# Validating the Knowledge Intensive Working Environment Survey Target 2.0 Latent Variable Measurement Interpretation.

Haakon Thorbergsen Haakstad.
Master's Thesis in Work and Organizational Psychology.
Department of Psychology.
Norwegian University of Science and Technology, Trondheim.
Spring 2017.

**Acknowledgements**

To my family; my mother, my father, and my sister – for their patience, their understanding, their faith, and their support. My success is predicated upon me having been blessed with a family possessing these virtues. Know that I recognize the fact that if not for your selfless support and unconditional positive regard, I would surely have been deprived of this opportunity at self-discovery and realization.

To the objective third-party person who believed in me, and helped me believe in myself when I needed it the most; directing me towards this path, encouraging me to take on this journey. I would not be where I now am, standing where I now stand, if not for you providing the impetus, pointing me in this direction.

To my mentors; who recognized my interests, encouraged me to pursue them, granted me the opportunities to develop them, and provided me the necessary conditions to succeed at them.

To my friends; who support and encourage me, and whom I experience as acknowledging, recognizing, and appreciating me. Know that I acknowledge, recognize, and appreciate you in turn. I can only hope you feel sufficiently reciprocated in terms of the encouragement and support I provide to you.

To my love; for your love.

**Abstract**

This thesis primarily concerns validation of the Knowledge Intensive Working Environment Survey Target (KIWEST) 2.0 measurement theory as a valid account of the observations generated by administering KIWEST 2.0 on a sample from its target population (N = 12170). A working model of validation is developed by combining validity- and latent variable theory. A distinction is drawn between latent variable measurement interpretations (weak claims) and identity interpretations (strong claims).

KIWEST consists of 119 items, and its measurement theory specifies between 27 to 33 latent variables to account for observed (co)variation, depending on whether its multifaceted constructs are represented by single or multiple factors. Following data integrity treatment, 7643 cases and 118 items were retained. The method employed was maximum likelihood confirmatory factor analysis, employing the alternative-models strategy of Jöreskog (1993), comparing the fit of- and selecting among 16 nested models accounting for item (co)variation.

The least parsimonious model was retained and subjected to evaluation of parameter estimates as evidence of validity. The results indicate that the model comprehensively account of the observations, but suffers from lack of parsimony. The discussion develops a number of suggestions for altering the interpretation to fit the observations (i.e., changes to the KIWEST theory; proximal remedies), and for altering the questionnaire to produce observations that fit the interpretation (i.e., changes to the KIWEST questionnaire; distal fixes). The conclusion of the thesis is that changes ought to be made to either the questionnaire or interpretation before proceeding with validation of the KIWEST latent variable identity interpretations.

*Keywords:* ARK, KIWEST, Working-environment, Working-climate, Academic sector, Validity, Validation

## Table of Contents

**Validating the KIWEST 2.0 Interpretation of the KIWEST 2.0 Questionnaire**

The purpose of this thesis is take on the task which Steven G. Sireci (2007, p. 477) describe as "*the ultimate challenge for a psychometrician*" – validation, addressing that which in the current Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014, p. 11) is proclaimed "*the most fundamental consideration in developing tests and evaluating tests*" – validity. Specifically, what will be subjected to validation is the Knowledge Intensive Working Environment Survey Target 2.0 survey measurement interpretation, offered by the ARK projects KIWEST measurement theory (Undebakke, Innstrand, Anthun, & Christensen, 2014; hereafter referred to as "the KIWEST theory"). KIWEST is a major component of the ARK project, which deals with working environment and climate research and interventions.

Measurement interpretations are the most basic sorts of claims in quantitative research, on which all subsequent research stand and fall. As such, the cost of faulty measurement can prove enormous (DeVellis, 2017), as invalid measurement interpretations render all further research assuming the validity of measurement interpretations invalid. For ARK specifically, any substantive claims about working climates and environments made based on research that makes use of data collected with KIWEST (and that subscribes to its default measurement interpretations) would be rendered invalid should the measurement interpretations on which those substantive claims build upon prove invalid. Thus, for this thesis, the primary research question is; "Does the KIWEST measurement theory offer a valid interpretation of the observations generated by administering the KIWEST questionnaire on its target population?"

By being principally concerned with a specific instance of validity and validation, the thesis naturally extends into validity and validation in a more general sense. That is, in order to accomplish its primary task, the thesis consults and builds on general literature of validity and validation. During the course of working on this thesis, it became apparent that there is a rather wide gap between the theory and practice of validity and validation; a gap that has not gone unnoticed by prominent validity theorists, and has been thoroughly documented (e.g., Cizek, Bowen, & Church, 2010; Cizek, Rosenberg, & Koons, 2007; Newton & Shaw, 2013, 2014; Shear & Zumbo, 2014).[1] Current common contemporary practice of validation appears

[1] Cizek et al. (2007), for example, found that from a sample of validation studies, only 9.5% explicitly consulted contemporary authoritative sources on validity when conducting validation research. Only 2.5% adopted the modern perspective on validity, and as many as 45.2% failed to make clear which conception of validity that was adopted. Cizek et al. (2007) considered this evidence that Frisbie (2005, p. 21) was correct in his remark that "*For a concept that is the foundation of virtually all aspects of our measurement work, it seems that the term* validity *continues to be one of the most misunderstood or widely misued of all.*" This thesis aims to avoid contributing to this common yet undesirable practice by making validity theory its very foundation.

at this point in time to be stuck at least eighteen years in the past in terms of its (non)usage of methods and terminology (i.e., appears to be stuck in a pre-AERA, APA, & NCME, 1999 era of thinking and talking about validity and validation; see Newton & Shaw, 2013, 2014).

Thus, by engaging with modern authoritative literature on validity and validation, the terminology employed might seem foreign to someone socialized into the old paradigm. As Kuhn (1962/2012) points out, the outcome of a rendezvous of paradigms might be confusion in communication across borders, as the old and the new framework of thinking might turn out to be fundamentally incompatible, and adherents to either paradigm – due to superficial similarities in their use of language – might not recognize the depth of the disconnect. In an attempt to minimize such confusion, this thesis includes an account of the old paradigm and reviews the new paradigm in light of the old. Additionally, the modern paradigm is reviewed in light of contemporary grievances with theory and practice in order to foster a self-critical view of the efforts of this thesis (as well as of the field more generally).

The method employed for the purposes of validation in this thesis is confirmatory factor analysis. This is a common method of validation within psychology and related fields, illustrated with the following PSYCNET query: Keywords: validation AND Peer-Reviewed Journals Only AND Methodology: Empirical Study. As of this time of writing (February 9th, 2017) fifteen of the first twenty-five articles (60%) made use of confirmatory factor analysis for the purposes of validation, either by itself or in conjunction with additional methods. As contemporary validation practice appears out of sync with modern validity theory (due mostly to widespread use of antiquated terminology), the matter of how contemporary validation practice interfaces with modern validity theory appears to be a relatively unexplored issue.

Thus, in addition to contributing specifically by validating KIWEST, the thesis aims to contribute more generally as an example by examining the interfacing between the method employed and modern validity theory. In light of the popularity of the method in validation research, it is argued that such an examination would constitute a contribution to the field more generally by potentially contributing to a realignment of theory and practice.

In the words of Ian Hacking (1983, p. 31): "*Science has two aims: theory and experiment. Theories try to say how the world is. Experiment and subsequent technology change the world. We represent and we intervene. We represent in order to intervene, and we intervene in the light of representations.*" This thesis is concerned with ensuring the quality of representations of psychosocial phenomena in the KIWEST questionnaire, as posited by the interpretation offered by the KIWEST measurement theory, for the purposes of securing the foundations for subsequent structural investigations and practical interventions. As put by the

esteemed statistician John Tukey (1969, p. 88): "*Clarity in the large comes from clarity in the medium scale; clarity in the medium scale comes from clarity in the small. Clarity always comes with difficulty.*"

**Research Questions and the Content and Structure of This Thesis**

The research question of this thesis was presented in the previous section. It is restated here for the sake of clarity, as well as to justify the way the thesis is structured:

- Does the KIWEST measurement theory offer a valid interpretation of the observations generated by administering the KIWEST questionnaire on its target population?

In order to investigate the primary research question, the natural theoretical point of departure is validity theory, which is supplemented and integrated with latent variable theory in order to integrate validity theory with the specific kinds of measurement interpretations that are validated in this thesis: latent variable measurement interpretations. For this purpose, the method employed for validation in this thesis is maximum likelihood confirmatory factor analysis (CFA), which is widely recognized as a powerful tool for latent variable modelling.

Following the presentation of the meta-theoretical foundation of KIWEST 2.0, validity theory and latent variable theory, the constructs positing specific latent variables measured by the KIWEST survey in the KIWEST theory are reviewed and presented in detail (in terms of their conceptual and operational definitions, and their individual theoretical foundations). The method section furthermore accounts for the data integrity analysis and treatment in detail to provide the necessary backing for the interpretations of the results (e.g., indices and parameter estimates) from the following CFA analysis. Furthermore, the theory and practice of CFA and how it integrates with validity theory for the purposes of validation (i.e., what the method can and cannot aid in determining) is accounted for.

Following the method section, the results from the CFA are accounted for in terms of how they relate to the validity of the proposed latent variable measurement interpretations. Drawing on literature pertaining to validity theory and latent variable theory, the resulting discussion concern alterations that could to be made to the KIWEST questionnaire (a distal fix) or to its latent variable measurement interpretation (a proximal remedy) in order to increase its validity in the short- or the long term.

## KIWEST, Validity, and Validation

The first standard in the Standards for Educational and Psychological Testing (AERA et al., 2014, p. 23) reads as follows: "*Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided.*" The purpose of the current segment is to comply with this standard by offering a brief-yet-comprehensive articulation of the intended uses and meta-theoretical interpretation of KIWEST. In short, the KIWEST questionnaire consists of 119 items, grouped by the KIWEST theory into sets that are postulated to constitute measures of between 27 to 33 psychosocial and working-environmental entities. An articulation of the intended interpretations of each individual test score is available in the methods section. The current account is based on the ARK report of Undebakke et al. (2014).

The theoretical and empirical backing for the selection of constructs to be included in- and measured by the KIWEST questionnaire is, in addition to the demands of the Norwegian working environment act (hereafter referred to as the "NWEA"; ASD, 2005), based on the meta-theoretical framework of- and findings made from the perspective of the Job Demands-Resources model (the JD-R model; Bakker, Demerouti, & Sanz-Vergel, 2014). The NWEA points to (but rarely defines) desirable properties of the working environment (see table 2.1). The JD-R model categorizes work-environmental conditions as either demands or resources (Tadić, Bakker, & Oerlemans, 2015), and suggest how they relate both to each other as well as to outcomes of interest (e.g., to the criteria set by the NWEA).

The purpose motivating the assembly of KIWEST has been to cover that which contemporary theory and research suggest are the most important psychosocial working environmental factors for the academic sector, in terms of the above-mentioned demands and resources. Schaufeli and Bakker (2004, p. 296) have defined demands as "*those physical, psychological, social, or organizational aspects of the job that require sustained physical and/or psychological (i.e., cognitive or emotional) effort and are therefore associated with*

Table 2.1.
*The Norwegian Working Environment Act §1-1: Preamble*

| Letter | The purpose of the law is: |
|---|---|
| a) | To secure a working environment that provides the basis for a health-promoting and meaningfull work situation, that provides full safety/security against physical and psychological harm, and with a welfare standard that is at any given time corresponds with the technological and social development of society. |
| b) | To safeguard conditions of employment and equal treatment at work and in working. |
| c) | To facilitate adjustments to working conditions tied to the circumstances of the employee. |
| d) | To provide the basis for the employer and the employees in the enterprises themselves to develop their working environment in cooperation with social partners, with the necessary guidance and control from public authority. |
| e) | To contribute to an inclusive working life. |

Note: Authors translation. Might not adequately represent the original intent of the law.

*certain physiological and/or psychological costs.*" Resources on the other hand, they (ibid.) defined as "*those physical, psychological, social, or organizational aspects of the job that either/or (1) reduce job demands and the associated physiological and psychological costs; (2) are functional in achieving work goals; (3) stimulate personal growth, learning and development.*" As such, KIWESTs construct theories specify expectations regarding how specific factors are supposed to relate to outcomes of interests (see figure 2.1).

The progenitors of the ARK project have for this purpose chosen to make use of freely available, standardized, and validated scales from known and renowned Nordic and European research initiatives, the most ambitious and comprehensive of which being the $QPS_{Nordic}$ (Dallner et al., 2000), the COPSOQ II (Pejtersen, Kristensen, Borg, & Bjorner, 2010), and the N-POP (Christensen et al., 2012). KIWEST exclusively contains whole scales, meaning that the respondents are asked to respond to at least three statements belonging to each scale. In composing the survey, the authors have made efforts to balance the included dimensions between: (1) the working climate level (individuals' perception of the collective experience of the working environment) and the individual level (individuals' perception of their own personal experience), (2) demands and resources, and (3) a focus on the individual, the group, the management, and the organization.

The KIWEST questionnaire is currently in its second working iteration, which is based on the results of the validation efforts of its first working iteration (KIWEST 1.0; Innstrand, Christensen, Undebakke, & Svarva, 2015). Since its previous iteration, a number of measures have been removed or replaced based on either apparent lack of relevance, or due to failure of factor convergence of items or discrimination between proposed factors. As such, KIWEST 2.0 contains measures of a number of factors previously featured in KIWEST 1.0, as well as a number of factor measures that were not present in the previous iteration. Thus, the thesis builds on previous theory and research, and is for this reason conducted from a primarily confirmatory and evaluative stance. The intended measurement interpretations of KIWEST 2.0 are known (i.e., predefined), and are the subjects of validation to determine their validity.

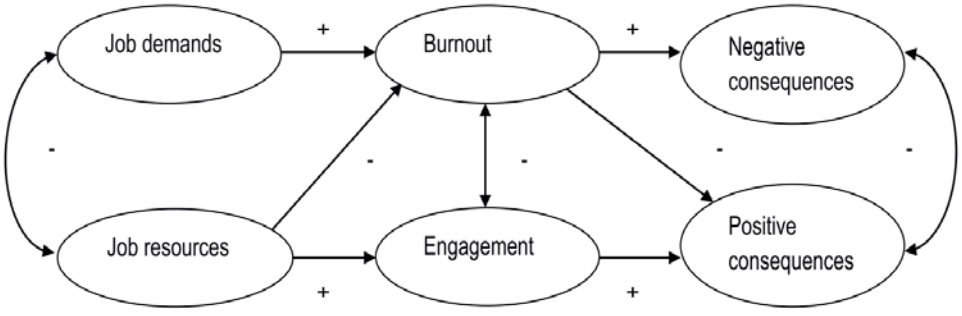


*Figure 2.1.* The Job Demands-Resources Model (JD-R). Adopted and adapted from Undebakke et al. (2014).

**Validity and Validation: Theory and Practice**

In the current Standards for Educational and Psychological Testing (hereafter referred to as "the Standards"), validity is defined as "*the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a tests*" (AERA et al., 2014, p. 225). This represents the current official definition of validity, which one might be justified in labeling the "consensus definition" (e.g., Newton, 2012a, 2012b; Newton & Shaw, 2014). Validation is defined as "*the process through which the validity of a proposed interpretation of test scores for their intended uses is investigated*" (AERA et al., 2014, p. 225), and is described as a process that "*involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations*" (AERA et al., 2014, p. 11).

The Standards is intended as a normative and authoritative document, instructing researchers and practitioners in how they *should* or *ought* to practice their craft. As articulated by Newton and Shaw (2013), the notion of "standards" captures the idea of consensus, within a community, concerning how its members ought to behave, and are essential to communities because they enable individuals to function collectively. In the most current standards, its stated purpose is to provide criteria for the development and evaluation of tests and testing practices, as well as guidelines for assessing the validity of interpretations of test scores for the intended test uses (AERA et al., 2014).

According to Newton (2012a), each iteration of the standards have been plagued by ambiguities and inconsistencies, revealing them as essentially representing a compromise- rather than a consensus position, attempting to simultaneously accommodate the particular foci and interests of several disciplines and sub-disciplines. As a consequence of the strain compromise exerts on the concept in order to unite several concerns one banner, the academic debate regarding the interrelated concerns of the appropriate meaning, scope, and focus of validity is to this day alive and well. The vitality of the debate surrounding validity and validation is demonstrated by the sheer number of journal special issues and edited books dedicated entirely to the subject leading up to the publication of the recent 2014 edition (e.g., Kane, 2013a; Lissitz, 2009; Lissitz & Samuelsen, 2007; McGrath, 2005a; Newton, 2012a; Slaney & Racine, 2013a; Zumbo & Chan, 2014).

By its nature as an official document of an organizing body claiming authority and demanding obedience, the standards serve as a natural focal point around which to structure an account of validity theory and the debate surrounding it. The position articulated in the Standards has evolved with each iteration (Eignor, 2013; Newton, 2012a; Newton & Shaw, 2013), simultaneously directing- and being directed by the developmental trajectory of

validity theory. Each new iteration represents an attempt at reflecting the consensus (some would say compromise) of its era. The Standards is currently in its sixth iteration (AERA et al., 2014), and both the orthodox and the unorthodox sides of the contemporary debate are rooted historically in its development.

The developmental trajectory of validity theory and the Standards can, according to Newton and Shaw (2014), be said to have oscillated between periods of fragmentation and crystallization, characterized by (according to Markus & Borsboom, 2013a) the interacting processes of expansion, unification, and partitioning. A practical consequence of the historical development of the Standards and validity theory is that it strictly speaking has become incorrect to use modifier labels and prefixes such as content-, criterion-, and construct validity (a principle that appear close to universally ignored and systematically violated in practice; Newton & Shaw, 2013). In fact, while still prevalent in common literature, this terminology (e.g., that of referring to content, criterion, and construct validity) was discarded in the 1999 Standards, and remains so in the 2014 Standards. What was once considered *types* of validity is now considered *sources of evidence* of *one* validity – construct validity.

Because of the well documented disconnect between validity theory as articulated by the Standards and the common contemporary practice and discourse of validity and validation (Cizek et al., 2010; Cizek et al., 2007; Newton & Shaw, 2013; Shear & Zumbo, 2014), it is deemed necessary to pre-empt objections by accounting for the history of validity theory and validation practice in some detail. In the words of philosopher of hermeneutics Hans-Georg Gadamer (1974/2004, p. 182): "*The breakdown of the immediate understandings of things in their truth is the motive for the detour into history.*" As evinced by Kuhn (1962/2012), understanding the historical development of theory and practice – to recognize our way of being and doing as historically situated and determined – can allow us to better understand where we are, and why we do the things we do the way we do them in the present. The current official position on validity as articulated by the standards will for this reason be reviewed both in the context of its historical development, as well as in the context of the contemporary debate surrounding it.

**A brief historical account of validity theory and the Standards.**

As stated by Sireci (2009, p. 20); "*Validity theory and the practice of validation are almost as old as the practice of testing itself, but not quite.*" According to Newton and Shaw (2014), the concept of validity has its origins in the "measurement movement," which emerged from the context of a number of advancements within testing, assessment, and statistical methodology

in the mid- to late 1800's. Among these were developments such as the proliferation of the standardized test, and the invention of statistical techniques such as the correlation coefficient and factor analysis. Obviously, the birth of the term validity predates the Standards, and by a good number of decades. The earliest attempts at establishing a standardized definition of validity is generally traced to around (first) the late 1890's and (second) the early 1920's, but is described by Newton and Shaw (2013, 2014) as for the most part unsuccessful.

The first official standards (generally and officially recognized as such; Eignor, 2013) did not see the light of day until 1954/1955, with revised editions published in 1966, 1974, 1985, 1999, and most recently, (as of this time of writing) 2014. Each subsequent edition of the standards represents a substantial revision of the official position, prompted – and necessitated, in order for there to even *be* a relevant unifying official stance on how to think and talk about validity (Newton & Shaw, 2013) – by the contemporary developments in theory and practice at the time. Since 1966, the Standards have been the product of a joint commission consisting of members from the American Educational Research Association (AERA) the American Psychological Association (APA), and the National Council of Measurement in Education (NCME).

The Standards have evolved from instructing practitioners to talk and think in terms of types of validities, to think and talk in terms of aspects of one validity. This doctrine is known as the "unitarian" conception of validity, which in the 1985 standards officially substituted the "trinetarian" or "tripartite" conception which had prevailed since the 1954 standards. In the subsequent 1999 standards, prior to which content-, criterion- and construct validity were considered the "pillars of validity" (representing something akin to a "holy trinity", hence "trinetarian"; Newton & Shaw, 2013, 2014), all of validity was subsumed under the banner of construct validity. Since 1999, all means of validation is to be performed in the service of construct validation. From the perspective of construct validity theory, validation of test score interpretations was to be performed relative to an amalgam of theoretical constructions.

This conception of validity and validation is termed "construct validity theory", and it represents, in the words of Slaney and Racine (2013b), the "methodological imperative" of contemporary validity theory. Although it has evolved since its introduction in terms of the ontology of constructs (Lovasz & Slaney, 2013; Maraun & Gabriel, 2013) and epistemology of validation (Kane, 2013a, 2013b) – construct validity owes its conceptual roots to the works of MacCorquodale and Meehl (1948) and Cronbach and Meehl (1955). A *construct* is in the current standards defined as "*the concept or characteristic that a test is designed to measure*" (AERA et al., 2014, p. 11). This definition appears to accommodate both realist and anti-

realist convictions regarding the ontological status of mental phenomena, an equating that some consider an illegitimate conflation (Markus, 2008; Slaney & Racine, 2013a, 2013b).

As construct validity now *is* de facto validity, the modifier "construct" in "construct validity" is, in principle, superfluous. An account of validity theory and the contemporary debate surrounding it thus naturally centers around the historical development of the concepts of construct validity and construct validation, and it is this history which will be reviewed in greater detail in the following segment.

***The genesis, ascension, evolution, and triumph of construct validity.***

Construct validity was introduced as a response- and an alternative to the neo-behavioristic concept of the "intervening variable" (Cronbach & Meehl, 1955; MacCorquodale & Meehl, 1948). Behaviorism was based on operationism as articulated by Bridgman (1927, p. 5), who stated: "*we mean by any concept nothing more than a set of operations.*" Thus, behaviorism defined its concepts in terms of how they were operationalized, and as such, operations were not considered proxies for accessing phenomena of interest. For operationists, phenomena such as "hunger" *was* "time since feeding" (e.g., Tolman, 1936, p. 384), and "experience" *was* "discriminatory reactions" (e.g., Stevens, 1935, p. 521).

Criterion validity, a brainchild of operationist behaviorism where the correlation coefficient was considered degree of equivalence, was concerned with finding operations that could stand for each other. If two different means of operationalization displayed perfect (or at least sufficiently strong) correlation, they were considered "*equated operations for the same entity*" (Boring, 1945, p. 244). As such, tests were not means to quantify some external phenomena; the phenomena was what the tests tested. An alternative operations yielding perfectly correlated scores with a criterion-reference would consequently be considered a valid measure of that concept (i.e., an equated operation). The validity of the criterion was never in question, as it was considered valid by definition. It was what was being measured.

Construct validity was introduced as an alternative to criterion validity for those who were not satisfied with defining their objects of study as the "how" rather than the "what" of measurement. Questioning what was measured would be meaningless to an operationist, but would constitute the essence of the matter for researchers engaged in construct validation. Construct validity was by Cronbach and Meehl (1955) explicitly tied to the philosophy of Carnap (e.g., 1959), and his concept of the nomological network. Construct validation was "*involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined.' The problem faced by the investigator is 'what constructs*

*account for variance in test performance?'*" (Cronbach & Meehl, 1955, p. 282). Construct validation and the issue of construct validity was thus invoked whenever the nature and identity of a phenomenon was in question (which it by definition could never be for an operationist). What was to be validated was the researchers interpretation of what exactly was being measured, and an understanding of the identity (or meaning) of the test-assessed quality or attribute was to be constructed from observing the behavior of the construct-test in terms of how it related to other construct-tests (i.e., investigating its nomological network).

Gaining popularity in the period between 1948 and 1955, construct validity briefly predates the release of the first edition of the Standards published in 1954/1955, within which it was included as a type of validity. The first Standards instructed practitioners not to confuse terms by addressing "validities" as aspects of one validity. The following excerpt from the first Standards provides evidence of this: "*When validity is reported, the manual should indicate clearly what type of validity is referred to. The unqualified term 'validity' should be avoided unless its meaning is clear from the context*" (APA & NCMUE, 1954, pp. 18-19; cited in Newton & Shaw, 2013). Thus, the first edition of the Standards maintained the partitioning of validity into types, and discouraged talk of aspects.

Despite these edicts of the first standards, the developmental trajectory of validity theory pointed decisively in the direction of unification. The official position followed suit – seemingly reluctantly, evidenced by the inconsistencies apparent in the subsequent iterations of the Standards. For example, the 1966 Standards shifted indecisively between addressing content-, criterion-, and construct as types and aspects of validity, evinced by the following excerpt: "*These three aspects of validity are only conceptually independent, and only rarely is just one of them important in a particular situation. A complete study of a test would normally involve information about all types of validity*" (APA, AERA, & NCME, 1966, p. 14; cited in Messick, 1987). The classic terminology was abandoned in the 1999 standards, where aspects of validity were replaced with types of evidence. The most recent edition of the standards (AERA et al., 2014, p. 14) states that "*sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept.*"

The unification of validity was spearheaded by Samuel Messick, "*whose ideas on validity*", according to Newton and Shaw (2014, p. 99), "*increasingly came to dominate the landscape [...], ultimately becoming the very zeitgeist of late 20$^{th}$-century thinking on validity.*" It was under the influence of Messick (Markus & Borsboom, 2013a; Newton & Shaw, 2013, 2014; Sireci, 2009) that construct validity came to be recognized as all of validity in the 1999 standards (admittedly backed by other influential figures at the time, such

as Robert Guion, 1980; and Jane Loevinger, 1957). Some scholars label the period leading up to the publication of the 1999 edition of the Standards the "Messick years" (e.g., Newton & Shaw, 2014), and describe the resulting 1999 standards as close to merely constituting an official reaffirmation of Messick's position.

In addition to subsuming all validity under a single header, the concept of construct validity came, under the influence of Messick (1975, 1987, 1995), to have its domain expanded to include ethical considerations addressing the social consequences of decisions made based on the application of tests. The factoring in of social consequences in the determination of validity and the meaning of constructs have been criticized from several fronts and for several reasons since it was canonized in the 1999 Standards (e.g., Borsboom, Mellenbergh, & van Heerden, 2004; Cizek, 2012), but the notion still garners support from some influential theorists (e.g., Kane, 2013a). Regardless of whether one accepts the inclusion of ethically relevant consequences in the concept of validity, the 1999 Standards did adopt the notion. As such, evidence pertaining to consequences of use came to be considered relevant to test score interpretations and uses, and the realms of evidence to examine during the process of validation expanded to include consequences of decisions based on test use.

At the same time as consequences of use was officially recognized as its own source category of evidence for validity in the 1999 Standards, the classical terminology employed in the 1985 Standards – that of referring to content-, criterion- and construct-related evidence of validity – was largely discarded in favor of four new categories rearranging and expanding on the classical terminology. These new categories of evidence were evidence based on (1) test content, (2) response processes, (3) internal structure, (4) relations with other variables, and (5) consequences of testing; a partitioning maintained by the 2014 standards (see figure 2.1 for an illustration of the evolution of the structural partitioning of validity). Validation was concerned with investigating these types of evidence, which were to be integrated in the form of an argument for or against the claim to validity. In the 1999 Standards, validity was defined as "*the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests*" (AERA et al., 1999, p. 9).

**A synopsis of the contemporary debate.**

The contemporary debate surrounding validity theory leading up to the publication of the 2014 Standards could be characterized, according to Newton and Shaw (2014), as a period of deconstruction, allegedly motivated by the concern that validity theory had become so broadly encompassing that it was no longer clear how to translate it into practice. When perusing the

various journal issues and books published on validity since the 1999 Standards, it is apparent that the concept of validity has been tackled from the perspective of each of the components of the relatively recently unified conception of validity – construct validity. Specifically, the

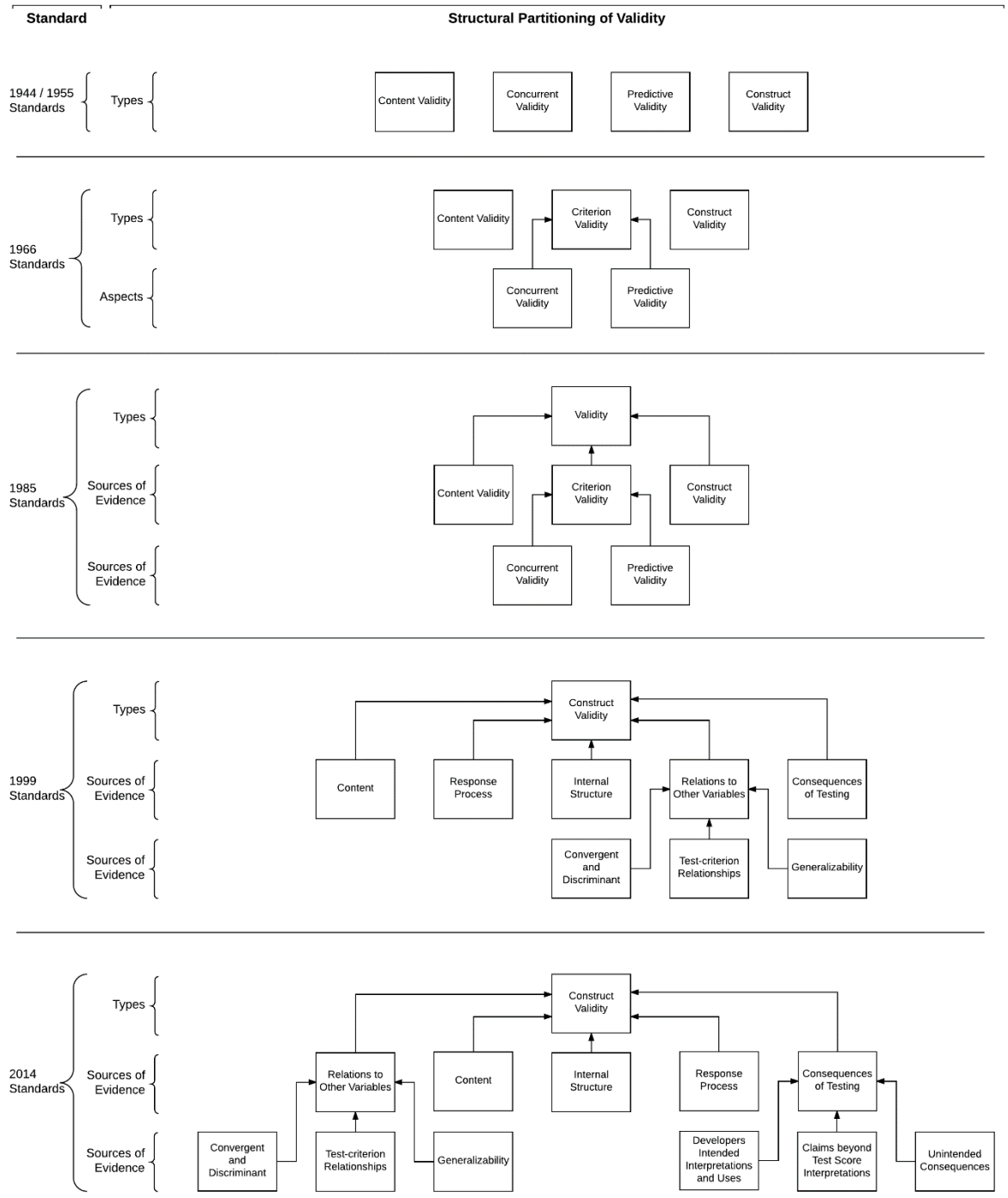


*Figure 2.1.* The evolution of the structural partitioning of validity according to the editions of the Standards. Reconstructed from a variety of primary (AERA et al., 1999, 2014) and secondary (e.g., Eignor, 2013; Kane, 2013a; Markus & Borsboom, 2013a; Messick, 1987; Newton & Shaw, 2013, 2014; Sireci, 2009; Sireci & Faulkner-Bond, 2016; Sireci & Sukin, 2013) sources. The figure illustrates how the partitioning of validity has not only altered with respect to *degree* (that is, towards increasing complexity in partitioning evinced by an increasing number of levels and categories), but also in *kind* by redefining the nature of the levels and categories (that is, by moving from types of validities to sources of evidence of one validity, and replacing categories).

debates leading up to the publication of the 2014 standards might be characterized as inquiries into the essences of the concepts of constructs, validity, and validation, where the position held for one inevitably bleeds into and mutually determines how one construes the others.

