, interval or ratio [43]. Nominal scales use numerals or other symbols that merely name or classify objects or events, without putting them in any order. Ordinal scales classify and ascribe a hierarchy to the objects, making operations such as "stronger than" or "larger than" meaningful. In multi-item questionnaires each item usually consists of a scale with several mutually exclusive response categories, so called ordinal variables. Usually, response categories are numerically coded, showing the magnitude, frequency etc. These values are rank-ordered, which means that each category has more of the attribute being measured than the previous category although, but the differences between the categories are unknown. Statements such as "twice as" are therefore not meaningful since the distance between the classes of objects is not defined and is not necessarily equal. Interval scales classify objects, ascribe a hierarchy and denote numerical differences that reflect the differences between the objects. The intervals between each value on a scale are equal, which means that besides "larger than", "twice as much" is also meaningful. Ratio scales are like interval scales but with a naturally occurring zero value, making all arithmetic operations meaningful.

Although it is tempting to use numerical coding of ordinal variables as numbers in statistical analysis, the numerals assigned to the response alternatives are arbitrary and can be changed as long as their ordering is preserved [43, 51]. A discussion about how statistical analysis of the ordinal data is to be performed has been the subject of an ongoing debate for a long time [52] and different solutions are offered within CTT and MTT. In applied research, many issues regarding the handling of ordinal data have been extensively discussed [53-57]. Statistical methods need to take into account the

non-metric properties of the ordinal data. Depending on the study design and the aim of the analysis, many methods exist that have been especially developed for ordinal data, for example many agreement measures for paired ordinal data [58-61]. A guidelines for statistical evaluation of ordinal data is provided by Svensson [62]. A review of methods for ordinal data is provided by Agresti and Liu [63, 64].

## 1.3 The validation process

The soundness of the data collected by means of questionnaires is judged by their measurement properties, i.e. *validity* and *reliability*, which are the key quality concepts. Validity refers to the ability of an instrument to measure what it is intended to measure. Reliability relates to the extent to which repeated measurements yield similar results. Reliability can be regarded as the quality of data and the validity as the quality of the decisions and inferences based on the questionnaire scores [50, 65]. Validity thus refers to the quality of decisions or inferences drawn from questionnaire data, and *validation* is a process in which evidence is collected to support the appropriateness, meaningfulness and usefulness of the decisions and inferences.

Validation is an ongoing process, and modified versions of the questionnaire at hand, or applied in new settings, or a new group of patient diagnoses, call for new evaluations [66]. The validity of data from questionnaires is a prerequisite for their applicability and involves accumulating evidence to provide a scientific basis to support study specific purposes [67, 68]. Validation practices vary across a number of academic disciplines. Within behavioural and social sciences, *psychometrics* has been developed as a speciality involving the measurement of unobservable phenomena. The terms *measurement properties* and *psychometric properties* are often used synonymously.

Moreover, *sensitivity* and *responsiveness* are also important and interrelated concepts. Sensitivity is the ability to detect differences between individuals or groups. Responsiveness refers to the ability to detect changes [69]. In addition, although not considered as a measurement property, interpretability of the scores is another important concept [70].

### 1.3.1 Construct validity

Historically, validity has been separated into content, criterion and construct validity, but the variation in terminology in the literature is extensive [71], and causes confusion. Several studies have shown that measurement property concepts such as validity and reliability are frequently misunderstood and misapplied [71-73]. The field of validation and questionnaire development

within epidemiology suffers from low status, and epidemiologists need to take the developments of research instruments and the validity of questionnaire data more seriously [74]. In contemporary conceptualisation, validity is a unitary concept and is referred to as *construct validity* [66, 75]. Multiple sources of construct validity evidence are required. These are: *content relevance*, *response process*, *relationship to other variables*, *internal structure* and *consequences*, explained below [71, 75-77]. The sources of validity evidence that need to be collected depend on the intended use and interpretation of assessment scores [66, 77].

The *content relevance*, also known as face validity, is an important source of validity, ensuring that the items represent the variable being measured, and is often based on judgements from experts in the specific field of research. Theoretical and empirical analysis of the *response process* is another important step in collecting the validity evidence. The *response process* is related to the quality control of all data flowing from assessments, such as ensuring that the items are understandable and recognisable to the respondents and eliminating errors associated with the questionnaire administration [76, 77]. The *relationship with other variables* is about convergent and divergent (or discriminant) evidence between variables, intended to assess similar and different constructs respectively [75, 78, 79].

*Internal structure* is related to reliability and item analysis. One aspect of item analysis is checking whether a particular item functions similarly for comparable groups of respondents (e.g. women and men), sometimes called differential item functioning (DIF). Another aspect of internal structure is checking whether a questionnaire designed to measure multiple constructs demonstrates heterogeneous responses in a pattern predicted by the construct. Similarly, a questionnaire designed to measure a single dimension, would require evidence of item homogeneity. The extent to which item interrelationships support the presumptions of the conceptual framework should be examined. Reliability refers to reproducibility or consistency of the scores over time and across groups and settings. The various types of reliability can be evaluated, each addressing the specific type of agreement, such as test-retest related to reproducibility or stability over time, parallel forms (different versions of an instrument) and inter-rater addressing agreement between different raters.

Finally, the *consequential* aspects of validity refer to the impact of assessment scores on the respondents. Some consequences follow directly from the interpretation of scores for the intended use, e.g. classifying symptom severity into low, moderate and high in order to differentiate between groups of patients who will receive a certain form of treatment. The process used to determine

cut-off points for global scores is related to this aspect of validity, since the scores in turn affect the decision-making processes [77].

### 1.3.2 Construction of global scores

To characterise a person's location on a latent construct, responses to individual items included in the questionnaire are combined into a single global score. In the literature, these scores are referred to as: *total, global, overall, aggregated, composite* or *raw scores*. In this thesis, the term global scores will be used. There are different ways of constructing global scores depending on the measurement paradigm and traditions within different research areas. In this thesis, four different ways of constructing global scores (mean, median, criterion-based and Rasch metric scores) will be presented and discussed in later sections. Firstly, certain properties and requirements for a scale construction will be explained.

*Unidimensionality* is a requirement for items responses to be combined into a global score. Unidimensionality is an important concept in the process of validation, and means that all items in the questionnaire must be indicative of the same underlying latent variable. *Interpretability* is mentioned as an important concept in the validation process [66, 70]. In theoretical job-stress models, some characteristics are described as being especially harmful to health. Taking the JDC as an example, the most stressful situation is identified by the combination of high job demands and low control. It is therefore important to be able to define which values are regarded as high demands on a global demand scale and which values are indicative of low control on a global control scale.

The usefulness of global scores is dependent on the properties of sensitivity and responsiveness. In other words, the scale needs to be sensitive enough to allow the question of whether two persons experience the same or different levels of latent construct (e.g. stress response) to be answered. Similarly, responsiveness implies the possibility to tell whether the level of the latent construct has been changed over time. Global scores can be constructed on a continuous scale or as categorical variable. If a continuous scale is applied, the unit of change on a global scale should be well defined and constant across the entire scales (equidistance scale categories), meaning that a one-unit change should reflect the same magnitude of change on a latent variable, regardless of the position on the global scale. Equidistance is implied by the properties of sensitivity and responsiveness.

*Sufficiency* is another prerequisite for global scores to be meaningful and useful. The concept of sufficiency is associated with how well the global scores represent the item responses. In other words, it should be sufficient to know the value of a global score to understand person's location on the latent

construct. In order to be regarded as a sufficient statistic, the global scores should contain all information about the latent construct captured by the item responses, i.e. no further information can be gained from responses to individual items. The global score is regarded as a proxy of a latent variable and the inference about a stress exposure for example should be the same regardless of whether the global score or the responses to individual stress items are recorded in data.

### 1.3.3 Classical and modern test theories

Various statistical methods are used for the evaluation of measurement properties and for the construction of global scores. Although there are some guidelines for what should be included in the quality evaluations of questionnaire data [66, 68, 74], there are no agreed standards for how this is to be evaluated statistically. The rules for the assignment of numerals to objects are usually based on statistical models for those data. Two main paradigms concerning measurements are classical test theory (CTT) [50, 68, 80] and modern test theory (MTT) [81-84].

To create construct-valid measurements certain criteria need to be fulfilled. Unidimensionality is an important concept in the process of validation in both MTT and CTT, and as mentioned above, a prerequisite for the construction of global scores. The main focus of CTT is on the global scores. CTT assumes a linear association between the latent variable and each item. An assumption within CTT is that the items are parallel, i.e. each items is an equally strong estimator of the latent variable. According to CTT, the actual state of a latent variable is its hypothetical true score, and the observed variable is a mixture of the true score and error. The observed score can be represented by the simple formula $X=T+E$, where X is the observed score, T is the true score and E is the error. A good item should yield a score that is relatively close to the true value. Errors are assumed to be random and their mean is assumed to be zero.

Within CTT, item reliability is established by means of inter-item correlations. Items that are more strongly correlated with each other are also assumed to be more correlated with the true score of the latent variables, and are thus better items. The greater the proportion of shared variation between the items, the more the items have in common and the more strongly they reflect a common true score. Furthermore, item reliability is extended to scale reliability. More items will yield higher scale reliability. The rationale behind this statement is that as more items are included in the scale, errors associated with each individual item are more likely to balance each other out and thus have a lesser effect on the total scale score. Under CTT a scale should be unidimensional and consist of multiple items that are highly correlated with each other. One measure for evaluating scale reliability is Cronbach's coefficient alpha, where

there is only one measurement at a time and not repeated measurement as is the case in test-retest studies. The higher the alpha value, the better the scale is considered to be. However, it is important to note that reliability indexes measure the precision of measurement, *given* unidimensionality. Unidimensionality is assessed by means of factor analysis. Many methods within CTT require normally distributed data, which is not the case with the ordinal data from questionnaires. Construction of global scores is usually done by creating sum or mean scores of item responses, and this requires interval level data.

In addition to the scale properties, performance of the individual items should also be investigated. In contrast to CTT, item response theory (IRT) or MTT stress the importance of item response models. One advantage of item response models is that no parallel items requirement is needed. Items in a questionnaire can vary in terms of difficulty. A collection of IRT models has been developed that are *stochastic models*, i.e. a person's item responses are assumed to be probabilistic. The probability of an item taking on a certain value is a function of two sets of parameters: the person's location on the latent variable, i.e. person parameter, and the characteristic of the item, i.e. item parameter. Consequently, the relationship between the locations of individuals on the latent construct (e.g. how stressful a certain situation feels) and the item responses can be explained using statistical models that describe the probability of an item response as a function of the latent variable.

An aspect considered in IRT, but not in the simple forms of CTT, is that items should function similarly between comparable groups, e.g. gender. Suppose, for instance, that an item asks how often you did have felt stressed during the past week, and it is measured on a scale with the response categories: *frequently, sometimes*, *rarely* and *never*. Do women and men interpret *frequently* in the same way? If not, then this is referred to as *differential item functioning* (DIF) and can be easily examined using IRT methods. In the presence of DIF, global stress scores would not be comparable between women and men. Another aspect of instrument validity is the category ordering of each item, i.e. whether the response categories work as expected. This aspect is easily examined using IRT methods but is not as straight forward using CTT. If the categories do not seem to have the intended ordering, this is categorised as a problem of *reversed thresholds.*

### 1.3.4 Rasch analysis

A special case in IRT is the Rasch model, named after the Danish mathematician Georg Rasch [82]. In contemporary use, the model is applied in the development and evaluation of measurement properties of multi-item questionnaires. A further purpose is to provide sufficient statistic, global score,

for the latent construct that is being measured by the questionnaire. In his original work, Rasch had a starting point in educational testing (student's reading ability) and he developed a model by making an analogy with the properties of physical measurement. Reading ability should thus be evaluated quantitatively, with positive real numbers defined as regularly as the measurement of height, and not through some arbitrary grading scale. In this way fundamental or objective measurement can be achieved. An important property of fundamental measurement is that it allows for arithmetic operations such as addition and subtraction.

The Rasch model operationalises the axioms of additive conjoint measurements, which are the requirements for the fundamental measurement construction [85-88]. The Rasch model for polytomous items [89, 90] was used in this thesis:

$$P\{x_{ni} = x\} = \frac{e^{-\tau_{1i} - \tau_{2i} \ldots - \tau_{xi} + x(\beta_n - \delta_i)}}{\sum_{x'=0}^{m_i} e^{-\tau_{1i} - \tau_{2i} \ldots - \tau_{x'i} + x'(\beta_n - \delta_i)}},$$

where $\beta_n$ is the location (stress level) of person $n$, $\delta_i$ is the difficulty of the item $i$, and $\tau_{xi}$; x=1,2…$m_i$ are the thresholds that partitioned the latent continuum of item $i$ into $m_i$+1 ordered categories. X is the score of the item.

A unique feature in Rasch analysis compared to all other approaches is that fitting the data to the Rasch model places both item and person estimates on the same log-odds units (logit) scale, and in the case of model fit these are independent parameters. The response structure required by the Rasch model is a stochastically consistent item order, i.e. a probabilistic Guttman pattern [91]. This means, for example, that persons who experience higher stress levels, are expected to assess more items with high stress categories, whereas persons with lower stress levels are expected to assess fewer items with high stress categories. Since this is a stochastic and not a deterministic model, there is room for random variation, which means that two persons with the same total score do not need to respond to all items in exactly the same way. However, for data to fit the model, this probability needs to be relatively low. The process of Rasch analysis is concerned with whether or not the data meets the model expectations. The adequacy of the fit is evaluated by means of multiple tests of summary fit statistics and, item and person statistics, as well as graphical examinations of fit.

Important concepts in the context of the Rasch analysis are invariance, unidimensionality, monotonicity, local independence and DIF. According to Rasch, using a ruler to measure height, for example, should have the same meaning regardless of whether it is a physical person or an object that is being

measured. This is known as a principle of invariance or objectivity. The invariance criterion implies that the items need to work in the same way (invariantly) across the whole continuum of the latent construct for all individuals. Given the same level of the latent trait (e.g. stress), the scale should also function in the same way for all comparable groups (e.g. gender). This is commonly known as differential item functioning (DIF). Monotonicity implies that the item responses are positively related to the latent variable.

The concept of *local dependency* is another important aspect. Construct validity requires that the latent variable explains all the correlation between the items otherwise the items are locally dependent. Local dependency is manifested in two ways – through response dependency and trait dependency. Response dependency is where items are linked in a way that the response to one item will depend on the response to another item. This may occur when a particular rating for one item implies logically the same rating for another item, e.g. two items reflecting reversed statements such as "I feel tired" and "I feel alert". Trait dependency is characterised by the presence of multidimensionality. Response dependency inflates the reliability and multidimensionality tends to decrease it [92], which is something Cronbach's alpha does not take into the account. Another disadvantage of Cronbach's alpha is that it is based on correlations computed for the item values in the sample and, there is thus a possibility that different samples with different variances will not yield equivalent values for this measure. In Rasch analysis, the Person Separation Index (PSI) is calculated instead of Chronbach's alpha, and is interpreted in a similar way, except that PSI is based on estimated person locations that are a non-linear transformation of the raw scores, which overcomes the above-mentioned drawbacks of alpha.

In Rasch analysis, local dependency is evaluated by means of factor analysis of item residuals and evaluation of residual correlations. The occurrence of any systematic relationship between residuals is interpreted as a violation of local independency. As opposed to traditional factor analysis, which is performed using the raw values of items, analysis of residuals takes into account both the item difficulty and the person locations. Conducting an analysis of residuals will reveal whether there are any systematic patterns among a subset of items after minimising the occurrence of difficulty factors. Whether it is multidimensionality or response dependency that is the source of violation is answered by the empirical design structure and the format of the questionnaire. Consequently, solid theoretical models underlying questionnaires are needed in order to understand the results of the Rasch analysis.