According to Newton and Shaw (2014), an account of the discourse of the period in between the 1999- and the 2014 standards can be structured along the lines of two reasonably distinct debates, which concerns each of the components of construct validity and validation. The first debate focuses on the nature of the concept of the construct, and whether it makes sense for construct validity and validation be considered all of validity and validation (e.g., Borsboom, Cramer, Kievit, Scholten, & Franić, 2009; Kane, 2012). The second debate can be construed as an extension of the first, and regards the scope of validity. As scope play a part in determining focus, the debate extends to the concern of what is to be considered relevant evidence for determining whether a particular gestalt is in possession of the property of validity – and thus how one should go about practicing validation. The strands of argument are here reviewed from the perspective of their implications for the concept of validity and the practice of validation, with a particular focus on the nature of constructs, and on the relatively recent emergence of a new orientation to validation: the argument based approach.

***The concept of (construct) validity.***

The construct concept is obviously central to the concepts of construct validity and validation, and the meaning of construct validity theory is fundamentally contingent on how the construct concept is construed and conceptualized. Stated somewhat more drastically, the very viability of construct validity theory as not only *the*-, but as *a* model for validity and validation hinges fundamentally on the viability of the construct concept itself. As such, opponents of construct validity theory have made several attempts at its life, mounting attacks from several fronts aimed squarely at the structural integrity of the construct concept.

A selection of examples of criticisms brought to bear on the construct concept include (but are not limited to) it having been conceived from the frameworks of philosophies of science which have since been dismissed as unviable (e.g., Borsboom et al., 2009; Borsboom et al., 2004; Michell, 2013), that the concept obscures important distinctions and illegitimately equates and conflates a number of substantially different concepts (e.g., Maraun & Gabriel, 2013; Slaney & Racine, 2013a), that it is not always necessary nor appropriate to invoke hypothetical constructs when interpreting and using test scores (e.g., Kane, 2012), and that insisting on invoking abstract multifaceted constructs have led to a lack of specificity with regards to what exactly it is that is being measured (e.g., Kagan, 2005; McGrath, 2005a).

A consequence of the contemporary discontent with the construct concept is that theorists are debating what should replace is. That is, if it isn't construct theories that should be considered valid or invalid (Haig, 2012), what is? Stances on this question can be ranked along a continuum ranging from less to more expansive. The unifying basis for the stances are scores generated by the applications of tests. What separates the stances are the range of particulars to which the property of validity can be assigned. The continuum stretches from the least (validity is a property of tests; e.g., Borsboom et al., 2004; Hood, 2012; Lissitz & Samuelsen, 2007), to the most (validity as a property of interpretations and uses of test scores; e.g., Bachman, 2005; Kane, 2013a) expansive.

Occupying the ranks between the end-points are positions advocating for validity to be considered a property of interpretations of test scores (disqualifying "uses" and "tests" as particulars that can be considered to possess the properties of validity or invalidity). This thesis can be considered party to this conception; it is not concerned with validating uses, and whether a test or its interpretation is what is valid is deemed inconsequential (i.e., as a matter of semantics). The common denominator appears to be that validity is at the least concerned with the adequacy of measurement interpretations[2] (what constitutes "adequacy" is, however, a matter of dispute; cf., Borsboom & Markus, 2013; Kane, 2013b). What seems to be the common to the otherwise divergent stances is a discontent with the "construct" concept.

***The practice of validation, and the argument based approach.***

As already mentioned, much of the discontent with validity theory as offered by the Standards is based on the sentiment that it is too broadly encompassing, and does not offer a sufficiently useful framework to guide validation practice (Collie & Zumbo, 2014; Lissitz & Samuelsen, 2007; Newton & Shaw, 2014). As such, much of the contributions to the debate has focused on narrowing the scope (e.g., Borsboom et al., 2004; Lissitz & Samuelsen, 2007) or clarifying the definition (e.g., Newton, 2012a) of validity, and in doing so pointing to particular sources of evidence (e.g., content-, response processes-, internal structure-, etc.) as foundational to the validity of whatever it is that is being validated (e.g., whether it is tests, interpretations, or uses that are being validated and can be considered valid).

[2] Disclaimer: This is a simplification. Adherents to the test-based approach to validation would not accept that it is interpretations that are being validated; tests are (though this appears to be an academic matter of semantics with no practical consequence. After all, does it matter whether one proclaims; "test X is a valid measure of latent variable Y", or "the interpretation that test X measures latent variable Y is valid"?). Conversely, those who advocate for use validation do not restrict validity to whether one achieves what one hopes to achieve by actions based on test scores, but whether the proposed actions are ethically justifiable; an issue that can be wholly independent of the validity of measurement interpretations.

While drawing attention to their preferred sources of validity evidence, scholars do not appear to offer much in the way of procedural descriptions for how to translate theory into practice. An exception is the work of Kane (2013a), who is considered the leading pioneer of the argument-based approach to validation. This approach has been developed over the last few decades as an alternative to the nomological networks approach initially advocated for by Cronbach and Meehl (1955), and is intended to offer a generalized framework for specific application-entailed interpretations of test scores.

The procedure of the approach as articulated by Kane (2013a) can be summed up in two points: (1) state the intended interpretations and uses of test scores (i.e., specifying an "Interpretation and Use Argument", or IUA for short), and (2) evaluate the plausibility of the IUA (evaluating a "Validity Argument", or VA for short). Developing the VA involves accumulating evidence for and against the IUA, and the specific evidence required for the VA depends on the claims contained in the IUA. "Claims" are inferences based on data (singular; "datum") that rely on "warrants" (e.g., statistical treatment of data), which are supported by "backing" (e.g., empirical and theoretical support for claims). When a claim attains sufficient backing to be considered justified, the claim can itself be treated as a datum (Kane, 2013a).

The argument-based approach has for the most part been favorably received, and is adopted by the Standards (AERA et al., 1999, 2014). However, besides terminology and the specification of sources of evidence, the Standards do not provide much in the way of guidance for how to proceed with argument-based validation. Kane (2013a) represents the most recent elaboration on the approach. There are criticisms levied against the articulation of approach however, and they mirror those echoed against contemporary canonical validity theory. Kane's articulation is intentionally broad, and accommodates the entire spectrum of perspectives on validity. Hence, it provokes the ire of those theorists who would deny non-measurement "interpretations" and "uses" the property of validity (e.g., Borsboom & Markus, 2013; Cizek, 2012).

**The 2014 Standards on validity and validation, a summary.**

The Standards specify 25 standards, partitioned into three clusters. The first cluster includes standards pertaining to establishing intended uses and interpretations, the second with issues regarding samples and settings used in validation, and the third with specific forms of validity evidence. (AERA et al., 2014). In addition to establishing a set of "standards," the Standards provide a theoretical framework of validity that include a comprehensive glossary of terms and definitions to guide validation practice (see table 2.1 for a selection of these).

Table 2.1.

*A Selection of Concepts from the Standards Deemed Relevant for the Current Research Effort.*

| Term. | Definition/description. |
|---|---|
| Consequences | The outcomes, intended and unintended, of using tests in particular ways in certain context and with certain populations. |
| Construct | The concept or characteristic that a test is designed to measure. |
| Construct domain | The set of interrelated attributes (e.g., behaviours, attitudes, values) that are included under a construct's label. |
| Construct-irrelevant variance | Variance in test-taker scores that is attributable to extraneous factors that distort the meaning of the scores and thereby decrease the validity of the proposed interpretation. |
| Construct underrepresentation | The extent to which a test fails to capture important aspects of the construct domain that the test is intended to measure, resulting in test scores that do not fully represent that construct. |
| Content-related evidence | Evidence based on test content that supports the intended interpretation of test scores for a given purpose. May address issue such as the fidelity of content to performance in the domain in question and the degree to which test content representatively samples a domain, such as a course curriculum or job. |
| Convergent evidence | Evidence based on the relationship between test scores and other measures of the same or related construct. |
| Differential item functioning (DIF) | For a particular item in a test, a statistical indicator of the extent to which different groups of test takers who are at the same ability level have different frequencies of correct responses or, in some cases, different rates of choosing various item options. |
| Discriminant evidence | Evidence indicating whether two tests interpreted as measures of different constructs are sufficiently independent (uncorrelated) that they do, in fact, measure two distinct constructs. |
| Effort | The extent to which a test taker appropriately participates in test taking. |
| Empirical evidence | Evidence based on some form of data, as opposed to that based on logic or theory. |
| Factor | Any variable, real or hypothetical, that is an aspect of a concept or construct. |
| Factor analysis | Any of several statistical methods of describing the interrelationships of a set of variables by statistically deriving new variables, called factors, that are fewer in number than the original set of variables. |
| Internal structure | In test analysis, the factorial structure of item responses or subscales of a test. |
| Item context effect | Influence of item position, other items administered, time limits, administration conditions, and so forth, on item difficulty and other statistical item characteristics. |
| Random error | A nonsystematic error; a component of test scores that appears to have no relationship to other variables. |
| Relevant subgroup | A subgroup of the population for which a test is intended that is identifiable in some way that is relevant to the interpretation of test scores for their intended purposes. |
| Reliability coefficient | A unit-free indicator that reflects the degree to which scores are free of random measurement error. |
| Reliability/precision | The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker. |
| Response bias | A test taker's tendency to respond in a particular way or style to items on a test that yields systematic, construct-irrelevant error in test scores. |
| Systematic error | An error that consistently increases or decreases the scores of all test takers or some subset of test takers, but is not related to the construct the test is intended to measure. |
| Validation | The process through which the validity of a proposed interpretation of test scores for their intended uses is investigated. |
| Validity | The degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. |
| Validity argument | An explicit justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores for their intended uses. |

Note: Reproduced with permission from the AERA. Adopted and adapted from the Standards' glossary (AERA et al., 2014, pp. 215-225).

The standards describe validation as a process that "*involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations*" (AERA et al., 2014, p. 11). As for the procedure, the standards states that validation "*logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation includes specifying the construct the test is intended to measure*" (AERA et al., 2014, p. 11). Validation should be guided by a conceptual framework, ideally indicating "*how the construct as represented is to be distinguished from other constructs and how it should relate to other variables*" (AERA et al., 2014, p. 11). The 2014 standards advocate for the argument-based approach to validation, stating that it "*can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use*" (AERA et al., 2014, p. 11).

Validation thus involves putting forth a validity argument; "*an explicit justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores for their intended uses.*" (AERA et al., 2014, p. 225). The kinds of evidence that are to be included in the validity argument depends on the specific proposed interpretations of tests for specific uses. An interpretation can be claimed to be valid when relevant evidence, accumulated through validation, is put forth that supports the interpretation. The standards specify two principal sources of invalidity. The first is *construct underrepresentation*, which refers to the degree to which a test fails to capture important aspects of the construct. The second is known as *construct-irrelevant variance*, and refers to the degree to which test scores are affected by processes that are extraneous to the intended purpose of the test. The specified sources of evidence investigated during validation are essentially concerned with these.

Like the 1999 standards, the 2014 standards does not partition validity into types and aspects. Rather, validity is conceived of as a unitary concept, established by investigating sources of (in)validity. The 2014 standards maintain the partitioning of sources of evidence introduced in the 1999 standards (i.e., those of content, response processes, internal structure, relationships with other variables, and consequences of testing), with some alteration in sub-partitioning (particularly concerning the "consequences" category of evidence). With the exception of the minor changes to the sub-partitioning of the sources of evidence, the chapter on validity remained largely unchanged in the 2014 edition. Ultimately, the Standards specify interpretations as the particulars that can possess the property of validity, and the process of validation involves their interrogation.

**Latent Variable Theory and Modelling**

The thesis now turns to accounting for the specific kind of measurement interpretations that are to be validated – latent variable measurement interpretations. According to Borsboom (2008), latent variable theory is a meta-theoretical framework for latent variable modelling. Latent variable theory is inextricably tied to entity realism (Borsboom, 2005), the central claim of which is that "*a good many theoretical entities really do exist*" (Hacking, 1983, p. 27). An example of an "entity" could be an employees opinion about- or experience with an aspect of the employees' own working environment, or mental states of employees such as burnout or work engagement.

Latent variables are considered "existential concepts," meaning that they have the status of postulated entities not (at least currently) directly observable (i.e., they are not necessarily unobservable in principle; Bollen, 2002; Markus & Borsboom, 2013b). By referring to objects with causal properties, a latent variable construct can be construed as "characteristic constructs" as opposed to "concept constructs" (cf., the Standards definition of "construct," table 2.1, p. 17)[3], by constituting a linguistic attempt to represent a non-linguistic constituent of natural reality (Maraun & Gabriel, 2013). This is the notion of a "construct" adopted in this thesis; latent variable construct interpretations claim the existence of entities that are the causes of observations, which in turn are considered manifestations of those latent variables that function as proxies by which latent variable entities can be indirectly assessed.

Several competing definitions (i.e., classification criteria) for latent variables exist, each of which determining whether or not a given entity should be considered latent or observed (Bollen, 2002). According to Borsboom (2008), a variable should be considered latent if there is any uncertainty at all associated with its measurement (i.e., observability is a matter of all-or-nothing). This contrasts with the position of Kane (2013a), who consider observability a matter of degree (i.e., variables can be more or less observable). Disagreement on this point is however of no consequence to the compatibility of the argument-based approach to validation and the validation of latent variable interpretations.

---

[3] This statement assumes that the Standards use of the term "characteristic" refers to that which Maraun and Gabriel (2013) denote as "constituents of natural reality", and the term "concept" to denote "constituent of language." The Standards treat these terms as "primitive," i.e., "*terms that are not defined and are assumed to be understood by the academic field*" (Wacker, 2004, p. 632). As such, the authors of Standards must have assumed that readers immediate tacit understanding of the terms are adequately precise (Polanyi, 1966/2009), as neither the definition nor concept of the construct is given any further clarification or treatment. In light of the debates leading up to the publication of the 2014 Standards, this seems odd, as the concept of the construct (or rather, the lack of a clear conceptualization of the "construct"), was clearly one of the main grievances with contemporary validity theory (e.g., Borsboom et al., 2009; Kane, 2012; McGrath, 2005b; Slaney & Racine, 2013a).

As latent variables are not directly observed, they are modelled reflectively (i.e., as causes of variation in their indicators). Latent variable models are intermediaries connecting observations with latent variable theories, and CFA is one tool for latent variable modelling. Such modelling is based on the "common factor model" (e.g., Spearman, 1961), stating that correlations between indicators should exhibit local independence; meaning that there is no residual covariation between indicators if the hypothesized common cause is held constant (Bollen, 2002; Borsboom, 2008; Borsboom, Mellenbergh, & van Heerden, 2003). Residual error covariance among indicators should be accidental, and by fixing residual error covariance to zero, one should effectively be filtering out random measurement error.

As latent variable interpretations of responses claim that observations are commonly caused by entities not directly observable, it subscribes to a causal theory of measurement asserting that "*an item measures a particular attribute only if differences on the attribute causes differences in the item scores*" (Markus & Borsboom, 2013b, p. 55). In the context of questionnaire-based research, the central claim is that there is something about respondents that cause them to check the boxes the way they do. When validating such an interpretation, the task is to examine whether what causes patterns of responses in tests is the targeted entities. "*Somewhere in the chain of events that occurs between item administration and item response, the measured attribute must play a causal role in determining what value the measurements outcomes will take*" (Borsboom et al., 2004, p. 1062).

According to Markus and Borsboom (2013b), from the perspective of a latent variable theory subscribing to a causal theory of measurement (CTM), the central undertaking when validating test interpretations involves (1) fixing the identity of the measured attribute, and (2) establishing a causal link between the attribute and the item responses. Causal interpretations are vulnerable to what is called the "reification fallacy" (Kline, 2016), which "*involves an inference from an observed regularity to the existence of some 'thing' that is the source of the regularity*" (Kane, 2013a, p. 19). To avoid falling victim to this fallacy, investigating causal evidence is primary, and all other types of evidence are secondary, relevant only to the extent that they are needed to establish such causal evidence.

**Validating latent variable Interpretations.**

A CFA can test whether a set of items on statistical grounds can be said to measure the same thing and can as such, according to Lissitz and Samuelsen (2007), be considered one way of formally approaching content validation. However, demonstrating that the variation in a set of items likely are manifested by the same entity does not constitute definitive evidence in favor

of the specific identity claim regarding which entity the modelled latent variable represents. Kline (2016, p. 300) calls this logical error the "naming fallacy," stating that "*just because a factor is named does not mean that the hypothetical construct is understood or even correctly labelled.*" Associated errors of reasoning are the "jingle fallacy", which occurs when two factors sharing the same name are taken to represent the same entity, and the "jangle fallacy", which occurs when two factors with different names are assumed to represent different entities. As such, two types of latent variable interpretations can be distinguished: (1) latent variable *measurement* interpretations (something is being measured), and (2) latent variable *identity* interpretations (what is being measured).

From the perspective of CTM however, content-related evidence is not necessarily relevant. Content evidence will primarily be useful initially in the developmental process of a test, and becomes increasingly less important as the process matures to the point of examining response-processes and internal structure categories of evidence. After all, that a test appears as if it *should* elicit the intended types of response processes by virtue of its content (weak evidence) is not nearly as important as that it appears as if it *does* elicit the intended types of response processes (strong evidence; Markus & Borsboom, 2013b). Once evidence for the claim that the test elicits appropriate response processes in the context in which it is employed appears to be established, examining its internal structure constitutes a test of the claim.

Investigating the matter of response processes is necessary to avoid committing the "begging-the-question fallacy," which occurs "*when some critical inference or assumption in an argument is simply taken for granted*" (Kane, 2013a, p. 18). Another applicable name for this phenomenon is "the psychologists fallacy" (James, 1890/2015; Markus & Borsboom, 2013a), which occurs when one simply assumes that the test constitutes equivalent stimuli to the test user and the test taker (e.g., settling for content-related evidence; Guion, 1977). To justifiably treat this assumption as a premise, the researcher must be reasonably confident that respondents comprehend the intended literal and pragmatic meanings of the questions that are presented to them (Schwarz, 1999). That is, when the targets of testing are attitudes and behaviors – and when these are to be assessed by means of self-reporting – respondents must understand the words and purpose of the questions in order for the test to systematically elicit the cognitive processes necessary to retrieve construct-relevant information (Embretson, 1998, 2007).

Demonstrating correspondence between the response-processes and internal-structure sources of evidence provide necessary and sufficient backing to justify at the very least a provisional identity claim, provisional as it might be refuted by evidence procured when

investigating relations with other variables. Ultimately, the process of validation boils down to examining evidence relevant to the claim that the sources of the observed variance are construct-relevant. The claim to valid measurement interpretations is supported if (1) sources of construct-relevant variance appears to be producing observations, and (2) sources of construct-irrelevant factors appears to not be producing observations.

Construct-relevant variance occurs when the outcome of the measurement of an entity is determined (partially or completely) by the targeted entity. Conversely, construct-irrelevant variance occurs when the measurement of an entity is contaminated (partially or completely) by entities other than the ones targeted. Such effects threatens the validity of the interpretation that one measures with the test that which one claims to measure with the test, and the causes of such disturbance can be categorized as arising from the test itself (test-related factors), or from the circumstances within which it is administered (context-related factors).

According to Podsakoff, MacKenzie, and Podsakoff (2012), method effects (cf., “Item Context Effects”, table 2.1, p. 17) are sources of construct-irrelevant variance not caused by the test per se (i.e., the respondent comprehends the literal and pragmatic meaning of the questions they respond to; Schwarz, 1999), but by factors of the context of test application. Method effects are at play in situations where ability-, motivational-, or task factors might cause biased responding even if the test itself should, if administered under appropriate circumstances, be of sufficient quality. As such, method effects refer to sources of construct-irrelevant variance caused by the test in the context of its application. Ability factors concern the “ableness” of respondents to provide relevant responses to items, and motivational factors their “willingness” (cf., “Effort”; table 2.1, p. 17). Task factors concerns whether the test conditions either does not facilitate or outright obstructs the engagement of targeted response processes (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Podsakoff et al., 2012).

Construct-irrelevant variance due to method effects can thus be summed up as being caused by the extent to which respondents are not willing nor able to respond in a manner intended to the test, as well as the tests capacity for producing the desired response processes in otherwise ideal circumstances (respondents being fully willing and able). The factors can be categorized in terms of how one would go about addressing them – that is, whether the test should be adapted to the circumstances or the circumstances adapted to the test, which does not exclude the possibility that both approaches are simultaneously appropriate in any given case. In other words, bringing about the desired response processes is a matter of test-context fit, which can be considered as a kind of tripartite fit between (1) the test, (2) the context within which the test is employed, and (3) the respondent to which the test is administered.

Test- and ability factors are both most naturally addressed by making adjustments to the test itself, and can be categorized as test-related factors (e.g., allowing foreign employees to respond to an adapted native-language version of the questionnaire). Motivation- and task-related factors are naturally addressed by altering the context within which the test is administered (e.g., freeing up time in the work schedule of the employee to respond to the questionnaire, ensuring anonymity), and can be categorized as context-related factors.

Thus, examining evidence related to the response-processes category of evidence should not be restricted to sterilized "laboratory" settings where the influence of "naturally occurring" construct-irrelevant ability-, motivational-, or task factors are absent or suppressed (i.e., examining the performance of the test under "ideal" circumstances). A complete study of evidence related to response-processes ought to include examining natural applications to ensure that contextual influences of the settings in which tests are employed can be examined (i.e., examining and assessing the adequacy of the test setting, or control for its detrimental effects).

Sources of construct-irrelevant variance are not mutually exclusive. To the extent that a factor is judged relevant, checks and balances ought to be included in the study design in order to determine the extent of their influence (table 2.2 provides descriptions of possible combinations of sources, building on the taxinomy of Podsakoff et al., 2012). For example, reading comprehension (an ability factor) is not a likely cause of construct-irrelevant variance when employing a survey in the academic sector, but accurate reporting of information that is deemed sensitive might be unduly influenced by absence of trust (a motivational factor).

Once the construct-irrelevant influence of a factor is identified, steps can be taken to alleviate its detrimental effect on measurement by means of statistical modelling ("statistical

Table 2.2.

*Potential Sources and Combinations of Construct Irrelevant Variance.*

| Test-Related Factors | | Context-Related Factors | | |
|---|---|---|---|---|
| Test | Ability | Task | Motivation | Description of combinations of factors producing construct-irrelevant variance. |
| 0 | 0 | 0 | 0 | Desired, ideal state. No sources of construct irrelevant variance. |
| 0 | 0 | 0 | 1 | Motivational factors (e.g., distrust) are contaminating measurement. |
| 0 | 0 | 1 | 0 | The circumstances (e.g., time pressure) are contaminating measurement. |
| 0 | 0 | 1 | 1 | Both circumstances and motivational factors are contaminating measurement. |
| 0 | 1 | 0 | 0 | Ability factors (e.g., language barriers) are contaminating measurement. |
| 0 | 1 | 0 | 1 | Respondents neither willing nor able to provide relevant responses. |
| 0 | 1 | 1 | 0 | Both circumstances and ability factors are contaminating measurement. |
| 0 | 1 | 1 | 1 | Circumstances, motivational-, and ability factors contaminating measurement. |
| 1 | 0 | 0 | 0 | Testing circumstances are prime, but the test itself is not fit for purpose. |
| 1 | 0 | 0 | 1 | Both the test and motivational factors are contaminating measurement. |
| 1 | 0 | 1 | 0 | Both the test and the testing circumstances are contaminating measurement. |
| 1 | 0 | 1 | 1 | Test, circumstances, and motivational factors are contaminating measurement. |
| 1 | 1 | 0 | 0 | Both the test and ability factors are at play in contaminating measurement. |
| 1 | 1 | 0 | 1 | Test, motivational-, and ability factors are contaminating measurement. |
| 1 | 1 | 1 | 0 | Test, circumstances, and ability factors are contaminating measurement. |
| 1 | 1 | 1 | 1 | Least desired state. Every potential source of construct irrelevant variance present. |

Note: 0 denotes the factor is not at play. 1 denotes that the factor is at play.

remedies"), as long the effect merely distort the test score. If the target latent variable plays no part in determining the outcome, the problem needs to be addressed by targeting the test or the context ("procedural remedies"; Podsakoff et al., 2012).

**Synthesizing Validity Theory and Latent Variable Theory: The LVIV Model**

Figure 2.2 (see appendix 1 for a scaled up version of the model with a legend) illustrates a working model of how one could go about examining the effects of the test and the context in producing construct-irrelevant variance, termed "the LVIV model" (acronym for Latent Variable Interpretation Validation). It makes use of the Standards articulation of validity theory (i.e., its terminology in terms of source categories of validity evidence), and applies it to latent variable theory.

In the spirit of Kane (2013a) – who considers validity a matter of justification for interpretative claims – the model is based on a principle of recursive confidence building by means of step-wise testing of provisional measurement claims. In the terminology of Kane (2013a), it offers a procedural framework for turning a latent variable "claim" into a "datum" for higher-order investigations (which in turn can feed back to influence the confidence of those claims). As latent variable identity interpretations – the claims regarding what is being

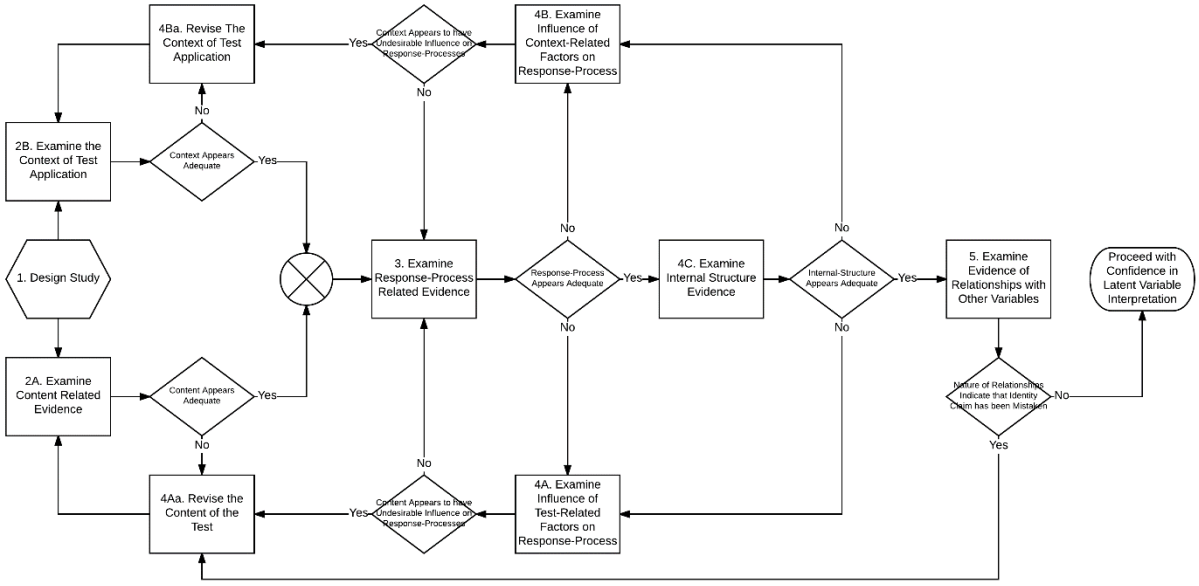


*Figure 2.2.* The LVIV Model for validating latent variable interpretations, a synthesis of validity theory and latent variable theory of a procedure for turning a "claim" into a "datum" (Kane, 2013a). The core of the process, the point to which the process returns should evidence sow doubt on the validity of measurement interpretations, is the accumulation and scrutiny of evidence pertaining to the response processes category of the Standards (i.e., examining sources of construct-relevant and irrelevant variance). The model is primarily inspired by the terminology and procedural descriptions of the Standards (AERA et al., 2014), Cronbach and Meehl (1955), Kane (2013a), Markus and Borsboom (2013b), and Podsakoff et al. (2012).

measured – are justified by demonstrating a causal link between the latent variable and the responses offered by respondents, the response process category of evidence is made the core of the model (step 3). For latent variable measurement interpretations – the claim that "something" is being measured – evidence regarding internal structure and relationships with other variables are generally considered adequate (steps 4C and 5 of the model).

The model recognizes that the test and the context within which it is administered are the proximal eliciting causes of response processes. As such, the model points to the test and the context as the primary targets of manipulation in order to bring about the kind of response processes that are taken as evidence in favor of the claim that the targeted latent variable is responsible for respondents' responses to the test (steps 4Aa and 4Ba). When validating latent variable identity claims, evidence suggesting an absence of construct relevant response processes leads to renewed interrogation of the test and its context of application, in order to examine what it is that is contributing to elicitation of construct irrelevant response processes.

If elicited response processes appear satisfactory (i.e., observed processes involved in the production of observations can be interpreted as arising from the targeted latent variable), investigations of internal structure (4C) provide evidence of generalizability; that it appears that the observed variation in the entity-associated items consistently are caused by a single latent variable. If the identity of the latent variable is not in question (i.e., one is exclusively interested in whether one measures *a* latent variable, not *the* latent variable), the investigation starts at the level of internal structure (i.e., investigating if or which items appear caused by a common and distinct latent variable, with no regard to what that latent variable represents).

Once an apparently satisfying internal structure is arrived at, investigating how factors relate to each other allows one to investigate theoretical claims regarding the causal properties of the latent variables (i.e., their antecedents and consequences). If theoretical expectations are not met, it might indicate one of two things. First, the structural theory regarding causal properties of the latent variable(s) are wrong, and that the latent variables construct theories ought to be re-examined and revised. Second, it can show that the tests used for measuring the target latent variables failed to represent adequately one or more of the allegedly assessed latent variables, which exposes the latent variable identity claims as invalid.

Once evidence is gathered that supports the proposed interpretation at every step of the model, a specific latent variable identity interpretation can be considered valid. For latent variable identity interpretations to be valid, presence of supporting favorable- and absence of contradicting unfavorable evidence for response processes, internal structure, and relations to other variables is required. In the absence of evidence in favor of specific response processes

and precise testable claims regarding how the target latent variable is supposed to relate to other variables, the most that can be accomplished in a validation study is establishing the validity or invalidity of latent variable measurement interpretations.

**Establishing Formal Definitions in Preparation for the Validity Argument**

As stated and elaborated in the previous section, the purpose of this thesis is to evaluate the adequacy of the KIWEST measurement theory as an account of the observations generated by administering the KIWEST survey. This is done by means of interrogating the justification for its hypothesized internal structure. In doing so, this thesis is primarily concerned with step 4C in the LVIV model (figure 2.2, p. 23; Appendix 1). The investigation extends briefly into cursory treatment of step 5 (relations to other variables) by examining evidence relating to convergence and discrimination, which conventionally constitute sources of evidence directly relevant as evidence pertaining to internal structure (see table 2.1, p. 17).

In the terminology of Kane (2013a), the claims made on the basis of investigating content- and process-related evidence should be sufficiently strong to justify treating the "claims" as "datums" for investigating evidence of internal structure. If one is reasonably confident that the items elicit construct-relevant response processes, a favorable outcome of this investigation will constitute further evidence supporting KIWEST latent variable identity claims. For this reason, the investigation of evidence of internal structure category serves as a test that has the potential to confirm or disconfirm the adequacy of provisional identity claims based on, principally, response-process related evidence.

To substantiate the choice of hypotheses it is prudent to account for the necessary and sufficient conditions for latent variable interpretations to be claimed "valid." While validation does not reduce to a formal calculus, formalization can according to Markus and Borsboom (2013a) further the practice by route of clarification. Thus, for latent variable measurement interpretations, validity can be formally expressed a function of the extent to which evidence support two competing criteria: comprehensiveness and parsimoniousness, which might be expressed in terms of predicate logic (Tomassi, 1999) as:

$$\forall_a[F_a \leftrightarrow (G_a \,\&\, H_a)] \qquad \text{(Formula 1)}$$

That is, for every particular *a* (latent variable measurement interpretation), *a* possesses the property *F* (validity) if and only if *a* possesses the property *G* (comprehensiveness) and the property *H* (parsimoniousness). What is established by these criteria is that a set of items appear to measure "something" (convergence), and that "something" is distinct from what is

measured with other item sets (discrimination). In turn, a latent variable identity interpretation is a more ambitious claim than a latent variable measurement interpretation, and requires as such additional evidence. A minimal argument for latent variable identity interpretations can be formally expressed as follows:

$$\forall_b[F_b \leftrightarrow (F_a \,\&\, F_c)] \qquad \text{(Formula 2)}$$

That is, for every particular *b* (latent variable identity interpretation), *b* possess property *F* (validity) if and only if the particular *a* (the latent variable measurement interpretation) possesses the property *F* (validity) and the particular *c* (response process interpretations) possesses the property *F* (validity). As such, criteria for the validity of response-process interpretations are required.

Based on validity theory, the adequacy of response-process interpretations depends on the extent to which it is determined that sources of construct-relevant variance are producing observed scores, and that sources of construct-irrelevant variance are not producing observed scores. This can be formally expressed as:

$$\forall_c[F_c \leftrightarrow (I_{de} \,\&\, \sim I_{df})] \qquad \text{(Formula 3)}$$

That is, for every particular *c* (response process interpretation), particular *c* possesses property *F* (validity) if and only if particular *d* (observed variance) stand in relation *I* (is determined by) to particular *e* (construct-relevant variance) and not to particular *f* (irrelevant variance).

These formal definitions provide the basis for evaluating the validity of measurement interpretations for stand-alone constructs within KIWEST. Formula 1 provides a framework for establishing justification for (and as such the validity of-) latent variable measurement interpretations (weak claims), while formula 2 provides a framework for establishing the justification (i.e., validity) for specific latent variable identity interpretations (strong claims for which formula 1 and formula 3 provide critical components). As such, two sets of criteria can be established for the KIWEST latent variable interpretations. These are:

1. The KIWEST latent variable measurement interpretation is valid if the criteria of formula 1 holds true for every factor (a weak claim that requires every factor to satisfy the criteria of convergence and discrimination).
2. The KIWEST latent variable identity interpretation is valid if the criteria for formula 2 holds true for all factors (a strong claim, which requires that the criteria of formula 1 and formula 3 holds true for every factor).

The basic claims made in this thesis prior to investigations are that the criteria for point number one hold true (i.e., that the claims of latent variable measurement interpretations are justified). The purpose of the investigation of thesis is to provide backing for these latent variable measurement claims by interrogating them and evaluating the extent to which (and whether) they stand to scrutiny.

**Hypotheses.**

Having established criteria- and formal definitions for valid latent variable measurement- and identity interpretations, hypotheses can be proposed that constitute evidence for or against the latent variable identity- and measurement interpretations proposed or implied by the KIWEST measurement theory. This thesis is not primarily concerned with specific latent variable interpretations, but the "omnibus," composite latent variable interpretation provided by the KIWEST theory.

This thesis is primarily concerned with evidence pertaining to the Standards "internal structure" category of evidence, thus examining the degree of justification for latent variable measurement claims, and not for latent variable identity claims. As argued previously, valid latent variable measurement interpretations are necessary (but by themselves insufficient) for valid latent variable identity interpretations. Having established what constitutes a valid measurement interpretation, the hypotheses for the KIWEST latent variable measurement interpretation as a whole (i.e., its internal structure) can be formulated as follows:

- H1: The KIWEST measurement theory adequately accounts for the observations generated by administering the KIWEST questionnaire on its target population.
    - H1a: The KIWEST measurement theory comprehensively accounts for the observations generated by administering the KIWEST questionnaire.
    - H1b: The KIWEST measurement theory parsimoniously accounts for the observations generated by administering the KIWEST questionnaire.

Furthermore, the KIWEST latent variable identity claims are subjected to cursory (though, incomplete) treatment, by examining evidence relating to the Standards' categories of "relations to other variables." The treatment of the "relations to other variables" category will be restricted to the "convergent and discriminant" subcategory of evidence. The claim of the KIWEST measurement theory is that every hypothesized latent variable should exhibit statistical convergence and discrimination. As such, a set of hypotheses can be formulated as:

- H2: All of the items included in the KIWEST questionnaire statistically converge on the factors to which they are assigned by the KIWEST measurement theory.
- H3: All of the factors specified in the KIWEST measurement model represents a unique latent variable, and as such should exhibit statistical discrimination with all other factors included in the KIWEST model.

It should be cautioned that the cursory treatment of evidence relevant to latent variable identity claims provided in this thesis are not sufficient to establish their validity. However, the evidence can feature as part of further investigations into the validity of specific identity claims. In order to support latent variable identity claims, fine-grained investigation into any conceivably context-relevant source of construct-relevant and irrelevant variance ought to be undertaken. Beyond circumstantial evidence (e.g., deriving meaning from relations with other variables; Borsboom et al., 2004; Cronbach & Meehl, 1955), there does not seem to exist enough readily available information to justifiably make such a determination.

That being said, the KIWEST measurement interpretation will be proclaimed "valid" to the extent to which the hypotheses are supported by the current investigation, and "invalid" to the extent to which they are not supported. According to contemporary canonical validity theory however (e.g., AERA et al., 2014; Kane, 2013a), validity is a matter of degree, and one can be more or less justified in interpreting observations in certain ways. In the words of Kane (2013a, p. 3): "*Interpretations and uses that make sense and are supported by appropriate evidence are considered to have high validity (or for short, to be valid), and interpretations or uses that are not adequately supported, or worse, are contradicted by the available evidence are taken to have low validity (or for short, to be invalid).*"

## Method

As validity bears on the justification of interpretations made based on test scores, the following sections will go to great lengths to account for- and justify the procedural and methodological decisions that have been made in this study. In order to secure the validity of parameter estimate interpretations following the confirmatory factor analysis, a great deal of effort is dedicated to data integrity analysis and treatment, which consists of missing value- and multivariate normality analysis and treatment.

As this thesis is part of an ongoing study, only the operations performed in this thesis are accounted for and presented in detail. The operations performed prior to this thesis (those decisions over which the current author has had no say in or control over, i.e., steps 1, 2A and 2B depicted in the LVIV model; figure 2.2, p. 24; appendix 1) are accounted for only in brief. Readers interested in the details concerning the rationale and procedure for the design and execution of ARK and KIWEST are referred to Undebakke et al. (2014). A stepwise depiction of the operations performed prior to- and in the current study is presented in figure 3.1.

### The Scales of the KIWEST Questionnaire

The presentation of the scales of KIWEST (Undebakke et al., 2014) will be structured in accordance with the foundations of this research. Being that theory play a fundamental role in

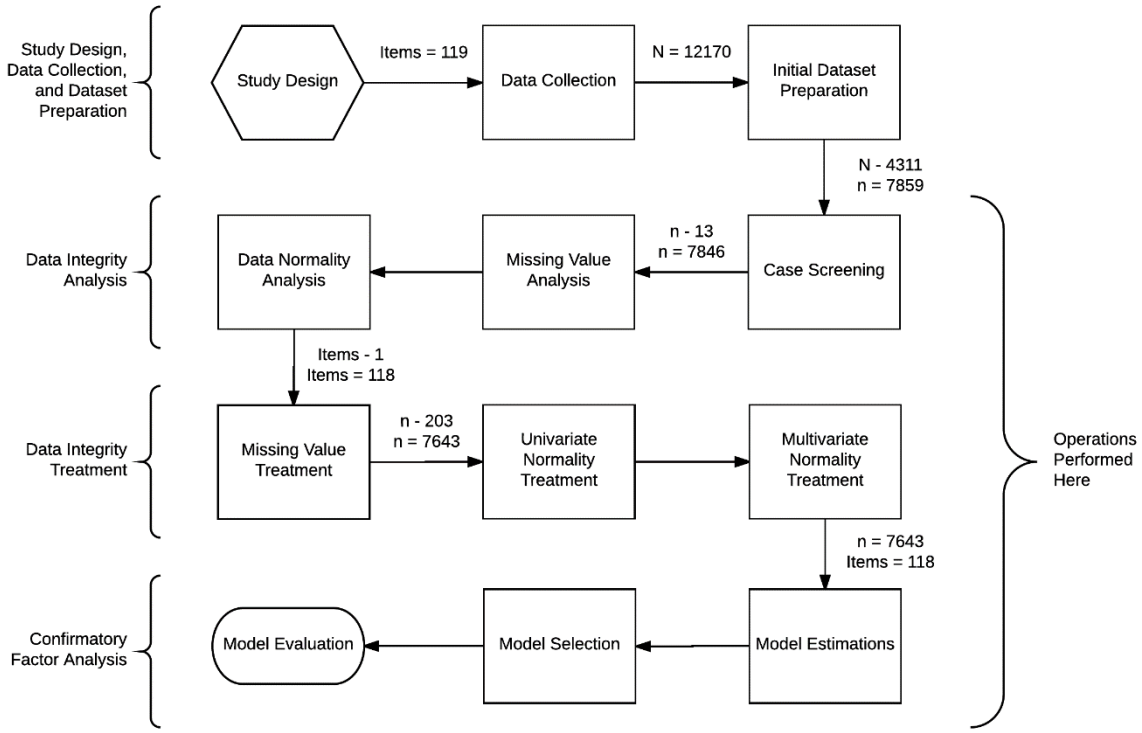


*Figure 3.1.* Chart depicting the methodological procedure of this thesis, including junctions of case and item exclusions.

both validity and CFA, efforts have been made to trace and explicate the theoretical origin of each construct. As measurement of constructs is based on the operationalization of concepts, and as CFA is ultimately built on these operationalizations (variation in observed variables are hypothesized to be commonly caused by specified latent factors; Bollen, 2002; Edwards, 2010; Edwards & Bagozzi, 2000; Markus & Borsboom, 2013b), the items operationalizing each construct (i.e., the operational definitions) are presented in table 3.1 (pp. 33-35).

Operationalizations should be based on good formal conceptual definitions to further theory development (Wacker, 1998, 2004). As such, efforts have here been made to trace the theoretical origins and the conceptual definitions of each construct included in KIWEST in order to comprehensively and yet concisely account for them, so as to provide for the basis to evaluate them as well as their corresponding operationalizations. This could in turn be useful for helping readers evaluate the adequacy of the items in terms of representing the constructs, as defined by their corresponding conceptual definitions, as well as the adequacy of the conceptual definitions themselves (e.g., to evaluate content-related evidence of validity).

In this thesis, it is assumed that the quality of the necessary conceptual groundwork as pertaining to constructs and items has been performed adequately prior to this study, and the constructs conceptual definitions will for this reason not be evaluated here. These should however be revisited should the CFA fail to demonstrate the convergence of their items or the discrimination of theoretically distinct factors. In tracing the theoretical, conceptual and operational foundations of the constructs included in KIWEST questionnaire however, it was made apparent that the constructs included in KIWEST vary greatly with respect to their theoretical elaboration. This is reflected in the following segment by the amount of space allotted to presenting each individual construct (i.e., the relative amount of space allotted to each construct is not based on any sort of “construct-favoritism” on part of the author).

***Cohesion in Work Teams.***

The Cohesion in Work Teams scale is modified from Carless and De Paola (2000) by Christensen et al. (2012). The stated interpretation of a high score is that “*the respondents experience good teamwork between colleagues at their own unit*” (Undebakke et al., 2014, p. 9). Operationalizing the construct are three statements on a 5-point likert scale, with responses ranging from “disagree” to “agree”.

Table 3.1.

*Construct Operationalizations.*

| Scale | Item | Operationalizing Statement | Scale type |
|---|---|---|---|
| Cohesion in Work Teams | v01_007 | This unit gives me ample opportunities to improve my personal performance | 5-point likert |
| | v02_003 | In our unit, we stand together in trying to reach our performance goals | |
| | v02_005 | I'm happy with my unit's level of task commitment | |
| Commitment to the Workplace | v09_001 | I am happy to tell others about my workplace | 5-point likert |
| | v09_008 | I would recommend a close friend to apply for a position at my workplace | |
| | v09_014 | I feel that my workplace is of great importance to me | |
| Competency Demands | v03_003 | I am expected to continually develop my competence | 5-point likert |
| | v03_009 | The nature of my work means I continually have to develop and think in new ways | |
| | v03_012 | I feel pressure to continually learn new things in order to manage my work tasks (reverse scored) | |
| DUWAS (Excessiveness) | v10_001 | I seem to be in a hurry and racing against the clock (reverse scored) | 4-point frequency |
| | v10_002 | I find myself continuing to work after my co-workers have called it quits (reverse scored) | |
| | v10_004 | I stay busy and keep many irons in the fire (reverse scored) | |
| | v10_006 | I spend more time working than on socializing with friends, on hobbies, or on leisure activities (reverse scored) | |
| | v10_008 | I find myself doing two or three things at one time, such as eating lunch and writing a memo, while talking on the telephone (reverse scored) | |
| DUWAS (Compulsiveness) | v10_003 | It is important to me to work hard even when I do not enjoy what I am doing (reverse scored) | 4-point frequency |
| | v10_005 | I feel that there's something inside me that drives me to work hard (reverse scored) | |
| | v10_007 | I feel obliged to work hard, even when it is not enjoyable (reverse scored) | |
| | v10_009 | I feel guilty when I take time off work (reverse scored) | |
| | v10_010 | It is hard for me to relax when I'm not working (reverse scored) | |
| Dysfunctional Support | v01_002 | People in my unit sometimes help me in a difficult situation, but do not support in a way that is matter-of-factly (reverse scored) | 5-point likert |
| | v01_005 | People in my unit sometimes help me in a difficult situation, but indicate that I should have dealt with the problem myself (reverse scored) | |
| | v01_006 | People in my unit sometimes help me in a difficult situation, but expect infinite thankfulness (reverse scored) | |
| | v01_011 | People in my unit sometimes help me in a difficult situation, but support me reluctantly (reverse scored) | |
| | v01_013 | People in my unit sometimes help me in a difficult situation, but do so with a reproachful tone or gaze (reverse scored) | |
| | v01_015 | People in my unit sometimes help me in a difficult situation, but combines this with reproaches (reverse scored) | |
| Empowering Leadership | v07_001 | My immediate superior contributes to the development of my skills | 5-point likert |
| | v07_002 | My immediate superior encourages me to participate in important decisions | |
| | v07_003 | My immediate superior encourages me to speak up when I have a different opinion | |
| Fairness of the Supervisor | v07_004 | My immediate superior treats the employees fairly | 5-point likert |
| | v07_007 | My immediate superior distributes work assignments fairly | |
| | v07_008 | My immediate superior treats the employees impartially | |
| Goal Clarity | v04_001 | What is expected of me at work is clearly expressed | 5-point likert |
| | v04_004 | I feel that the objectives of my job are diffuse and unclear (reverse scored) | |
| | v04_009 | I have a clear understanding of which tasks constitute my job | |
| Illegitimate Tasks | v04_003 | I must carry out work which I think should be done by someone else (reverse scored) | 5-point likert |
| | v04_008 | I must carry out work that put me into awkward positions (reverse scored) | |
| | v04_011 | I must carry out work that I think it is unfair that I should do (reverse scored) | |
| | v04_014 | I must carry out work which I feel demands more of me than is reasonable (reverse scored) | |
| Inclusiveness and Social Responsibility | v02_001 | Men and women are treated as equals in my unit | 5-point likert |
| | v02_002 | In my unit, there is room for employees of a different ethnic background or religion | |
| | v02_004 | In my unit, there is room for older employees | |
| | v02_006 | In my unit, there is room for employees with various illnesses or disabilities | |

Table 3.1. (continued).

*Construct Operationalizations.*

| Scale | Item | Operationalizing Statement | Scale type |
|---|---|---|---|
| Innovation | v04_002 | My unit is constantly evolving to meet the employees' needs | 5-point likert |
| | v04_006 | My unit is open-minded and adapts to changes | |
| | v04_010 | In my unit, no one listens to new suggestions and ideas (reverse scored) | |
| | v04_012 | My unit is flexible and constantly adapts to new ideas | |
| | v04_015 | My unit strives to retain status quo rather than to change (reverse scored) | |
| Interpersonal Conflicts | v01_003 | My work is hampered by power struggles and territorial thinking in my unit (reverse scored) | 5-point likert |
| | v01_014 | In my unit, intrigues impair the work climate (reverse scored) | |
| | v01_017 | In my unit, there is a great deal of tension due to prestige and conflicts (reverse scored) | |
| Job Autonomy | v04_005 | I have a sufficient degree of influence in my work | 5-point likert |
| | v04_007 | I can make my own decisions on how to organize my work | |
| | v04_013 | There is room for me to take my own initiatives at work | |
| | v04_016 | I manage my work situation in the direction I want | |
| Meaning of Work | v09_003 | I feel motivated and involved in my work | 5-point likert |
| | v09_006 | My work is meaningful | |
| | v09_016 | I feel that the work I do is important | |
| Recognition | v06_005 | I am treated fairly by my unit management | 5-point likert |
| | v06_007 | My work is recognized and appreciated by my unit management | |
| | v06_010 | I am respected by my unit management | |
| Reliability of Management - Next Administrative Level | v08_001 | I can expect the management of the next administrative level to treat me in a consistent and predictable way | 5-point likert |
| | v08_002 | The management of the next administrative level is always reliable | |
| | v08_003 | The management of the next administrative level is open and honest with me | |
| | v08_004 | I am confident that I can trust the management of the next administrative level | |
| | v08_005 | I have complete confidence in the management of the next administrative level | |
| Reliability of Management - Own Unit | v06_001 | My unit management is always reliable | 5-point likert |
| | v06_003 | I can expect my unit management to treat me in a consistent and predictable way | |
| | v06_006 | My unit management is open and honest with me | |
| | v06_009 | I have complete confidence in my unit management | |
| | v06_012 | I am confident that I can trust my unit management | |
| Role Conflict | v03_002 | I am often given assignments without adequate resources to complete them (reverse scored) | 5-point likert |
| | v03_006 | I frequently receive incompatible requests from two or more people (reverse scored) | |
| | v03_007 | My job involves tasks that are in conflict with my personal values (reverse scored) | |
| | v03_011 | I have to do things that I feel should be done differently (reverse scored) | |
| Role Overload | v03_004 | It happens quite often that I have to work under heavy time pressure (reverse scored) | 5-point likert |
| | v03_010 | I frequently have too much to do at work (reverse scored) | |
| | v03_013 | I am given enough time to do what is expected of me in my job | |
| Social Climate | v01_008 | The climate in my unit is distrustful and suspicious (reverse scored) | 5-point likert |
| | v01_016 | The climate in my unit is encouraging and supportive | |
| | v01_018 | The climate in my unit is relaxed and comfortable | |

Table 3.1. (continued).

*Construct Operationalizations.*

| Scale | Item | Operationalizing Statement | Scale type |
|---|---|---|---|
| Social Community at Work | v01_001 | I feel that I am part of a community at my unit | 5-point likert |
| | v01_009 | There is a good atmosphere between me and my colleagues | |
| | v01_010 | There is a good sense of fellowship between the colleagues at my unit | |
| Social Support from Supervisors | v07_005 | My immediate superior talks with me about how well I carry out my work | 5-point likert |
| | v07_006 | My immediate superior listens to me when I have problems at work | |
| | v07_009 | My immediate superior gives me the help and support I need from her/him | |
| Task Completion Ambiguity | v03_001 | I know when a task is completed | 5-point likert |
| | v03_005 | I determine when my work assignments are completed | |
| | v03_008 | It is up to me to assess when I have completed a work assignment | |
| Trust Regarding Management | v06_002 | I can trust information from my unit management | 5-point likert |
| | v06_004 | My unit management withholds important information from the employees (reverse scored) | |
| | v06_008 | It is possible for the employees at my unit to express their views | |
| | v06_011 | My unit management trusts the employees to do their work well | |
| UWES (Vigor) | v11_001 | At my work, I feel bursting with energy | 7-point frequency |
| | v11_002 | At my job, I feel strong and vigorous | |
| | v11_003 | When I get up in the morning, I feel like going to work | |
| UWES (Dedication) | v11_004 | I am enthusiastic about my job | 7-point frequency |
| | v11_005 | My job inspires me | |
| | v11_006 | I am proud on the work that I do | |
| UWES (Absorption) | v11_007 | I feel happy when I am working intensely | 7-point frequency |
| | v11_008 | I am immersed in my work | |
| | v11_009 | I get carried away when I'm working | |
| Work-Family Conflict | v09_002 | Job worries or problems distract me when I am at home (reverse scored) | 5-point likert |
| | v09_007 | My job reduces the effort I can give to activities at home (reverse scored) | |
| | v09_009 | Stress at work makes me irritable at home (reverse scored) | |
| | v09_015 | My job makes me feel too tired to do the things that need attention at home (reverse scored) | |
| Work-Family Facilitation | v09_004 | The things I do at work help me deal with personal and practical issues at home | 5-point likert |
| | v09_005 | The things I do at work make me a more interesting person at home | |
| | v09_011 | The skills I use at work are useful for things I have to do at home | |
| | v09_013 | Having a good day at work makes me a better companion when I get home | |
| Work-SOC (Comprehensibility) | v12_001 | Manageable - Unmanageable (How do you feel about your present job and workplace in general?) | 7-point semantic differential |
| | v12_003 | Structured - Unstructured | |
| | v12_006 | Clear - Unclear | |
| | v12_009 | Predictable - Unpredictable | |
| Work-SOC (Manageability) | v12_004 | Easy to influence - Impossible to influence | 7-point semantic differential |
| | v12_007 | Controllable - Uncontrollable | |
| Work-SOC (Meaningfulness) | v12_002 | Meaningless - Meaningful | 7-point semantic differential |
| | v12_005 | Insignificant - Significant | |
| | v12_008 | Unrewarding - Rewarding | |

***Commitment to the Workplace.***

The Commitment to the Workplace scale, also referred to as Organizational Commitment, was adopted from Christensen et al. (2012), where it was adopted and adapted from Pejtersen et al. (2010) from where in turn it had been further developed from COPSOQ I (Kristensen, Hannerz, Høgh, & Borg, 2005). None of the reports seem to make clear the exact theoretical origin of the construct being operationalized. The stated preferred interpretation of a high score in KIWEST is that it indicates that "*the respondent experience having positive ties to their place of work*" (Undebakke et al., 2014, p. 12). Three statements on a 5-point likert scale operationalize the construct, with responses ranging from "disagree" to "agree".

***Competency Demands.***

The Competency Demands scale is intended to capture the sense that work tasks demand that one learn new knowledge, and that the nature of work requires continuous training. It has been retrieved from Näswall et al. (2010), where it was originally adopted from van der Vliet and Hellgren (2002). The stated interpretations of a high score is that "*employees have the sense that their work tasks demand learning of new knowledge, and that the nature of work requires continuous training*" (Undebakke et al., 2014, p. 10). Operationalizing the concept are three statements on a 5-point likert scale, responses ranging from "disagree" to "agree".

***Dutch Workaholism Scale (DUWAS).***

The Dutch Workaholism Scale is adopted from Schaufeli, Shimazu, and Taris (2009a), which is a measure based on a proposed two-factor structure of work addiction; excessiveness and compulsiveness. (Schaufeli, Shimazu, & Taris, 2009b, p. 322) define workaholism as "*the tendency to work excessively hard (the behavioral dimension) and being obsessed with work (the cognitive dimension), which manifests itself in working compulsively.*" Undebakke et al. (2014, p. 12) state the interpretation of a high score as indicating "*little addiction to the work.*" Two times five statements on a 4-point scale operationalize the construct with responses ranging from "(almost) never" to "(almost) always".

***Dysfunctional Support.***

The stated preferred interpretation of a high score on the Dysfunctional Support scale, the source of which being cited in Undebakke et al. (2014) as Semmer, Amstad, and Elfering (2006, a paper presented at a conference), is that "*the respondents experience a low degree of dysfunctional support*" (Undebakke et al., 2014, p. 11). Operationalizing the construct are six statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

### *Empowering Leadership.*

The Empowering Leadership scale is adopted from Dallner et al. (2000), which does not appear to make clear from where or what the scale is based on. They cite Thomas and Velthouse (1990) in defining psychological empowerment as "*intrinsic motivation manifested in four cognitions reflecting an individual's orientation to his or her work role: meaning, competence, self-determination, and impact*" (Dallner et al., 2000, p. 34). Undebakke et al. (2014, p. 9), however, define empowerment as "*assigning or transferring power to another person, and to enabling someone to do something*", citing Stang (2003). A high score is taken to indicate that "*employees perceive management to be empowering*" (Undebakke et al., 2014, p. 9). The concept is operationalized with three statements on a 5-point likert scale ranging from "disagree" to "agree".

### *Fairness of the Supervisor.*

The Fairness of the Supervisor scale is retrieved from Dallner et al. (2000), where it is linked to the theory of organizational justice generally, and procedural justice specifically. Dallner et al. (2000, p. 35) state that the "*perceived fairness of the decision-making process is a key factor in procedural justice*", citing Tyler (1989). In KIWEST, the preferred interpretation of a high score is stated as "*the respondent experience that management is fair*" (Undebakke et al., 2014, p. 10). The concept is operationalized by means of three statements on a 5-point likert scale, with responses ranging from "disagree" to "agree".

### *Goal Clarity.*

The Goal Clarity scale is adopted and adapted from Näswall et al. (2010), where it was developed by adapting items from scales reflecting goal ambiguity which according to them were originally developed by Rizzo, House, and Lirtzman (1970) and Caplan (1971). The original goal clarity measure from Näswall et al. (2010) contained four items intended to measure "*the extent to which the purpose of one's work tasks is clear*" (Näswall et al., 2010, pp. 8-9). The stated interpretation of a high test score is "*the respondent has a clear picture of the purpose of his or her own work*" (Undebakke et al., 2014, p. 10). Operationalizing the concept are three statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

### *Illegitimate Tasks.*

The Illegitimate Tasks scale is adopted from Semmer, Tschan, Meier, Facchin, and Jacobshagen (2010), where it is stated as being specifically tied to role theory generally and

role expectations specifically (citing Ilgen & Hollenbeck, 1991; Katz & Kahn, 1978; Sheldon & Burke, 2000), as well as identity theory (citing Thoits, 1991) and the concept of "feeling offended." Tasks are legitimate "*to the extent that they conform to norms about what can reasonably be expected from a given person*", and they are illegitimate "*to the extent that they violate such norms*" (Semmer et al., 2010, p. 72). Illegitimate tasks are conceived of as offending one's professional identity, and thus, the self. The construct is additionally tied to the concept of counterproductive work behavior, defined as "*behavior intended to hurt the organization or other members of the organization*" (Semmer et al., 2010, p. 71, citing; Spector & Fox, 2002). The stated preferred interpretation of a high score is that "*the respondents experience that they have a low degree of illegitimate work tasks, in other words tasks that are perceived as being outside one's area of responsibility and seen as something that should have been performed by someone else*" (Undebakke et al., 2014, pp. 10-11). The construct is operationalized with four statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

***Inclusiveness and Social Responsibility.***

The Inclusiveness and Social Responsibility scale is retrieved from Pejtersen et al. (2010), which was a new scale developed for and incorporated into the COPSOQ II. The stated preferred interpretation of a high score is that "*inclusion and social responsibility are generally taken care of*" (Undebakke et al., 2014, p. 10). The concept is operationalized by means of four statements on a 5-point likert scale ranging from "disagree" to "agree".