An advantage of the Rasch model over CTT methods is that the ordinal data can be used as there is no assumption of normal distribution. In addition, more detailed information about the items, persons and response categories is

obtained in a more feasible way. Given that data fit the Rasch model, construct valid and objective measurement is achieved and the total score is a sufficient statistic. In case that data does not fit the model, this is interpreted as an indication that the questionnaire does not have the good enough measurement properties and hence needs to be revised and improved.

## 1.4 Longitudinal associations

In occupational stress research, in-depth knowledge about the causal process between stress exposures, stress responses and stress outcomes is of interest. Theoretical stress models offer explanations and suggest mechanisms that need to be tested empirically. Cross-sectional studies do not provide the opportunity to explore causal relationships. To obtain such knowledge, longitudinal studies are needed where the same variables are measured at least twice across time (at least two waves) for the same sample of individuals.

In longitudinal studies, repeated observations of one individual over time are not independent of each other. For example, strain levels at one time point may have an influence on the strain levels at a later time point. Moreover, some individuals may react to an increase or decrease in stressor levels with an immediate change on the level of strain, whereas another take much longer time to react. Consequently, in an analysis of longitudinal data it is necessary to apply statistical methods that take into account a dependent structure of repeated observations and allows for individual variation.

## 1.5 Rationales for the thesis

The Swedish Stress-Energy Questionnaire (SEQ) for assessment of affective stress response at work [27] was included in a longitudinal cohort study of health-care and social insurance workers. To our knowledge, no analysis of the psychometric properties of the SEQ using modern analytical techniques has been published to verify the use of the global stress and energy scores. For the purpose of the cohort study, a modified version of the SEQ was also constructed, to measure perceived affective stress outside work, henceforth called the SEQ during leisure time (SEQ-LT). Modified questionnaires require an evaluation of validity for intended use.

Theoretical stress models, such as the job demand-control (JDC) model [6, 7] and the effort-reward imbalance (ERI) model [8], as well as the theory behind the SEQ, define the risk groups for adverse health effects. It is necessary to bring theoretical knowledge back into defining these risk groups, in order to increase the interpretability and usefulness of global scores from questionnaires.

Although accumulated evidence points to a relationship between unfavourable psychosocial working conditions and mental health problems, there are several methodological limitations in the existing evidence. For instance, a recent review examining the association between psychosocial working conditions and burnout, only identified six methodologically adequate longitudinal studies [39]. The evidence presented for many risk factors is based on just a few studies for each factor [36]. Moreover, there is a lack of studies where both the JDC and the ERI are evaluated simultaneously with regards to their associations with burnout. Although the importance of systematic studies of how stressor-strain relationships unfold in time was highlighted in the beginning at the millennium [93], there is still only a limited number of methodologically adequate, high-quality longitudinal studies, particularly studies with multiple time intervals, i.e. more than two waves [94].

# 2 AIM

The aim of this thesis was to increase knowledge about validity evaluation and interpretability of a multi-item self-report questionnaire used in occupational health and stress research, and to investigate longitudinal associations between the psychosocial work environment and symptoms of burnout.

Specific aims in Papers I-IV:

I) To find a method for constructing global scores from the Stress-Energy Questionnaire that will define high stress and low energy risk groups

II) To evaluate the construct validity of the Stress-Energy Questionnaire at work

III) To evaluate the construct validity of the Stress-Energy Questionnaire during leisure time

IV) To investigate longitudinal associations between psychosocial work environment and burnout, adjusted for affective stress responses at work and during leisure time.

# 3 MATERIAL AND METHODS

## 3.1 Data material

Data in all papers comes from a four-wave cohort study of employees in two human service organisations in Western Sweden. The cohort study covers a range of topics with the aim of longitudinally studying psychosocial working conditions, stress, health and well-being. The baseline data (T1) was collected in 2004 through a postal questionnaire sent to a random sample (n = 5,300) of 48,600 employees of the Region Västra Götaland, a large public healthcare organisation, and a random sample (n = 700) of 2,200 social insurance office workers in the same geographical area. An inclusion criterion of at least one year of employment (at least 50% of full-time employment) was applied. Three follow-ups were carried out with a time lag of two years, i.e. in 2006 (T2), 2008 (T3) and 2010 (T4). Social insurance workers were followed only on the first three occasions (T1-T3). The total response rate at baseline was 62% (n = 3,717). Response rates at follow-ups of those eligible (still employed and participated in a previous wave) were at T2 85% (n = 3,136), T3 83% (n = 2,233) and T4 72% (n = 1,422). Detailed information about questionnaires used in this thesis is presented in the next section.

Due to the selection criteria, the participants were mainly employed in the healthcare sector (86%). Approximately 85% were women. The three most common professions were nurse, assistant nurse and physician and the mean age was 48 years. Further demographic and study-specific details are available in published studies [95-97]. More detailed information about the datasets and inclusion criteria in each paper is shown in Table 1.

*Table 1. Subjects included in Papers I-IV.*

| Paper | Study population | Measures | Selection criteria |
|---|---|---|---|
| **Paper I**<br>Construction of a global score from multi-item questionnaires in epidemiological studies | T1<br>n=2,817 | SEQ | Complete items on all SEQ items at baseline. |
| **Paper II**<br>Internal construct validity of the Stress-Energy Questionnaire in a working population, a cohort study | T1<br>n=880 | SEQ | Complete items on all SEQ items at baseline (N=2,817). Balanced dataset regarding gender was required. Eligible and include were 439 men and 441 women randomly selected from a total of 2,378 women. |
| **Paper III**<br>Affective stress responses during leisure time- validity of a modified version of the Stress-Energy Questionnaire | T1<br>n=952 | SEQ-LT | Complete items on all SEQ-LT items at baseline and balanced dataset regarding gender. Eligible and included were 476 men and 476 women randomly selected from total of 2,755 women. |
| **Paper IV**<br>A longitudinal study of the impact of psychosocial job stressors on symptoms of burnout; synchronous and delayed effects | T1 n=3,209<br>T2 n=2,665<br>T3 n=1,970<br>T4 n=1,422 | SEQ<br>SEQ-LT<br>DCQ<br>ERI<br>SMBQ | Included were all participants employed in the Region Västra Götaland. |

SEQ=Stress-Energy Questionnaire, LT = Leisure Time, DCQ = Demand-Control Questionnaire, ERI = Effort-Reward Imbalance, SMBQ = Shirom-Melamed Burnout Questionnaire.

## 3.2 Measurements

### 3.2.1 Stress-energy questionnaire

The Stress-Energy Questionnaire (SEQ) is an adjective checklist developed to describe two critical aspects of mood at work [27, 28]. The original overall question to be answered through the checklist is: "How do you usually feel at the end of a normal working day?" In a modified version of the SEQ, the time perspective was changed to "during the past week" [98, 99]. Based on the theory of allostatic overload [100], we postulated that the dominant level of arousal during the past week rather than at the end of a working day would be more closely related to long-term stress exposure and consequently modified version of SEQ was used.

The SEQ is based on Russell's model of affect [24]. According to this model, stress and energy represent bipolar dimensions. Hence, the stress dimension ranges from positively evaluated low activation to negatively evaluated high activation. The energy dimension ranges from negatively loaded low activation to positively loaded high activation. Each dimension is operationalised using three positively oriented items (stress: *rested, relaxed, calm*; energy: *active, energetic, focused*) and three negatively oriented items (stress: *tense, stressed, pressured*; energy: *dull, inefficient, passive*). The response alternatives are: *not at all, hardly, somewhat, fairly, much* and *very much*. The interpretation of response categories goes in opposite directions for positive and negative items. For positively loaded items, *very much* implies the lowest stress level and the highest energy level (the most favourable response), while *not at all* is the least favourable response. The opposite is true for negatively loaded items.

Response categories are coded numerically (0-5) so that 0 always indicates the lowest stress and energy levels and 5 always indicates the highest (see Tables 2 and 3). Usually, a global score is calculated as a mean of the item responses to represent the latent dimension being measured. In previous studies, a mean value of 2.4 was proposed as neutral point (neither stressed nor calm) for the stress scale. The corresponding value for the energy scale is 2.7 [28]. However, due to the non-metric properties of the ordinal data, mean scores cannot be assumed to be valid without further investigation of the measurement properties. In this thesis transformed Rasch scores are used as global scores for stress and energy. These scores ranged from 0 to 5, with 0 being the lowest stress and energy levels, and 5 being the highest level. According to work by Kjellberg and Wadman, the most unfavourable condition is characterized by the combination of high stress and low energy [28]. A criterion-based approach (CBA) was used to define groups of persons with high and low levels of stress and energy.

*Table 2. Numerical coding of the response categories for the stress items*

| | Not at all | Hardly | Somewhat | Fairly | Much | Very much |
|---|---|---|---|---|---|---|
| Stressed | 0 | 1 | 2 | 3 | 4 | 5 |
| Pressured | 0 | 1 | 2 | 3 | 4 | 5 |
| Tense | 0 | 1 | 2 | 3 | 4 | 5 |
| Relaxed | 5 | 4 | 3 | 2 | 1 | 0 |
| Rested | 5 | 4 | 3 | 2 | 1 | 0 |
| Calm | 5 | 4 | 3 | 2 | 1 | 0 |

*Table 3. Numerical coding of the response categories for the energy items*

| | Not at all | Hardly | Somewhat | Fairly | Much | Very much |
|---|---|---|---|---|---|---|
| Active | 0 | 1 | 2 | 3 | 4 | 5 |
| Energetic | 0 | 1 | 2 | 3 | 4 | 5 |
| Focused | 0 | 1 | 2 | 3 | 4 | 5 |
| Passive | 5 | 4 | 3 | 2 | 1 | 0 |
| Inefficient | 5 | 4 | 3 | 2 | 1 | 0 |
| Dull | 5 | 4 | 3 | 2 | 1 | 0 |

### 3.2.2 Stress-Energy Questionnaire for leisure time

In the cohort study, the SEQ was used in a new way: for assessing affective response during leisure time. This modified version was called SEQ during leisure time (SEQ-LT). In the SEQ-LT, the overall question asked about feelings "*during the past week, when you were not working*". Otherwise, the SEQ-LT consists of the same 12 adjectives as the original SEQ. The response alternatives, the interpretation and the numerical coding of the items are also the same. Global scores for each dimension are calculated by means of Rasch scores. Since this was the first time the scale was used in its present form, the values on the stress and energy scales that identify high and low levels needed to be determined.

### 3.2.3 Job Demand-Control Questionnaire

JDC was measured using the Demand-Control Questionnaire (DCQ), which consists of five demand items and six control items [101]. In the present study, all the demand items and the two decision authority items, a sub-dimension of control, were used. The sub-dimension skill discretion was considered difficult to interpret in the context of this study since demands related to skills and learning are nowadays inherent in highly professional work such as healthcare

and are therefore expected. All the items were expressed as questions with four frequency-based response options (*often, sometimes, seldom, never*). The classification into high, medium and low levels of demand and decision authority was done using the criterion based approach (CBA) [102] and was computed in collaboration with experts on the subject (including Professor Töres Theorell, personal communication). Details of the classification are given in Paper IV.

### 3.2.4 Effort-reward imbalance questionnaire

The effort dimension of the Effort-Reward Imbalance (ERI) questionnaire consists of six items. One item regarding physical load is usually excluded when evaluating white-collar workers, which was also the case in the present study. The reward dimension was operationalised using 11 items, divided into three sub-dimensions: esteem (five items), promotion (four items) and job security (two items). All items were formulated as statements describing typical experiences at work, and were responded to in a two-step procedure. Firstly, subjects agree or disagree with an item statement. Secondly, if they agree, subjects are asked to evaluate on a four-point Likert scale the extent to which they feel distressed by the statement (*not at all distressed/somewhat distressed/distressed/very distressed*). The global scores for each dimension of the ERI were defined by the CBA in collaboration with experts (including Professor Johannes Siegrist, personal communication) and are described in details in Paper IV.

### 3.2.5 Symptoms of burnout

The Shirom-Melamed Burnout Questionnaire (SMBQ) was used to measure symptoms of burnout [41]. Important to note is that the SMBQ is measuring symptoms of burnout and not the clinical burnout. The SMBQ originally contained 22 items with four subscales: physical fatigue (eight items), cognitive weariness (six items), tension (four items) and listlessness (four items). All items are expressed as statements and are rated using a seven-point response scale (*almost never* to *almost always*). In the present study, a revised 18-item version (tension excluded) was used and proved to have good construct validity [97]. Instead of the mean score of the 18 items, a recommended transformed score was calculated [97]. This score ranges from 18 to 126, with higher values indicating a high degree of burnout symptoms.

## 3.3 Statistical analysis

In all the papers, descriptive statistics were given in percentages for categorical variables, and means and standard deviation (SD) for continuous variables. In

Papers II and III construct validation of the SEQ and SEQ-LT was evaluated by means of Rasch analysis. The criterion-based approach (CBA) was developed in Paper I and was used along with the median approach to define groups of individuals with high and low stress and energy levels. The CBA is also applied to DCQ and ERI in Paper IV. Longitudinal associations in Paper IV were analysed using mixed effects regression models with random intercept [103]. An overview of the papers and methods is given in Table 4. See each paper for detailed descriptions of the statistical methods.

### 3.3.1 Rasch analysis

The overall fit to the model was evaluated using the item-trait interaction ($\chi^2$ statistic), and mean person/item fit residuals. A statistically non-significant value of the $\chi^2$ statistic reflects the property of invariance across the trait. The mean person and item fit residuals are expected to be close to zero with a standard deviation (SD) of one. The reliability of the scale is reported as a Person Separation Index (PSI). Values of 0.7 and 0.9 are indicative of sufficient reliability for group and individual use respectively [104].

The fit of an individual item was evaluated using $\chi^2$ statistic of the item, the ability of the item to discriminate (item fit residuals are expected to be within the range ±2.5), the appropriateness of the response categories (threshold ordering), response independence relative to other items (residual correlations >0.2 above the average residual correlation) and the absence of DIF for gender and age.

DIF was tested by conducting an ANOVA of standardised residuals, which enables separate estimations of misfit along the latent trait, uniform and non-uniform DIF. Detection of DIF can be dealt with by splitting a misfitted item into two items, one item for women, with missing values for men, and the other for men, leaving women with non-responses [105]. In order to understand the nature and magnitude of DIF, the initial and resolved analysis can be compared in terms of parameter estimates, given the fit to the model [106, 107].

Trait dependency was tested using Smith's test of unidimensionality [108]. For this test, items loading positively and negatively on the first principal component of the residuals are used to make independent person estimates, and were contrasted through a series of independent t test [108]. Less than 5% of such tests would support unidimensionality of the scale. A 95% binomial confidence interval of proportions was used to show that the lower limit of the observed proportion was below the 5% level [108]. Possible local dependency can be accounted for by combining correlated items into testlets and comparing the model fit with the fit provided by the initial analysis [109]. Evaluation of the items and persons targeting in the sample were examined graphically using

a person item distribution graph. In the case of good fit, Rasch person estimates, which are logits, can be transformed into a convenient range (henceforth referred to as a metric score) [110].

### 3.3.2 Measure of disorder

Svensson's measure of disordered pairs (D) [111, 112] was calculated for comparison between different global scores. This measure is built up as the excess of concordant pairs over discordant pairs adjusted for tied observations. To calculate this measure, the pairs of observations are first arranged in a ($m_1$ x $m_2$) contingency table, with the main diagonal of increasing values oriented from the lower-left corner to the upper-right corner. Then the measure D is defined as follows:

$$D = \frac{\sum_{i=1}^{m_1} \sum_{j=1}^{m_2} x_{ij}\left(x_{ij}^{ul} + x_{ij}^{lr}\right)}{n\,(n-1) - t}$$

Where $x_{ij}$ is the number of individual classified to the *i:th* and *j:th* category respectively, $x_{ul}$ is $x_{lr}$ is the number of observations in the upper-left and lower-right region relative the *ij:th* cell (i.e. disordered pairs), respectively, and *t* is the correction factor for tied observations defined as:

$$t = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} x_{ij}\left(x_{ij} - 1\right).$$

The measure of disorder (D) is the proportion of disordered pairs among all possible combinations of pairs. Possible values of D range from 0 (complete ordering) to 1 (complete disorder). In the case of complete ordering D = 0 and no pairs are found in the upper-left or lower-right regions relative to the cells.

### 3.3.3 Mixed effect regression with random intercept

Mixed effect regression with random intercept was used to analyse longitudinal associations. A general form of random coefficient analysis of the longitudinal relationship between a continuous outcome variable Y and several predictor variables can be described as:

$$Y_{it} = \beta_{0i} + \sum_{j=1}^{J} \beta_{1j} X_{itj} + \beta_2 t + \varepsilon_{it},$$

Where $Y_{it}$ are observations for subject *i* at time *t*, $\beta_{0i}$ is the random intercept, $X_{ijt}$ is the independent variable *j* for subject *i* at time *t*, and $\beta_{1j}$ is the regression coefficient for independent variable *j*, *J* is the number of independent variables, $\beta_2$ is the regression coefficient for indicator of time *t* and $\varepsilon_{it}$ is the "error" for subject *i* at time *t*.

In this model the coefficients of interest are $\beta_{1j}$, as these regression coefficients show the magnitude of the relationship between the longitudinal development

of the outcome variable ($Y_{it}$) and the development of the predictor variables ($X_{ijt}$). This analysis combines a within-subject relationship and a between-subject relationship into a single regression coefficient [103]. The between-subjects relationships provides information about the relationship between absolute values at each time-point. The interpretation of the regression coefficient regarding between-subjects relationship is that a difference between two subjects in 1 unit in the predictor variable X, is associated with a difference of β units in the outcome variable Y. The within-subjects interpretation indicates that a change within one subject of 1 unit in the predictor variable X, is associated with a change of β units in the outcome variable Y.

In an autoregressive model, the value of Y at time point *t-1* is also included in the model. In an autoregressive model the value of the outcome variable *Y* at time point *t* is defined to be related not only to the value of the predictor variable *X* at time *t*, but also to the value of the outcome variable at *t-1*. The underlying idea behind the autoregressive model is that the value of an outcome variable at each time-point is influenced by the value of this variable one measurement earlier. To estimate the "real" influence of the predictor variables on the outcome variable, the model should correct for the value of the outcome variable at time-point *t-1*. A simple form of the autoregressive model is:

$$Y_{it} = \beta_{0i} + \sum_{j=1}^{J} \beta_{1j} X_{itj} + \beta_2 t + \beta_3 Y_{it-1} + \varepsilon_{it},$$

Where $\beta_3$ is the regression coefficient for outcome *Y* at time *t-1*, and all other parts of the model as described above. With the autoregressive model the between-subject part of the analysis is more or less removed from the analysis [103].

In Paper IV, longitudinal association between psychosocial work stressors (demands, decision authority, effort and reward) and symptoms of burnout (SMBQ) were analysed, with regard to two time aspects. The first analysis was called short-term effect, where both the workplace factors and the outcome were measured at the same time point on three occasions. A simplified model showing only the outcome Y (SMBQ), random intercept and workplace stressors **X** is shown here:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \beta_{0i} + \boldsymbol{\beta_1} \begin{pmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \boldsymbol{X_3} \end{pmatrix} + \cdots$$

Where $Y_{1,2,3}$ is the SMBQ at time 1 (2004), time 2 (2006) and time 3 (2008) respectively, $\boldsymbol{X_{1\text{-}3}}$ are workplace stressors at time 1-3, $\boldsymbol{\beta_1}$ are regression

coefficients associated with each stressor, $\beta_{0i}$ is the random intercept for subject *i*.

The second analysis was called the delayed effects model, where the workplace factors were measured two years before the outcome, and the simplified model is:

$$\begin{pmatrix} Y_2 \\ Y_3 \\ Y_4 \end{pmatrix} = \beta_{0i} + \boldsymbol{\beta_1} \begin{pmatrix} \boldsymbol{X_1} \\ \boldsymbol{X_2} \\ \boldsymbol{X_3} \end{pmatrix} + \cdots$$

Where *Y* (SMBQ) is measured at time 2-4 i.e. years 2006, 2008 and 2010 and workplaces factors at times 1-3, i.e. years 2004, 2006 and 2008. Autoregressive models were used for longitudinal for both short-term and delayed effects.

*Table 4. Overview of the papers included in the thesis*

| Paper | Aim | Subjects | Methods and measures |
|---|---|---|---|
| **Paper I**<br>Construction of a global score from multi-item questionnaires in epidemiological studies | To present approaches to the construction of global scores that take into account the non-metric properties of ordinal data and propose new approach for SEQ | Complete items on all SEQ items at baseline, n = 2,817. | Criterion-based approach, median approach<br>SEQ |
| **Paper II**<br>Internal construct validity of the Stress-Energy Questionnaire in a working population, a cohort study | To evaluate construct validity of the SEQ by means of modern test theory. | Complete items on all SEQ items at baseline (n=2,817). A balanced dataset regarding gender was required. Eligible and included were 439 men and 441 women randomly selected from a total of 2,378 women. | Rasch analysis<br>SEQ |
| **Paper III**<br>Affective stress responses during leisure time-validity of a modified version of the Stress-Energy Questionnaire | To investigate the internal construct validity of the SEQ-LT and to define cut-off points that could indicate high and low levels of stress and energy respectively. | Complete items on all SEQ-LT items at baseline and balanced dataset regarding gender. Eligible and included were 476 men and 476 women randomly selected 476 from a total of 2,755 women. | Rasch analysis, criterion based approach<br>SEQ-LT |
| **Paper IV**<br>A longitudinal study of the impact of psychosocial job stressors on symptoms of burnout; synchronous and delayed effects | To study longitudinal associations between workplace factors (demands, decision authority, effort and reward) and symptoms of burnout, and whether possible associations existed also in absence of affective stress responses. | Included were all those employed in Region Västra Götaland who participated in a four-wave cohort study, T1 n = 3,209, T2 n = 2,665, T3 n = 1,970, T4 n = 1,422. | Mixed effects regression models with random intercept<br>SEQ, SEQ-LT, DCQ, ERI, SMBQ |

SEQ = Stress-Energy Questionnaire, DCQ = Demand-Control Questionnaire, ERI = Effort-Reward Imbalance, SMBQ = Shirom-Melamed Burnout Questionnaire.

# 4 RESULTS

## 4.1 Paper I

The main finding of this study is that a new approach – the criterion-based approach (CBA) - was proposed for defining high and low values on stress and energy scales. The basic idea behind the CBA is that the theories underlying questionnaires should be brought back into defining of the risk groups. The theoretical and empirical knowledge of the experts in the field were considered when deciding cut-off values for the scales. Another approach for the SEQ proposed in this paper was the median score. Both approaches take into account the non-metric properties of the ordinal data. The mean score approach was the most commonly used method. Limitations of the mean score approach were commented on briefly.

### 4.1.1 Mean scores

When applying the CTT, the most common way of creating the global scores is to simply total or average the responses, which was also the procedure applied to the SEQ in the previous studies. In that case, all items are considered equally important and equidistance between scale categories is assumed. Taking the item *stressed* in the SEQ as an example, it is implied that the distance between *not at all stressed* and *hardly stressed* is the same as the distance between *somewhat* and *fairly stressed*, or the distance between any other two adjacent scale categories.

Responses from multi-item questionnaires can be assessed in many different ways, resulting in different response profiles. Since each item of the SEQ is assessed on a scale consisting of six ordered categories, there is a total of $6^6=46,656$ possible permutations for each dimension of the SEQ. The data, consisting of $n$ respondents can be presented in a matrix as shown in Table 5, where in this case stress items are put in the columns and the individual responses of each respondent are in the rows.

*Table 5. Schematic view of individual responses to stress items for a total of n respondents*

| Respondents | Stressed | Pressured | Tense | Relaxed | Rested | Calm |
|---|---|---|---|---|---|---|
| [1] | 4 | 2 | 2 | 0 | 1 | 0 |
| [2] | 2 | 0 | 2 | 4 | 0 | 1 |
| [3] | 2 | 2 | 1 | 4 | 0 | 0 |
| . | . | . | . | . | . | . |
| . | . | . | . | . | . | . |
| [n] | 0 | 3 | 1 | 1 | 4 | 5 |

Since it is the mean values that are of interest, the ordering of the items is not important. Consequently, the first three outcomes in Table 5 can be treated as equal. If we let $X_i$, i=1,2,…6 represent the six stress items, where $X_i$'s are discrete variables taking the values 0,1,…5, we can instead rewrite these three outcomes by arranging the responses from the lowest to the highest levels as shown below:

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
|---|---|---|---|---|---|
| **0** | 0 | 1 | 2 | 2 | 4 |

The total number of distinct response combinations can then be calculated by counting according to unordered sampling with replacement, and is reduced to: $\binom{6+6-1}{6} = 462.$

We can let $Y$ represent the mean score. Then $Y = \frac{\sum_{i=1}^{6} X_i}{6}$ is a discrete variable with 31 possible outcomes. To be regarded as a sufficient statistic, the mean scores for stress or energy assessments should contain all the information captured by the raw data and the inference about the stress/energy levels should be the same regardless of whether the mean score or the individual items, $X_i$ *for i=1,…6,* are recorded in the data. Respondents sharing the same mean score should be experiencing the same magnitude of the measured construct. However, the sufficiency of the mean scores may not always hold, due to the fact that many response combinations can results in the same value.

### 4.1.2 Criterion based approach

An alternative approach for the construction of global scores in the SEQ was introduced and recommended instead of the mean scores. This method is based

on previous work by Svensson, where two items measuring pain in a Short-Form-36 Health Survey were combined into a single global score [113]. In this paper, the method was extended to situations with more than two items and was termed a criterion-based approach (CBA).

The CBA scores are defined by experts in the particular field of interest on the basis of theoretical knowledge. The scores are based on the frequency distribution of the item responses into predefined response combinations. In this study, the criterion-based global scores for stress and energy measured by the SEQ were defined in collaboration with experts in the stress research field. For the sake of simplicity, only the global stress scores are presented here. The same rationale was applied to the energy scores.

The method of defining the stress scores according to the CBA was done in a sequence comprising several steps. Firstly, the six categories for each stress item were grouped into low, medium and high stress responses. For the items *stressed, pressured* and *tense,* the low stress responses were the categories *not at all* or *hardly* and the high stress responses were the categories *much* or *very much*. The reverse was the case for the items *relaxed, rested* and *calm.* The categories *somewhat/fairly* were defined as medium stress responses for all six items.

The frequency distribution of the item responses in the three predefined stress responses was then arranged into a matrix as shown in Figure 1. The matrix is arranged in such a way that the number of items with high stress responses is shown in rows and the number of medium responses in columns. With the six items and three response levels, there were 28 different response profiles. The first cell represents the response profile with zero high response and zero medium responses. In other words, all six items were responded with low stress responses. Expert judgement was applied to each cell in the matrix, and they were classified as high stress, medium stress and low stress. The CBA score for the response profile in the first cell was *low stress*.

| # High | # Medium responses: 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| **0** | Low | Low | Low | Low | Medium | Medium | Medium |
| **1** | Low | Low | Low | Medium | Medium | Medium | |
| **2** | Medium | Medium | Medium | Medium | Medium | | |
| **3** | Medium | Medium | Medium | High | | | |
| **4** | High | High | High | | | | |
| **5** | High | High | | | | | |
| **6** | High | | | | | | |

*Figure 1. Number of high stress responses (rows) and medium stress responses (columns) out of the six stress items, and the criterion-based scores for each response combination.*

Lastly, the possible response combinations for the six stress items were listed and presented in Table 6. Taking the response profile [26] as an example, none of the six responses were found in the lowest response categories (A), two items were assessed as either *somewhat* or *fairly* (B) and four items were found for the response categories defined as *high* (C). The CBA score for the response profile [26] was *highly stressed*.

The criteria for the high stress according to the CBA were: assessing at least four high stress responses out of a total of six items, or a combination of three high and three medium stress response categories. The criteria for a low level of stress were: at least four out of six responses fall into low stress categories while a maximum of one response falls into high stress categories, or three low stress responses in combination with three medium stress responses. All other response combinations were considered to belong to the medium stress level. The same rationale was applied to the energy scores.

*Table 6. The low, medium and high levels of stress according to median approach (defined as the third or the fourth of the six ordered responses) and the criterion approach (CBA) based on the number of responses found in response categories A, B and C out of six stress items in the Stress-Energy Questionnaire. For the items stress, pressured and tense: A= not at all/hardly, B= somewhat/fairly C= much/very much. For the items rested, relaxed and calm: A= much/very much, B= somewhat/fairly, C= not at all/hardly.*

| **Response profiles** | **Number of responses** | | | **CBA** | **Median** | **Median** |
|---|---|---|---|---|---|---|
| | **A** | **B** | **C** | | **3rd response** | **4th response** |
| [1] | 6 | 0 | 0 | Low | Low | Low |
| [2] | 5 | 1 | 0 | Low | Low | Low |
| [3] | 5 | 0 | 1 | Low | Low | Low |
| [4] | 4 | 2 | 0 | Low | Low | Low |
| [5] | 4 | 1 | 1 | Low | Low | Low |
| [6] | 3 | 3 | 0 | Low | Low | Medium |
| [7] | 4 | 0 | 2 | Medium | Low | Low |
| [8] | 3 | 2 | 1 | Medium | Low | Medium |
| [9] | 3 | 1 | 2 | Medium | Low | Medium |
| [10] | 3 | 0 | 3 | Medium | Low | High |
| [11] | 2 | 4 | 0 | Medium | Medium | Medium |
| [12] | 2 | 3 | 1 | Medium | Medium | Medium |
| [13] | 2 | 2 | 2 | Medium | Medium | Medium |
| [14] | 2 | 1 | 3 | Medium | Medium | High |
| [15] | 1 | 5 | 0 | Medium | Medium | Medium |
| [16] | 1 | 4 | 1 | Medium | Medium | Medium |
| [17] | 1 | 3 | 2 | Medium | Medium | Medium |
| [18] | 1 | 2 | 3 | Medium | Medium | High |
| [19] | 0 | 6 | 0 | Medium | Medium | Medium |
| [20] | 0 | 5 | 1 | Medium | Medium | Medium |
| [21] | 0 | 4 | 2 | Medium | Medium | Medium |
| [22] | 2 | 0 | 4 | High | High | High |
| [23] | 1 | 1 | 4 | High | High | High |
| [24] | 1 | 0 | 5 | High | High | High |
| [25] | 0 | 3 | 3 | High | Medium | High |
| [26] | 0 | 2 | 4 | High | High | High |
| [27] | 0 | 1 | 5 | High | High | High |
| [28] | 0 | 0 | 6 | High | High | High |

### 4.1.3 Median approach

Another approach that was tested in the paper was the median approach. The median is defined as the category, θ, such that P(X< θ) and P(X> θ) are both less than or equal to half. The median score for the variables stress and energy, as measured by the SEQ, was calculated for each individual by ordering the responses of the six items from the lowest to the highest stress levels. As the interpretation of the positively and negatively loaded items goes in opposite directions, the items *stressed, pressured* and *tense* were ordered from *not at all* to *very much* and the items *relaxed, rested* and *calm* from *very much* to *not at all*. The energy items *active, energetic* and *focused* were ordered from *not at all* to *very much* and the items *passive, ineffective* and *dull* were ordered the other way around. For the sake of simplicity, it was decided that for all items the lowest levels would be called *not at all* and the highest *very much.*

The median scores for the possible response combinations on the stress scale are shown in Table 6. Since the variables stress and energy are each measured using six items, both the third and the fourth ordered response could serve as the median level. When these two responses differ, the decision about which category is to be regarded as the median should then be made on theoretical grounds, based on previous research. The median global scores of *much* or *very much* were regarded as *high stress* or *high energy*. The median score corresponding to *not at all* or *hardly* was regarded as *low stress* or *low energy*.