***Innovation.***

This scale is adopted and adapted from Mellor, Mathieu, and Swim (1994). While it was originally intended for investigating the conditions of unions, it is in KIWEST modified to investigate the organizational culture for innovation (or improvement) more generally, and this modification has not as of yet been validated. The stated preferred interpretation of a high score is that "*the respondents experience that there is a culture for continuous improvement in the unit*" (Undebakke et al., 2014, p. 10). Operationalizing the construct are five statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

***Interpersonal Conflicts.***

The Interpersonal Conflicts scale is retrieved from Näswall et al. (2010), where it was adapted from Hovmark and Thomsson (1995) with the intended purpose being to measure the extent to which work is negatively affected by conflicts between employees. The stated preferred

interpretation of a high score is that it indicates that "*the respondents to a little degree are negatively influenced by conflicts between colleagues*" (Undebakke et al., 2014, p. 11). Operationalizing the construct are three statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

***Job Autonomy.***

The Job Autonomy scale is intended to measure the extent of autonomy and influence over how the work is carried out. It is adopted from Näswall et al. (2010), from which it in turn has been adapted from Sverke and Sjoberg (1994), which in turn is based on the works of Walsh, Taber, and Beehr (1980) and Hackman and Oldham (1975). Undebakke et al. (2014, p. 9) state the desired interpretation of a high score as indicating that "*the employees feel they have autonomy and influence on how the work are to be carried out.*" Operationalizing the concept are four statements on a 5-point likert scale ranging from "disagree" to "agree".

***Meaning of Work.***

The Meaning of Work scale is adopted from Pejtersen et al. (2010), where it in turn was adopted from the COPSOQ I (Kristensen et al., 2005) without any changes being made. The stated interpretation of a high score is that it indicates that "*the respondent experience to a high degree that their work is meaningful*" (Undebakke et al., 2014, p. 11). Operationalizing the construct are three statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

***Recognition.***

The Recognition scale is adopted from Pejtersen et al. (2010), from which it originally was adopted and adapted from the effort-reward imbalance model of Siegrist (1996). The stated interpretation of a high score is that "*employees feel to a high degree that they are recognized and appreciated for their efforts*" (Undebakke et al., 2004, p. 10). Operationalizing the concept are three statements on a 5-point likert scale ranging from "disagree" to "agree".

***Reliability of Management (Own Unit and Next Administrative Level).***

The Trust scale is adopted from Näswall et al. (2010), which in turn adapted it off of four items from Robinson (1996). The measure is based on the theory of the psychological contract (Rousseau, 2011), and scores are supposed to reflect employee perceptions of the employers' trustworthiness. Undebakke et al. (2014, p. 10) state the desired interpretation of a high score as indicating that the respondent "*experience to a high degree that management is reliable*

*and trustworthy*." In KIWEST the scale is applied twice, once for the management of the respondents own unit, and once for the next administrative level of the respondents. The concept is operationalized by means of two times five statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

***Role Conflict.***

The Role Conflict scale is adopted from Dallner et al. (2000), which is tied to role theory and the concept of role expectations (Cook, Hepworth, Wall, & Warr, 1981; citing Kahn, Wolfe, Quinn, Snoek, & Rosenthal, 1964; Kelloway & Barling, 1990; and Rizzo et al., 1970). Role conflict is said to occur when role expectations are in conflict. Three types of role conflict are intrasender conflict (conflicting messages from one person), intersender conflict (conflicting messages from two or more persons), and interrole conflict (when one person has two or more conflicting roles). Role conflict may be due to an excess of- or difficult functions. The stated interpretation of a high score is that it indicates that "*the respondents perceive little conflict between their different roles*" (Undebakke et al., 2014, p. 11). Unclear roles or the experience of conflicts different roles, can concern both differing expectations from different people, and of a tension between the employees own expectations and the expectations of others. Role Conflicts might lead to stress for the individual and to conflicts with others. Operationalizing the construct are four statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

***Role Overload.***

The Role Overload scale is adopted and adapted from Näswall et al. (2010), where it was built on three items from Beehr, Walsh, and Taber (1976). The scale is reversed in KIWEST, with the stated preferred interpretation of a high score being that the respondent "*to a little extent experience having too much to do in too little time*" (Undebakke et al., 2014, p. 11). The construct is operationalized by three statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

***Social Climate.***

The Social Climate scale is adopted from Dallner et al. (2000), who, citing Moran and Volkwein (1992), define organizational climate as "*those behavioral and attitudinal characteristics of people that are accessible to external observers*" (Dallner et al., 2000, p. 35). Undebakke et al. (2014, p. 10) state the preferred interpretation of a high score as

indicating "*a good social climate.*" Operationalizing the construct are three statements on a 5-point likert scale, responses ranging from "disagree" to "agree".

***Social Community at Work.***

The Social Community at Work scale is adopted from Pejtersen et al. (2010), which was directly incorporated into the COPSOQ II from the COPSOQ I questionnaire (Kristensen et al., 2005; in which it is labeled "Sense of Community"). The stated preferred interpretation by Undebakke et al. (2014, p. 10) is that a high score indicates that "*the respondents experience a high degree of social community with colleagues in their own unit.*" Operationalizing the concept are tree statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

***Social Support from Supervisors.***

The Social Support from Supervisors scale is adopted and adapted from Pejtersen et al. (2010), the development of which was based on the results from validating the initial COPSOQ I (Kristensen et al., 2005). The interpretation of a high score is stated as indicating that "*the respondent feels a high degree of support from his/her closest superior*" (Undebakke et al., 2004, p. 9). Operationalizing the construct are three statements on a 5-point likert scale ranging from "disagree" to "agree".

***Task Completion Ambiguity.***

The Task Completion Ambiguity scale is intended to capture the extent to which employees themselves can, or have to, determine when their tasks are completed. In KIWEST 2.0 its adopted and adapted from Näswall et al. (2010), from which it in turn has been adopted from Hellgren, Sverke, and Näswall (2008). According to the authors from which it is adapted, high scores indicate that "the individual feels she has a sense of what her tasks entails, and when the tasks can be considered completed". The preferred interpretation of a high score is that "*the employees themselves can, or have to, determine when their tasks are completed*" (Undebakke et al., 2014, p. 9). Operationalizing the construct are three statements on a 5-point likert scale ranging from "disagree" to "agree".

***Trust Regarding Management.***

The Trust Regarding Management scale is adopted from Pejtersen et al. (2010), where it is stated as being inspired by- and having its foundations in theories of social capital (citing Coleman, 1988) and trust (citing Cook & Wall, 1980; and Nooteboom, 2003). The stated

desired interpretation of a high score in KIWEST 2.0 is that it indicates "*a high degree of perceived trust in management*" (Undebakke et al., 2014, p. 10). Operationalizing the construct are four statements on a 5-point likert scale ranging from "disagree" to "agree".

***Utrecht Work Engagement Scale (UWES-9).***

The Utrecht Work Engagement Scale was originally developed as a 17-item measure by Schaufeli, Salanova, González-romá, and Bakker (2002, also known as UWES-17), on which the short-form questionnaire originating in Schaufeli, Bakker, and Salanova (2006, also known as UWES-9) was built. It is this short-form version that is employed in KIWEST.

UWES has its roots in the theory of job-burnout in general, and the Maslach Burnout Inventory in particular (MBI; Maslach, Schaufeli, & Leiter, 2001). Schaufeli et al. (2002, pp. 74-75) define engagement as "*a positive, fulfilling, work-related state of mind that is characterized by vigor, dedication, and absorption. Rather than a momentary and specific state, engagement refers to a more persistent and pervasive affective-cognitive state that is not focused on any particular object, event, individual, or behavior. Vigor is characterized by high levels of energy and mental resilience while working, the willingness to invest effort in one's work, and persistence even in the face of difficulties. Dedication is characterized by a sense of significance, enthusiasm, inspiration, pride, and challenge. [...] The final dimension of engagement, absorption, is characterized by being fully concentrated and deeply engrossed in one's work, whereby time passes quickly and one has difficulties with detaching oneself from work*". In KIWEST, the stated preferred interpretation of a high score is that it indicates that "*the respondents experience a high degree of work engagement*" (Undebakke et al., 2014, p. 11). The construct is operationalized by means of nine statements on a 7-point frequency scale, measuring the frequencies of occurrences ranging from "never" to "every day".

***Work-Family Balance: Conflict and Facilitation.***

The two scales measuring work-family balance are adopted from Innstrand, Langballe, Falkum, Espnes, and Aasland (2009), where it was adapted from Frone (2003) for use in Norway. Taken together, the stated interpretation of a high score on the scales is that it indicates that "*work has little negative impact on family life and that the job has a positive impact on the home situation*" (Undebakke et al., 2014, p. 11-12).

The framework of Frone (2003) suggest that work-family balance consists of four dimensions, where each domain (work and family) can affect the other positively (facilitation) or negatively (conflict). According to Innstrand et al. (2009), the dominant theoretical perspective used to explain the relationship between work and family has been role theory

with conflict and facilitation perspectives (citing Hanson, Hammer, & Colton, 2006; Voydanoff, 2002). The conflict perspective, also known and role strain theory, suggests that work-family conflict occurs when demands associated with one domain are incompatible with demands associated with the other domain (citing Perrewé, Hochwarter, & Kiewitz, 1999). In contrast to the conflict perspective, the facilitation perspective proposes that occupying multiple roles can be beneficial and even outweigh the cost of multiple role involvements (citing Sieber, 1974). Facilitation is said to occur when participation in one role is made better or easier due to participation in the other role (citing Wayne, Musisca, & Fleeson, 2004). Operationalizing the construct are two times four statements on a 5-point likert scale with responses ranging from "disagree" to "agree".

#### *Work-SoC – Sense of Coherence in Work.*

The Sense of Coherence in Work scale is adopted from Vogt, Jenny, and Bauer (2013) and Bauer and Jenny (2007), and is based on the concept and perspective of salutogenesis (health promotion) from Aaron Antonovsky (1979), which is intended to serve as the conceptual opposite of the pathogenesis perspective (sickness prevention). Antonovsky (1979) theorized that an individuals' "sense of coherence" (SOC) constitute an important salutogenic factor, and it is conceived of as a cognitive concept composed of three (supposedly) closely intertwined themes; comprehensibility, manageability, and meaningfulness.

According to Vogt et al. (2013, p. 2), comprehensibility describes "*the extent to which a work situation is perceived as structured, consistent and clear.*" Manageability describes "*the extent to which an employee perceives that adequate resources are available to cope with the demands in the workplace.*" Finally, meaningfulness describes "*the extent to which a situation at work is seen as worthy of commitment and involvement.*" The perception of comprehensibility, manageability and meaningfulness is supposed to be influenced by the interaction between individual characteristics (an employee's personality and experiences) and the characteristics of the working environment (work-related structures and processes). In KIWEST, the stated interpretation of a high score is that it indicates that "*the respondents experience to a high degree that their workplace is health promoting*" (Undebakke et al., 2014, p. 12). Operationalizing the construct are nine statements on a 7-point semantic differential scale, with responses ranging from one end of the extreme to the other.

### Sampling and Data Collection

Sampling and data collection was performed prior to this study. The information pertaining to data collection and sampling from the ARK research platform website (ARK, 2016) reads as

Table 3.2.
*Initial Sample Description, Sorted by Type of Position and Gender.*

| | Gender | | | |
|---|---|---|---|---|
| Type of position | Female | Male | Other | Total |
| Academic | 1441 | 1817 | 1 | 3259 |
| Doctoral research fellow | 829 | 600 | 0 | 1429 |
| Technical/administrative | 1688 | 1116 | 0 | 2804 |
| Unit leader, level 1 | 1 | 2 | 0 | 3 |
| Unit leader, level 2 | 14 | 14 | 0 | 28 |
| Unit leader, level 3 | 84 | 135 | 0 | 219 |
| Unit leader, level 4 | 41 | 59 | 0 | 100 |
| Unit leader, level 5 | 4 | 13 | 0 | 17 |
| Total | 4102 | 3756 | 1 | 7859 |

Note: Category membership pertaining to type of position and gender is retreived from registry data.

follows: "*Data collection with KIWEST#2 was done at Norwegian universities and university colleges in the period from October 2013 to December 2015. Employees with regular payroll for minimum 20% position were invited to participate. They received an e-mail with a link to the online questionnaire. The online data collection was conducted by the IT department at the Faculty of Social Sciences and Technology Management, NTNU (SVT-IT), using the SelectSurvey.NET software package from ClassApps (www.classapps.com).*"

Concerning response rates, 12170 out of an unspecified total number of employees responded to the survey, whereof 6527 of respondents were women, and 5642 were men. The data file supplied for this study had been subject to some cursory treatment based on missing value patterns. That is, cases missing more than 50% of responses on any scale were excluded from the set (4275 cases). The data thus included 7895 cases (see table 3.2 for description of the sample in the dataset supplied for this study, sorted by type of position and gender).

**Data Integrity Analysis**

Data integrity analysis consists of manual case screening (i.e., screening cases for response patterns indicating lack of motivation, weeding out cases in which the target constructs clearly played no causal part in producing the responses), missing value analysis, and data normality analysis. Data screening consisted of computing case standard deviations of responses and examining those that were low. The mean standard deviation of responses was 1.4 (SD = .24). Cases were sorted by their standard deviations of responses, and based on visual inspection of response patterns, cases exhibiting standard deviations below .6 (13 cases in total) appeared to exhibit obviously irrelevant patterns of responses, and were excluded from further analyses. Beyond this point, response patterns did not appear to reflect obviously irrelevant variance.

**Missing value analysis: Little's test for MCAR.**

Analysis of missing entries in the dataset addresses to the "response process" category of validity evidence in the standards, as factors that are at play when respondents do not respond to certain items are part of processes involved in producing the missing entries in the data set (Allison, 2002; Enders, 2010). Identifying processes at play in producing missing values plays part in determining how they are to be treated – that is, whether the problem of missing values is ignorable or non-ignorable. The question of whether missing entries are random or non-random (i.e., systematic; cf., table 2.1, p. 19) concerns whether there are processes at play in producing the missing entries, and thus how missing entries should be interpreted. Analyses of the causes of missing entries are thus of relevance to validity in measurement, as treating non-random causes of missing entries as random can introduce construct irrelevant variance.

A popular classification scheme providing terminology for describing processes involved in the production of missing entries has its origin in- and is generally attributed to Rubin (1976). Within this framework, processes involved in producing missing entries are abstractly categorized as MCAR (missing completely at random), MAR (missing at random), or NMAR (not missing at random). Missing entries can be considered MCAR if they are not caused by other observed variable in the data set nor by itself (i.e., no discernable pattern), and MAR if they are predictable by other observed variables in the data set but not by itself (e.g., variables such as gender or type of position in predicting disclosure of trust). Entries are NMAR if they cause themselves to be missing (i.e., the items are themselves the causes, e.g., if employees' state of trust impacts their willingness to disclose trust; APA, 2010).

If entries are NMAR, the problem is non-ignorable and non-treatable after-the-fact in absence of additional information (i.e., factors that can predict instances of missing entries), as the issue points to problems at the level of data collection. In the case of NMAR, the cause of "missingness" ought to be examined and dealt with at the level of survey design and/or administration. In questionnaire-based research, there are several possible causes for why an item would consistently produces missing entries. These reasons are not only of potential relevance to the question of validity, but also of quality (i.e., effectiveness and efficiency). If, for example, examination of the cause reveals that the item causes non-responding due to respondents perceiving it to be irrelevant, removing the item from future applications of the questionnaire can increase the efficiency of the survey. Alternatively, the item could be replaced with a relevant and informative one to improve effectiveness at no cost to efficiency.

If missing entries are MCAR or MAR, there are procedures available that allow one to replace the missing entries with informed value estimates without introducing bias, and thus

without threats to the validity of interpretations. Among these procedures there are limited-information approaches such as single or multiple imputation (MI), or full-information approaches such as full-information maximum likelihood (FIML). Among these, FIML is considered the most potent and appropriate, followed by MI, followed by single stochastic regression imputation, followed by conditional mean replacement, etc. (Enders, 2010).

Little's test for MCAR (R. J. A. Little, 1988) is a test devised to aid in determining whether missing entries are missing completely at random or not. It does this by evaluating mean differences across subgroups of cases that share the same missing data pattern, providing a test statistic that constitutes a weighted sum of the standardized differences between subgroup means and the grand mean (Enders, 2010). In STATA14, a user-written module ("mcartest") exists for performing Little's MCAR test (Li, 2013).

***Outcome of missing value analysis.***

When Little's test for MCAR is administered to the dataset as a whole, the results indicate that there are discernable non-random patterns of missingness, as the test yields a $\chi^2(101501) = 108440.1$, $p < 0.001$ ($N = 7859$). In an attempt to localize and isolate cases of systematically caused missingness, the MCAR test was applied to groups of items intended to constitute a scale or subscale. The outcome of this lower-order analysis indicate that 11 of the 33 possible scales demonstrate patterns of missing entries that appear to not be completely random (i.e., attains a p-value < 0.05). Results are available for inspection in table 3.3 (pp. 46-47).

**Univariate and multivariate normality analysis: G and $E_p$.**

Analysis of multivariate normality, in contrast to analysis of missing entries, does not directly address any of the categories of validity evidence specified in the standards, but it constitutes a precondition for examining evidence related to the "internal structure" category, as ML-based CFA assumes multivariate normality. According to Hu and Bentler (1999, p. 8), "*violation of the multivariate normality assumption can seriously invalidate normal-theory test statistics.*" As such, for applications- and interpretations of ML-based CFA to be valid, it is necessary to diagnose (and, if necessary, to treat) deviations from normality.

Several tests for the purposes of detecting and quantifying the severity of deviations have been devised (e.g., D'Agostino & Belanger, 1990; Doornik & Hansen, 2008; Joanes & Gill, 1998; Mardia, 1970). Most scholars caution against strictly adhering to them however, the reason being that significance testing deviations will detect even trivial deviations if the sample size is large enough, and fail to detect severe violations if the sample size is too small. For this reason, scholars instruct researchers to exercise judgement when deciding if variables

exhibit problematic amounts of deviation from normality, and visual inspection of distribution plots in order to inform decisions is encouraged. That being said, some rules of thumb have been proposed, such as skewness or kurtosis values $\geq 1$ or $\leq -1$ of measures such as *b*, *g*, or *G* (Joanes & Gill, 1998), as indicating severe deviation from univariate normality.

If non-normally distributed data is an issue, one can attempt to remedy the issue by means of data transformations that tend to normalize distributions. If data transformations fail to bring the distribution within acceptable bounds one need to either, (1) consider alternative methods of estimation that do not require normally distributed data, or (2) exclude offending measures from the analysis. According to Hu and Bentler (1999) most of the ML-based fit indices outperform those obtained from general least squares and asymptotic distribution free estimation, and should for this reason be preferred indicators for evaluating model fit. Thus, ML as a means of estimation is not on the table, and items that cause severe violation of the normality assumption are for this reason jettisoned from the analysis.

***Outcome of normality analysis.***

Analysis of multivariate normality was conducted making use of the Doornik-Hansen $E_p$ test (Doornik & Hansen, 2008) in STATA14. The results are available in table 3.3 (pp. 48-49), along with the results from the analysis of missing values. The tests revealed that every scale deviate significantly from perfectly normal distributions. According to Hair, Black, Babin, and Anderson (2013), most cases of multivariate non-normality are caused by univariate non-normality. For this reason, univariate normality is also included in table 3.3. The analyses revealed that a good deal of items exceed the values generally considered acceptable bounds for normality. The “Joint” column represents an attempt to quantify severity of deviations, in terms of the sum of skewness- ($G_1$) and kurtosis ($G_2$) deviations from univariate normality. Furthermore, visual inspection of the univariate distributions revealed that item v10_010 (pertaining to the DUWAS scale) exhibit a roofing effect (i.e., the mode value is an end-point value). As such, the item does not discriminate well between respondents with a high standing on the variable (DeVellis, 2017; Penfield, 2013), and is thus excluded from further analyses.

**Data Integrity Treatment**

Data integrity treatment consisted of the replacement of missing entries by means of single ordinal regression imputation, and the normalizing of distributions by means of square- and square-root variable transformations – subsequently finding (with a $k^3$ matrix approach) the combinations of transformed- and non-transformed variables most closely approximating multivariate normal distributions.

Table 3.3.
*Data Integrity Analysis: Missing Values and Normality.*

| | Multivariate (scale level) | | | | | | | Univariate (item level) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing entries Little's MCAR | | | Normality $E_p$ | | | | Missing entries N = 7846 | | Normality | | |
| | | | | | | | | | | Skewness | Kurtosis | Joint |
| **Scale** | $\chi^2$ | df | p | $\chi^2$ | df | p | **Item** | # | % | $G_1$ | $G_2$ | $\sqrt{(G_1^2)} + \sqrt{(G_2^2)}$ |
| Cohesion in Work Teams | 16,90 | 6 | 0,010 | 1394,96 | 6 | 0,000 | v01_007 | 31 | 0,4 | -0,565 | -0,015 | 0,580 |
| | | | | | | | v02_003 | 55 | 0,7 | -0,448 | -0,394 | 0,842 |
| | | | | | | | v02_005 | 109 | 1,4 | -0,679 | 0,297 | 0,976 |
| Commitment to the Workplace | 1,15 | 6 | 0,979 | 2700,17 | 6 | 0,000 | v09_001 | 11 | 0,1 | -0,892 | 0,852 | 1,744 |
| | | | | | | | v09_008 | 14 | 0,2 | -0,899 | 0,644 | 1,543 |
| | | | | | | | v09_014 | 44 | 0,6 | -0,835 | 1,390 | 2,225 |
| Competency Demands | 6,35 | 6 | 0,386 | 1342,05 | 6 | 0,000 | v03_003 | 10 | 0,1 | -0,503 | -0,193 | 0,696 |
| | | | | | | | v03_009 | 25 | 0,3 | -0,802 | 0,727 | 1,529 |
| | | | | | | | v03_012 | 39 | 0,5 | 0,087 | -0,746 | 0,833 |
| DUWAS (Compulsive) | 43,94 | 45 | 0,517 | 2854,25 | 10 | 0,000 | v10_003 | 30 | 0,4 | -0,486 | -0,280 | 0,766 |
| | | | | | | | v10_005 | 74 | 0,9 | -0,008 | -0,582 | 0,590 |
| | | | | | | | v10_007 | 38 | 0,5 | -0,606 | -0,211 | 0,817 |
| | | | | | | | v10_009 | 24 | 0,3 | -0,696 | -0,301 | 0,997 |
| | | | | | | | v10_010 | 30 | 0,4 | -1,060 | 0,518 | 1,578 |
| DUWAS (Excessive) | 21,69 | 29 | 0,833 | 1468,75 | 10 | 0,000 | v10_001 | 7 | 0,1 | -0,202 | -0,513 | 0,715 |
| | | | | | | | v10_002 | 47 | 0,6 | -0,269 | -0,488 | 0,757 |
| | | | | | | | v10_004 | 33 | 0,4 | 0,066 | -0,623 | 0,689 |
| | | | | | | | v10_006 | 11 | 0,1 | -0,691 | -0,365 | 1,056 |
| | | | | | | | v10_008 | 31 | 0,4 | -0,364 | -0,602 | 0,966 |
| Dysfunctional Support | 72,59 | 50 | 0,020 | 6544,72 | 12 | 0,000 | v01_002 | 50 | 0,6 | -0,306 | -0,526 | 0,832 |
| | | | | | | | v01_005 | 16 | 0,2 | -0,697 | -0,027 | 0,724 |
| | | | | | | | v01_006 | 22 | 0,3 | -0,881 | 0,583 | 1,464 |
| | | | | | | | v01_011 | 28 | 0,4 | -0,623 | -0,066 | 0,689 |
| | | | | | | | v01_013 | 49 | 0,6 | -0,884 | 0,455 | 1,339 |
| | | | | | | | v01_015 | 69 | 0,9 | -0,886 | 0,457 | 1,343 |
| Empowering Leadership | 77,25 | 54 | 0,813 | 1370,86 | 6 | 0,000 | v07_001 | 55 | 0,7 | -0,677 | -0,033 | 0,710 |
| | | | | | | | v07_002 | 58 | 0,7 | -0,735 | 0,007 | 0,742 |
| | | | | | | | v07_003 | 29 | 0,4 | -0,686 | -0,014 | 0,700 |
| Fairness of the Supervisor | 22,42 | 6 | 0,001 | 2261,07 | 6 | 0,000 | v07_004 | 40 | 0,5 | -0,905 | 0,587 | 1,492 |
| | | | | | | | v07_007 | 356 | 4,5 | -0,517 | 0,071 | 0,588 |
| | | | | | | | v07_008 | 83 | 1,1 | -0,652 | -0,123 | 0,775 |
| Goal Clarity | 4,72 | 6 | 0,581 | 1574,55 | 6 | 0,000 | v04_001 | 12 | 0,2 | -0,460 | -0,523 | 0,983 |
| | | | | | | | v04_004 | 18 | 0,2 | -0,542 | -0,258 | 0,800 |
| | | | | | | | v04_009 | 32 | 0,4 | -0,912 | 0,878 | 1,790 |
| Illegitimate Tasks | 16,16 | 12 | 0,184 | 2279,24 | 8 | 0,000 | v04_003 | 25 | 0,3 | -0,231 | -0,785 | 1,016 |
| | | | | | | | v04_008 | 26 | 0,3 | -0,583 | -0,235 | 0,818 |
| | | | | | | | v04_011 | 39 | 0,5 | -0,784 | 0,348 | 1,132 |
| | | | | | | | v04_014 | 49 | 0,6 | -0,508 | -0,171 | 0,679 |
| Inclusiveness and Social Responsibility | 36,70 | 12 | 0,000 | 6199,74 | 8 | 0,000 | v02_001 | 133 | 1,7 | -1,183 | 1,137 | 2,320 |
| | | | | | | | v02_002 | 394 | 5,0 | -1,166 | 2,093 | 3,259 |
| | | | | | | | v02_004 | 59 | 0,8 | -1,221 | 2,188 | 3,409 |
| | | | | | | | v02_006 | 640 | 8,2 | -0,687 | 0,498 | 1,185 |
| Innovation | 41,73 | 35 | 0,201 | 1613,95 | 10 | 0,000 | v04_002 | 32 | 0,4 | -0,274 | -0,424 | 0,698 |
| | | | | | | | v04_006 | 38 | 0,5 | -0,728 | 0,350 | 1,078 |
| | | | | | | | v04_010 | 23 | 0,3 | -0,864 | 0,761 | 1,625 |
| | | | | | | | v04_012 | 20 | 0,3 | -0,414 | -0,059 | 0,473 |
| | | | | | | | v04_015 | 39 | 0,5 | -0,532 | -0,053 | 0,585 |
| Interpersonal Conflicts | 13,19 | 6 | 0,040 | 1194,42 | 6 | 0,000 | v01_003 | 17 | 0,2 | -0,518 | -0,757 | 1,275 |
| | | | | | | | v01_014 | 41 | 0,5 | -0,628 | -0,545 | 1,173 |
| | | | | | | | v01_017 | 38 | 0,5 | -0,561 | -0,418 | 0,979 |
| Job Autonomy | 18,82 | 12 | 0,093 | 4108,39 | 8 | 0,000 | v04_005 | 35 | 0,4 | -0,908 | 0,801 | 1,709 |
| | | | | | | | v04_007 | 29 | 0,4 | -0,970 | 1,765 | 2,735 |
| | | | | | | | v04_013 | 71 | 0,9 | -1,121 | 2,565 | 3,686 |
| | | | | | | | v04_016 | 35 | 0,4 | -0,528 | 0,159 | 0,687 |
| Meaning of Work | 1,84 | 6 | 0,934 | 3248,90 | 6 | 0,000 | v09_003 | 54 | 0,7 | -1,031 | 1,407 | 2,438 |
| | | | | | | | v09_006 | 24 | 0,3 | -0,897 | 2,013 | 2,910 |
| | | | | | | | v09_016 | 26 | 0,3 | -0,882 | 1,711 | 2,593 |
| Recognition | 5,56 | 6 | 0,474 | 2546,26 | 6 | 0,000 | v06_005 | 14 | 0,2 | -0,901 | 0,821 | 1,722 |
| | | | | | | | v06_007 | 25 | 0,3 | -0,738 | 0,235 | 0,973 |
| | | | | | | | v06_010 | 23 | 0,3 | -0,905 | 0,925 | 1,830 |

Note: Joint represents the sum of absolute deviation from normality in terms of skewness and kurtosis values.

Table 3.3 (continued).
*Data Integrity Analysis: Missing Values and Normality.*

| | Multivariate (scale level) | | | | | | | Univariate (item level) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Missing entries Little's MCAR | | | Normality $E_p$ | | | | Missing entries N = 7846 | | Normality | | |
| | | | | | | | | | | Skewness | Kurtosis | Joint |
| **Scale** | $\chi^2$ | df | p | $\chi^2$ | df | p | **Item** | # | % | $G_1$ | $G_2$ | $\sqrt{(G_1^2)} + \sqrt{(G_2^2)}$ |
| Reliability of Management (Next Administrative Level) | 62,60 | 47 | 0,063 | 6190,38 | 10 | 0,000 | v08_001 | 84 | 1,1 | -0,629 | 0,300 | 0,929 |
| | | | | | | | v08_002 | 61 | 0,8 | -0,411 | -0,017 | 0,428 |
| | | | | | | | v08_003 | 227 | 2,9 | -0,423 | 0,249 | 0,672 |
| | | | | | | | v08_004 | 49 | 0,6 | -0,434 | -0,058 | 0,492 |
| | | | | | | | v08_005 | 42 | 0,5 | -0,426 | -0,157 | 0,583 |
| Reliability of Management (Own Unit) | 33,86 | 26 | 0,139 | 4359,54 | 10 | 0,000 | v06_001 | 10 | 0,1 | -0,755 | 0,165 | 0,920 |
| | | | | | | | v06_003 | 54 | 0,7 | -0,831 | 0,314 | 1,145 |
| | | | | | | | v06_006 | 26 | 0,3 | -0,776 | 0,518 | 1,294 |
| | | | | | | | v06_009 | 20 | 0,3 | -0,733 | 0,016 | 0,749 |
| | | | | | | | v06_012 | 38 | 0,5 | -0,746 | 0,184 | 0,930 |
| Role Conflict | 7,41 | 12 | 0,829 | 3448,33 | 8 | 0,000 | v03_002 | 22 | 0,3 | -0,346 | -0,665 | 1,011 |
| | | | | | | | v03_006 | 27 | 0,3 | -0,477 | -0,279 | 0,756 |
| | | | | | | | v03_007 | 16 | 0,2 | -1,286 | 1,941 | 3,227 |
| | | | | | | | v03_011 | 43 | 0,5 | -0,120 | -0,726 | 0,846 |
| Role Overload | 5,43 | 6 | 0,490 | 676,47 | 6 | 0,000 | v03_004 | 18 | 0,2 | 0,388 | -0,558 | 0,946 |
| | | | | | | | v03_010 | 19 | 0,2 | 0,423 | -0,373 | 0,796 |
| | | | | | | | v03_013 | 15 | 0,2 | -0,023 | -0,892 | 0,915 |
| Social Climate | 4,22 | 8 | 0,837 | 2208,07 | 6 | 0,000 | v01_008 | 20 | 0,3 | -1,002 | 0,407 | 1,409 |
| | | | | | | | v01_016 | 48 | 0,6 | -0,737 | 0,407 | 1,144 |
| | | | | | | | v01_018 | 28 | 0,4 | -0,610 | -0,031 | 0,641 |
| Social Community at Work | 19,10 | 6 | 0,004 | 3077,08 | 6 | 0,000 | v01_001 | 14 | 0,2 | -1,058 | 1,125 | 2,183 |
| | | | | | | | v01_009 | 45 | 0,6 | -1,023 | 1,693 | 2,716 |
| | | | | | | | v01_010 | 65 | 0,8 | -0,781 | 0,398 | 1,179 |
| Social Support from Supervisor | 23,64 | 6 | 0,001 | 1954,29 | 6 | 0,000 | v07_005 | 42 | 0,5 | -0,426 | -0,618 | 1,044 |
| | | | | | | | v07_006 | 192 | 2,4 | -1,016 | 0,973 | 1,989 |
| | | | | | | | v07_009 | 83 | 1,1 | -0,782 | 0,196 | 0,978 |
| Task Completion Ambiguity | 5,02 | 6 | 0,541 | 2697,22 | 6 | 0,000 | v03_001 | 25 | 0,3 | -1,054 | 2,028 | 3,082 |
| | | | | | | | v03_005 | 36 | 0,5 | -0,626 | 0,006 | 0,632 |
| | | | | | | | v03_008 | 26 | 0,3 | -0,663 | 0,219 | 0,882 |
| Trust Regarding Management | 33,10 | 12 | 0,001 | 3574,71 | 8 | 0,000 | v06_002 | 21 | 0,3 | -0,937 | 0,685 | 1,622 |
| | | | | | | | v06_004 | 10 | 0,1 | -0,634 | -0,023 | 0,657 |
| | | | | | | | v06_008 | 18 | 0,2 | -1,176 | 1,812 | 2,988 |
| | | | | | | | v06_011 | 32 | 0,4 | -0,943 | 1,409 | 2,352 |
| UWES (Absorption) | 7,86 | 9 | 0,548 | 3640,89 | 6 | 0,000 | v11_007 | 37 | 0,5 | -1,393 | 2,104 | 3,497 |
| | | | | | | | v11_008 | 36 | 0,5 | -1,187 | 1,134 | 2,321 |
| | | | | | | | v11_009 | 43 | 0,5 | -0,660 | -0,382 | 1,042 |
| UWES (Dedication) | 10,57 | 9 | 0,306 | 4883,79 | 6 | 0,000 | v11_004 | 19 | 0,2 | -1,475 | 2,274 | 3,749 |
| | | | | | | | v11_005 | 23 | 0,3 | -1,263 | 1,455 | 2,718 |
| | | | | | | | v11_006 | 40 | 0,5 | -1,335 | 1,574 | 2,909 |
| UWES (Vigor) | 9,69 | 9 | 0,376 | 5906,71 | 6 | 0,000 | v11_001 | 16 | 0,2 | -1,294 | 1,875 | 3,169 |
| | | | | | | | v11_002 | 44 | 0,6 | -1,319 | 1,998 | 3,317 |
| | | | | | | | v11_003 | 19 | 0,2 | -1,606 | 2,492 | 4,098 |
| Work-Family Conflict | 24,72 | 12 | 0,016 | 405,79 | 8 | 0,000 | v09_002 | 31 | 0,4 | 0,201 | -0,992 | 1,193 |
| | | | | | | | v09_007 | 14 | 0,2 | 0,171 | -0,820 | 0,991 |
| | | | | | | | v09_009 | 15 | 0,2 | -0,027 | -0,866 | 0,893 |
| | | | | | | | v09_015 | 24 | 0,3 | -0,091 | -0,715 | 0,806 |
| Work-Family Facilitation | 9,65 | 12 | 0,647 | 577,06 | 8 | 0,000 | v09_004 | 26 | 0,3 | -0,023 | -0,203 | 0,226 |
| | | | | | | | v09_005 | 12 | 0,2 | -0,251 | -0,195 | 0,446 |
| | | | | | | | v09_011 | 32 | 0,4 | -0,259 | -0,381 | 0,640 |
| | | | | | | | v09_013 | 16 | 0,2 | -0,624 | 1,317 | 1,941 |
| Work SoC (Comprehens-ibility) | 41,54 | 18 | 0,001 | 2878,79 | 8 | 0,000 | v12_001 | 31 | 0,4 | 0,960 | 0,210 | 1,170 |
| | | | | | | | v12_003 | 42 | 0,5 | 0,382 | -0,634 | 1,016 |
| | | | | | | | v12_006 | 42 | 0,5 | 0,440 | -0,646 | 1,086 |
| | | | | | | | v12_009 | 57 | 0,7 | 0,382 | -0,610 | 0,992 |
| Work SoC (Manageability) | 8,55 | 2 | 0,014 | 769,28 | 4 | 0,000 | v12_004 | 42 | 0,5 | 0,432 | -0,395 | 0,827 |
| | | | | | | | v12_007 | 67 | 0,9 | 0,468 | -0,414 | 0,882 |
| Work SoC (Meaningfulness) | 6,13 | 9 | 0,727 | 7434,38 | 6 | 0,000 | v12_002 | 48 | 0,6 | -1,407 | 1,803 | 3,210 |
| | | | | | | | v12_005 | 44 | 0,6 | -1,328 | 1,822 | 3,150 |
| | | | | | | | v12_008 | 46 | 0,6 | -1,260 | 1,281 | 2,541 |

Note: Joint represents the sum of absolute deviation from normality in terms of skewness and kurtosis values.

**Missing value treatment: single ordinal logistic regression imputation.**

Missing values were treated with single ordinal logistic regression. The reasoning behind this choice of treatment is the following: while FIML and MI are considered the most potent and as such the desirable means of replacing missing entries with informed estimates, certain technical and practical limitations prevented their application. First, attempting to estimate the models using FIML revealed that every model would take approximately a week to estimate using this method. As such, the potential gain in effectiveness was judged inconsequentially small relative to its severe lack of efficiency. Second; STATA does not allow for the use of MI with SEM, and was therefore simply not available as an option.

As the most potent treatment techniques proved either unavailable or unfeasible for technical reasons, as well as the fact that the scales to which respondents had to represent their experiences were ordinal, the choice was made to employ SI using ordinal logistic regression to estimate actual possible responses. This choice was made in order for the technique to reproduce method-associated measurement error (i.e., construct irrelevant variance). The variables included in the estimations were the remaining manifest variables postulated to belong to the entities that the target manifest variables are intended to measure, as well as the group memberships of each respondent (i.e., type of employment and gender). This was done for the purposes of reproducing possible differential item functioning across subgroups of the targeted population (cf., table 2.1, p. 17; Millsap, 1997; Millsap & Everson, 1993; Rogers & Swaminathan, 2016). In other words, the intent underlying the treatment of missing values has been to maintain, not to obscure or compensate for, instances of poor item functioning and thus sources of poor model fit.

The treatment succeeded at imputing estimates for most missing entries. However, failures of imputations were observed for 52 of the 118 items, belonging to 13 of the 33 possible scales. Some items and scales exhibited greater frequencies of imputation failures than others did. The mean number of failed imputations for specific items was 5.6 (SD = 12.2). Drawing an arbitrary distinction of two standard deviations from a value of zero (which corresponds to a failure frequency of 24.4), four scales include items for which the frequency of imputation failure proved relatively large. These are the Reliability of Management (Next Level) scale, and all of the Work-SoC subscales.

**Multivariate normality treatment.**

Multivariate normality treatment consisted of multivariate normality diagnostics of the transformed and non-transformed variables composing each scale. The Doornik-Hansen test

for multivariate normality was conducted on every possible combination of square-, square-root-, and non-transformed variables making up a scale (2813 tests in total). The results of each test pertaining to each scale were compared to find which combination of transformed and non-transformed variables most closely approximated a multivariate normal distribution, in terms of attaining the lowest $\chi^2$ value. All significance tests were significant at the $p < .001$ level, meaning that none of the combinations of variables conforms exactly to a multivariate normal distribution.

The outcome of the multivariate normality treatment are presented in table 3.4, where the best combination of variables (i.e., the combinations of variables attaining the lowest $\chi^2$ values) is presented in the univariate column, and compared against the baseline (i.e., non-transformed) combinations in the multivariate column. It is difficult to assess the severity of violations by means of the $\chi^2$ statistic (whether the extent of misfit is trivial or non-trivial), as it is sensitive to sample size and complexity. However, the closer the $\chi^2$ value approximates a value of zero, the better the fit between the observed distribution and an ideal distribution. To illustrate of the extent of the pre- to post-treatment improvement, the pre-post treatment difference is provided in the right-most column of the multivariate column. "Δ" represents the absolute improvement in $\chi^2$ value (pre-treatment $\chi^2$ value minus post-treatment $\chi^2$ value), while "%" represents the relative improvement as a result of the treatment (100 minus post-treatment $\chi^2$ value percent of pre-treatment $\chi^2$ value).

Overall, the statistics indicate that the treatment yielded strong improvements to the multivariate normality of the scales, with a mean absolute $\chi^2$ value improvement of 2272.57, which corresponds to a mean percentage improvement in the $\chi^2$ value of 76.5%. The mean post-treatment $\chi^2$ value was 717, with a standard deviation of 790. As mentioned, it is difficult to assess by means of the $\chi^2$ statistic whether the multivariate normality of a set of items represents severe violations of the assumption. For the present purposes however, one can make some assessment based on the performance of a set of items relative to the remaining set of items in terms of their distance from a $\chi^2$ value of zero.

Drawing an arbitrary distinction of two standard deviations from a value of zero (corresponding to a post-treatment $\chi^2$ value of 1580), four of the scales appear to perform poorly relative to the remaining scales. These scales are: Dysfunctional Support, Reliability of Management (Next Level), Reliability of Management (Own Unit), and UWES (Vigor). No further attempts were made at normalizing the multivariate normal distribution of the items making up each of these scales. For this reason, parameter estimates pertaining to these scales in subsequent analyses must be interpreted with caution.

Table 3.4.
*Post Treatment Data Integrity Analysis: Multivariate Normality.*

| | Multivariate (scale level) | | | | | | | | | Univariate (item level) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Non-transformed $E_p$ | | | Transformed $E_p$ | | | $\chi^2$ Improvement | | | Employed variable transformations | | |
| **Scale** | $\chi^2$ | df | $\chi^2$/df | $\chi^2$ | df | $\chi^2$/df | Δ | % | **Item** | None | Square | Square-root |
| Cohesion in Work Teams | 1416,14 | 6 | 236,0 | 266,618 | 6 | 44,4 | 1149,52 | 81,17 | v01_007 | | x | |
| | | | | | | | | | v02_003 | | x | |
| | | | | | | | | | v02_005 | | x | |
| Commitment to the Workplace | 2743,49 | 6 | 457,2 | 185,504 | 6 | 30,9 | 2557,98 | 93,24 | v09_001 | | x | |
| | | | | | | | | | v09_008 | | x | |
| | | | | | | | | | v09_014 | | x | |
| Competency Demands | 1362,99 | 6 | 227,2 | 264,313 | 6 | 44,1 | 1098,68 | 80,61 | v03_003 | | x | |
| | | | | | | | | | v03_009 | | x | |
| | | | | | | | | | v03_012 | x | | |
| DUWAS (Compulsive) | 1689,17 | 8 | 211,1 | 1224,63 | 8 | 153,1 | 464,54 | 27,50 | v10_003 | x | | |
| | | | | | | | | | v10_005 | x | | |
| | | | | | | | | | v10_007 | | x | |
| | | | | | | | | | v10_009 | | x | |
| DUWAS (Excessive) | 1495,10 | 10 | 149,5 | 874,031 | 10 | 87,4 | 621,06 | 41,54 | v10_001 | x | | |
| | | | | | | | | | v10_002 | x | | |
| | | | | | | | | | v10_004 | x | | |
| | | | | | | | | | v10_006 | | x | |
| | | | | | | | | | v10_008 | x | | |
| Dysfunctional Support | 6544,72 | 12 | 545,4 | 2220,846 | 12 | 185,1 | 4323,88 | 66,07 | v01_002 | | x | |
| | | | | | | | | | v01_005 | | x | |
| | | | | | | | | | v01_006 | | x | |
| | | | | | | | | | v01_011 | | x | |
| | | | | | | | | | v01_013 | | x | |
| | | | | | | | | | v01_015 | | x | |
| Empowering Leadership | 1398,49 | 6 | 233,1 | 62,503 | 6 | 10,4 | 1335,98 | 95,53 | v07_001 | | x | |
| | | | | | | | | | v07_002 | | x | |
| | | | | | | | | | v07_003 | | x | |
| Fairness of the Supervisor | 2392,36 | 6 | 398,7 | 312,646 | 6 | 52,1 | 2079,71 | 86,93 | v07_004 | | x | |
| | | | | | | | | | v07_007 | | x | |
| | | | | | | | | | v07_008 | | x | |
| Goal Clarity | 1595,83 | 6 | 266,0 | 88,821 | 6 | 14,8 | 1507,01 | 94,43 | v04_001 | | x | |
| | | | | | | | | | v04_004 | | x | |
| | | | | | | | | | v04_009 | | x | |
| Illegitimate Tasks | 2362,31 | 8 | 295,3 | 390,912 | 8 | 48,9 | 1971,40 | 83,45 | v04_003 | | x | |
| | | | | | | | | | v04_008 | | x | |
| | | | | | | | | | v04_011 | | x | |
| | | | | | | | | | v04_014 | | x | |
| Inclusiveness and Social Responsibility | 7345,41 | 8 | 918,2 | 1598,603 | 8 | 199,8 | 5746,81 | 78,24 | v02_001 | | x | |
| | | | | | | | | | v02_002 | | x | |
| | | | | | | | | | v02_004 | | x | |
| | | | | | | | | | v02_006 | | x | |
| Innovation | 1653,04 | 10 | 165,3 | 581,717 | 10 | 58,2 | 1071,32 | 64,81 | v04_002 | x | | |
| | | | | | | | | | v04_006 | x | | |
| | | | | | | | | | v04_010 | | x | |
| | | | | | | | | | v04_012 | x | | |
| | | | | | | | | | v04_015 | | x | |
| Interpersonal Conflicts | 1204,03 | 6 | 200,7 | 227,748 | 6 | 38,0 | 976,28 | 81,08 | v01_003 | | x | |
| | | | | | | | | | v01_014 | | x | |
| | | | | | | | | | v01_017 | | x | |
| Job Autonomy | 4233,07 | 8 | 529,1 | 262,063 | 8 | 32,8 | 3971,01 | 93,81 | v04_005 | | x | |
| | | | | | | | | | v04_007 | | x | |
| | | | | | | | | | v04_013 | | x | |
| | | | | | | | | | v04_016 | | x | |
| Meaning of Work | 3284,64 | 6 | 547,4 | 292,032 | 6 | 48,7 | 2992,61 | 91,11 | v09_003 | | x | |
| | | | | | | | | | v09_006 | | x | |
| | | | | | | | | | v09_016 | | x | |
| Recognition | 2563,70 | 6 | 427,3 | 217,385 | 6 | 36,2 | 2346,32 | 91,52 | v06_005 | | x | |
| | | | | | | | | | v06_007 | | x | |
| | | | | | | | | | v06_010 | | x | |

Note: For Δ, a positive value represents an improvement in terms of multivariate normality as measured by the $\chi^2$ statistic.
An "x" denotes the version of the variable that is employed in the analysis of multivariate normality.

Table 3.4 (continued).
*Post Treatment Data Integrity Analysis: Multivariate Normality.*

| | Multivariate (scale level) | | | | | | | | | Univariate (item level) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Non-transformed $E_p$ | | | Transformed $E_p$ | | | $\chi^2$ Improvement | | | Employed variable transformations | | |
| **Scale** | $\chi^2$ | df | $\chi^2$/df | $\chi^2$ | df | $\chi^2$/df | Δ | % | **Item** | None | Square | Square-root |
| Reliability of Management (Next Administrative Level) | 6292,24 | 10 | 629,2 | 3502,237 | 10 | 350,2 | 2790,01 | 44,34 | v08_001 | | x | |
| | | | | | | | | | v08_002 | x | | |
| | | | | | | | | | v08_003 | x | | |
| | | | | | | | | | v08_004 | | x | |
| | | | | | | | | | v08_005 | | | x |
| Reliability of Management (Own Unit) | 4392,21 | 10 | 439,2 | 1903,358 | 10 | 190,3 | 2488,85 | 56,67 | v06_001 | | x | |
| | | | | | | | | | v06_003 | | x | |
| | | | | | | | | | v06_006 | | x | |
| | | | | | | | | | v06_009 | x | | |
| | | | | | | | | | v06_012 | | x | |
| Role Conflict | 3511,79 | 8 | 439,0 | 1307,991 | 8 | 163,5 | 2203,79 | 62,75 | v03_002 | | x | |
| | | | | | | | | | v03_006 | | x | |
| | | | | | | | | | v03_007 | | x | |
| | | | | | | | | | v03_011 | x | | |
| Role Overload | 684,23 | 6 | 114,0 | 145,97 | 6 | 24,3 | 538,26 | 78,67 | v03_004 | | | x |
| | | | | | | | | | v03_010 | | | x |
| | | | | | | | | | v03_013 | | | x |
| Social Climate | 2220,92 | 6 | 370,2 | 396,889 | 6 | 66,1 | 1824,03 | 82,13 | v01_008 | | x | |
| | | | | | | | | | v01_016 | | x | |
| | | | | | | | | | v01_018 | | x | |
| Social Community at Work | 3136,59 | 6 | 522,8 | 282,382 | 6 | 47,1 | 2854,21 | 91,00 | v01_001 | | x | |
| | | | | | | | | | v01_009 | | x | |
| | | | | | | | | | v01_010 | | x | |
| Social Support from Supervisor | 2038,49 | 6 | 339,7 | 195,522 | 6 | 32,6 | 1842,97 | 90,41 | v07_005 | | x | |
| | | | | | | | | | v07_006 | | x | |
| | | | | | | | | | v07_009 | | x | |
| Task Completion Ambiguity | 2734,25 | 6 | 455,7 | 119,766 | 6 | 20,0 | 2614,49 | 95,62 | v03_001 | | x | |
| | | | | | | | | | v03_005 | | x | |
| | | | | | | | | | v03_008 | | x | |
| Trust Regarding Management | 3605,30 | 8 | 450,7 | 249,121 | 6 | 41,5 | 3356,18 | 93,09 | v06_002 | | x | |
| | | | | | | | | | v06_004 | | x | |
| | | | | | | | | | v06_008 | | x | |
| | | | | | | | | | v06_011 | | x | |
| UWES (Absorption) | 3674,18 | 6 | 612,4 | 540,406 | 6 | 90,1 | 3133,77 | 85,29 | v11_007 | | x | |
| | | | | | | | | | v11_008 | | x | |
| | | | | | | | | | v11_009 | | x | |
| UWES (Dedication) | 4969,25 | 6 | 828,2 | 1401,773 | 6 | 233,6 | 3567,47 | 71,79 | v11_004 | | x | |
| | | | | | | | | | v11_005 | | x | |
| | | | | | | | | | v11_006 | | x | |
| UWES (Vigor) | 5929,81 | 6 | 988,3 | 1755,552 | 6 | 292,6 | 4174,26 | 70,39 | v11_001 | | x | |
| | | | | | | | | | v11_002 | | x | |
| | | | | | | | | | v11_003 | | x | |
| Work-Family Conflict | 412,97 | 8 | 51,6 | 249,69 | 6 | 41,6 | 163,28 | 39,54 | v09_002 | | | x |
| | | | | | | | | | v09_007 | | | x |
| | | | | | | | | | v09_009 | x | | |
| | | | | | | | | | v09_015 | x | | |
| Work-Family Facilitation | 579,88 | 8 | 72,5 | 219,077 | 6 | 36,5 | 360,80 | 62,22 | v09_004 | x | | |
| | | | | | | | | | v09_005 | x | | |
| | | | | | | | | | v09_011 | x | | |
| | | | | | | | | | v09_013 | | x | |
| Work SoC (Comprehensibility) | 2911,93 | 8 | 364,0 | 466,262 | 8 | 58,3 | 2445,67 | 83,99 | v12_001 | | | x |
| | | | | | | | | | v12_003 | | | x |
| | | | | | | | | | v12_006 | | | x |
| | | | | | | | | | v12_009 | | | x |
| Work SoC (Manageability) | 779,61 | 4 | 194,9 | 83,742 | 4 | 20,9 | 695,87 | 89,26 | v12_004 | | | x |
| | | | | | | | | | v12_007 | | | x |
| Work SoC (Meaningfulness) | 7458,87 | 6 | 1243,1 | 1764,953 | 4 | 441,2 | 5693,92 | 76,34 | v12_002 | | x | |
| | | | | | | | | | v12_005 | | x | |
| | | | | | | | | | v12_008 | | x | |

Note: For Δ, a positive value represents an improvement in terms of multivariate normality as measured by the $\chi^2$ statistic.
An "x" denotes the version of the variable that is employed in the analysis of multivariate normality.

**Confirmatory Factor Analysis**

CFA (Brown, 2015; Brown & Moore, 2012; Jöreskog, 1969) is a particular application of SEM (Bollen & Long, 1993; Hoyle, 2012; Kline, 2016; MacCallum & Austin, 2000) and is a frequently employed method in the service of validating measurement interpretations. It is a hypothesis-driven statistical modelling tool for testing measurement theories positing that observations are manifestations of hypothetical entities (i.e., for latent variable modelling; Bollen, 2002; Bollen & Lennox, 1991; Borsboom, 2008; Edwards & Bagozzi, 2000). As such, it constitutes the initial specification and the subsequent evaluation of the measurement model of a SEM analysis, specifying which observed variables constitute measures of which latent variables, based on theoretical expectations or prior research.

With CFA, where data is analyzed based on theory and evidence, one investigates whether a priori theories comprehensively and parsimoniously account for observations. As such, CFA conforms with the Standards by allowing researchers to investigate the degree of justification for pre-specified measurement interpretations based on prior theory and research. CFA allows researchers to examine several of the sources of validity-evidence specified in the standards relevant to measurement interpretations – most readily that of "internal structure." Depending on the design of the study, it can furthermore be used to examine evidence pertaining to the "response process" category by modelling theoretical and hypothesized sources of systematic error in measurement, allowing one to more closely examine evidence pertaining to the issue of construct irrelevant variance (e.g., Podsakoff et al., 2003; Podsakoff et al., 2012). In short, CFA allows one to examine sources of evidence the Standards label as response processes, internal structure, and relations to other variables, and thus to validate claims for which establishing these sources of evidence are necessary and sufficient.

The outcome of a CFA is evaluated primarily by means of global fit indices, providing indications of how well the model does at reproducing observations. Inadequate fit constitutes falsification of the measurement theory, and adequate fit can conversely be taken as evidence in favor of the proposed measurement interpretation. Secondary to examining global fit comes the examination of what Brown (2015) succinctly label "localized areas of strain" within the model, which involves investigating specific instances of poor fit. As a specified CFA model constitutes a specification of a preferred interpretation, it can be seen as a formal specification of Kane's (2013a) IUA. An estimated model can in turn be considered a formal specification of the VA, and the IUA can formally be considered valid to the extent that the parameter estimates of the model satisfy the necessary criteria.

**Assessing overall model fit by means of global fit indices.**

The $\chi^2$ statistic is the most widely used summary statistic for evaluating model fit, assessing the discrepancy between the sample and fitted covariance matrices. Due to the restrictive nature of the $\chi^2$ statistic as a test of exact-fit sensitive to sample size and model complexity (it becomes increasingly unlikely to attain a good fit as sample size or number of observed variables increases), a number of ancillary "approximate" (or descriptive) indices have been developed to supplement the $\chi^2$ (Scherermelleh-Engel, Moosbrugger, & Müller, 2003).

Many scholars would consider models to be simplifications or approximations of reality, and would for this reason consider the null-hypothesis that the model fits the data exactly as dismissible a priori (e.g., Jones & Tukey, 2000; Mislevy, 2009). Tests of exact fit will treat even the slightest deviations of fit as certain as the sample size grows large enough (in a sense merely confirming what is already know to be true), and does not provide useful information regarding a models degree of misfit. This sentiment is captured by the following quote from Box and Draper (1987, p. 74; cited in Mislevy, 2009, p. 84): "*All models are wrong; the practical question is how wrong do they have to be to not be useful.*" Approximate fit indices provide information regarding the degree of model fit with the observed data, and is not associated with the classic hypothesis-testing approach of exact fit.

Approximate fit indices can be classified along a range of dimensions that describe and determine their properties, and different indices thus supply researchers with different information regarding aspects of model fit. As stated by Kline (2016, p. 264); "*there is no such thing as a magical, single-number summary that says everything worth knowing about model fit.*" As different indices have different properties, it is considered good practice to make use of- and report several different indices with differing properties when evaluating model fit, as they provide different information regarding the fit of the model.

When presenting fit indices, most scholars appear to make use of the absolute vs. relative dimension as the primary category of classification (e.g., Hu & Bentler, 1999; West, Taylor, & Wu, 2012), while others would consider whether indices are parsimony-adjusted as an additional notable attribute (e.g., Brown, 2015; Hair et al., 2013). Another attribute of indices which some authors include as noteworthy is whether an index is population or sample based (e.g., Kline, 2016, albeit in special cases). Still others find it prudent to mention whether an index is scaled as a goodness or badness of fit (whether higher and lower values indicate better or worse fit; e.g., West et al., 2012).

Absolute approximate fit indices are similar to the classical $\chi^2$ test in that they assess how well an a priori model reproduces the sample data. Relative (also known as incremental

or comparative) indices measure the proportionate improvement in fit by comparing a target model with a restricted nested baseline model where one assumes no relationship between observed variables (i.e., the worst model imaginable). Parsimony adjusted indices are formed in such a manner that they tend to favor models with fewer free parameters by considering fit relative to model complexity.

The rationale underlying parsimony adjustment is twofold; (1) parsimony is generally considered a desirable attribute of models that should be encouraged, and (2) fit as measured by non-parsimony-adjusted indices tend to improve by adding parameters to the model, and the allure of adding parameters to a model simply to improve its fit is considered a practice to be discouraged (e.g., MacCallum, Roznowski, & Necowitz, 1992; Steiger, 1990). Predictive (population based) indices estimate model fit in hypothetical replication samples of the same size randomly drawn from the samples' population. A list of a variety of property-determining dimensions fit indices can belong to is available in table 3.5.

***Choosing fit indices.***

According to Hair et al. (2014), a researcher is faced with two basic questions in selecting measures of model fit, the answers to which, supposedly, are neither simple nor straight-forward. The first of these is what constitutes the best fit indices to objectively reflect the fit of a model, and the second being which cutoff values suggest a good model fit for any given index. A set of indices have emerged as generally accepted due to them having been proved to be "well-behaved" across a variety of circumstances, as well as due possessing desirable

Table 3.5.
*Definitions of a Selection of Dimensions along which Fit Indices can Vary.*

| Dimension. | Definition. |
|---|---|
| 1. Population vs. sample based | Population-based fit indices estimate a known population parameter; sample-based fit indices describe the data-model fit in the observed sample at hand. |
| 2. Simplicity vs. complexity | Fit indices that favor simple models penalize models in which many parameters are estimated; fit indices that do not employ such a correction do not penalize for model complexity. |
| 3. Normed vs. non-normed | Fit indices that are normed are constructed to lie within an approximate (0, 1) range; non-normed fit indices do not necessarily lie in this range. |
| 4. Absolute vs. relative | Relative fit indices are defined with respect to a specific model that serves as an anchor for subsequent model comparisons; absolute fit indices do not employ such a comparison anchor. |
| 5. Estimation method free vs. estimation method specific | Estimation method-free fit indices provide characterizations of model fit that are unaffected by the choice of a specific estimation method; estimation method-specific fit indices provide different fit summaries across different methods of estimation. |
| 6. Sample size independent vs. sample size dependent | Sample-size-independent fit indices are not affected by sample size, either directly or indirectly; sample-size-dependent fit indices vary as a function of observed sample size. |
| 7. Goodness of fit scaled vs. badness of fit scaled. | Greater values indicate better model fit for indices scaled as goodness-of-fit, while greater values indicate poorer fit for indices scaled as badness-of-fit. |

Note: Adopted from Tanaka (1993, p. 16), and adapted to include the dimension whether an index is scaled as goodness or badness of fit.

properties. A set of contemporarily popular indices are presented and described in table 3.6 in terms of how they relate to the dimensions presented in table 3.5.