## 4.2 Paper II

The main finding of this study was that the stress and energy scales of the SEQ have good psychometric properties, having accommodated for local dependency through the use of testlets. The results suggest that the stress scale can be used for assessment of work-related stress on the group level and the individual level. The energy scale is suitable for group evaluations only.

Summary statistics for both scales are shown in Table 7. For both scales the fit to the Rasch model was achieved after combining the positively and negatively oriented items into two testlets, see Table 7, analyses 2 and 4, for stress and energy respectively. Locations and fit statistics for individual items are presented in Table 8. All stress items had standardised residual fit values within the predefined range of ±2.5 and a non-significant $\chi^2$. All items had ordered thresholds and no DIF for gender was observed. The transformed Rasch scores (henceforth called metric scores) were provided and recommended for use in statistical analyses instead of the raw mean scores. The score range is set at 0 to 5 for convenience reasons, with 0 indicating the lowest stress level and 5 the highest.

As regards energy items, as seen in Table 8, out-of-bound fit residuals were found for the items *passive* and *inefficient.* However, these values were not statistically significant after the Bonferroni adjustment (adj. for six items p-values <0.0002). The item *dull* had disordered thresholds (Figure 2). The ordering of the thresholds suggests problems discriminating the first three categories. As seen in the Figure 2, the ordering of the 0 and 1, representing the categories *very much* and *much,* were reversed. Additional analysis was performed by rescoring the item *dull* into five categories, i.e. by collapsing the second and third response categories. This solution produced ordered thresholds for all items. However, the overall fit to the model was not improved. The change in individual person location from this additional analysis (mean 1.90, SD 1.49) compared with those from the original analysis (mean 1.87, SD 1.48) was marginal (mean difference -0.03, 95% CI -0.17; 0.11). Consequently, the rescoring did not seem justified. Alternative rescoring procedures were also checked and did not result in ordered thresholds.

*Table 7. Fit to the Rasch model.*

| | Item residual | | Person residual | | Chi square | | Unidimensionality | |
|---|---|---|---|---|---|---|---|---|
| **Analysis name** | **Mean** | **SD** | **Mean** | **SD** | **Value** | **p** | **PSI** | **Test %(95%CI)** |
| 1 Stress, 6 items | 0.48 | 1.30 | -049 | 1.19 | 52.79 | 0.52 | 0.92 | 10.5 (9.1;12.0) |
| 2 Stress 2 testlets | 0.31 | 0.56 | -0.62 | 1.03 | 13.57 | 0.75 | 0.87 | 4.4 (3.0;5.9) |
| 3 Energy, 6 items | -0.008 | 2.05 | -0.43 | 1.09 | 70.61 | 0.06 | 0.80 | 8.4 (7.0;9.9) |
| 4 Energy 2 testlets | 0.13 | 0.33 | -0.49 | 0.83 | 23.20 | 0.18 | 0.70 | 3.3 (2.2;5.1) |
| ***Ideal values*** | ***0.0*** | ***<1.4*** | ***0.0*** | ***<1.4*** | | ***>0.05*** | ***>0.7*** | ***(LCI <5%)*** |

*Table 8. Individual item fit.*

| | Location | Fit residual | Chi square | p-value |
|---|---|---|---|---|
| **Stress items** | | | | |
| Rested | -1,25 | 1,26 | 9,31 | 0,40 |
| Relaxed | -0,85 | -2,04 | 8,25 | 0,51 |
| Calm | 0,13 | 1,56 | 9,23 | 0,42 |
| Stressed | 0,29 | 0,33 | 6,19 | 0,72 |
| Pressured | 0,47 | 0,90 | 13,62 | 0,14 |
| Tense | 1,20 | 0,88 | 6,19 | 0,72 |
| **Energy items** | | | | |
| Passive | -1,45 | **-2,53** | 22,67 | 0,01 |
| Dull | -0,42 | 0,96 | 7,08 | 0,63 |
| Inefficient | -0,11 | **3,50** | 24,14 | 0,01 |
| Active | -0,09 | -1,01 | 7,18 | 0,62 |
| Focused | 0,70 | -0,32 | 4,42 | 0,88 |
| Energetic | 1,36 | -0,64 | 5,11 | 0,83 |

DIF for gender was detected for the item *passive*. Given the same level of energy, women rated slightly higher for this item compared to men. Consequently, additional analysis was done by splitting the item passive for women and men, showing almost no change in the fit to the model compared to the initial analysis.

Both scales indicated problems with local dependency, as fit residuals of positively and negatively oriented items clustered together. Cluster of items were in line with the theoretical foundation of the model used for development of the SEQ. According to this theory, positively and negatively loaded items are seen as bipolar dimensions with each scale. The problem of local dependency was resolved by forming testlets of negatively and positively oriented items.

As regards the energy scale, the items were not well targeted to the persons and as most of the participant reported very high levels of energy on all of the items, the variation in energy levels was limited. The lowest energy levels were not observed. Consequently, estimations of the Rasch scores for the lower part

of the scale might be unstable. However, the results indicated that the scale had satisfactory properties for the other parts. Transformation of the mean scores into Rasch metric scores is shown in the first two columns in Table 9. Based on the results in this study, we only recommend the use of the energy scale for values 0.9 or higher on the metric scale. Hence, for samples located in the middle or upper part of the scale, the energy scale works satisfactorily and can be used.

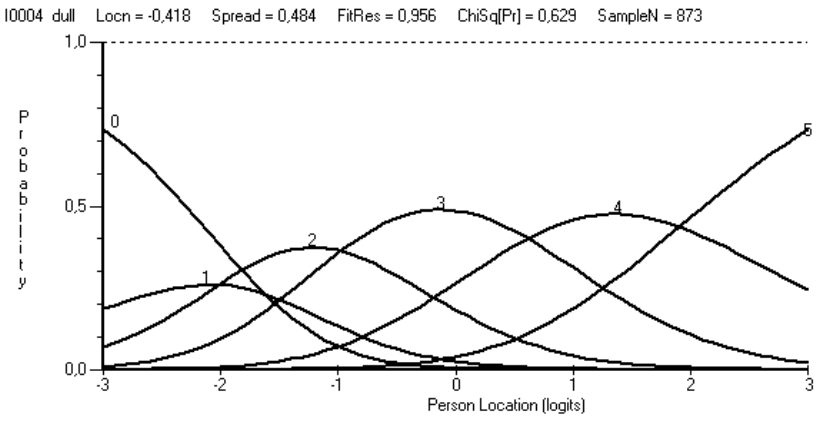


*Figure 2. Category probability curves item dull.*

*Table 9. Frequency distribution of the mean values, Rasch metric scores and criterion-based scores (CBA), energy scale of the Stress-Energy Questionnaire.*

| | Rasch | CBA | | | |
|---|---|---|---|---|---|
| **Mean** | **Metric score** | **Low** | **Medium** | **High** | **Total** |
| *1* | *0.90* | 1 | | | **1** |
| *1.17* | *0.98* | 1 | | | **1** |
| *1.33* | *1.06* | 1 | | | **1** |
| *1.50* | *1.15* | 3 | | | **3** |
| *1.67* | *1.23* | 2 | 1 | | **3** |
| *1.83* | *1.32* | 2 | 2 | | **4** |
| *2.00* | *1.41* | | 7 | | **7** |
| *2.17* | *1.51* | | 8 | | **8** |
| *2.33* | *1.61* | | 5 | | **5** |
| *2.50* | *1.71* | | 15 | | **15** |
| *2.67* | *1.82* | | 21 | 1 | **22** |
| *2.83* | *1.93* | | 27 | | **27** |
| *3.00* | *2.05* | | 33 | 2 | **35** |
| *3.17* | *2.17* | | 44 | 12 | **56** |
| *3.33* | *2.29* | | 7 | 26 | **70** |
| *3.50* | *2.42* | | 2 | 79 | **86** |
| *3.67* | *2.55* | | | 87 | **89** |
| *3.83* | *2.69* | | | 81 | **81** |
| *4.00* | *2.84* | | | 108 | **108** |
| *4,17* | *3.02* | | | 86 | **86** |
| *4.33* | *3.24* | | | 78 | **78** |
| *4.50* | *3.53* | | | 56 | **56** |
| *4.67* | *3.90* | | | 22 | **22** |
| *4.83* | *4.38* | | | 9 | **9** |
| *5.00* | *5* | | | 7 | **7** |
| ***Total*** | | **10** | **216** | **654** | **880** |

### 4.2.1 Comparison between different global scores

Additional analysis (not published) was done in order to determine the cut-off value that will indicate high and low stress levels on a metric score. The classification into high, medium and low stress levels was done using the criterion-based approach (CBA) defined in Paper I. According to the CBA, 28% (n = 248) of the participants reported low stress at work and 18% (n=162) reported high stress. The mean score on the metric scale was 2.9 (SD 0.68). A previously recommended neutral point on a stress scale, described as neither stressed nor calm, is a mean value of 2.4, which corresponds to the interval 2.97 to 3.07 on the Rasch metric scale. Approximately 49% of participant had scores above the neutral point.

Comparison between CBA scores and Rasch metric scores is shown in Table 10. The Rasch metric scale range corresponding to low stress was 0 to 2.66. The corresponding values for medium and high stress were 2.28 to 3.70 and 3.34 to 5 respectively. As indicated by the grey zones in the table, some overlap between the cut-off values was observed. Metric values in the range 2.28 to 2.66 corresponded to both low and medium stress levels according to the CBA. Overlap between medium and high values was found in the range 3.34 to 3.70. The measure of disorder was negligible (D = 0.002), which means that most of the pairs have the same ordering according to metric and CBA scores. The corresponding cut-off values for the energy scale are shown in Table 9. The measure of disorder was D = 0.004.

*Table 10. Frequency distribution of the Rasch metric scores and criterion-based scores (CBA), on the stress scale of the Stress-Energy Questionnaire. Areas with an overlap between Rasch metric scores and CBA scores are marked in grey.*

| Rasch metric score | CBA Low | CBA Medium | CBA High | Total |
|---|---|---|---|---|
| *0* | 3 | | | **3** |
| *0.61* | 2 | | | **2** |
| *1.05* | 3 | | | **3** |
| *1.38* | 12 | | | **12** |
| *1.62* | 19 | | | **19** |
| *1.82* | 20 | | | **20** |
| *1.99* | 30 | | | **30** |
| *2.14* | 37 | | | **37** |
| *2.28* | 48 | 2 | | **50** |
| *2.41* | 60 | 2 | | **62** |
| *2.54* | 10 | 34 | | **44** |
| *2.66* | 4 | 61 | | **65** |
| *2.77* | | 50 | | **50** |
| *2.87* | | 49 | | **49** |
| *2.97* | | 36 | | **36** |
| *3.07* | | 49 | | **49** |
| *3.16* | | 36 | | **36** |
| *3.25* | | 46 | | **46** |
| *3.34* | | 51 | | **54** |
| *3.43* | | 29 | 4 | **33** |
| *3.52* | | 19 | 18 | **37** |
| *3.61* | | 4 | 34 | **38** |
| *3.70* | | 2 | 25 | **27** |
| *3.80* | | | 18 | **18** |
| *3.90* | | | 19 | **19** |
| *4.01* | | | 11 | **11** |
| *4.13* | | | 9 | **9** |
| *4.26* | | | 10 | **10** |
| *4.43* | | | 5 | **5** |
| *4.66* | | | 3 | **3** |
| *5* | | | 3 | **3** |
| **Total** | **248** | **470** | **162** | **880** |

## 4.3 Paper III

The main finding of this study was that, having accommodated for local dependency stress and energy scales of the SEQ-LT showed good psychometric properties. The initial analysis revealed certain problems for both scales, as the lower limit of the CI was not below 5%. The residual correlation matrix gave an indication of the response dependency between the stress items: *rested* and *relaxed*, *stressed* and *pressured*, *relaxed* and *calm*, and the energy items: *active* and *energetic*. The PCA performed on residuals indicated that positively and negatively oriented items within each scales clustered into two different groups. An attempt to resolve these issues was made by combining these items into two testlets within each scale, which resulted in a satisfactory fit to the model, as shown in Table 11.

Looking at the individual item fit, all items had ordered thresholds. Individual item fit for both scales is shown in Table 12. For the stress items the locations ranged between -0.95 (*rested*) and 0.95 (*tense*). Corresponding figures for the energy scale were -0.80 (*passive*) to 1.05 (*energetic*). A fit residual value slightly outside the optimal range was observed for the stress item *rested* (-2.51) and the energy item *focused* (3.04). All other items were within the range ±2.5.

*Table 11. Fit to the Rasch model.*

| | **Item residual** | | **Person residual** | | **Chi square** | | **Unidimensionality** | |
|---|---|---|---|---|---|---|---|---|
| **Analysis name** | **Mean** | **SD** | **Mean** | **SD** | **Value** | **p** | **PSI** | **Test %** |
| Stress 2 testlets | 0.20 | 0.16 | -0.57 | 0.94 | 13.95 | 0.73 | 0.88 | 3.6 (2.2;5.0) |
| Energy 2 testlets | 0.27 | 0.88 | -0.59 | 1.00 | 14.06 | 0.72 | 0.79 | 3.3 (1.9;4.6) |
| ***Ideal values*** | ***0.0*** | ***<1.4*** | ***0.0*** | ***<1.4*** | | ***>0.05*** | ***>0.7*** | ***(LCI <5%)*** |

*Table 12. Individual item fit.*

| | Location | Fit residual | Chi square | p-value |
|---|---|---|---|---|
| **Stress items** | | | | |
| Rested | -0,95 | **-2,51** | 11,37 | 0,25 |
| Relaxed | -0,67 | 0,78 | 4,76 | 0,85 |
| Calm | -0,52 | 0,05 | 2,52 | 0,98 |
| Stressed | 0,52 | -0,51 | 6,22 | 0,72 |
| Pressured | 0,67 | 1,35 | 6,83 | 0,65 |
| Tense | 0,95 | 0,20 | 7,10 | 0,63 |
| **Energy items** | | | | |
| Passive | -0,80 | -2,13 | 14,90 | 0,094 |
| Inefficient | -0,66 | 0,35 | 10,09 | 0,34 |
| Dull | -0,42 | 0,51 | 8,89 | 0,45 |
| Active | 0,08 | -0,38 | 8,39 | 0,50 |
| Focused | 0,76 | **3,04** | 11,90 | 0,22 |
| Energetic | 1,05 | -1,71 | 16,37 | 0,06 |

No DIF for gender was seen. Uniform DIF for age was observed for the energy items *active* ($F = 15.92$, $df = 1,951$, $p < 0.0001$) and *dull* ($F = 19.73$, $df = 1,951$, $p < 0.0001$). There was a difference in the direction of the DIF. Given the same energy level, participants >48 years rated lower for the item *active* compared to those ≤48 years, while the reverse was seen for the item *dull*, as seen in Figure 3 and Figure 4 respectively. Additional analysis was done by creating a testlet of these two items and keeping all other items the same. This resulted in a DIF cancellation, no further actions was taken and all the items were kept within the scale.