There appears to be close to universal agreement that the SRMR, the RMSEA, and the CFI constitute a bare minimum of indices that should be reported (e.g., Brown, 2015; Kline, 2016). The AIC and the BIC are considered useful when comparing nested, nonhierarchical models. These indices are not intuitively meaningful on their own regarding the degree of model fit or misfit unless compared across models, leading some scholars to label them "model selection indices" rather than "approximate fit indices" (e.g., West et al., 2012).

***Evaluating models based on global fit indices.***

With regard to Hair et al.'s (2013) second question – which cutoff values suggest a good model fit for any given index – there is less of a consensus to be found than regarding which particular indices that should be employed, and is subject to much greater controversy. Approximate indices are measures of closeness of fit, and by employing them one has already conceded that the model does not perfectly account for the data and thus, in a sense, is false. The question then becomes which value on any given index indicates an unacceptable level of model misfit. Concerning the consensus with regard to which range of values represent good, adequate, or bad fit, one can distinguish between pre- and post- Hu and Bentler (1999).

Before Hu and Bentler (1999) conducted their simulation analyses, the consensus was that an SRMR between .05 and .1 constituted adequate fit, and that values < .05 indicated good fit. For incremental indices such as CFI and TLI, values > .9 was considered to indicate good fit. For RMSEA values < .05 indicated close fit, and values < .08 indicated adequate fit. In hoping to establish an empirical basis for the selection of cutoff criteria, Hu and Bentler

Table 3.6.
*Category Membership of a Selection of Contemporarily Popular Fit Indices as Pertaining to a Selection of Dimensions.*

| **Index** | Parsimony-adjusted | Normed to approx. (0-1) interval | Relative (vs. absolute) | Sample size sensitive | Scaled as goodness (vs. badness) of fit |
|---|---|---|---|---|---|
| $\chi^2$ | | | | X | |
| $\chi^2$/df | X | | | X | |
| RMSEA | X | X | | | |
| TLI* | X | X | X | | X |
| CFI | X | X | X | | X |
| AIC | X | | | X | |
| BIC* | X | | | X | |
| SRMR | | X | | X | |

Note: Based on tables from Tanaka (1993) and West et al. (2012). An X denotes that the index possesses the column attribute.
* Particularly strong parsimony adjustment relative to similar indices (e.g., greater adjustment for TLI than CFI, and for BIC than AIC).

(1999) performed a series of simulation studies in order to investigate which cutoff criteria for different fit indices resulted in the most consistent acceptance of "true" models and rejection of "false" models. Based on this study, they wound up suggesting that researchers employ a two-index strategy employing combinational rules with a cutoff value of ≥ .95 for CFI or TLI combined with a cutoff value of < .09 for SRMR for model acceptance. Combination rules of RMSEA ≤ .05 and an SRMR ≤ .06 resulted in "acceptable" type II error rates for both simple and complex misspecified models under both robustness and nonrobustness conditions, and combinational rules of an RMSEA ≤ .06 and an SRMR ≤ .09 resulted in the least sum of type I and type II error rates (Hu & Bentler, 1999).

Since the publication of Hu and Bentlers (1998; 1999) influential articles, these cutoff criteria have been widely adopted as rules of thumb. However, the blind acceptance of a set of thresholds for model acceptance and rejection has been criticized on a number of points. For example, it implicitly reintroduces the notion of a "true" model in a strict sense of the word. As noted by Marsh, Hau, and Wen (2004, p. 322) in commenting on Hu and Bentlers studies, if this was appropriate, "*for normally distributed data, a traditional maximum likelihood (ML) chi-square test would have outperformed all of their GOF indexes in relation to their stated purpose of optimally identifying a misspecified model.*" West et al. (2012) furthermore points out that Hu and Bentlers accept-reject criteria varied dramatically as a function of the type of misspecification and data characteristics. This lead them to conclude that the proposals of Hu and Bentler (1999) should only be interpreted as rough guidelines.

Acknowledging the limitations of relying exclusively on a set of pre-specified cutoff criteria for fit indices in model rejection or acceptance, a number of alternative strategies for model evaluation has been proposed. Among these are investigating the fit of the components that make up the model (e.g., investigating localized areas of strain, standardized residuals, modification indices, and identifying unnecessary parameters; Brown, 2015), and comparing the relative fit of alternative theoretically plausible models in accounting for the same data (i.e., comparing nested models; Brown, 2015; Jöreskog, 1993; Kline, 2016; West et al., 2012)

**Beyond overall model fit: convergence and discrimination of factors.**

Secondary to evaluating the overall fit of the model comes the more conventional procedures of construct validation, where one examine the performance of each individual construct measure included in the CFA. This approach to validation has its roots in the work of Fornell and Larcker (1981), where classical construct validity theoretical thinking and terminology (e.g., Campbell & Fiske, 1959; Cronbach & Meehl, 1955) is applied and adapted to SEM.

This evaluation and validation procedure includes examining convergent and discriminant evidence of validity for the constructs included in the model.

Evidence of convergence is investigated by examining the reliability of indicators assumed to measure the same latent variable (i.e., internal consistency), as well as examining the amount of variance explained in the indicators by the specified entity (i.e., communality, or the average variance extracted). Evidence of discrimination is investigated by examining the degree of correlation between factors, and whether the amount of variance accounted for in the indicators by their latent factors exceeds the amount of variance in a latent variable is accounted for by other estimated latent variables (i.e., when convergence ≥ discrimination; Fornell & Larcker, 1981; Hair et al., 2013; Mehmetoglu & Jakobsen, 2017).

According to the Standards, the reliability of test scores have implications for validity as it "*ultimately bears on the generalizability or dependability of the scores and/or the consistency of classifications of individuals derived from the scores*" (AERA et al., 2014, p. 34). In CFA, reliability constitutes convergent evidence of validity, as it represents evidence that the measures converge on the same latent variable. Reliability in CFA is not examined by means of measures such as the traditional Cronbachs α, but rho (ρ) based measures that do not assume tau (τ) equivalence (that all indicators load equally on their common factor). The reason for this is that if the assumption of τ equivalence is violated, α will underestimate reliability, while ρ based measures will provide more accurate estimates (Raykov, 1997). The composite reliability measure (CR; Fornell & Larcker, 1981; Hair et al., 2013) or Raykov's rho (Raykov, 1997) are examples of such measures.

By convention, a measure of a latent variable is considered reliable (or internally consistent) when the ρ score is, like α, either equal to or greater than a value of .7 (see Lance, Butts, & Michels, 2006 for a historical review and critique of this convention). Convergent evidence of validity is further established if a modelled latent variable achieves an "average variance extracted" (AVE) score from its indicator variables ≥ .5, and is calculated as the mean variance extracted for the items loading on a latent variable and is as such a summary convergence indicator. This criterion is generally more conservative than the ρ ≥ .7 criterion, and the rationale behind this threshold is that if it is satisfied, the latent variable factor explains more variance in its corresponding manifest variables than it leaves unexplained. An acceptable AVE is achieved when the average factor loadings (the standardized regression coefficient) of a constructs indicators approach a value of .71 (i.e., $\overline{\beta} \geq .5^2$).

Discriminant evidence of validity is concerned with the extent to which a factor represents something distinct from other factors, and is conventionally considered to be

established if the AVE of a latent variable is higher than its "maximum shared variance" (MSV) score, which is the factors' maximum squared correlation with any other factor included in the CFA. This means that the factor in question shares more variance with its associated indicators than with any other factor included in the CFA (Fornell & Larcker, 1981; Hair et al., 2013). The AVE ≥ MSV decision rule is a very conservative criterion for discrimination demarcation. From the perspective of construct validity theory (as opposed to criterion validity theory), strong correlations are not necessarily to be interpreted as degrees of equivalence, but that does not exclude equivalence as a possible explanation (Borsboom et al., 2004; Campbell & Fiske, 1959). According to Brown (2015), inter-factor correlations exceeding a value of .8 (corresponding to MSV > .64) is often considered indicative of poor substantive (as opposed to statistical) discrimination.

**Model Specification**

The models were specified in the statistical software program STATA14. Due to the ambiguous specifications of some of the latent variables intended to be measured by the KIWEST instrument – where some of the constructs are specified as multifaceted (i.e., DUWAS, UWES, Work Family Balance and Work-SoC) – the measurement interpretation offered by the KIWEST theory was deemed ambiguous (Kagan, 2005; McGrath, 2005a). For this reason, a somewhat exploratory strategy of analysis (i.e., the "alternative models" strategy of Jöreskog, 1993) was adopted.

Sixteen latent variable models (i.e., reflective measurement models), each representing an interpretation of the KIWEST theory, were specified so that they could be estimated and

Table 3.7.
*Overview of model factor solutions.*

| | Multifaceted constructs included in the KIWEST 2.0 survey | | | |
|---|---|---|---|---|
| **Model.** | DUWAS | UWES | Work-Family Balance | Work SoC |
| Model 1 | | | | |
| Model 2 | | | | X |
| Model 3 | | | X | |
| Model 4 | | | X | X |
| Model 5 | | X | | |
| Model 6 | | X | | X |
| Model 7 | | X | X | |
| Model 8 | | X | X | X |
| Model 9 | X | | | |
| Model 10 | X | | | X |
| Model 11 | X | | X | |
| Model 12 | X | | X | X |
| Model 13 | X | X | | |
| Model 14 | X | X | | X |
| Model 15 | X | X | X | |
| Model 16 | X | X | X | X |

Note: The remaning twenty-three constructs retain their proposed one-factor solution across all models. Blank spots represent one-factor solutions of the multifaceted constructs factor structure. X's represent multi-factor structure for the respective constructs (2 vs 1 factors for DUWAS and Work-Family Balance, 3 vs 1 factors for UWES and Work SoC).

compared against each other (in combination representing all possible permutations of single- and multifactor solutions to the hypothesized constructs included in KIWEST deemed vague with respect to their factor structures; see table 3.7). An example of one of the specified models (model 16) is illustrated in figure 3.2 (p. 62). The first item pertaining to each factor was specified as marker indicators (e.g., the uppermost indicators of each factor in figure 3.2), none of the error terms were permitted to correlate, and no factor cross-loadings were specified. Besides constraining the loadings of the marker indicators to a value of one, no additional constraints were specified.

**Model Estimation**

The remaining sample following data integrity analysis and treatment consisted of 7643 cases and 118 items, and it constitutes the sample employed for the estimations of the sixteen latent variable measurement models. A description of the post-treatment sample in terms of category memberships is available in table 3.8, where the effect of the integrity treatment on the sample size as pertaining to category memberships is described with Δ, representing the pre- minus post-data integrity treatment sample sizes. Due to imputation failure, 217 cases were excluded from estimations, meaning that the CFA is a post-imputation full-case analysis. The sixteen specified models were estimated with STATA14, making use of ML employing the sample variance-covariance matrices. Due to the general apparent success of multivariate normality treatment (see table 3.4, pp. 52-53), ML was judged an available option for model estimation, and thus the appropriate choice as it is generally considered to be the superior method (Hu & Bentler, 1999).

Table 3.8.
*Estimation Sample Description, Sorted by Type of Position and Gender.*

| | Gender | | | |
|---|---|---|---|---|
| Type of position | Female (Δ) | Male (Δ) | Other (Δ) | Total (Δ) |
| Academic (Δ) | 1408 (-33) | 1776 (-41) | 0 (-1) | 3185 (-75) |
| Doctoral research fellow (Δ) | 804 (-25) | 582 (-18) | 0 (-0) | 1386 (-43) |
| Technical/administrative (Δ) | 1629 (-59) | 1085 (-31) | 0 (-0) | 2714 (-90) |
| Unit leader, level 1 (Δ) | 1 (-0) | 2 (-0) | 0 (-0) | 3 (-0) |
| Unit leader, level 2 (Δ) | 14 (-0) | 14 (-0) | 0 (-0) | 28 (-0) |
| Unit leader, level 3 (Δ) | 82 (-2) | 133 (-2) | 0 (-0) | 215 (-4) |
| Unit leader, level 4 (Δ) | 40 (-1) | 55 (-4) | 0 (-0) | 95 (-5) |
| Unit leader, level 5 (Δ) | 4 (-0) | 13 (-0) | 0 (-0) | 17 (-0) |
| Total (Δ) | 3952 (-120) | 3660 (-96) | 0 (-1) | 7643 (-217) |

Note: Category membership pertaining to type of position and gender is retreived from registry data.
(Δ) represents the difference between pre- and post data integrity treatment sample size.

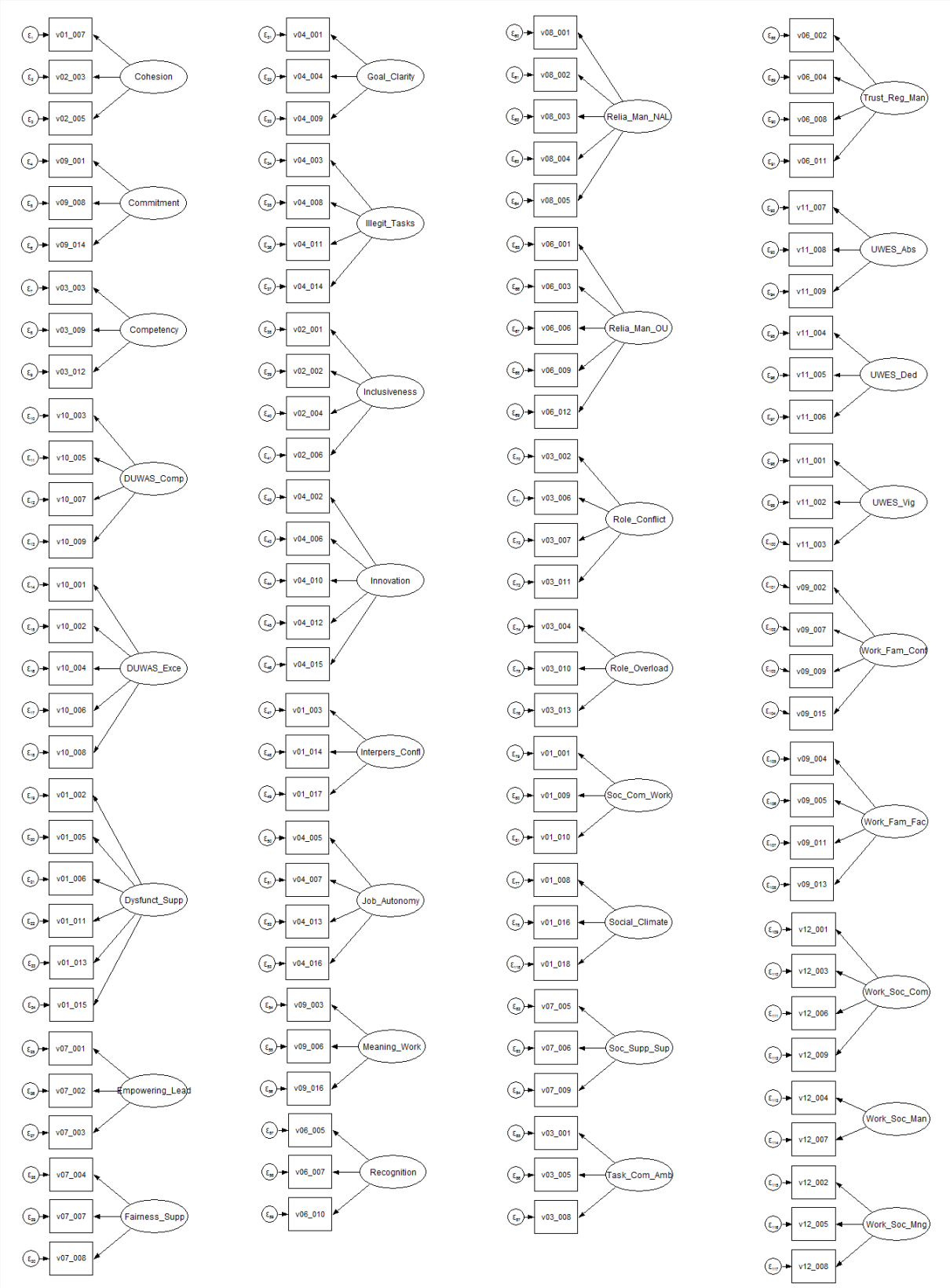


*Figure 3.2.* Diagram presentation of model 16 as an example. Paths representing inter-factor correlations have been omitted in order for the figure not to assume the appearance of a toddlers rendition of ovis aries. In the specified models, all of the specified latent variables and none of the error terms are correlated. Circles represent error terms, rectangles represent observed variables (i.e., items), and ovals represent factors.

## Results

Table 4.1 summarizes the outcome of all of the analyses. Little's MCAR test indicated that there are systematic causes of missing entries at play – and thus sources of construct irrelevant variance. More fine-grained analyses implied that subgroups of respondents pleaded the 5th on items of 11 of the 33 possible scales included in KIWEST (see table 3.3, pp. 48-49; and table 4.1). Missing entries were treated by means of ordinal logistic regression imputation, with no guarantee that this would not introduce bias in the data. Attempts were made to reduce bias by including category memberships of respondents in the estimations. However, the possibility that bias has been introduced remains, reducing confidence in interpretations involving the affected items and factors, as it might lead to overestimation of fit.

Data normality analysis revealed that steps had to be taken to prepare the data for ML estimation. For the most part, treatments appeared to successfully normalize the multivariate distributions. However, some scales proved more resistant to normalization than others did (see table 3.5, pp. 52-53; and table 4.1), the consequence being reduced confidence in any interpretations involving them. Specifically, violations of this assumption might artificially

Table 4.1.
*Summary of Scale-Level Issues of Factors Included in the Retained Model.*

| Factor | Systematic Missing Values ($p < .05$) | Multivariate Non-Normality ($\chi^2 \geq 1580$) | Insufficient Convergence ($\rho \leq .7$) | Insufficient Convergence (AVE $\leq .5$) | Insufficient Discrimination (MSV $\geq .64$) | Insufficient Discrimination (AVE < MSV) |
|---|---|---|---|---|---|---|
| Cohesion in Work Teams | X | | | X | | X |
| Commitment to the Workplace | | | | | | |
| Competency Demands | | | X | X | | |
| DUWAS (Compulsive) | | | X | X | | X |
| DUWAS (Excessive) | | | X | X | | X |
| Dysfunctional Support | X | X | | | | |
| Empowering Leadership | | | | | | X |
| Fairness of the Supervisor | X | | | | X | |
| Goal Clarity | | | | | | |
| Illegitimate Tasks | | | | X | X | X |
| Inclusiveness and Social Responsibility | X | | | X | | |
| Innovation | | | | | | |
| Interpersonal Conflicts | X | | | | X | X |
| Job Autonomy | | | | | | |
| Meaning of Work | | | | | | |
| Recognition | | | | | X | X |
| Reliability of Management (Own Unit) | | X | | | X | X |
| Reliability of Management (Next Level) | | X | | | | |
| Role Conflict | | | X | X | X | X |
| Role Overload | | | | | | |
| Social Climate | | | | | X | X |
| Social Community at Work | X | | | | X | X |
| Social Support from Supervisor | X | | | | X | X |
| Task Completion Ambiguity | | | | X | | |
| Trust Regarding Management | X | | | | X | X |
| UWES (Absorption) | | | | | | |
| UWES (Dedication) | | | | | | |
| UWES (Vigor) | | X | | | | |
| Work-Family Conflict | X | | | | | |
| Work-Family Facilitation | | | X | X | | |
| Work-SoC (Comprehensibility) | X | | | | X | X |
| Work-SoC (Manageability) | X | | | | X | X |
| Work-SoC (Meaningfulness) | | X | | | | |

Note: An X denotes that the factor demonstrates potential problems with the issue.

inflate the $\chi^2$ value, causing underestimates of goodness-of-fit (several approximate fit indices make use of it in their calculations; Andreassen, Lorentzen, & Olsson, 2006).

Building on the data integrity analysis and treatment, the CFA was conducted, which suggested (from an exploratory point of view) that model 16 appeared to represent the most appropriate interpretation of the KIWEST measurement theory. Further analyses revealed that out of the 33 factors representing theoretically distinct latent variables, five failed to converge according to the most lenient criterion (i.e., the factor lacks comprehensiveness in accounting for observed item variation). Of the factors that managed to attain convergence, twelve failed to satisfy the most lenient criterion of discrimination (i.e., the factors demonstrate lack of parsimony in accounting for observed item variation). The thesis now turn to accounting for model selection and analyses of factor convergence and discrimination in more detail.

**Analysis and Evaluation of Overall Model Fit, and Model Selection**

The estimation of models was executed without problems relating to the convergence of solutions. Table 4.2 displays how each of the models fared in terms of their overall goodness-of-fit. We see that none of the models satisfy the strictest criteria, which is a non-significant $\chi^2$ test. It was not expected that any model would satisfy this criteria, as the models are complex and the sample size is large. As for the second most strict criteria of fit, none of the models measure up to those proposed by Hu and Bentler (1999). That is, none of the models achieve a conservative index value combination of SRMR < .06, RMSEA ≤ .05, and CFI or TLI ≥ .95.

Specifically, relatively poor performances on the comparative fit indices appear to hold back the fit of the models, indicating that the models leaves something to be desired in terms of their fit relative to a baseline model assuming no relationships. The RMSEA value is

Table 4.2.
*Model Evaluation and Comparison.*

| | $\chi^2$ | | | Absolute fit | | Incremental fit | | Information criteria | |
|---|---|---|---|---|---|---|---|---|---|
| **Model.** | $\chi^2$ | df | $\chi^2$/df | SRMR | RMSEA | CFI | TLI | AIC | BIC |
| Model 1 | 87838,91 | 6434 | 13,65 | 0,065 | 0,041 | 0,871 | 0,861 | 4397000 | 4401000 |
| Model 2 | 76029,45 | 6379 | 11,92 | 0,062 | 0,038 | 0,889 | 0,880 | 4385000 | 4390000 |
| Model 3 | 82480,40 | 6407 | 12,87 | 0,060 | 0,039 | 0,879 | 0,870 | 4391000 | 4396000 |
| Model 4 | 70654,53 | 6350 | 11,13 | 0,057 | 0,036 | 0,898 | 0,889 | 4379000 | 4385000 |
| Model 5 | 81203,71 | 6379 | 12,73 | 0,065 | 0,039 | 0,881 | 0,871 | 4390000 | 4395000 |
| Model 6 | 69352,15 | 6320 | 10,97 | 0,062 | 0,036 | 0,900 | 0,891 | 4378000 | 4384000 |
| Model 7 | 75828,11 | 6350 | 11,94 | 0,060 | 0,038 | 0,890 | 0,880 | 4385000 | 4390000 |
| Model 8 | 63962,23 | 6289 | 10,17 | 0,057 | 0,035 | 0,908 | 0,899 | 4373000 | 4379000 |
| Model 9 | 86136,26 | 6407 | 13,44 | 0,064 | 0,040 | 0,873 | 0,863 | 4395000 | 4400000 |
| Model 10 | 74295,89 | 6350 | 11,52 | 0,061 | 0,037 | 0,892 | 0,883 | 4383000 | 4389000 |
| Model 11 | 80777,11 | 6379 | 12,66 | 0,058 | 0,039 | 0,882 | 0,872 | 4390000 | 4395000 |
| Model 12 | 68919,45 | 6320 | 10,90 | **0,055** | 0,036 | 0,901 | 0,891 | 4378000 | 4384000 |
| Model 13 | 79473,25 | 6350 | 12,51 | 0,063 | 0,039 | 0,884 | 0,874 | 4388000 | 4394000 |
| Model 14 | 67586,35 | 6289 | 10,75 | 0,061 | 0,036 | 0,903 | 0,893 | 4377000 | 4382000 |
| Model 15 | 74097,52 | 6320 | 11,72 | 0,058 | 0,037 | 0,892 | 0,882 | 4383000 | 4389000 |
| Model 16 | **62195,53** | 6257 | **9,94** | **0,055** | **0,034** | **0,911** | **0,902** | **4371000** | **4377000** |

Note: All $\chi^2$ tests of exact fit significant at the $p < 0,001$ level. Bolded values represent the best fit-value attained for a given index.

within acceptable limits for all of the models (with model 16 demonstrating the best fit on the index), and the SRMR value is acceptable for six of the models (model 4, 8, 11, 12, 15, and 16, where model 12 and 16 are tied for the lead). None of the models attain values on the CFI and TLI indices satisfying Hu and Bentlers criteria however. Turning to the most liberal criteria – the pre-Hu and Bentler criteria of CFI and TLI values ≥ .9 – five of the models (6, 8, 12, 14 and 16) attains a CFI within acceptable bounds (model 16 once again demonstrating the greatest performance). Only model 16 attains a TLI value satisfying this relatively liberal criteria (model 8 does however come close). On the model selection indices (i.e., information criteria), model 16 outperforms every other model on both the AIC and the BIC indices.

As such, it appears (from an exploratory perspective) that model 16 best represents the latent variable measurement interpretation of the KIWEST theory, followed most closely by model 8 and 12. Model 16 represents the least parsimonious KIWEST theory latent variable interpretation, yet it outperforms every other model on even the most strongly parsimony-adjusted indices (i.e., on the TLI and the BIC indices). Furthermore, model 1 – representing the most parsimonious interpretation – unequivocally demonstrated the worst fit out of all the models, closely followed by model 9 and 5.

Keeping this in mind – that some more parsimonious solutions come close to rivaling the fit of model 16 – the thesis proceeds with evaluating model 16 in more detail. In moving on, we keep in mind that the model demonstrated SRMR and RMSEA values that satisfy even the conservative thresholds of Hu and Bentler (1999), though not uniquely so. It was however the only model satisfying the more lenient, classic criteria for the CFI and TLI indices, though not exclusively for CFI, and only by a hairs breadth for TLI. Furthermore, fit might have been artificially inflated by imputation treatment, and deflated by multivariate non-normality. Overall, the retained model can be considered provisionally acceptable, and thus, hypothesis 1a and 1b can be considered supported. The results do however indicate that improvements might be made to the model, and we keep this in mind as we turn to more fine-grained evaluation of localized areas of misfit in the model.

**Retained Model Convergent and Discriminant Evidence of Validity**

The evaluation of convergent and discriminant evidence was done in accordance with the procedure sketched pages 58-60. Scale ρ values were computed employing the user written STATA module “relicoef” (Mehmetoglu & Jakobsen, 2017), which computes Raykov’s ρ (Raykov, 1997). Scale AVE scores were in turn computed making use of the user written

Table 4.3.
*Convergent and Discriminant Evidence of Validity.*

| | Scale level | | | | | Item level | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Convergent | | Discriminant | | | | | 95% Confidence Interval | |
| **Scale** | ρ | AVE | MSV | AVE - MSV | **Item** | β | SE | Lower | Upper |
| Cohesion in Work Teams | 0,733 | 0,482 | 0,503 | -0,021 | v01_007 | 0,474 | 0,010 | 0,455 | 0,494 |
| | | | | | v02_003 | 0,815 | 0,005 | 0,804 | 0,826 |
| | | | | | v02_005 | 0,746 | 0,005 | 0,733 | 0,758 |
| Commitment to the Workplace | 0,816 | 0,589 | 0,517 | 0,072 | v09_001 | 0,828 | 0,005 | 0,819 | 0,838 |
| | | | | | v09_008 | 0,805 | 0,005 | 0,795 | 0,815 |
| | | | | | v09_014 | 0,658 | 0,007 | 0,644 | 0,672 |
| Competency Demands | 0,653 | 0,380 | 0,004 | 0,376 | v03_003 | 0,616 | 0,011 | 0,594 | 0,637 |
| | | | | | v03_009 | 0,742 | 0,011 | 0,721 | 0,762 |
| | | | | | v03_012 | 0,458 | 0,013 | 0,434 | 0,483 |
| DUWAS (Compulsive) | 0,661 | 0,418 | 0,448 | -0,030 | v10_003 | 0,707 | 0,007 | 0,693 | 0,722 |
| | | | | | v10_005 | 0,547 | 0,010 | 0,528 | 0,567 |
| | | | | | v10_007 | 0,752 | 0,007 | 0,738 | 0,766 |
| | | | | | v10_009 | 0,554 | 0,009 | 0,536 | 0,572 |
| DUWAS (Excessive) | 0,626 | 0,439 | 0,448 | -0,009 | v10_001 | 0,764 | 0,006 | 0,752 | 0,775 |
| | | | | | v10_002 | 0,708 | 0,007 | 0,695 | 0,721 |
| | | | | | v10_004 | 0,635 | 0,008 | 0,620 | 0,650 |
| | | | | | v10_006 | 0,626 | 0,008 | 0,610 | 0,641 |
| | | | | | v10_008 | 0,563 | 0,009 | 0,547 | 0,580 |
| Dysfunctional Support | 0,895 | 0,603 | 0,266 | 0,337 | v01_002 | 0,507 | 0,009 | 0,489 | 0,524 |
| | | | | | v01_005 | 0,742 | 0,006 | 0,730 | 0,753 |
| | | | | | v01_006 | 0,770 | 0,005 | 0,760 | 0,780 |
| | | | | | v01_011 | 0,777 | 0,005 | 0,767 | 0,787 |
| | | | | | v01_013 | 0,900 | 0,003 | 0,895 | 0,906 |
| | | | | | v01_015 | 0,897 | 0,003 | 0,891 | 0,902 |
| Empowering Leadership | 0,894 | 0,738 | 0,741 | -0,003 | v07_001 | 0,843 | 0,004 | 0,835 | 0,850 |
| | | | | | v07_002 | 0,882 | 0,003 | 0,876 | 0,889 |
| | | | | | v07_003 | 0,851 | 0,004 | 0,844 | 0,859 |
| Fairness of the Supervisor | 0,872 | 0,695 | 0,691 | 0,004 | v07_004 | 0,887 | 0,003 | 0,881 | 0,894 |
| | | | | | v07_007 | 0,820 | 0,003 | 0,812 | 0,829 |
| | | | | | v07_008 | 0,790 | 0,005 | 0,781 | 0,801 |
| Goal Clarity | 0,773 | 0,529 | 0,324 | 0,205 | v04_001 | 0,728 | 0,007 | 0,715 | 0,742 |
| | | | | | v04_004 | 0,787 | 0,006 | 0,775 | 0,800 |
| | | | | | v04_009 | 0,660 | 0,008 | 0,645 | 0,675 |
| Illegitimate Tasks | 0,780 | 0,472 | 0,748 | -0,276 | v04_003 | 0,661 | 0,008 | 0,646 | 0,675 |
| | | | | | v04_008 | 0,621 | 0,008 | 0,605 | 0,636 |
| | | | | | v04_011 | 0,745 | 0,007 | 0,732 | 0,757 |
| | | | | | v04_014 | 0,714 | 0,007 | 0,701 | 0,728 |
| Inclusiveness and Social Responsibility | 0,767 | 0,452 | 0,224 | 0,228 | v02_001 | 0,693 | 0,008 | 0,677 | 0,708 |
| | | | | | v02_002 | 0,681 | 0,008 | 0,665 | 0,696 |
| | | | | | v02_004 | 0,682 | 0,008 | 0,666 | 0,700 |
| | | | | | v02_006 | 0,633 | 0,009 | 0,616 | 0,650 |
| Innovation | 0,742 | 0,572 | 0,503 | 0,069 | v04_002 | 0,706 | 0,006 | 0,693 | 0,719 |
| | | | | | v04_006 | 0,838 | 0,004 | 0,830 | 0,847 |
| | | | | | v04_010 | 0,723 | 0,006 | 0,711 | 0,735 |
| | | | | | v04_012 | 0,829 | 0,004 | 0,820 | 0,837 |
| | | | | | v04_015 | 0,672 | 0,007 | 0,659 | 0,686 |
| Interpersonal Conflicts | 0,865 | 0,684 | 0,711 | -0,027 | v01_003 | 0,790 | 0,005 | 0,780 | 0,800 |
| | | | | | v01_014 | 0,836 | 0,004 | 0,828 | 0,845 |
| | | | | | v01_017 | 0,853 | 0,004 | 0,845 | 0,861 |
| Job Autonomy | 0,805 | 0,507 | 0,224 | 0,283 | v04_005 | 0,751 | 0,006 | 0,739 | 0,763 |
| | | | | | v04_007 | 0,683 | 0,007 | 0,669 | 0,697 |
| | | | | | v04_013 | 0,716 | 0,007 | 0,703 | 0,729 |
| | | | | | v04_016 | 0,696 | 0,007 | 0,682 | 0,709 |
| Meaning of Work | 0,830 | 0,619 | 0,517 | 0,102 | v09_003 | 0,804 | 0,005 | 0,794 | 0,813 |
| | | | | | v09_006 | 0,807 | 0,005 | 0,797 | 0,816 |
| | | | | | v09_016 | 0,748 | 0,006 | 0,737 | 0,760 |
| Recognition | 0,895 | 0,742 | 0,805 | -0,063 | v06_005 | 0,865 | 0,003 | 0,859 | 0,872 |
| | | | | | v06_007 | 0,822 | 0,004 | 0,813 | 0,830 |
| | | | | | v06_010 | 0,896 | 0,003 | 0,890 | 0,901 |

Note: All factor loadings significant at the $p < 0,001$ level.

Table 4.3 (continued).
*Convergent and Discriminant Evidence of Validity.*

| | Scale level | | | | | Item level | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Convergent | | Discriminant | | | | | 95% Confidence Interval | |
| **Scale** | ρ | AVE | MSV | AVE - MSV | **Item** | β | SE | Lower | Upper |
| Reliability of Management (Next Administrative Level) | 0,914 | 0,824 | 0,107 | 0,717 | v08_001 | 0,851 | 0,004 | 0,845 | 0,858 |
| | | | | | v08_002 | 0,928 | 0,002 | 0,925 | 0,932 |
| | | | | | v08_003 | 0,908 | 0,002 | 0,904 | 0,913 |
| | | | | | v08_004 | 0,930 | 0,001 | 0,926 | 0,933 |
| | | | | | v08_005 | 0,919 | 0,002 | 0,915 | 0,923 |
| Reliability of Management (Own Unit) | 0,940 | 0,792 | 1,006 | -0,214 | v06_001 | 0,882 | 0,003 | 0,876 | 0,887 |
| | | | | | v06_003 | 0,864 | 0,003 | 0,858 | 0,870 |
| | | | | | v06_006 | 0,864 | 0,003 | 0,870 | 0,888 |
| | | | | | v06_009 | 0,875 | 0,002 | 0,903 | 0,912 |
| | | | | | v06_012 | 0,920 | 0,002 | 0,917 | 0,924 |
| Role Conflict | 0,677 | 0,396 | 0,748 | -0,352 | v03_002 | 0,657 | 0,008 | 0,643 | 0,672 |
| | | | | | v03_006 | 0,645 | 0,008 | 0,630 | 0,661 |
| | | | | | v03_007 | 0,557 | 0,009 | 0,539 | 0,574 |
| | | | | | v03_011 | 0,653 | 0,008 | 0,638 | 0,668 |
| Role Overload | 0,809 | 0,587 | 0,445 | 0,142 | v03_004 | 0,766 | 0,006 | 0,754 | 0,778 |
| | | | | | v03_010 | 0,796 | 0,006 | 0,785 | 0,807 |
| | | | | | v03_013 | 0,735 | 0,007 | 0,722 | 0,748 |
| Social Climate | 0,731 | 0,612 | 0,722 | -0,110 | v01_008 | 0,782 | 0,005 | 0,772 | 0,791 |
| | | | | | v01_016 | 0,791 | 0,005 | 0,781 | 0,801 |
| | | | | | v01_018 | 0,773 | 0,005 | 0,763 | 0,784 |
| Social Community at Work | 0,844 | 0,647 | 0,724 | -0,077 | v01_001 | 0,747 | 0,006 | 0,736 | 0,759 |
| | | | | | v01_009 | 0,823 | 0,005 | 0,814 | 0,832 |
| | | | | | v01_010 | 0,839 | 0,005 | 0,830 | 0,848 |
| Social Support from Supervisor | 0,869 | 0,692 | 0,741 | -0,049 | v07_005 | 0,776 | 0,005 | 0,766 | 0,786 |
| | | | | | v07_006 | 0,837 | 0,004 | 0,829 | 0,845 |
| | | | | | v07_009 | 0,879 | 0,005 | 0,873 | 0,886 |
| Task Completion Ambiguity | 0,712 | 0,467 | 0,090 | 0,377 | v03_001 | 0,363 | 0,012 | 0,341 | 0,387 |
| | | | | | v03_005 | 0,804 | 0,008 | 0,789 | 0,820 |
| | | | | | v03_008 | 0,788 | 0,008 | 0,773 | 0,803 |
| Trust Regarding Management | 0,850 | 0,591 | 1,006 | -0,415 | v06_002 | 0,873 | 0,003 | 0,867 | 0,879 |
| | | | | | v06_004 | 0,682 | 0,007 | 0,670 | 0,695 |
| | | | | | v06_008 | 0,741 | 0,005 | 0,730 | 0,752 |
| | | | | | v06_011 | 0,767 | 0,005 | 0,757 | 0,777 |
| UWES (Absorption) | 0,858 | 0,672 | 0,516 | 0,156 | v11_007 | 0,828 | 0,006 | 0,817 | 0,838 |
| | | | | | v11_008 | 0,845 | 0,005 | 0,835 | 0,856 |
| | | | | | v11_009 | 0,786 | 0,006 | 0,774 | 0,798 |
| UWES (Dedication) | 0,908 | 0,769 | 0,516 | 0,253 | v11_004 | 0,901 | 0,003 | 0,896 | 0,906 |
| | | | | | v11_005 | 0,916 | 0,002 | 0,911 | 0,921 |
| | | | | | v11_006 | 0,810 | 0,004 | 0,801 | 0,818 |
| UWES (Vigor) | 0,879 | 0,727 | 0,605 | 0,122 | v11_001 | 0,903 | 0,003 | 0,897 | 0,909 |
| | | | | | v11_002 | 0,924 | 0,003 | 0,919 | 0,929 |
| | | | | | v11_003 | 0,716 | 0,006 | 0,703 | 0,728 |
| Work-Family Conflict | 0,717 | 0,518 | 0,169 | 0,349 | v09_002 | 0,725 | 0,007 | 0,711 | 0,738 |
| | | | | | v09_007 | 0,687 | 0,007 | 0,672 | 0,702 |
| | | | | | v09_009 | 0,721 | 0,007 | 0,707 | 0,734 |
| | | | | | v09_015 | 0,745 | 0,006 | 0,732 | 0,758 |
| Work-Family Facilitation | 0,232 | 0,347 | 0,158 | 0,189 | v09_004 | 0,700 | 0,009 | 0,683 | 0,717 |
| | | | | | v09_005 | 0,715 | 0,008 | 0,698 | 0,731 |
| | | | | | v09_011 | 0,588 | 0,010 | 0,569 | 0,606 |
| | | | | | v09_013 | 0,199 | 0,013 | 0,174 | 0,224 |
| Work SoC (Comprehens-ibility) | 0,818 | 0,531 | 0,895 | -0,364 | v12_001 | 0,667 | 0,007 | 0,653 | 0,681 |
| | | | | | v12_003 | 0,716 | 0,006 | 0,703 | 0,729 |
| | | | | | v12_006 | 0,837 | 0,005 | 0,828 | 0,846 |
| | | | | | v12_009 | 0,684 | 0,007 | 0,670 | 0,697 |
| Work SoC (Manageability) | 0,714 | 0,558 | 0,895 | -0,337 | v12_004 | 0,662 | 0,008 | 0,648 | 0,677 |
| | | | | | v12_007 | 0,823 | 0,006 | 0,811 | 0,835 |
| Work SoC (Meaningfulness) | 0,880 | 0,709 | 0,412 | 0,297 | v12_002 | 0,833 | 0,004 | 0,824 | 0,841 |
| | | | | | v12_005 | 0,827 | 0,004 | 0,818 | 0,836 |
| | | | | | v12_008 | 0,865 | 0,004 | 0,858 | 0,873 |

Note: All factor loadings significant at the $p < 0,001$ level.

STATA module "condisc" (Mehmetoglu & Jakobsen, 2017). MSV scores and AVE-MSV differential scores were computed manually, making use of output from the condisc module. Univariate item parameters (regression coefficients, standard errors and confidence intervals) were estimated with- and provided by the STATA14 built-in SEM module.

The output of the analysis in its entirety is available in table 4.3 (pp. 66-67). In brief, the analysis yielded evidence of failure of convergence for nine of the factors, and indicated that every factor is in some way involved in failures of discrimination. A Heywood case was also encountered (i.e., $\beta > 1$ for the relationship "Reliability of Management – Own Unit" ↔ "Trust Regarding Management"), an occurrence that formally invalidates the model (Brown, 2015; Wothke, 1993). Fine-grained analyses thus yield a picture that is less flattering when compared to global fit, as only 14 of the 33 factors (42.4%) satisfy all of the criteria.

**Convergent evidence of validity.**

In the methods section, two criteria for convergence were established, whereof one is generally more liberal ($\rho \geq .7$), and the other more conservative (AVE ≥ .5). Among the factors that failed to satisfy the conservative criterion, we find nine factors: Cohesion in Work Teams, Competency Demands, DUWAS (Excessive and Compulsive), Illegitimate Tasks, Inclusiveness and Social Responsibility, Role Conflict, Task Completion Ambiguity, and Work-Family Facilitation. Of these, six failed the liberal criterion: Competency Demands, DUWAS (Excessive and Compulsive), Inclusiveness and Social Responsibility, Role Conflict, and Work-Family Facilitation (see table 4.1, p. 63; and table 4.3, pp. 66-67). The "offending items" (i.e., those with $\beta$ values > .7) can also be seen in table 4.3. As hypothesis 2 required that all of the factors demonstrate convergence, it is not supported.

**Discriminant evidence of validity.**

As is the case with convergence, two criteria for factor discrimination were set; one generally conservative (AVE > MSV), and one generally liberal (MSV < .64, or r < .8). Output from the condisc module output indicate that every factor is in some way involved in failure of factor discrimination. That is, while a single factor might satisfy both of the criteria for convergence and discrimination, another factor is more strongly accounted for by that factor than its own indicators. Fifteen factors fail to measure up to the conservative criterion. These are the ones in table 4.3 (pp. 66-67) that exhibit negative AVE-MSV values. Twelve failed to satisfy the liberal criterion (see table 4.4, pp. 69-70 for the complete inter-factor correlation matrix). As hypothesis 3 required that all of the factors demonstrate discrimination, it is not supported.

Table 4.4.
*Inter-Factor Correlation Matrix for Model 16.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Cohesion in Work Teams | **0,694** | | | | | | | | | | | | | | | | |
| 2. Commitment to the Workplace | 0,465 | **0,767** | | | | | | | | | | | | | | | |
| 3. Competency Demands | 0,033 | 0,097 | **0,616** | | | | | | | | | | | | | | |
| 4. DUWAS (Compulsion) | 0,053 | 0,028 | 0,122 | **0,647** | | | | | | | | | | | | | |
| 5. DUWAS (Excessive) | 0,021 | 0,000 | 0,195 | 0,669 | **0,663** | | | | | | | | | | | | |
| 6. Dysfunctional Support | 0,256 | 0,240 | 0,005 | 0,062 | 0,014 | **0,777** | | | | | | | | | | | |
| 7. Empowering Leadership | 0,415 | 0,380 | 0,063 | 0,015 | 0,000 | 0,175 | **0,859** | | | | | | | | | | |
| 8. Fairness of the Supervisor | 0,422 | 0,358 | 0,025 | 0,026 | 0,006 | 0,209 | 0,750 | **0,834** | | | | | | | | | |
| 9. Goal Clairity | 0,423 | 0,358 | 0,016 | 0,067 | 0,025 | 0,207 | 0,280 | 0,291 | **0,727** | | | | | | | | |
| 10. Illegitimate Tasks | 0,279 | 0,317 | 0,001 | 0,182 | 0,131 | 0,342 | 0,195 | 0,297 | 0,451 | **0,687** | | | | | | | |
| 11. Inclusiveness and Social Resp. | 0,473 | 0,329 | 0,033 | 0,019 | 0,001 | 0,304 | 0,308 | 0,385 | 0,215 | 0,252 | **0,672** | | | | | | |
| 12. Innovation | 0,709 | 0,465 | 0,048 | 0,057 | 0,013 | 0,271 | 0,445 | 0,407 | 0,391 | 0,313 | 0,368 | **0,756** | | | | | |
| 13. Interpersonal Conflicts | 0,481 | 0,320 | 0,002 | 0,085 | 0,040 | 0,516 | 0,247 | 0,339 | 0,270 | 0,378 | 0,334 | 0,448 | **0,827** | | | | |
| 14. Job Autonomy | 0,332 | 0,465 | 0,112 | 0,033 | 0,000 | 0,259 | 0,370 | 0,338 | 0,369 | 0,380 | 0,349 | 0,439 | 0,267 | **0,712** | | | |
| 15. Meaning of Work | 0,259 | 0,719 | 0,168 | 0,002 | 0,041 | 0,143 | 0,241 | 0,191 | 0,321 | 0,207 | 0,183 | 0,249 | 0,132 | 0,400 | **0,787** | | |
| 16. Recognition | 0,456 | 0,488 | 0,033 | 0,041 | 0,003 | 0,293 | 0,548 | 0,568 | 0,357 | 0,319 | 0,404 | 0,525 | 0,370 | 0,473 | 0,292 | **0,861** | |
| 17. Reliability of Management (NL) | 0,244 | 0,248 | 0,013 | 0,027 | 0,009 | 0,109 | 0,202 | 0,231 | 0,174 | 0,184 | 0,166 | 0,242 | 0,180 | 0,173 | 0,140 | 0,286 | **0,908** |
| 18. Reliability of Management (OU) | 0,468 | 0,425 | 0,022 | 0,036 | 0,007 | 0,258 | 0,491 | 0,589 | 0,314 | 0,303 | 0,365 | 0,527 | 0,393 | 0,375 | 0,206 | 0,889 | 0,327 |
| 19. Role Conflict | 0,373 | 0,372 | 0,000 | 0,173 | 0,116 | 0,403 | 0,252 | 0,327 | 0,569 | 0,865 | 0,291 | 0,388 | 0,466 | 0,447 | 0,253 | 0,362 | 0,226 |
| 20. Role Overload | 0,054 | 0,018 | 0,169 | 0,346 | 0,667 | 0,024 | 0,015 | 0,050 | 0,072 | 0,306 | 0,017 | 0,033 | 0,071 | 0,029 | 0,001 | 0,033 | 0,035 |
| 21. Social Climate | 0,706 | 0,516 | 0,019 | 0,083 | 0,029 | 0,495 | 0,395 | 0,444 | 0,340 | 0,379 | 0,452 | 0,651 | 0,843 | 0,424 | 0,236 | 0,543 | 0,223 |
| 22. Social Community at Work | 0,615 | 0,463 | 0,018 | 0,057 | 0,011 | 0,361 | 0,319 | 0,330 | 0,292 | 0,248 | 0,372 | 0,509 | 0,540 | 0,320 | 0,234 | 0,430 | 0,166 |
| 23. Social Support from Supervisor | 0,436 | 0,391 | 0,029 | 0,029 | 0,004 | 0,190 | 0,861 | 0,831 | 0,337 | 0,245 | 0,320 | 0,431 | 0,273 | 0,353 | 0,239 | 0,611 | 0,209 |
| 24. Task Completion Ambiguity | 0,036 | 0,070 | 0,013 | 0,014 | 0,000 | 0,048 | 0,033 | 0,040 | 0,112 | 0,079 | 0,050 | 0,043 | 0,037 | 0,300 | 0,097 | 0,063 | 0,015 |
| 25. Trust Regarding Management | 0,484 | 0,457 | 0,029 | 0,036 | 0,005 | 0,323 | 0,511 | 0,603 | 0,339 | 0,342 | 0,440 | 0,581 | 0,438 | 0,430 | 0,239 | 0,897 | 0,323 |
| 26. UWES (Vigor) | 0,129 | 0,271 | 0,012 | 0,013 | 0,012 | 0,058 | 0,087 | 0,070 | 0,172 | 0,091 | 0,061 | 0,123 | 0,061 | 0,136 | 0,339 | 0,123 | 0,083 |
| 27. UWES (Dedication) | 0,182 | 0,468 | 0,085 | 0,003 | 0,024 | 0,066 | 0,145 | 0,112 | 0,201 | 0,120 | 0,085 | 0,166 | 0,084 | 0,211 | 0,605 | 0,169 | 0,113 |
| 28. UWES (Aborption) | 0,093 | 0,280 | 0,123 | 0,024 | 0,094 | 0,023 | 0,074 | 0,058 | 0,105 | 0,051 | 0,041 | 0,076 | 0,026 | 0,128 | 0,434 | 0,080 | 0,063 |
| 29. Work Family Conflict | 0,194 | 0,199 | 0,041 | 0,398 | 0,301 | 0,152 | 0,089 | 0,131 | 0,214 | 0,411 | 0,117 | 0,165 | 0,214 | 0,143 | 0,104 | 0,160 | 0,108 |
| 30. Work Family Facilitation | 0,202 | 0,398 | 0,041 | 0,029 | 0,004 | 0,040 | 0,158 | 0,116 | 0,142 | 0,097 | 0,082 | 0,181 | 0,063 | 0,174 | 0,362 | 0,167 | 0,106 |
| 31. Work SoC (Comprehensibility) | 0,297 | 0,276 | 0,002 | 0,129 | 0,090 | 0,157 | 0,162 | 0,198 | 0,436 | 0,337 | 0,153 | 0,259 | 0,224 | 0,226 | 0,177 | 0,237 | 0,161 |
| 32. Work SoC (Manageability) | 0,334 | 0,323 | 0,002 | 0,117 | 0,066 | 0,174 | 0,238 | 0,248 | 0,382 | 0,362 | 0,201 | 0,350 | 0,254 | 0,393 | 0,198 | 0,302 | 0,200 |
| 33. Work SoC (Meaningfulness) | 0,224 | 0,493 | 0,081 | 0,022 | 0,008 | 0,141 | 0,181 | 0,149 | 0,249 | 0,173 | 0,143 | 0,231 | 0,131 | 0,277 | 0,642 | 0,226 | 0,125 |

Note: Bolded values represent the factors' square-root AVE values. Underlined values represents strong evidence for failure of discrimination ($r \geq .8$).

Table 4.4 (Continued).
*Inter-Factor Correlation Matrix for Model 16.*

| | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Cohesion in Work Teams | | | | | | | | | | | | | | | | |
| 2. Commitment to the Workplace | | | | | | | | | | | | | | | | |
| 3. Competency Demands | | | | | | | | | | | | | | | | |
| 4. DUWAS (Compulsion) | | | | | | | | | | | | | | | | |
| 5. DUWAS (Excessive) | | | | | | | | | | | | | | | | |
| 6. Dysfunctional Support | | | | | | | | | | | | | | | | |
| 7. Empowering Leadership | | | | | | | | | | | | | | | | |
| 8. Fairness of the Supervisor | | | | | | | | | | | | | | | | |
| 9. Goal Clairity | | | | | | | | | | | | | | | | |
| 10. Illegitimate Tasks | | | | | | | | | | | | | | | | |
| 11. Inclusiveness and Social Resp. | | | | | | | | | | | | | | | | |
| 12. Innovation | | | | | | | | | | | | | | | | |
| 13. Interpersonal Conflicts | | | | | | | | | | | | | | | | |
| 14. Job Autonomy | | | | | | | | | | | | | | | | |
| 15. Meaning of Work | | | | | | | | | | | | | | | | |
| 16. Recognition | | | | | | | | | | | | | | | | |
| 17. Reliability of Management (NL) | | | | | | | | | | | | | | | | |
| 18. Reliability of Management (OU) | **0,890** | | | | | | | | | | | | | | | |
| 19. Role Conflict | 0,363 | **0,629** | | | | | | | | | | | | | | |
| 20. Role Overload | 0,038 | 0,278 | **0,766** | | | | | | | | | | | | | |
| 21. Social Climate | 0,527 | 0,477 | 0,070 | **0,782** | | | | | | | | | | | | |
| 22. Social Community at Work | 0,382 | 0,316 | 0,023 | 0,851 | **0,804** | | | | | | | | | | | |
| 23. Social Support from Supervisor | 0,551 | 0,297 | 0,037 | 0,420 | 0,327 | **0,832** | | | | | | | | | | |
| 24. Task Completion Ambiguity | 0,046 | 0,130 | 0,008 | 0,053 | 0,044 | 0,041 | **0,683** | | | | | | | | | |
| 25. Trust Regarding Management | 1,003 | 0,398 | 0,031 | 0,582 | 0,422 | 0,556 | 0,056 | **0,769** | | | | | | | | |
| 26. UWES (Vigor) | 0,096 | 0,120 | 0,002 | 0,113 | 0,124 | 0,094 | 0,051 | 0,098 | **0,853** | | | | | | | |
| 27. UWES (Dedication) | 0,132 | 0,159 | 0,000 | 0,160 | 0,158 | 0,142 | 0,052 | 0,135 | 0,610 | **0,877** | | | | | | |
| 28. UWES (Aborption) | 0,065 | 0,067 | 0,006 | 0,063 | 0,071 | 0,070 | 0,038 | 0,066 | 0,447 | 0,718 | **0,820** | | | | | |
| 29. Work Family Conflict | 0,142 | 0,392 | 0,388 | 0,225 | 0,156 | 0,125 | 0,041 | 0,144 | 0,143 | 0,109 | 0,026 | **0,720** | | | | |
| 30. Work Family Facilitation | 0,130 | 0,116 | 0,039 | 0,171 | 0,151 | 0,158 | 0,026 | 0,119 | 0,192 | 0,315 | 0,204 | 0,129 | **0,589** | | | |
| 31. Work SoC (Comprehensibility) | 0,237 | 0,441 | 0,176 | 0,293 | 0,236 | 0,212 | 0,063 | 0,247 | 0,154 | 0,167 | 0,086 | 0,308 | 0,133 | **0,729** | | |
| 32. Work SoC (Manageability) | 0,292 | 0,438 | 0,160 | 0,353 | 0,268 | 0,271 | 0,087 | 0,303 | 0,151 | 0,183 | 0,092 | 0,280 | 0,165 | 0,946 | **0,747** | |
| 33. Work SoC (Meaningfulness) | 0,178 | 0,242 | 0,001 | 0,225 | 0,208 | 0,177 | 0,054 | 0,198 | 0,308 | 0,521 | 0,361 | 0,121 | 0,266 | 0,315 | 0,312 | **0,842** |

Note: Bolded values represent the factors' square-root AVE values. Underlined values represents strong evidence for failure of discrimination ($r \geq .8$).

## Discussion

The end of the theory section of this thesis established a set of formal criteria for valid latent variable measurement interpretations and a set of hypotheses were derived from those criteria. Specifically, the first set of criteria specified that the model ought to be comprehensive and parsimonious, criteria which were assessed at the global level (hypothesis 1) by employing model-level fit indices, and the local level (hypotheses 2 and 3) by employing factor-level indices of convergence and discrimination for each individual factor.[4] The discussion that follows will constitute the informal "validity argument" demanded by the Standards (AERA et al., 2014, p. 21; see also table 2.1, p. 17 of this thesis), which "*integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses.*"

While hypothesis 1 managed to satisfy lenient criteria (thus appearing supported by liberal evidential standards), hypotheses 2 and 3 (stricter hypotheses than hypothesis 1) were not supported, as only 11 out of the 33 factors in the retained model satisfied 100% of the criteria (see table 4.1, p. 64). As such, it appears that while KIWESTs' interpretation seems to fall within some acceptable bounds on global indices, local indices indicate that there is room for improvement. The results pertaining hypothesis 1 does not unequivocally indicate a good fit of the interpretation however, and the claim that it does can be challenged on both construct-theoretical as well as latent variable-theoretical grounds.

In keeping with the Standards terminology, discussions pertaining to the evidential source category of "internal structure" will deal with the analysis of global model fit, while discussion to the source category of "relations to other variables" will deal with convergence and discrimination of individual factors hypothesized to constitute a common cause of the variance observed in sets of individual items. As evidence regarding the detailed nature of test-takers responding (i.e., response processes) is not available, the sources of evidence that will constitute the basis of the following validity argument will be evidence pertaining to content (see table 3.1, pp. 33-35), internal structure (see table 4.2, p. 64), and relations to other variables (see table 4.3, pp. 66-57 and table 4.4, pp. 69-70).

---

[4] There is a disconnect between the use of terms between latent variable theory and validity theory on this point. From the perspective of latent variable theory, evidence pertaining to convergence and discrimination is directly relevant to the question of internal structure by considering each item a fallible test in and of itself. In a sense, the means by which convergence and discrimination is examined is reminiscent of the logic behind criterion validity theory that considered tests "equated operations of the same entity" to an extent proportional with their correlation coefficient (which was also known as the "validity coefficient"). The heavy-handed empiricism of criterion-validity theory is tempered in construct-validity theory by bringing substantial theoretical predictions and explanations of the correlations into the mix.

The purpose of this thesis is to evaluate and improve upon the validity of the omnibus KIWEST interpretation. However, global fit is ultimately a consequence of consistent local fit, and the discussion concerning lack of fit will as such focus on the local factor level for the purposes of improving fit at the global level. As such, the analytical questions the sources of evidence will inform – and which will guide the discussion – are the following. First; in the case of failures of convergence, why does the observed variation in the items hypothesized to share a common cause appear as if they in fact do not? Second; in the case of failures of discrimination, why do items hypothesized to constitute indicators of distinct latent variables appear as if they share a common cause?

As such, the following discussion will focus on examining and attempting to explain the observed failures of convergence and discrimination pertaining to the specific factors specified to explain the observed (co)variation in the items constituting the KIWEST survey. Necessary information required to examine context effects does not appear available, and examination will for this reason be restricted to step 4A of the LVIV model (see appendix 1), and as such offer suggestions for step 4Aa. Suggestions will however be offered for potential alterations that can be made to the KIWEST survey, which will allow it to gather information useful for examining and controlling for detrimental context-related effects on measurement (Podsakoff et al., 2003; Podsakoff et al., 2012; see also table 2.1, p. 17 and table 2.2, p. 23).

From the perspective of latent variable theory, examining probable causes of failures of convergence and discrimination based on item content analysis constitutes theorizing about the response processes of test takers, and thus involves theorizing about the identities of latent variable entities. The conclusions ought not to be considered valid claims about the response processes that are at play at producing the observations. The discussion should be considered theory-driven speculation regarding the causal processes connecting stimuli (test content) and responses (response behavior), which result in more or less plausible answers as to why items fail to converge or factors to discriminate, from which testable hypotheses might be derived.

Treating the conclusions as valid claims rather than as a basis for testable hypotheses in absence of relevant supporting evidence would be committing “the psychologists fallacy” (Markus & Borsboom, 2013a). It would be simply assuming that the test generally constitutes stimuli that is equivalent to test-takers and test-users (e.g., a test factor), that test-takers generally retrieve information as envisioned by the test-user (e.g., an ability factor), as well as generally being willing to provide honest responses (e.g., a motivational factor; Podsakoff et al., 2012). Additionally, it would constitute a case of the “begging-the-question fallacy,” as it would be treating claims as datums without necessary backings and warrants (Kane, 2013a).

**Examining Evidence Relating to Internal Structure**

The internal structure of KIWEST was examined as a whole by comparing the fit of different plausible modelled interpretations of the KIWEST theory on generally accepted global indices of model fit. The accepted model attained fit within conservative criteria on the absolute fit indices (i.e., SRMR and RMSEA), but only barely managed to fit by liberal criteria on the relative fit indices (i.e., CFI and TLI). As such, hypothesis 1 was supported, but one must concede that the evidence in favor of the validity of the omnibus KIWEST measurement interpretation is not rock solid.

What follows are two conceivable objections against the claim to validity from the perspective of latent variable theory and construct validity theory. Overall, it appears fair to say that the KIWEST measurement interpretation is strong (relatively speaking). However, it exhibits some issues that should be addressed, such as whether the results indicate successful discrimination or failure of convergence for the multifaceted constructs included in KIWEST.

**The latent variable-theoretical objection against the global claim to validity.**

From a latent variable-theoretical perspective, one could conceivably argue that the fit of the model is inadequate on two grounds; absolute and relative. On absolute grounds, the criteria by which the interpretation is deemed adequate are admittedly liberal, and judges of more conservative persuasions could argue that they simply do not serve as sufficient evidential standards. From a relative point of view, the strategy that was employed (the alternative-models approach of Jöreskog, 1993, where fit of nested models are examined and compared) revealed that it was only the least parsimonious of the modeled KIWEST interpretations that demonstrated fit values falling within specified acceptable bounds. Hence, judges could conceivably argue that only a parsimonious modelled interpretation of a theory could be considered to fit adequately by the liberal criteria, and that one ought to expect the more comprehensive nested interpretations of the same observations – such as the current modelled interpretation – to satisfy the conservative criteria in order to be deemed sufficient.