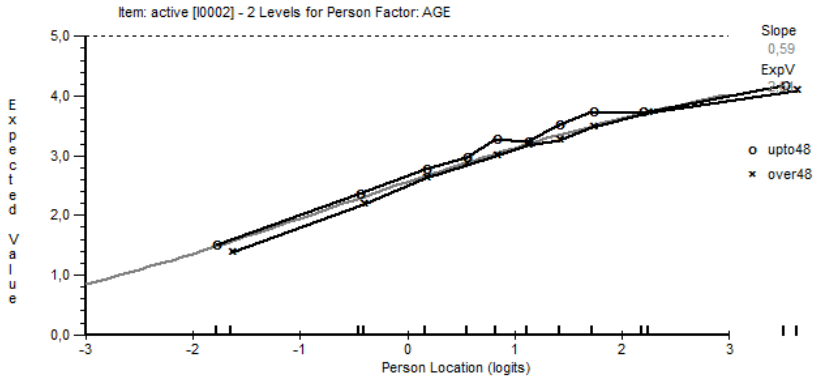


*Figure 3. Item characteristic curves item active.*

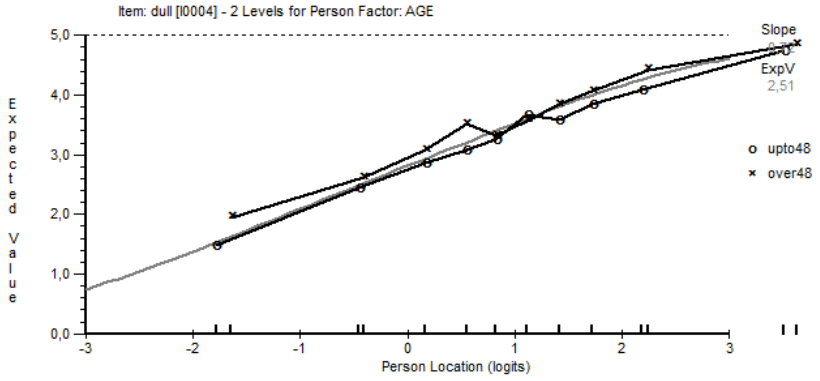


*Figure 4. Item characteristic curves item dull.*

The targeting of the persons and items was reasonable for both the stress and energy scales. Given fit to the model, Rasch scores were transformed into a metric score with a range of 0 to 5, indicating the lowest and highest scale values respectively. Comparison between metric and mean scores on the stress and energy scales is shown in Figure 5. As seen in Figure 5, increase of one unit in the mean score is not constant across the scales.

The proposed cut-off point the low stress is in the interval 2.45 to 3.02 for the Rasch metric score, and for high stress the cut-off point was in the interval 3.65 to 3.90. This means that in the interval 2.45 to 3.02 on the metric scale, individuals could be classified as either low or medium stressed according to the CBA (disordered pairs). Correspondingly, classification of individual scores in the interval 3.65 to 3.90 could be either medium or high stress. However, the measure of disorder was negligible (D = 0.002).

The suggested cut-off points for low and high energy levels are values 1.73; and 1.97, and 2.66 and 3.08 respectively. The measure of disorder was D = 0.005. According to the CBA, 8% (n = 75) of the participants rated high stress and 43% (n=413) rated low stress. As regards energy, 5% (n = 45) were found in the low energy group and 43% (n=413) in the high energy group (CBA). The mean value on the metric scale was 2.8 (SD=0.81) and 2.8 (SD=0.59) for the stress and energy scale respectively.

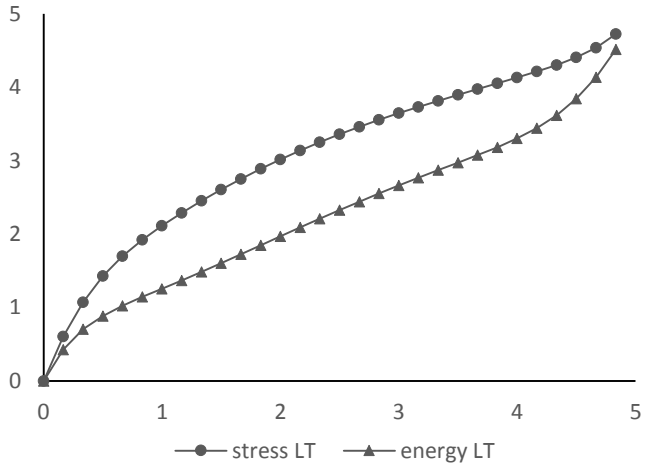


*Figure 5. Comparison of metric and mean scores on the stress and energy scales.*

## 4.4 Paper IV

A simplified theoretical model describing associations between workplace factors and symptoms of burnout is shown in Figure 6. As seen in the figure, a direct association between unfavourable working conditions and development burnout can be assumed. Moreover, it can also be assumed that workplace factors have an impact on the affective mood responses at work and outside work, which in turn affects the outcome. There are of course factors that can confound the relationship between workplace factors and the symptoms of burnout, such as lifestyle factors, e.g. physical activity, social support at work and outside the work, level of educational and many other factors.

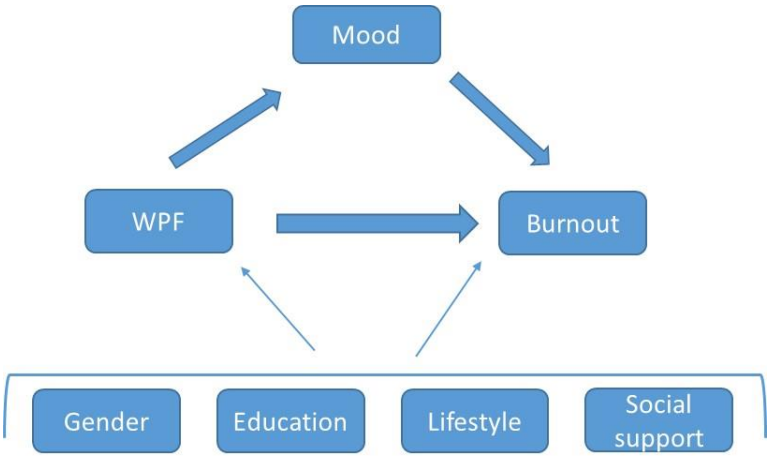


*Figure 6. Simple theoretical model of associations between workplace factors (WPF), affective mood and symptoms of burnout.*

The workplace factors investigated were: demands, decision authority, effort and reward. In addition, symptoms of burnout measured at a previous time i.e. *t-1* (SMBQ-lag) was included in a model (autoregressive model), recognising that this could influence both workplace factors and SMBQ at time *t*. In an additional step, regression models also included affective stress responses at work and during leisure time. The regression models were adjusted for potential confounders (physical activity, length of education, gender and social support outside work) when appropriate. Interactions, i.e. joint effects of job demands and decision authority as well as effort and reward on burnout, were investigated by creating combined variables of the two factors.

The main finding was that demand and decision authority, as well as effort and reward, were independently associated with symptoms of burnout in the final

model. As seen in Table 13, high demands increased the burnout score by 1.81 points compared to low demands. The corresponding figure for medium demands was 1.14, indicating a dose-response relationship. The same patterns were observed for all other workplace factors. No joint effects of demands and decision authority, or effort and reward on burnout symptoms were observed. The association between all the factors, except job demands, remained even after adjusting for stress and energy levels at work and outside work.

Evidence of long-term effects of workplace factors on burnout symptoms was limited only to reward (see Table 14). The effect of reward on the burnout score was also present after additional adjustment was made for stress and energy at work and during leisure time (Table 14, model 2). Joint effects were not investigated since reward was the only workplace factor that showed a delayed effect on burnout.

*Table 13. Longitudinal analysis showing the short-term effect between workplace factors and burnout symptoms measured using SMBQ[1], adjusted for confounders (Model 1), plus affective stress response at work and during leisure time (Model 2); regression coefficients (Coeff), 95% confidence intervals (CI) and p-values.*

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| | **Coeff (95% CI)** | **p-value** | **Coeff (95% CI)** | **p-value** |
| **Intercept** | 26.6 | <0.0001 | 51.7 (50.00;54.43) | <0.0001 |
| **Time** | | | | |
| T1 | 1.8 (0.84;2.60) | <0.0001 | 0.9 (0.19;1.54) | 0.012 |
| T2 | 0.3 (-0.30;0.80) | 0.374 | 0.1 (-0.49;0.52) | 0.951 |
| T3 | 0 | | 0 | |
| **Demand**[2] | | | | |
| High | 1.8 (0.77;2.85) | 0.001 | | |
| Medium | 1.1 (0.39;1.89) | 0.003 | | |
| Low | 0 | | | |
| **Decision**[2] | | | | |
| Low | 1.9 (1.30;2.54) | <0.0001 | 1.3 (0.69;1.87) | <0.0001 |
| Medium | 1.1 (0.43;1.78) | 0.001 | 0.8 (0.19;1.44) | 0.010 |
| High | 0 | | 0 | |
| **Effort**[3] | | | | |
| High | 3.4 (2.52;4.29) | <0.0001 | 2.2 (1.44;3.02) | <0.0001 |
| Medium | 1.0 (0.28;1.66) | 0.006 | 0.4 (-0.19;1.04) | 0.174 |
| Low | 0 | | 0 | |
| **Reward**[3] | | | | |
| Low | 4.9 (3.53;6.22) | <0.0001 | 2.7 (1.43;3.97) | <0.0001 |
| Medium | 2.9 (2.02;3.70) | <0.0001 | 1.3 (0.50;2.03) | 0.001 |
| High | 0 | | 0 | |
| **Physical act.**[4] | | | | |
| Sedentary | 5.6 (4.75;6.46) | <0.0001 | 2.9 (2.10;3.74) | <0.0001 |
| Light | 1.7 (1.17;2.31) | <0.0001 | 1.5 (0.96;2.03) | <0.0001 |
| Moderate/intense | 0 | | 0 | |
| **Social support**[5] | | | | |
| No | 3.5 (2.42;4.64) | <0.0001 | 1.3 (0.28;2.32) | 0.012 |
| Yes | 0 | | 0 | |
| **SMBQ-lag** | 0.5 (0.50;0.54) | <0.0001 | 0.2 (0.22;0.26) | <0.0001 |
| **Stress-W**[8] | | | 3.7 (3.19;4.18) | <0.0001 |
| **Energy-W**[8] | | | -3.4 (-3.85;-2.95) | <0.0001 |
| **Stress-LT**[7] | | | 2.8 (2.39;3.13) | <0.0001 |
| **Energy-LT**[7] | | | -4.0 (-4.52;-3.40) | <0.0001 |

[1]Shirom-Melamed Burnout Questionnaire, [2]Demand-control Questionnaire, decision authority=subscale of the control dimension, [3]Effort-Reward Questionnaire, [4]Level of physical activity, [5]Social support outside the workplace, [6]Stress-Energy Questionnaire at work, [7]Stress-Energy Questionnaire during leisure time

*Table 14. Longitudinal analysis showing delayed effects between workplace factors and burnout symptoms measured using SMBQ[1], adjusted for confounders (Model 1), plus affective stress response at work and during leisure time (Model 2); regression coefficients (Coeff), 95% confidence intervals (CI) and p-values.*

| | **Model 1** | | **Model 2** | |
|---|---|---|---|---|
| | **Coeff (95% CI)** | **p-value** | **Coeff (95% CI)** | **p-value** |
| **Intercept** | 21.6 (20.08;23.02) | <0.0001 | 26.8 (23.70;30.05) | <0.0001 |
| **Time** | | | | |
| T2 | -0.7 (-1.30;-0.08) | 0.027 | -0.7 (-1.39;-0.03) | 0.039 |
| T3 | -0.9 (-1.55;-0.25) | 0.007 | -1.0 (-1.72;-0.26) | 0.008 |
| T4 | 0 | | 0 | |
| **Reward-lag[2]** | | | | |
| Low | 1.1 (-0.06;2.26) | 0.063 | 1.2 (-0.15;2.48) | 0.081 |
| Medium | 1.5 (0.71;2.18) | <0.0001 | 1.2 (0.43;2.06) | 0.003 |
| High | 0 | | 0 | |
| **Physical act.[3]** | | | | |
| Sedentary | 4.1 (3.36;4.92) | <0.0001 | 3.6 (2.73;4.49) | <0.0001 |
| Light | 1.3 (0.81;1.85) | <0.0001 | 1.4 (0.81;1.97) | <0.0001 |
| Moderate/intense | 0 | | 0 | |
| **Social support[4]** | | | | |
| No | 2.7 (1.66;3.72) | <0.0001 | 2.9 (1.73;4.01) | <0.0001 |
| Yes | 0 | | 0 | |
| **SMBQ-lag** | 0.7 (0.65;0.69) | <0.0001 | 0.6 (0.57;0.64) | <0.0001 |
| **Stress-W-lag[5]** | | | 0.4 (-0.09;0.97) | 0.107 |
| **Energy-W-lag[5]** | | | -0.8 (-1.33;-0.33) | 0.001 |
| **Stress-LT-lag[6]** | | | 0.8 (0.40;1.22) | <0.0001 |
| **Energy-LT-lag[6]** | | | -0.7 (-1.31;-0.07) | 0.029 |

[1]Shirom-Melamed Burnout Questionnaire, [2]Reward dimension of the Effort-Reward Questionnaire, [3]Level of physical activity, [4]Social support outside the workplace, [5]Stress-Energy Questionnaire at work, [6]Stress-Energy Questionnaire during leisure time

### 4.4.1 Criterion based approach for DCQ and ERI

In this paper a new way of defining the risk groups of demands, decision authority effort and reward was proposed that is consistent across different populations and over time.

The CBA scores of demand and decision authority were calculated as follows. The response alternatives for each item were first classified as low, medium or high responses for demands and decision authority. As regards the demand items, *never* and *seldom* were considered to be low responses, and *often* a high response. The direction of the responses for one demand item was reversed (*Do you have sufficient time for all your work tasks?*). For that item and for the two decision authority items, *often* was considered a low response and *never* or *seldom* high responses. The response alternative *sometimes* was considered to be a medium response for all the items. Finally, a CBA score (low, medium, high) was calculated for each dimension based on the frequency distribution of the items into predefined categories. The CBA scores for decision authority are shown in Table 15, and for demands in Table 16.

*Table 15. The low, medium and high levels of decision authority according to the criterion approach (CBA) based on the number of responses found in the response categories often, sometimes, never and seldom, out of two decision authority items in the Demand-Control Questionnaire.*

| Response profiles | Never or seldom | Sometimes | Often | CBA score |
|---|---|---|---|---|
| [1] | 2 | 0 | 0 | Low |
| [2] | 1 | 1 | 0 | Low |
| [3] | 1 | 0 | 1 | Medium |
| [4] | 0 | 2 | 0 | Medium |
| [5] | 0 | 1 | 1 | High |
| [6] | 0 | 0 | 2 | High |

*Table 16. The low, medium and high levels of demands according to the criterion approach (CBA) based on the number of responses found in the response categories often, sometimes, never and seldom, out of five demand items in the Demand-Control Questionnaire.*

| **Response profiles** | **Often** | **Sometimes** | **Never or seldom** | **CBA score** |
|---|---|---|---|---|
| [1] | 5 | 0 | 0 | High |
| [2] | 4 | 1 | 0 | High |
| [3] | 4 | 0 | 1 | High |
| [4] | 3 | 2 | 0 | High |
| [5] | 3 | 1 | 1 | High |
| [6] | 3 | 0 | 2 | High |
| [7] | 2 | 3 | 0 | Medium |
| [8] | 2 | 2 | 1 | Medium |
| [9] | 2 | 1 | 2 | Medium |
| [10] | 2 | 0 | 3 | Medium |
| [11] | 1 | 4 | 0 | Medium |
| [12] | 1 | 3 | 1 | Medium |
| [13] | 1 | 2 | 2 | Medium |
| [14] | 1 | 1 | 3 | Medium |
| [15] | 1 | 0 | 4 | Medium |
| [16] | 0 | 5 | 0 | Medium |
| [17] | 0 | 4 | 1 | Medium |
| [18] | 0 | 3 | 2 | Medium |
| [19] | 0 | 2 | 3 | Low |
| [20] | 0 | 1 | 4 | Low |
| [21] | 0 | 0 | 5 | Low |

The CBA scores for the ERI were defined as follows. The response alternatives *distressed* and *very distressed* were considered to be high effort and low reward responses, and s*omewhat distressed* was considered to be a medium response. If answering *yes* in the first step led to the evaluation of distress, then the response alternatives *no* and *yes, but not at all distressed* were considered to be low effort and high reward responses. If *no* required evaluation then *yes* and *no, but not at all* distressed were considered to be high reward responses.

CBA scores for effort are shown in Table 17. The three sub-dimensions of reward were combined into a total reward score. At least two high levels out of the three sub-dimensions were considered to be a high level of reward and vice versa for low levels. The high levels of esteem were achieved by having three or more high responses for esteem items and no more than one low response. Three or more low responses corresponded to low esteem. At least two high responses and a maximum of one low response resulted in high promotion levels, and at least two low responses resulted in low promotion levels. At least one high response and no low response defined high security levels. The reversed was true for low security levels.

*Table 17. The low, medium and high levels of effort according to the criterion approach (CBA) based on the number of responses found in the response categories low, medium and high out of five effort items in the Effort-Reward Questionnaire. Low effort responses = No or Yes, but not at all distressed, medium effort response =somewhat distressed high effort responses = distressed or very distressed.*

| **Response profiles** | **Low effort responses** | **Medium response** | **High effort responses** | **CBA score** |
|---|---|---|---|---|
| [1] | 5 | 0 | 0 | Low |
| [2] | 4 | 1 | 0 | Low |
| [3] | 4 | 0 | 1 | Low |
| [4] | 3 | 2 | 0 | Low |
| [5] | 3 | 1 | 1 | Low |
| [6] | 3 | 0 | 2 | Medium |
| [7] | 2 | 3 | 0 | Medium |
| [8] | 2 | 2 | 1 | Medium |
| [9] | 2 | 1 | 2 | Medium |
| [10] | 2 | 0 | 3 | Medium |
| [11] | 1 | 4 | 0 | Medium |
| [12] | 1 | 3 | 1 | Medium |
| [13] | 1 | 2 | 2 | Medium |
| [14] | 1 | 1 | 3 | High |
| [15] | 1 | 0 | 4 | High |
| [16] | 0 | 5 | 0 | Medium |
| [17] | 0 | 4 | 1 | Medium |
| [18] | 0 | 3 | 2 | Medium |
| [19] | 0 | 2 | 3 | High |
| [20] | 0 | 1 | 4 | High |
| [21] | 0 | 0 | 5 | High |

# 5 DISCUSSION

## 5.1 Main findings

One of the main findings of this thesis is that the measurement properties of the SEQ for assessing affective stress responses during the past week in a working situation were found to be satisfactory. Another finding was that the SEQ that was developed to assess stress and energy levels outside work, i.e. during leisure time (SEQ-LT), also showed good measurement properties. Both scales can be used in the present form to compare different groups of workers, e.g. in research studies or in occupational health services investigations. The stress scales from both SEQ and SEQ-LT, can also be used for evaluations at the individual level, which can be useful in clinical situations.

An important finding was that that the mean scores for the stress and energy scales in SEQ and SEQ-LT were not linear. An increase of one unit in the middle of the scale did not correspond to the same increase in stress or energy level compared to an increase of one unit at the ends of the scales. The use of the transformed Rasch scores is proposed instead of using mean scores. Two other approaches for the construction of global scores were also presented: the median and the criterion-based approach (CBA). One of the proposed uses of SEQ is to identify the risk groups for adverse health effects, i.e. individuals with high stress and low energy levels. The CBA developed in this thesis offers a simple solution for this purpose. In a broader perspective, this approach can also be applied to other questionnaires. In this thesis, for example, the CBA was applied to define the risk groups on the job demands, decision authority, effort and reward scales.

The fact that both the SEQ and the SEQ-LT worked well for group comparisons, made it possible to include the questionnaires in a study investigating longitudinal associations between workplace factors and symptoms of burnout. The question of whether associations were present even in the absence of affective stress response could be answered. The main finding was that job demands, decision authority, effort and reward, were associated with short-term effects on symptoms of burnout. The associations remained even after adjustment of affective stress responses for all workplace factors with the exception of job demands. Evidence of delayed effects was limited to rewards.

## 5.2 Validity aspects

Logically, validation of questionnaire data begins with an explicit statement of the proposed interpretation of questionnaire scores and with a rationale for the

relevance of the interpretation to the proposed use. It is important to remember that it is the interpretation of the questionnaire scores that are evaluated, not the questionnaire itself [75]. When scores are interpreted in more than one way, each intended use and interpretation need to be validated [66]. For instance, a questionnaire measuring stress exposure may be used to assess a current working situation at a certain workplace, which would help to make decisions about possible interventions at the workplace level. In that case the questionnaire is used for assessment at a group level. Another intended use may be in clinical settings, where the interpretation of the questionnaire score would be of help in making a clinical decision (individual level). In this thesis, measurement properties of the SEQ at work were evaluated for use at both the group and the individual level.

All modifications of existing questionnaires, and uses for purposes other than those specified by the developers of the scale, need to be the subject of a new validation process [66]. Consequently, the validity of SEQ for assessment of affective mood during leisure time (SEQ-LT) also needed to be evaluated. In applied research however, reporting of validation process is often inadequate and the meaning of validity is usually unclear, even in studies where the questionnaires are reported as validated [74]. The fact that the validation is an ongoing process needs to be emphasised more often.

A solid theoretical framework is a prerequisite for the Rasch analysis and it enables interpretation and understanding of the results. In this thesis, problems of local dependency were found for both SEQ and SEQ-LT, although, this observed response behaviour was not a surprise. As mentioned before, the SEQ is based on Russell's model of affect [24], whereby stress and energy are seen as bipolar dimensions. If persons respond to the items in a manner that is logically consistent, then some local dependency between the items can be expected. For instance, reporting *very much active*, *energetic* and *focused* and at the same time *very much passive*, *inactive* and *dull* would not be a logically consistent response pattern. Hence, given the theoretical background of the SEQ and the content of the positively and negatively loaded items, observed local dependency could be explained and handled by forming testlets of these items.

Logical response behaviour is a basic assumption behind all kind of questionnaires. In the initial phase of questionnaire design, many scale developers choose to include both negatively and positively oriented items [50, 69]. The goal of including both types of items is to minimise possible response bias, such as tendency to choose extreme responses or a tendency to choose positive responses, also known as acquiescence. The response process is an important aspect of the validation process whereby the response behaviour can be evaluated. Statistical methods within classical test theory (CTT), such as

factor analysis, do not address logical response behaviour explicitly [114]. An advantage of the Rasch analysis is that logical consistency of the responses is tested empirically in various ways and is implied by the model fit. It should be noted however, that the Rasch model is stochastic and the Rasch scores are not deterministic. A stochastic model implies that even if certain response combinations are logically inconsistent, they can still be present in the sample. Nevertheless, the probability of irrational response patterns should be reasonably low, provided the data fit the Rasch model expectations.

Although not used for the results in this thesis, another possibility offered by the Rasch analysis is that persons with irregular or logically inconsistent response patterns can be identified. These cases can then be examined further, to see whether other explanations can be found for response inconsistency that cannot be attributed to the scale items and response categories not working properly. The same possibility is also afforded by the CBA. Using this approach, the experts are given the opportunity, based on theoretical knowledge and logical arguments, to define what they considered to be an irrational response combination. A huge responsibility is thus assigned to the researchers that requires a great deal of previous knowledge regarding the construct of interest, and which can prove time-consuming. On the other hand, an advantage is that the possible response inconsistencies are not data-driven, but are based on theoretical grounds. Criteria for logically inconsistent response profiles can therefore be applied for comparable populations and for comparison over time.

## 5.3 Global scores

Defining a construct theoretically and realising it through a measuring instrument, i.e. a questionnaire, is an initial step in the measurement process. Making use of the measurement is the next step. After looking at face validity the Rasch analysis is a suitable next step in a validation process, as it checks the functioning of items and in the end the metric properties of the scale. The global scores can then be constructed. The global scores are assumed to measure the location of persons on the latent construct. Although there are other approaches, four different ways of constructing global scores were considered in this thesis: mean scores, median scores, criterion-based approach (CBA) scores and Rasch scores.

Although mean scores are by far the most used for construction of global scores [50, 113], a discussion about how to treat ordinal data from questionnaires, and whether or not the mean scores can be used, has been an ongoing debate for a long time [115-117]. However, uncritical and unreflective use of mean scores for ordinal data is associated with a certain degree of risk [118]. Firstly, the

linearity of the mean scores should be investigated. Mean scales tend to be more sensitive at the centre than at the extremes (ends) of the scale [80]. That is, although mean scores for some scales seem linear in the middle of the scale, as was the case with the stress and energy scales of SEQ and SEQ-LT, an increase of one unit towards the ends of the scale may not imply the same increase or decrease in the latent variable. Non-linearity of the scale compromises the properties of responsiveness and sensitivity. This may have an effect on the results and interpretation of the study. Mean scores also require parallel items, a condition which may not always hold.

The median approach is more appropriate for ordinal data, since this approach does not require equidistant scale categories. However, the median also implies parallel items. Another requirement, which also applies to the mean score, is having the same number of response categories for each item. A drawback of the median is that the number of possible values it can take is limited to the number of response categories in the items. So an advantage of the mean score over the median score would be increased precision, with many more scale values. On the other hand, the advantage of the median approach is that the interpretation of the median scores is straightforward. Furthermore, as was shown theoretically in Paper I, the same mean score can be a result of different response patterns implying different levels of the measured latent construct, thus calling into question the sufficiency of the mean scores. To some extent the same problem is shared by the median approach. Although no assumption about the data properties is required, there can still be sufficiency problems, since many different response profiles can result in the same median.

Stress and energy measured using the SEQ consist of six items each. However, it is not unusual for questionnaires to consist of 15-20 items or more with a varying number of response categories. One example is SMBQ, originally consisting of 22 items each consisting of seven response categories [41]. In view of the fact that the number of distinct response profiles increases with the increasing numbers of response categories and items, the problem of heterogeneous response profiles resulting in the same mean or median value becomes even more noticeable. Counting all the possible response profiles that may occur and result in the same mean or median can be viewed as a deterministic way. Although theoretically possible, not all response profiles could be a result of a logically consistent response behaviour, as discussed in a previous section.

The CBA, suggested in this thesis, is a method that has good flexibility and interpretability properties. In our application of CBA to the SEQ, parallel items were assumed for reasons of simplicity, in order to demonstrate the method. However, the condition of parallel items is not a requirement for this approach. The criteria can be formulated in such a way that responses to some items are

given more weight. The CBA scores can be regarded as a sufficient statistic. The same information about the latent construct is obtained regardless of whether the individual items or the global scores are recorded in the data material. It should be noted however, that even according to CBA, different response profiles also result in the same global level. Applied to the SEQ, and later also to demands, decision authority, effort and reward scales, many different response profiles were combined in the same category on the CBA scales, i.e. low, medium and high. However, by using this approach, experts are given the opportunity to qualitatively evaluate the meaning of each response profile for the intended use. Although some response profiles may imply more or less stress or energy or some other measured construct, these response profiles are adjudged by the experts to belong to the same risk group.

The last type of global scores discussed in this thesis is the Rasch score, obtained from the Rasch analysis. Rasch score is a sufficient statistic provided that data fit the expectation of the Rasch model. The Rasch scores are originally on a logit scale, but can be standardised into metric scores that reflect the values of the original scale. To make interpretation of the scores easier, transformation to a convenient range can be offered [110], as was done for SEQ and SEQ-LT. These values can be applied to similar population. As was shown in this thesis, the CBA scores can be used to help interpret the metric scale values. This can be especially useful when applied to a new scale, or when the scale is used for different populations or in a new context.

To summarise, the bottom line is that increased awareness is needed of the different ways of constructing global scores. Each approach has its advantages and limitations. In addition to evaluation of the validity of the scale for intended use, a decision is needed about how the global score will be constructed. Which approach is best suited to the intended use of the scale, is something that requires careful consideration.

## 5.4 Applications in stress research

Previously, the SEQ has been used in many Scandinavian epidemiological studies, most frequently as a risk factor in multivariate models that analyse associations with various health outcomes [29-35]. One of the goals of these and similar studies could be to identify groups with high stress and low energy levels, as these were hypothesised as potential risks for adverse health effects. Hence, values of the stress and energy scales that would define the low and the high levels on a global scale need to be defined. In epidemiological studies, this can be done in variety of ways. When using means as a global score, one possibility is to calculate the group mean based on the individual mean scores and use this as a cut-off point for low and high levels. Using this method, an

individual with a higher mean score than the mean score of the group is classified as being highly stressed. Another way is to calculate quartiles based on the distributions of the individual mean scores, and identify respondents above the median value or within the range of the highest quartile as a high stress group. This procedure is of course not unique for the SEQ, but is applied for many other questionnaires as well, such as JDC [119]. However, these cut-off points are data-derived and are not comparable across the different groups or occasions. Data-derived scores based on a sample from one study population are not necessarily applicable in the context of another population or even in a sample from the same population drawn for other study purposes.

An advantage of the SEQ is that so called neutral points are proposed, which will identify the values that indicate neither stressed nor calm for the stress scale and neither passive nor active for the energy scale [27, 30]. Neutral points are not sample dependent and can be used for comparison over time and between populations, provided the mean scale is linear and that neutral point is stable across different populations. For example whether the same neutral point would apply for the SEQ-LT is a matter for discussion and is not considered in this thesis. Rasch analysis showed that the mean scales of stress and energy for SEQ and SEQ-LT were not linear. In this thesis high and low stress and energy levels were defined using the CBA and compared with the Rasch metric scores. In contemporary validity thinking, Rasch analysis is proposed as a golden standard for scale evaluation, and is proposed for use whenever there is an intention to create a global score from item responses [105, 114]. Performing the Rasch analysis on SEQ and SEQ-LT and being able to relate the cut-off values defined by CBA (and neutral points for SEQ) to the metric scale, is certainly a strength.

There is of course always certain risk of losing important information by grouping the continuous variable into a small number of categories [120]. Nevertheless, depending on the context and intended use of the scores, combining the response profiles into smaller number of categories, using neutral points or the CBA, may be appropriate. One advantage of the CBA scores is that low, medium and high scores can be easily described in words, thus providing the opportunity for other researchers to assess the soundness of the applied criteria, or to modify criteria for the purposes of future studies.

In this thesis, the CBA was also defined for the demands, decision authority, effort and reward scales. Considering the risk of non-linearity and not having metric properties, mean scores for these scales were not considered in this thesis. The CBA classification also better fitted the purpose of the study, which was to contrast known risk groups with regard to their effect on the outcome. This was a trade-off and the ideal scenario would of course be to first perform

the Rasch analysis in order to verify the measurement properties for these scales, but this was beyond the scope of this thesis.

Nonetheless, including demands and decision authority, as well as effort and reward in the same model to investigate associations with symptoms of burnout, as part of longitudinal design with three waves, has to our knowledge not been done previously. Although to some extent overlapping, the models behind the JDC and ERI are considered to be complementary and to have independent effects on mental distress [121]. In this thesis, independent effects were found in the risk of symptoms of burnout, which may be seen as an interesting contribution in the field of stress research.

## 5.5 Longitudinal analysis

Time is an important aspect in longitudinal studies, and the appropriate length of time lags in longitudinal studies is a crucial issue in research methodology [103]. Studies are needed that aim to investigate how stress effects unfold in time, and which will assess in more detail the duration and intensity of exposure required for developing ill-health [93]. In stress research, however, taking into account the relevance of time lags has to date been a more or less an overlooked research issue [25, 94, 122].

When studying longitudinal effects, some stressors have an immediate effects on strain, and other stressors take longer time to react. Observed time lags between waves in a longitudinal study need to be in accord with the causal time lag. If the observed time lag is too short, occupational stressors may not have sufficient time to affect the outcome variable. Conversely, if the observed time lag is too long, the effects of being exposed to certain unfavourable conditions may already have disappeared.

In the longitudinal studies of working conditions and symptoms of burnout, the length of follow-up varied between eight months and three years [39]. To our knowledge there is no clear evidence of a suitable follow-up length with regard to the development of burnout. In this thesis, two time aspects were considered in relations to the associations between the psychosocial stressors and symptoms of burnout. For an investigation of delayed effects, the time lag between measurements was two years, which could be regarded as too long and more studies are needed to shed a light on this issue.

Sometimes, a time lag has already been taken into account in the way a certain predictor variable is measured [103]. Both the JDC and the ERI were developed to measure conditions in work environments that are long-lasting. This means that it could also be argued that the time-lag has already been taken into account in these measures, and the time lag models, such as delayed analysis in this thesis, may imply longer time aspects than two years. Similarly,

the models used in the thesis for short-term effect, may indicate a longer time period than is implied by the name given to this analysis.

## 5.6 Limitations

As regards the energy scale of the SEQ, the lowest energy levels were not observed in our sample. This means that targeting between the items and persons was not optimal. It might be the case that the energy levels in this working population of health care workers are indeed very high, otherwise it would be difficult to perform optimally at work. In order to improve targeting between the levels of energy expressed by the items and the actual levels of energy among the persons being examined, one possibility is to adjust the energy scale by including items that would be better targeted at perceived levels in this type of high-performing working population.

Each measurement is associated with certain measurement errors, and that is the case with the values in the transformation tables of the Rasch metric scores. The measurement error is dependent on the sample size and on the variability of the construct in the sample. The measurement error is also dependent on targeting between persons and items. Taking the energy scale of the SEQ as an example, given that most of the persons in our sample reported very high energy levels, we did not have large variability to ensure that the estimated Rasch scores would be sufficiently stable for the lower part of the scale. This may also be a possible explanation for disordered thresholds found for item *dull*. There is also a possibility of response bias, referred to as social desirability. Do we really admit to being *very much passive* or *not at all active* at work? However, the energy scale did fit the expectations of the Rasch model after accounting for local dependency. The scale can therefore be used in its present form, although it would be desirable to test the scale on a more targeted sample. As regards methodological aspects, the stability of the Rasch metric scores is something that can be evaluated in more detail in future studies.

Moreover, the results regarding the validity of the SEQ and SEQ-LT for use for group and individual comparison is limited to similar populations and in Scandinavian studies. For use in other countries, further investigations are needed. The SEQ was developed for Swedish conditions, and due to cultural and regional differences, comprehensions of the item contents and the construct of affective stress response could perhaps be defined differently in other countries. As regards the presented CBA scores, it should be noted that the criteria were defined by an expert panel of stress researchers. There is a possibility that the criteria would be defined differently by clinicians for use in clinical settings, or by professionals working in the occupational health services. This stresses once again the importance of the fact that the validation

is an ongoing process and that it needs to be confirmed for each intended use of the scale.

Appropriateness of the length of time lags between waves in a longitudinal study is highlighted as important but is a somewhat neglected issue in stress research [93, 94, 122]. When investigating the effects of psychosocial working conditions on symptoms of burnout, it can be argued that a time lag of two years is too long [123]. Methodological ideals are of course important in order to maintain the quality of the research. On the other hand, there is always a trade-off between the scientific ideal and what is perhaps possible financially and or pragmatically. For example, asking participants to reply to surveys too often might result in low response rates, and thus give rise to other methodological considerations.

## 5.7 Practical implications

A methodological implication regarding the practice of validity evaluation is that in addition to the requirement to evaluate the measurement properties of questionnaires for each intended use, a decision needs to be taken about which approach will be used for the construction of global scores. This decision-making process needs to be based on the intended use and proposed interpretation of global scores, and not necessarily on traditions within a certain research field or current practices.

A practical implication for stress research, is that with good psychometric properties for use on the group level in a sample of health care workers, both the SEQ and SEQ-LT can be used for employees within similar populations. These can be used either as stand-alone questionnaires, or together in order to obtain an indication of the respondents' work-life balance and the degree of recovery from work-related stressors.

Moreover, the findings of this thesis suggest that it is important to prevent all sorts of unfavourable working conditions (high demands, low decision authority, high effort and low reward), even if those are not consciously appraised by the employees or perceived as stressful. The practical implication of these results is that it is important to systematically monitor, identify and minimise unfavourable working conditions through for example workplace surveys. In this way, the results can be used for an early discovery of employees who are at increased risk of developing burnout.

# 6 CONCLUSION

The SEQ and SEQ-LT are valid and useful tools for assessing work-related and non-work-related affective stress responses respectively. Each modification of the questionnaire, or use other than those proposed by the developers of the scale, requires new validations. Measurement properties of scales should be taken into account before the construction of the global scores, in order to provide a useful basis for inference procedure. Rasch analysis is proposed for evaluation of measurement properties. The validity and reliability of questionnaires are prerequisite for their usefulness.

There must be increased awareness of the fact that there are several different ways to construct a global score of a multi-item questionnaire. To simply use a mean score because it the tradition within a specific field does not hold in a contemporary validity thinking. Critical evaluation of measurement properties is required. The CBA can be used for identification of the groups at risk for adverse health effects, as defined by the theoretical foundations of the questionnaires, provided there are good measurement properties defined by the Rasch model.

Longitudinal associations were found between workplace factors (demand, decision authority, effort and reward) and the symptoms of burnout. This finding was confirmed even in the absence of the affective stress response. Results can be used in workplace investigations, to obtain an indication of early signs of future burnout cases, which is important to prevent the high burden of mental and somatic symptoms due to burnout.

# 7 FUTURE PERSPECTIVES

Although the energy scale of the SEQ did fit the Rasch model, certain problems were identified. As a further development of the scale, it would be of interest to test the energy scale on a well-targeted sample and on other populations to see whether the problems remain.

In this thesis, the risk groups for ERI and JDC (more precisely the demands and decision authority, a subscale of decision) where defined using CBA. In previous studies, concerns were raised regarding the comparability of the different versions of the JDC [119], computation of the sum scores and the ERI index using a version of the questionnaire with a two-step procedure [124], and the way the exposure or risk groups are defined in different studies [125]. Since the ERI and the JDC are the two most commonly used measures of psychosocial stress exposure, there is a need to test the measurement quality of the scale by means of Rasch analysis, which is in accordance with the modern conceptualisation of validity.

A further methodological aspects is exploring in more details the stability of the Rasch person scores. In this thesis, Rasch metric scores for SEQ and SEQ-LT were introduced. Recognising the fact that the Rasch model is stochastic, it would be interesting to see if and how targeting and sample size affect the stability of the person estimates. What is the acceptable measurement error and how can this be accounted for when presenting transformation tables?

# ACKNOWLEDGEMENTS

I would like to sincerely acknowledge the help and support from everyone who has been involved in my PhD and the writing of this thesis. I would like to thank the following in particular:

Anna Grimby-Ekman, my main supervisor, for sharing your fantastic expertise and knowledge of statistics and research in general. Thank you for many rewarding and inspiring discussions, for your incredible involvement in my work, and for always finding time to meet and write despite a busy schedule. For all your positive and encouraging comments, which have guided me through many moments of doubt and helped me to do my best.

Åsa Lundgren-Nilsson, my co-supervisor, for introducing me to the fields of measurement and Rasch analysis, for sharing your insights into the way research functions, and for all the occasions that you have taken the time to listen and help me to formulate my thoughts and ideas. I always left your office with a smile on my face and full of inspiration.

Gunnar Ahlborg Jr., my co-supervisor and former manager at the Institute of Stress Medicine, for generously sharing your knowledge and experience over the years, for encouraging me to start this journey, and for believing in my ideas and work ever since I started at ISM. Thank you for giving me the freedom to choose the ideas for both the licentiate and PhD theses. Without your support these projects would not have been possible.

Elisabeth Svensson, my former supervisor for the licentiate thesis in statistics, for introducing me to the analysis of ordinal data, for your incredible help during the licentiate project.

Ingibjörg Jonsdottir, for being a great manager and a great researcher, for generously sharing your knowledge, ideas and Islandic candies, and most of all for being such a warm and caring person.

Carita Håkansson, co-author, for your rewarding collaboration and for broadening my awareness of issues such as work-life balance and gender.

Jonas Björk and Anders Pousette for your valuable comments on my work and excellent discussions prior to my defence.

Cecilia Andreasson for administrative help.

All of my dear present and former colleagues at the Institute of Stress Medicine for making the ISM a wonderful workplace, and for all the laughs around the coffee table.

Agneta Lindegård Andersson, a wonderful colleague and mentor in science and life in general, for all the great discussions in the corridor about “upsetting” and not so upsetting topics.

Sandra Pettersson and Nina Engen for being great persons and for incredible help and support over the years.

Lisa Björk for brainstorming about the title of the thesis and Kristina Glise for burnout discussions.

Anette Johansson and Ingrid Lundin for helping me with the Swedish summary, for taking me out for lunch walks and making sure that I kept breathing fresh air, and for increasing my knowledge in trees, frogs and toads.

Former and present PhD colleagues, especially Fredrik Bååthe and Lars Rödjer – during the final intensive months of this PhD journey, it was marvellous to share moments of frustration and happiness with someone in the same situation.

Eva Sahlin, a former PhD student and friend, for your great support and encouraging messages and phone calls, which meant a lot to me.

I would like to thank all my wonderful friends and family. Thank you for putting up with me, especially during the past few months, and for cheering me up and helping me in any way possible, especially Seka, Adisa, Meca, Alma, Nermin and Denis. Love you guys!

Fatima for being the best sister in the world and my best friend, and her family Almir, Benjamin and Alma. Thank you for your love and support and many moments of joy.

Finally, my parents Azemina and Smail, my biggest supporters in everything I do, for letting me grow up in the belief that nothing is impossible and encouraging me to follow my own path, but most of all for your unconditional love and always being there for me. Volim vas puno!

# REFERENCES


1. EU-OSHA Ea: Eurofound and EU-OSHA (2014), Psychosocial risks in Europe: Prevalence and strategies for prevention. In. Edited by Flintrop J, Vargas O. Luxembourg; 2013.
2. Stefansson CG: Chapter 5.5: major public health problems - mental ill-health. *Scand J Public Health Suppl* 2006, 67:87-103.
3. Commision E: Report on the implementation of the European social partners' Framework Agreement on Work-related Stress. In: *Commission staff working paper*. Brussels: European commision; 2011.
4. Commission E: Guidance on work related stress - Spice of life or kiss of death? In: *Office for Official Publications of the European Communities.* Luxembourg; 2000.
5. Levi L: Occupational stress: Spice of life or kiss of death? *American Psychologist* 1990, 45(10):1142-1145.
6. Karasek R: Job Demands, Job Decision Latitude, and Mental Strain: Implications for Job Redesign. *Adm Sci Q* 1979, 24(2):285 - 308.
7. Karasek R, Theorell T: Healthy work : stress, productivity, and the reconstruction of working life. New York, N.Y.: Basic Books; 1990.
8. Siegrist J: Adverse health effects of high-effort/low-reward conditions. *J Occup Health Psychol* 1996, 1(1):27-41.
9. van Vegchel N, de Jonge J, Bosma H, Schaufeli W: Reviewing the effort–reward imbalance model: drawing up the balance of 45 empirical studies. *Soc Sci Med* 2005, 60(5):1117-1131.
10. Beauregard N, Marchand A, Blanc M-E: What do we know about the non-work determinants of workers' mental health A systematic review of longitudinal studies. *BMC Public Health* 2011, 11(1):439.
11. Clark C, Pike C, McManus S, Harris J, Bebbington P, Brugha T, Jenkins R, Meltzer H, Weich S, Stansfeld S: The contribution of work and non-work stressors to common mental disorders in the Adult Psychiatric Morbidity Survey. *Psychol Med* 2007, 42(04):829 - 842.
12. Cole D, Ibrahim S, Shannon H, Scott F, Eyles J: Work and life stressors and psychological distress in the Canadian working population: a structural equation modelling approach to analysis of the 1994 National Population Health Survey. *Chronic Dis in Canada* 2002, 23(3):91 - 99.
13. Hasselberg K, Jonsdottir I, Ellbin S, Skagert K: Self-reported stressors among patients with Exhaustion Disorder: an exploratory study of patient records. *BMC Psychiatry* 2014, 14(1):66.
14. Wilcock AA, Chelin M, Hall M, Hamley N, Morrison B, Scrivener L, Townsend M, Treen K: The relationship between occupational balance and health: a pilot study. *Occup Ther Int* 1997, 4(1):17-30.

15. Wagman P, Hakansson C: Exploring occupational balance in adults in Sweden. *Scand J Occup Ther* 2014, 6:1-6.
16. Hakansson C, Bjorkelund C, Eklund M: Associations between women's subjective perceptions of daily occupations and life satisfaction, and the role of perceived control. *Aust Occup Ther J* 2011, 58(6):397-404.
17. van Veldhoven MJ, Sluiter JK: Work-related recovery opportunities: testing scale properties and validity in relation to health. *Int Arch Occup Environ Health* 2009, 82(9):1065-1075.
18. Håkansson C, Ahlborg G, Jr.: Perceptions of Employment, Domestic Work, and Leisure as Predictors of Health among Women and Men. *Journal of Occupational Science* 2010, 17(3):150-157.
19. McEwen BS: Stress, Adaptation, and Disease: Allostasis and Allostatic Load. *Ann N Y Acad Sci* 1998, 840(1):33-44.
20. Hobfoll SE: Conservation of resources: A new attempt at conceptualizing stress. *Am Psychol* 1989, 44(3):513-524.
21. Ursin H, Eriksen HR: The cognitive activation theory of stress. *Psychoneuroendocrinology* 2004, 29(5):567-592.
22. Cohen S, Kessler RC, Gordon LU: Strategies for measuring stress in studies of psychiatric and physical disorders. In: *Measuring stress: a guide for health and social scientists.* edn. Oxford; New York, N.Y: Oxford University Press; 1995: 3-26.
23. Nixon A, Mazzola J, Bauer J, Krueger J, Spector P: Can work make you sick A meta-analysis of the relationships between job stressors and physical symptoms. *Work Stress* 2011, 25(1):1 - 22.
24. Russell JA: A circumplex model of affect. *J Pers Soc Psychol* 1980(39):1161-1178.
25. Ford MT, Matthews RA, Wooldridge JD, Mishra V, Kakar UM, Strahan SR: How do occupational stressor-strain effects vary with time? A review and meta-analysis of the relevance of time lags in longitudinal studies. *Work Stress* 2014, 28(1):9-30.
26. Cohen S, Williamson G: Perceived Stress in a Probability Sample of the United States. Newbury Park, CA: Sage; 1988.
27. Kjellberg A, Iwanowski A: Stress/energy formuläret: Utveckling av en metod för skattning av sinnesstämning i arbetet [The Stress/Energy Questionnaire: Development of an Instrument for Measuring Mood at Work]. In. Solna, Sweden: National institute of occupational health; 1989.
28. Kjellberg A, Wadman C: Subjektiv stress och dess samband med psykosociala arbetsförhållanden och hälsobesvär. En prövning av Stress-Energi modellen. [Subjective stress and its relation to psychosocial work conditions and health complaints. A test of the Stress-Energy model]. In: *Arbete och Hälsa.* vol. 2002. Stockholm: National Institute for Working Life; 2002.

29. Wahlstrom J, Lindegard A, Ahlborg G, Jr., Ekman A, Hagberg M: Perceived muscular tension, emotional stress, psychological demands and physical load during VDU work. *Int Arch Occup Environ Health* 2003, 76(8):584-590.
30. Kjellberg A, Wadman C: The role of the affective stress response as a mediator of the effect of psychosocial risk factors on musculoskeletal complaints--Part 1: Assembly workers. *Int J Ind Ergon* 2007a, 37(4):367-374.
31. Kjellberg A, Toomingas A, Norman K, Hagman M, Herlin RM, Tornqvist EW: Stress, energy and psychosocial conditions in different types of call centres. *Work* 2010, 36(1):9-25.
32. Eklöf M, Ingelgård A, Hagberg M: Is participative ergonomics associated with better working environment and health? A study among Swedish white-collar VDU users. *Int J Ind Ergon* 2004, 34(5):355-366.
33. Hansen AM, Blangsted AK, Hansen EA, Sogaard K, Sjogaard G: Physical activity, job demand-control, perceived stress-energy, and salivary cortisol in white-collar workers. *Int Arch Occup Environ Health* 2009, 83(2):143-153.
34. Kristiansen J, Mathiesen L, Nielsen PK, Hansen AM, Shibuya H, Petersen HM, Lund SP, Skotte J, Jorgensen MB, Sogaard K: Stress reactions to cognitively demanding tasks and open-plan office noise. *Int Arch Occup Environ Health* 2009, 82(5):631-641.
35. Nabe-Nielsen K, Garde AH, Diderichsen F: The effect of work-time influence on health and well-being: a quasi-experimental intervention study among eldercare workers. *Int Arch Occup Environ Health* 2010.
36. Nieuwenhuijsen K, Bruinvels D, Frings-Dresen M: Psychosocial work environment and stress-related disorders, a systematic review. *Occup Med* 2010, 60(4):277-286.
37. Netterstrom B, Conrad N, Bech P, Fink P, Olsen O, Rugulies R, Stansfeld S: The relation between work-related psychosocial factors and the development of depression. *Epidemiol Rev* 2008, 30:118 - 132.
38. Stansfeld S, Candy B: Psychosocial work environment and mental health-a meta-analytic review. *Scand J Work Environ Health* 2006, 32(6):443 - 462.
39. Seidler A, Thinschmidt M, Deckert S, Then F, Hegewald J, Nieuwenhuijsen K, Riedel-Heller S: The role of psychosocial working conditions on burnout and its core component emotional exhaustion - a systematic review. *J Occup Med Toxicol* 2014, 9(1):10.
40. Maslach C: Burned-out. *Hum Behav* 1976, 5:16 - 22.
41. Melamed S, Kushnir T, Shirom A: Burnout and risk factors for cardiovascular diseases. *Behav Med* 1992, 18(2):53-60.

42. Glise K, Ahlborg G, Jonsdottir I: Course of mental symptoms in patients with stress-related exhaustion: does sex or age make a difference? *BMC Psychiatry* 2012, 12(1):18.
43. Stevens SS: On the Theory of Scales of Measurement. *Science* 1946, 103(2684):677-680.
44. Bridgman P: The Logic of Modern Physics. New york: Macmillan; 1922.
45. Hand DJ: Statistics and the Theory of Measurement. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1996, 159(3):445-492.
46. Hand DJ: Measurement theory and practice: the world through quantification Edward Arnold. ; 2004.
47. Luce RD, Suppes P: Representational Measurement Theory. In: *Stevens' Handbook of Experimental Psychology*. 3rd edn.: John Wiley & Sons, Inc; 2001: 1-42.
48. Michell J: Measurement scales and statistics: A clash of paradigms. *Psychol Bull* 1986, 100(3):398-398-407.
49. Michell J: An Introduction to the Logic of Psychological Measurement. Hllsdale: Erlbaum; 1990.
50. DeVellis RF: Scale development: theory and applications. Newbury Park: Sage; 2003.
51. Stevens SS: On the averaging of data. *Science* 1955, 121(3135):113-116.
52. Kampen J, Swyngedouw M: The Ordinal Controversy Revisited. *Quality & Quantity* 2000, 34(1):87-102.
53. Erik von E, Matthias E: The scandal of poor epidemiological research. *BMJ* 2004, 329(7471):868-869.
54. Rushton L: Reporting of occupational and environmental research: use and misuse of statistical and epidemiological methods. *Occup Environ Med* 2000, 57(1):1-9.
55. Pocock SJ, Timothy JC, Kimberley JD, Bianca LdS, Marlene BG, Leslie AK, Linda EK, Valerie AM: Issues in the reporting of epidemiological studies: a survey of recent practice. *BMJ* 2004, 329(7471):883.
56. Baccaglini L, Shuster JJ, Cheng J, Theriaque DW, Schoenbach VJ, Tomar SL, Poole C: Design and statistical analysis of oral medicine studies: common pitfalls. *Oral Dis* 2010, 16(3):233-241.
57. Novack L, Jotkowitz A, Knyazer B, Novack V: Evidence-based medicine: assessment of knowledge of basic epidemiological and research methods among medical doctors. *Postgrad Med J* 2006, 82(974):817-822.
58. Cohen J: A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* 1960, 20(1):37-46.
59. Siegel S, Castellan NJ: Nonparametric statistics for the behavioral sciences, 2nd ed. edn. New York: McGraw Hill; 1988.

60. Svensson E: Application of a rank-invariant method to evaluate reliability of ordered categorical assessments. *Journal of Epidemiology and Biostatistics* 1998, 3(4):403-409.
61. Svensson E: Ordinal invariant measures for individual and group changes in ordered categorical data. *Stat Med* 1998, 17(24):2923-2936.
62. Svensson E: Guidelines to statistical evaluation of data from rating scales and questionnaires. *J Rehabil Med* 2001, 33(1):47-48.
63. Agresti A: Categorical Data Analysis, 2nd edn. New York: Wiley 2002.
64. Liu I, Agresti A, Tutz G, Simonoff JS, Kateri M, Lesaffre E, Loughin TM, Svensson E, Aguilera AM: The analysis of ordered categorical data: An overview and a survey of recent developments. *Test* 2005, 14(1):1-73.
65. Chan EKH: Standards and Guidlines for Validation Practices: Development and Evaluation of Measurement Instruments. In: *Validity and Validation in Social, Behavioral, and Health Sciences. Volume Social Indicators Research Series*, edn. Edited by Zumbo BD, Chan EKH. Switzwerland: Springer International Publishing; 2014.
66. American Educational Research A, American Psychological A, National Council on Measurement in E, Joint Committee on Standards for E, Psychological T: Standards for educational and psychological testing; 2014.
67. Messick S: Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. Research Report RR-94-45. In.; 1994.
68. Nunnally JC: Psychometric theory. New York: McGraw-Hill; 1994.
69. Fayers P, Machin D: Quality of Life: The Assessment, Analysis and Interpretation of Patient-reported Outcomes, 2 edn. Chichester: John Wiley & Sons; 2007.
70. Mokkink L, Terwee C, Patrick D, Alonso J, Stratford P, Knol D, Bouter L, de Vet HW: The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010, 19(4):539-549.
71. Cook AD, Beckman TJ: Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American journal of medicine* 2006, 119(2):166.e167-166.e116.
72. Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN: How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med* 2004, 19(9):971-977.
73. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, Savoy SM, Kostas-Polston E: A psychometric toolbox

for testing validity and reliability. *J Nurs Scholarsh* 2007, 39(2):155-164.
74. Olsen Jr, Group IEAEQ: Epidemiology deserves better questionnaires. *Int J Epidemiol* 1998, 27(6):935.
75. Messick S: Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol* 1995, 50(9):741-749.
76. American Educational Research Association, American Psychological Association, Education NCoMi: Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
77. Downing SM: Validity: on the meaningful interpretation of assessment data. *Med Educ* 2003, 37(9):830-837.
78. Svensson E: Analysis of systematic and random differences between paired ordinal categorical data Gothenburg, Sweden: University of Gothenburg; 1993.
79. Campbell DT, Fiske DW: Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 1959, 56:81-105.
80. DeVellis RF: Classical Test Theory. *Med Care* 2006, 44(11):S50-S59.
81. Lord FM: Applications of item response theory to practical testing problems. New Jersey: Erlbaum; 1980.
82. Rasch G: Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press; 1960.
83. Hambleton RK, Swaminathan H, Rogers HJ: Fundamentals of Item Response Theory. Newbury Park: Sage; 1991.
84. Lord FM, Novick MR: Statistical theories of mental test scores: Addison-Wesley Pub. Co.; 1968.
85. Luce RD, JW T: Simultaneous conjoint measurement: A new type of fundamental measurement. *J Math Psychol* 1964, 1:1-27.
86. Van Newby A, Conner GR, Bunderson CV: The Rasch model and additive conjoint measurement. *J Appl Meas* 2009, 10:348–354.
87. Perline R, Wright BD, H W: The Rasch model as additive conjoint measurement. *Appl Psycho Meas* 1997, 3:237-256.
88. Karabatos G: The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *J Appl Meas* 2001, 2:389-423.
89. Masters G: A Rasch model for partial credit scoring. *Psychometrika* 1982, 47:149-174.
90. Andrich D: Rating formulation for ordered response categories. *Psychometrika* 1978(43):561-573.
91. Guttman L: The basis for Scalogram analysis. In Studies in social psychology in World War II: Vol. 4. Measurement and Prediction. Princeton: Princeton University Press; 1950.

92. Marais I, Andrich D: Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of applied measurement* 2008, 9(2):105-124.
93. Garst H, Frese M, Molenaar PC: The temporal factor of change in stressor-strain relationships: a growth curve model on a longitudinal study in east Germany. *J Appl Psychol* 2000, 85(3):417-438.
94. Tang K: A reciprocal interplay between psychosocial job stressors and worker well-being? A systematic review of the "reversed" effect. *Scand J Work Environ Health* 2014, 40(5):441-456.
95. Glise K, Hadzibajramovic E, Jonsdottir IH, Ahlborg G, Jr.: Self-reported exhaustion: a possible indicator of reduced work ability and increased risk of sickness absence among human service workers. *Int Arch Occup Environ Health* 2010, 83(5):511-520.
96. Jonsdottir IH, Rodjer L, Hadzibajramovic E, Borjesson M, Ahlborg G, Jr.: A prospective study of leisure-time physical activity and mental health in Swedish health care workers and social insurance officers. *Prev Med* 2010, 51(5):373-377.
97. Lundgren-Nilsson A, Jonsdottir IH, Pallant J, Ahlborg G, Jr.: Internal construct validity of the Shirom-Melamed Burnout Questionnaire (SMBQ). *BMC Public Health* 2012, 12:1.
98. Folkhälsokommittén: Stress och utmattning i Västra Götaland. [Stress and exhaustion in Western Sweden]. In: *Folkhälsan i Västra Götaland, del 2*. Public Health Committee; 2002.
99. Grimby-Ekman A, Hagberg M: Simple neck pain questions used in surveys, evaluated in relation to health outcomes: a cohort study. *BMC Res Notes* 2012, 5(1):587.
100. McEwen BS: Central effects of stress hormones in health and disease: Understanding the protective and damaging effects of stress and stress mediators. *Eur J Pharmacol* 2008, 583(2-3):174-185.
101. Theorell T, Perski A, Akerstedt T, Sigala F, Ahlberg-Hulten G, Svensson J, Eneroth P: Changes in job strain in relation to changes in physiological state. A longitudinal study. *Scand J Work Environ Health* 1988, 14(3):189-196.
102. Hadzibajramovic E: Methodological apsects of the analysis of psychosocial work environment. Örebro, Sweden: Örebro University School of Business; 2013.
103. Twisk JWR: Applied Longitudinal Data Analysis for Epidemiology, p. 77-90. Cambridge: Cambridge Univerisity Press; 2003.
104. Bland JM, Altman DG: Statistics notes: Cronbach's alpha, vol. 314; 1997.
105. Tennant A, Conaghan PG: The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care Res* 2007, 57(8):1358-1362.

106. Andrich D, Hagquist C: Real and Artificial Differential Item Functioning in Polytomous Items. *Educ Psychol Meas* 2015, 75(2):185-207.
107. Hagquist C, Andrich D: Is the Sense of Coherence-instrument applicable on adolescents? A latent trait analysis using Rasch-modelling. *Pers Individ Dif* 2004, 36(4):955-968.
108. Smith EV, Jr.: Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of applied measurement* 2002, 3(2):205-231.
109. Hagquist C, Bruce M, Gustavsson JP: Using the Rasch model in nursing research: an introduction and illustrative example. *Int J Nurs Stud* 2009, 46(3):380-393.
110. Smith EV, Jr.: Metric development and score reporting in Rasch measurement. *Journal of applied measurement* 2000, 1(3):303-326.
111. Svensson E: Concordance between ratings using different scales for the same variable. *Stat Med* 2000, 19(24):3483-3496.
112. Claesson L, Svensson E: Measures of order consistency between paired ordinal data: application to the Functional Independence Measure and Sunnaas index of ADL. *J Rehabil Med* 2001, 33(3):137-144.
113. Svensson E: Construction of a single global scale for multi-item assessments of the same variable. *Stat Med* 2001, 20(24):3831-3846.
114. Kreiner S: Validity and objectivity: Reflections on the role and nature of Rasch models. *Nordic Psychology* 2007, 59(3):268-298.
115. Coste J, Bouyer J, Job-Spira N: Construction of Composite Scales for Risk Assessment in Epidemiology: An Application to Ectopic Pregnancy. *Am J Epidemiol* 1997, 145(3):278-289.
116. Coste J, Fermanian J, Venot A: Methodological and statistical problems in the construction of composite measurement scales: A survey of six medical and epidemiological journals. 1995, 14(4):331-345.
117. Coste J, Guillemin F, Pouchot J, Fermanian J: Methodological approaches to shortening composite measurement scales. *J Clin Epidemiol* 1997, 50(3):247-252.
118. Grimby G, Tennant A, Tesio L: The use of raw scores from ordinal scales: time to end malpractice? *J Rehabil Med* 2012, 44(2):97-98.
119. Fransson E, Nyberg S, Heikkila K, Alfredsson L, Bacquer DD, Batty GD, Bonenfant S, Casini A, Clays E, Goldberg M *et al*: Comparison of alternative versions of the job demand-control scales in 17 European cohort studies: the IPD-Work consortium. *BMC Public Health* 2012, 12(1):62.
120. Senn S, Julious S: Measurement in clinical trials: a neglected issue for statisticians? *Stat Med* 2009, 28(26):3189-3209.

121. Calnan M, Wadsworth E, May M, Smith A, Wainwright D: Job strain, effort--reward imbalance, and stress at work: competing or complementary models? *Scand J Public Health* 2004, 32(2):84-93.
122. Taris TW, Kompier MAJ: Cause and effect: Optimizing the designs of longitudinal studies in occupational health psychology. *Work Stress* 2014, 28(1):1-8.
123. Shirom A, Melamed S, Toker S, Berliner S, Shapira I: Burnout and Health Review: Current Knowledge and Future Research Directions. In: *International Review of Industrial and Organizational Psychology 2005*. edn.: John Wiley & Sons, Ltd; 2006: 269-308.
124. Tsutsumi A, Iwata N, Wakita T, Kumagai R, Noguchi H, Kawakami N: Improving the measurement accuracy of the effort-reward imbalance scales. *Int J Behav Med* 2008, 15(2):109-119.
125. Choi B, Schnall P, Landsbergis P, Dobson M, Ko S, Gomez-Ortiz V, Juarez-Garcia A, Baker D: Recommendations for individual participant data meta-analyses on work stressors and health outcomes: comments on IPD-Work Consortium papers. *Scand J Work Environ Health* 2015.