Because the interpretation in question is the least parsimonious of a number of less comprehensive and more parsimonious nested alternatives – none of which demonstrate adequate fit with the observations – it must be conceded that it is from a very liberal point of view that the current measurement interpretation of the KIWEST theory can be claimed to demonstrate “acceptable” fit. Thus, it appears that we are dealing with a borderline case of fit, and whether or not we are justified in introducing “*G*” and “*H*” as premises to arrive at “*F*” in formula 1 (p. 26) is for this reason not a simple matter. In light of the knowledge that

introducing these premises would be giving the interpretation a strong benefit of the doubt, it is tempting to defer to the saying “if there is any doubt, there is no doubt,” and claim that the results indicate that the KIWEST latent variable measurement interpretation is neither comprehensive nor parsimonious. At the very least, the observations suggest that alterations can be made to individual test score interpretations within KIWEST that would render the default composite interpretation (i.e., the KIWEST measurement theory) more valid.

**The construct-theoretical objection against the global claim to validity.** The fact that only the least parsimonious of the proposed KIWEST theory interpretations managed to satisfy these criteria is potentially consequential from a construct-theoretical perspective, as whether the modelled interpretation constitutes a legitimate translation of the theory can be disputed on from the perspective of the construct theories of each of the multi-faceted constructs. These observations could be taken to suggest that the facets of the multifaceted constructs included in KIWEST (i.e., DUWAS, UWES, WFB, and Work-SoC) fail to demonstrate convergence on their hypothesized higher-order factors, while those of another judge could suggest that the facets successfully discriminate.

As such, how these observations ought to be interpreted is not a cut-and-dry matter. Reviewing table 3.7 (p. 60) and comparing it with table 4.2 (p. 64), the results suggests that the interpretations of the multifaceted constructs included in KIWEST are best interpreted as multidimensional (at least from a statistical point of view). The least appears to be gained by considering DUWAS multi- as opposed to unidimensional (cf., model 1 and 9, and 8 and 16), followed by UWES (cf. model 1, 5 and 9, and 8, 12 and 16). Statistically, it is clear from the results that the most is gained by considering Work-Family Balance as multidimensional (cf., model 1, 2 and 3, and 14, 15 and 16), followed by Work-SoC (ibid.). The KIWEST manual appears to suggest that all of these constructs ought to be interpreted as unidimensional, which the current observations would seem to suggest is inappropriate.

The concerns voices by McGrath (2005a) appear relevant to the current case. McGrath argue that conceptually complex constructs such as hierarchically ordered constructs and sub-constructs lead to complex scales measuring disparate phenomena, which “*compromise the potential for accurate and precise characterization of psychosocial phenomena*” (p. 112). Indeed, the results from the current examination appears to support the notion that this is correct; that the facets are better considered entities in their own right and not as principally caused by the higher-order construct, as the least parsimonious of the models exhibits superior performance on even the most strongly parsimony-adjusted indices such as TLI and BIC. If

the hypothesized common causes of variance in the lower-order facets is indeed the targeted higher-order latent variables, they appear best considered as relatively distal rather than proximal causes of those covariations. Therefore, alternative means of measurement ought to be considered that might get at the targeted latent variables more directly (Kagan, 2005).

## Examining Evidence Relating to Relations to Other Variables

The above section examined and considered some potential objections and counter-arguments against the claim to global fit and found it suspect, thus sowing doubt on the validity of the default KIWEST measurement interpretation. As evidence pertaining to internal structure suggests that improvements can be made to the default interpretation of the survey-generated observations, the discussion now turns to examining localized areas on strain (Brown, 2015) within the model by examining evidence pertaining to convergence and discrimination, which the Standards (AERA et al., 2014) categorize as pertaining to "relations to other variables."

### Examining failures of convergence.

Hypothesis 2 require that all factors demonstrate convergence on the local factor-level in order to support the composite KIWEST latent variable measurement interpretation. The observations failed to provide support for this hypothesis. The frequencies and magnitudes of convergence failure does not appear severe. In terms of frequencies, nine factors fail to satisfy the conservative criterion: Cohesion in Work Teams, Competency Demands, DUWAS (Excessive and Compulsive), Illegitimate Tasks, Inclusiveness and Social Responsibility, Role Conflict, Task Completion Ambiguity, and Work-Family Facilitation. Of these factors, six fail to meet the liberal criterion: Competency Demands, DUWAS (Excessive and Compulsive), Inclusiveness and Social Responsibility, Role Conflict, and Work-Family Facilitation. As for magnitude, with two exceptions (specifically, Competency Demands and Work-Family Facilitation), most factors that failed to satisfy the conservative criterion does not do so by a very large margin, and convergence is likely attained by item exclusions.

Examinations of failures of convergence will take the form of operational analyses, primarily of correspondence between items (conceptual-operational convergence of items), to assess why statistical analyses suggest why the items do not seem to be caused by the same property of the same particular. The results appear after all to provide conclusive evidence that whatever the items measure, it is not the same entity, and the measurement interpretation should thus be more comprehensive. That is, the conjecture that the variance observed in the specified items share a common cause has been refuted (Popper, 1963/2002). What's more, it is also conceivable that the apparently "offending" items are in fact the "non-offenders," and

that it is the “well-behaved” items that evoke construct-irrelevant variance. Excluding the “poorly-behaved” items in favor of the “well-behaved” items would consequently invalidate identity interpretations, rendering substantive conclusions of higher-order investigations employing the measure invalid.

***Cohesion in Work Teams.***

The factor representing the latent variable “cohesion in work teams” exhibited weak failure of convergence (AVE < .5). Investigating item factor loadings revealed that item v01_007 attained a standardized factor loading of .47 (see table 4.3), while the two remaining items display standardized factor loadings well above the recommended threshold of .7 (.82, .75). This suggests that the item with the low factor loading does not share a proximal common cause with the remaining two items whom appears on statistical grounds as if they do.

Reviewing the content of the items (see table 3.1), this observation does not appear particularly confounding. The offending item refers to the particular that is “the unit” which has the property of “giving opportunity to improve ones personal performance.” There is no reference to the property of cohesion in the wording of the item, and it is conceivable that a unit can contribute to improvement of ones personal performance in several different ways – cohesion being only one conceivable avenue. As such, it comes as no surprise that the item performs poorly as an indicator of cohesion, as cohesion is likely to constitute a distal cause of variation in the item, if at all. From a post hoc perspective it appears as if it would in fact be more surprising if the item would have exhibited a strong factor loading (assuming that the covariation in the remaining items is indeed commonly caused by cohesion).

***Competency Demands.***

The factor representing the latent variable “competency demands” exhibited strong failure of convergence, failing to satisfy both conservative and liberal criteria ($\rho < .7$) of convergence. Investigating standardized item factor loadings revealed that two of the three items, v03_003 ($\beta = .62$) and v03_012 ($\beta = .46$) exhibited values less than .7, which suggests that variations observed in the items do not share a proximal common cause. As none of the items converge, it does not appear that the test as a measure of the targeted latent variable can be salvaged, as the variation observed in each item appear to be systematically caused by different entities.

Reviewing the content of the items reveal that all of the items refer to the properties of competence, development and learning. However, the particular to which the property is attributed differs from item to item. For example, item v03_003 employs the phrase “I am expected to …” which implies a social dimension, while item v03_009 employs the phrase

"The nature of my work …" which implies a structural dimension. Furthermore, the item v03_012 employs the phrase "I feel pressure to …", which is ambiguous with regards to the origin of the demand for competency development (i.e., whether the origin of the demand is structural or social in nature). As such, the items appear to be measuring the same property, however, the property is of different particulars (i.e., social, structural, or unspecific). Keeping this in mind it comes at no surprise from an ad hoc perspective that the variations in the items do not appear to share a common cause, and it appears as if this ought indeed to be the case.

***DUWAS (Excessiveness and Compulsiveness).***

It was not clear whether DUWAS ought to best be interpreted as best represented by a single or multiple factors, based neither on theory nor on global fit indices (the least gain on global fit indices was observed by splitting the two facets into separate factors, seen relative to the remaining multifaceted constructs). Neither of the factors representing the constructs' facets managed to attain strong nor weak convergence according to our criteria. As such, they are examined in unison due to the theoretical and empirical ambiguousness of their structures.

Both of the factors attains convergence scores within a value of .1 for being within acceptable ranges on both the weak as well as the strong criteria for convergence (i.e., $\rho > .6$ and AVE > .4), and the summary indicators of convergence seem to suggest that violations of the criteria are not particularly severe for either facet. Examining items, we see that two items exceed the $\beta \geq .7$ criterion for both the Compulsiveness and the Excessiveness factors, with two items (50%) failing to meet this criterion for Compulsiveness (v10_003 and v10_009), and three items (60%) did not satisfy this criterion for Excessiveness (v10_004, v10_006, and v10_008). None of the factor loadings for the "offending items" fall below .5, and two of the five offending values exhibit $\beta$ exceeding a .6 value.

Reviewing the content of the items, it appears that the "well behaved" items on the Compulsiveness scale indeed does refer to a "compulsiveness" dimension. For example, v10_005 refers to the mind-internal, contextually-external, and self-external notion of "drive" that causes one to work hard. The two "poorly performing" items refer to internalized pressures in the form of, perhaps, a sense of duty (v10_003) and guilt (v10_009), which can perhaps be interpreted as sharing outside socialized and social causes of the behavior rather than compulsions. According to the original authors of the scale (Schaufeli et al., 2009b, p. 322), work addicts: "*Rather than being motivated by […] external or contextual factors, [are typically] motivated by a strong internal drive that cannot be resisted.*" It is puzzling that they still included items referring to factors that can be construed as external and contextual.

As for the items of the Excessiveness scale, it is not immediately obvious why any of them should describe *excessive* work behavior. In fact, the items with the lowest loadings appear as if they could just as easily be interpreted as manifestations of the hypothesized facets of one conceptualization of work engagement (i.e., UWES' Vigor, Dedication, and Absorption facets; Schaufeli et al., 2006; Schaufeli et al., 2002). The two items that attain the strongest factor loadings appear in turn to be related to perceived time pressure, and thus attributable once again to external and contextual factors. These two items might in turn be interpreted as potential indicators of the "role overload" construct (Näswall et al., 2010).

To summarize, the DUWAS construct is confounding and it is difficult to make sense of it. It seems to represent a psychometric borderline case in terms of its internal structure, indicating at the very least that the phenomenon of workaholism (if indeed it is what the items converge on) is not a particularly proximal common cause of the observed variation in the items. In terms of content considerations, it appears as if the authors have not stayed true to their own conceptual definition of workaholism when construing the items, and it appears as if the content of the items substantially overlap with related constructs included in KIWEST, most immediately apparently those of UWES and Role Overload.

In conclusion, the validity of the intended interpretation of scores pertaining to the DUWAS instrument as measuring "workaholism" is rendered suspect. To improve the validity of these test score interpretations, the items should be closely scrutinized to examine whether they ought to be reinterpreted, and if so, how exactly they are to be interpreted (i.e., if it is not workaholism they measure, then what?). Conceptual analysis and response-process theorizing suggests that the items might best serve as indicators for other modelled latent variables included in KIWEST (specifically, Work Engagement and Role Overload).

***Illegitimate Tasks.***

The "Illegitimate Tasks" factor exhibited weak failure of convergence (AVE = .47), while managing to satisfy the liberal criterion ($\rho = .78$). As such, the factors failure to converge does not appear particularly severe. Inspecting the $\beta$ estimates of the items reveals that half of them (item v04_003 and v04_008) fail to achieve factor loadings of .7 (yet remains within a value of .1 of doing so). The other half (v04_011 and v04_014) exceed this threshold by a decent margin (i.e., a lower bound of the 95% CI > .7).

Reviewing the content of the items, they appears to conform reasonably well to the constructs' conceptual definition. Item v04_008 does however refer to "awkward positions." It is not appear immediately obvious that tasks that put one into awkward positions invariably

must be experienced as illegitimate. Rather, it appears likely that one can experience tasks as awkward, yet still consider them legitimate parts of the job. This item is the "worst behaving" item in the set ($\beta = .62$), and one might consider removing it and perhaps group it with the items intended to constitute measures of "Role Conflict." It is likely that doing so will allow the scale to achieve strong convergence (i.e., an AVE $\geq .5$, thus improving the validity of the measurement interpretation), as well as the validity of the identity interpretation.

***Inclusiveness and Social Responsibility.***

The factor representing the latent variable "Inclusiveness and Social Responsibility" exhibit weak failure of convergence (AVE = .45), yet retaining weak convergence ($\rho = .77$). None of the items' factor loadings reaches the $\beta \geq .7$ threshold. The factor presents an interesting case as the confidence intervals around the parameter estimates – tight as they are due to the large sample size – includes the threshold values, and as such arguably do not significantly differ from them. The 95% confidence interval of two of the items (v02_001 and v02_004) includes .7 as a value, with v02_002 coming very close (upper bound of 95% CI = .696). The most poorly performing item (v02_006) exhibits a $\beta$ of .63, with an upper bound 95% CI of .65. As such, if one were to round the values at two decimals, none of the factor loadings would differ significantly from a value of .7. Nevertheless, they do differ, and we are for this reason dealing with a borderline case of convergence.

Reviewing the content of the items, what seems striking is that there is no apparent a priori reason for the items to reflect a "general sense" of inclusiveness. Rather, the items each attach the property of "inclusiveness" to different particulars (i.e., groups towards which the unit can be inclusive towards. To the extent that item scores agree, one might legitimately interpret the test as tapping a general sense of inclusiveness towards specific minority groups. To the extent that the items disagree, they might indicate differential experiences concerning inclusiveness with regard to specific groups.

As such, it might be more appropriate to treat the items as an index rather than a scale, thus assessing it formatively rather than reflectively. This would in turn invalidate the interpretation of the items as measuring a latent variable, as the stated causal relationship between items and factor would be reversed (Edwards, 2010; Edwards & Bagozzi, 2000). Any particular units score on inclusiveness would as such try to get at anything "real" but should be treated as an indicator. In order to maintain the latent variable interpretation, each "facet" of inclusiveness should probably be modelled as a reflectively measured latent variable in its own right (see Markus & Borsboom, 2013b).

One might furthermore ask whether asking everyone to report on whether they are inclusive towards minority groups is the best way to go about measuring inclusiveness, as complacency or a lack of awareness of the majority might mask legitimate grievances of minority groups about the units' inclusiveness towards them. Statistically analyzing whether any particular unit is "generally" inclusive should probably include analyses of latent mean differences between the unit in general (the particular to which the property of inclusiveness can be assigned), and the stakeholder groups to which the particular unit can be inclusive towards. As such, both statistical, conceptual, and practical concerns sow doubt on the validity of the default interpretation that this particular test score can at face value legitimately be claimed to capture the extent to which "*inclusion and social responsibility are generally taken care of*" (Undebakke et al., 2014, p. 10) in any given working environment.

***Role Conflict.***

The factor representing the latent variable "Role Conflict" fails to meet the criteria for both weak ($\rho = .68$) and strong (AVE = .4) convergence, indicating that the observed variance in the items constituting the scale do not share a proximal common cause. None of the items reach the desired factor loading threshold, with item v03_007 attaining the lowest value ($\beta = .56$). Rounding up, the remaining items exhibit $\beta$ coefficients between .65 and .66 with none of the 95% CI's including .7 as a probable population value.

Reviewing the operational definitions (item content) of the construct in light of its conceptual definitions, it is (with the exception of item v03_006) not obvious how one ought to categorize the items in terms of the facet they are hypothesized to belong to (i.e., whether a particular item should be classified as tapping intrasender-, intersender-, or interrole conflict). For example, item v03_002 asks whether the employee lacks the necessary resources to complete their tasks. Unless lack of resources can be attributed to incompatibility of roles, it is difficult to imagine why this item should have anything to do directly with role conflict as it is defined here. Items v03_007 and v03_011 on the other hand appears as if they could just as easily on qualitative grounds be considered indicators of the "Illegitimate Tasks" factor.

In conclusion, it appears that the constructs operational definition might be lacking items that directly address the facets of the indicators' hypothesized common cause. In retrospect, it is almost surprising that the factor performs as well as it does, even though it does not live up to expectations. A possible remedy to bring the factor within acceptable bounds in terms of convergence criteria might simply be to drop item v03_007 (perhaps grouping it with the items intended to measure Illegitimate Tasks).

According to the criteria set in this thesis, this would provide evidence in favor of the validity of the factors latent variable measurement interpretation. Based on the content of the items however, the identity interpretation would still be suspect in the absence of compelling response-process related evidence linking the latent variable with response patterns, and thus invalid due to lack of justification. Should one wish to use the scale one ought to undertake further investigations to examine what exactly it is that the items in unison appear to measure.

***Task Completion Ambiguity***

The factor representing the latent variable "Task Completion Ambiguity" fails to attain strong convergence (AVE = .47) but manages to achieve weak convergence (ρ = .71). Inspecting the items factor loadings, it is apparent that item v03_001 proves to be detrimental to factor convergence, as it exhibits a factor loading well below the desired threshold of .7 (β = .36). The two remaining items, v03_005 and v03_008 exhibit strong factor loadings (β's = .8 and .79 respectively). As such, the variance observed in item v03_001 appear primarily caused by sources distinct from those causing variance in items v03_005 and v03_008.

Inspecting the contents of the items, it appears as if it is actually v03_001 that exhibits the most convincing content-related evidence of validity relative to the construct label. It reads: "I know when a task is completed," which ironically does not appear ambiguous. The remaining items appears in contrast as if they would serve better as indicators for the Job Autonomy factor (Näswall et al., 2010), as it is conceivable that it can be up to oneself to determine when a task is completed, yet still experience certainty about whether it actually is. That is, autonomy in decisions does not entail uncertainty in decisions. The interpretation of the test score specified by Undebakke et al. (2014, p. 9) does however read: "*the employees themselves can, or have to, determine when their tasks are completed.*" The strongly loading items do indeed appear to conform to this interpretation, and the identity claim does as such appear strong. Whether the stated interpretation conforms to the label appears debatable.

As such, while items v03_005 and v03_008 actually exhibit the strongest factor loadings (i.e., strong evidence for latent variable measurement interpretation), it appears as if it is actually item v03_001 exhibits, in and of itself, the strongest evidence in favor of the latent variable identity interpretation. Nevertheless, it is the measurement interpretation that is the primary object of validation in this thesis, and as such it appears as if it is item v03_005. and v03_008 that share a proximal common cause, which appears distal for item v03_001.

In conclusion, the latent variable measurement interpretation appears partially valid as two of the items appear to converge strongly on "something," and (based on item content) this

"something" appears to match the stated interpretation of the test score. However, it seems questionable whether the labelling of the interpretation is appropriate, as it is rendered suspect by its corresponding conceptual and operational definitions. The default identity interpretation ought perhaps to be considered invalid in absence of response-process evidence suggesting otherwise. In fact, the "offending" item is probably in this case the one most likely to elicit the response-processes necessary to get at something one would intuitively consider deserving of the label "Task Completion Ambiguity" rather than, perhaps, "Task Completion Autonomy."

***Work-Family Facilitation.***

The factor representing the latent variable "Work-Family Facilitation" exhibits the most extreme failure of convergence out of all the latent variable factors hypothesized to account for the observed (co)variance in the KIWEST questionnaire ($\rho = .23$, AVE = .35). Item v09_013 demonstrates a particularly low factor loading ($\beta = .2$), and as such proves highly detrimental to factor convergence. Of the remaining items, two satisfy the $\beta \geq .7$ criterion (v09_004 and v09_005; however, the lower bounds of their 95% CI's includes values < .7), while item v09_011 falls short of this criterion ($\beta = .6$).

Reviewing the content of the items making up the scale, it is not particularly surprising that item v09_013 performs so poorly relative to the rest. The remaining items appear to do with skills acquired at work that are applicable to situations at home, while v09_013 appears as if it has more to do with an individual difference variable in terms of how having "a good day at work" affects ones capacity to be a good companion at home. It does not appear as if this necessarily needs to have anything at all to do with "roles." As such, v09_013 deals with quite a different way in which work can facilitate home than the remaining items, and as such does not necessarily have to share a common cause (i.e., the cause might be more to do with the individual person than the intrinsic nature of their work). As for the remaining items, it appears that they intend to tap either practical (v09_011) or personal (v09_005) transferable benefits from work to home, or both (v09_004). As such, the nature of the facilitating effects of work are different across items, which might explain their failures to converge, even though the purpose of item v09_004 appears to be to connect items v09_011 and v09_005.

In conclusion, it appears reasonable to consider all of the items as tapping aspects of work family facilitation. However, based on conceptual and statistical analyses, it does not appear as if one is justified in interpreting the variation in the items as sharing a common cause. It appears as if the items tap reasonably distinct ways in which work can facilitate family life, and should perhaps be considered distinct latent variables. Alternatively the factor

can be specified and assessed formatively, in which case the latent variable interpretation of work family facilitation would be invalidated. As it stands, the interpretation that the items share a proximal common cause appears conclusively refuted, suggesting that a less parsimonious and more comprehensive means of conceptualizing and operationalizing work-family facilitation is appropriate.

**Examining failures of discrimination.**

Having accounted for failures of convergence, the thesis now turns to accounting for and addressing failures of discrimination. Meaningful discrimination between factors by the conservative criterion (AVE ≥ MSV) requires convergence, and factors are evaluated relative to this criterion only if they satisfied at least one criterion for convergence. If a factor failed to achieve convergence, it will only be evaluated with regards to discrimination if it failed to satisfy the liberal criterion ($r \geq .8$). This proved to be the case only for the Role Conflict factor. In addition to Role Conflict, thirteen factors that managed to converge failed to discriminate by either the conservative or liberal criterion.

The results indicate that the primary cause of concern with the KIWEST measurement interpretation is failures of discrimination, as the "condisc" STATA14 module flagged all of the factors as being somehow involved in such failures on the conservative criterion. Fifteen of the 33 factors are "recipients" of such failures (the factors with negative AVE – MSV values in table 4.3, pp. 66-67), while 13 factors fail to satisfy the liberal criterion. Thus, in terms of numbers, the frequencies and magnitudes of discrimination failures between factors representing theoretically distinct latent variables in KIWEST can be described as "severe." Specific instances are marked in the inter-factor correlation matrix (table 4.4, pp. 69-70).

In order to reduce the amount of information and guide examination of discrimination failures, a second-order exploratory principal factor analysis was performed on the factor scores estimated through the CFA that were "recipients" of discrimination failure (14 in total). This was done to break the results up in manageable chunks, and to facilitate the exploration of which of the factors that cluster together. The outcome of this analysis is available for inspection in table 5.1. The analysis suggested that the failures of discrimination can best be described as forming three fairly strongly correlated yet reasonably distinct clusters of factors. Each of these clusters are examined to explore the failures of discrimination in more detail.

It bears mentioning once again that theorizing about why it is that some factors fail to satisfy criteria involves speculating about their identities. The investigations of discrimination failure will here be based on the dictum of Tukey (1969); that clarity at the large scale flows

Table 5.1.

*Second-Order Exploratory Principal Factor Analysis Describing the Clustering of Converging Non-Discriminating First-Order Factors.*

| | Multivariate | | | Univariate | |
|---|---|---|---|---|---|
| **Variable** | **Factor1** | **Factor2** | **Factor3** | **Uniqueness** | **Commonality** |
| Cohesion in Work Teams | 0,29 | **0,61** | -0,10 | 0,19 | 0,81 |
| Empowering Leadership | **0,97** | -0,05 | 0,01 | 0,13 | 0,87 |
| Fairness of the Supervisor | **0,90** | 0,02 | -0,04 | 0,11 | 0,89 |
| Illegitimate Tasks | 0,01 | 0,29 | **-0,66** | 0,24 | 0,76 |
| Interpersonal Conflicts | -0,05 | **0,95** | -0,04 | 0,11 | 0,89 |
| Recognition | **0,77** | 0,21 | -0,04 | 0,11 | 0,89 |
| Reliability of Management (Own Unit) | **0,76** | 0,21 | -0,02 | 0,14 | 0,86 |
| Role Conflict | 0,02 | 0,33 | **-0,67** | 0,14 | 0,86 |
| Social Climate | 0,12 | **0,90** | -0,02 | 0,01 | 0,99 |
| Social Community at Work | 0,08 | **0,91** | 0,05 | 0,13 | 0,87 |
| Social Support from the Supervisor | **0,98** | -0,08 | -0,06 | 0,08 | 0,92 |
| Trust Regarding Management | **0,72** | 0,27 | -0,03 | 0,12 | 0,88 |
| Work SoC (Comprehensibility) | 0,00 | 0,08 | **1,00** | 0,10 | 0,90 |
| Work SoC (Manageability) | -0,10 | 0,07 | **0,94** | 0,09 | 0,91 |

| **Inter-factor Correlation Matrix** | | | | **Variance** | **Proportion** |
|---|---|---|---|---|---|
| Factor 1 | 1,00 | | | 8,04 | 0,65 |
| Factor 2 | 0,68 | 1,00 | | 7,52 | 0,61 |
| Factor 3 | -0,54 | -0,60 | 1,00 | 5,57 | 0,45 |

Note: As per the suggestions of Mehmetoglu and Jakobsen (2017), factor loadings ≥ .4 is taken to indicate practical significance and thus marked by boldface. Factoring performed by means of Principal Factor Analysis. Three factors retained based on the eigenvalues point of inflection. Factors obliquely rotated (promax) to maximize factor loadings.

from clarity of the medium scale, which in turn flows from clarity at the small scale. Here, this will take the form of first examining the conceptual discrimination of the constructs' definitions (medium scale) to assess whether discrimination failure can be attributed to lack of conceptual distinctiveness. Second, should the conceptual examinations fail to provide a reasonable account for discrimination failure, the investigation turns to examining whether the extreme correlations can be attributed to lack of operational distinctiveness (small scale; comparative item content analysis). If the analysis of conceptual operations in turn fail to reasonably account for failure of statistical discrimination, it might be concluded that the correlation is "legitimate" in absence of further evidence to the contrary (i.e., demonstrating that the correlation is contaminated by construct-irrelevant covariance).

### *Cluster 1: Experiences, feelings, and perceptions towards superiors.*

Cluster 1 (Factor1 in table 5.1) points to six of the factors as showing strong relationships with each other. Each of these factors concerns attitudes towards or experiences with management. This factor contains the "Haywood case" observed in the initial CFA; that is, the relationship between "Reliability of Management (Own Unit)" and "Trust Regarding Management" which exhibited a $\beta$ value > 1 (an impossible and thus impermissible outcome of the CFA).

Table 5.2.

*Correlation Matrix for the Factors Loading Strongly on Factor-Cluster 1.*

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1. Empowering Leadership | 1,00 | | | | | |
| 2. Fairness of the Supervisor | **0,92** | 1,00 | | | | |
| 3. Recognition | 0,78 | **0,81** | 1,00 | | | |
| 4. Reliability of Management (Own Unit) | 0,74 | **0,81** | **0,97** | 1,00 | | |
| 5. Social Support from the Supervisor | **0,97** | **0,95** | **0,82** | 0,78 | 1,00 | |
| 6. Trust Regarding Management | 0,75 | **0,81** | **0,97** | **0,99** | 0,78 | 1,00 |

Note: Inter-factor correlations ≥ .8 marked by boldface.

Examining a correlation matrix composed of the strong-loading factors of this cluster (here defined as loadings ≥ .4, following the recommendations of Mehmetoglu & Jakobsen, 2017). Specifically, the factors of "Empowering Leadership" and "Social Support from the Supervisor" form a pair discriminating from "Reliability of Management (Own Unit)" and "Trust Regarding Management" as a pair. However, these sub-clusters are in turn connected by "Fairness of the Supervisor" (which exhibits universal failure of discrimination), and the "Recognition" factor (which fails to discriminate with all factors except "Empowering Leadership"; see table 5.2 and figure 5.1).

As the factor labels and conceptual definitions refer to particulars such as "leadership," "management" and "supervisor," and the stated interpretations of scores refer to particulars such as "perceptions," "experiences" and "feelings," it seems at the face of it reasonable to label this second-order factor "experiences, feelings, and perceptions towards superiors." The question then becomes; "why do these factors posited to assess differing properties assigned to related particulars appear (statistically) as if they converge on a single property of a single particular?" After all, reviewing the correlation matrix reveals that if all of these factors were

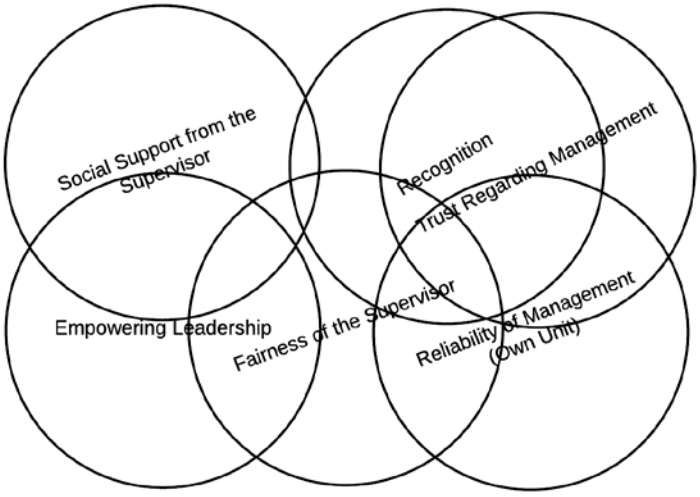


*Figure 5.1.* Diagram illustrating factor clustering on the liberal criterion for discrimination for factor-cluster 1. The circles represents individual factors, and overlapping circles represents factor correlations ≥ .8 while non-overlapping circles represent factor correlations < .8. The figure illustrates how all of the factors are connected by way of "Fairness of the Supervisor," and second-most strongly by the "Recognition" factor which only manages to discriminate on the liberal criterion with "Empowering Leadership."

assigned as indicators of a higher-order factor, they would achieve AVE and reliability values far above the recommended thresholds for convergence (specifically, AVE = .82, ρ = .96).

The obvious place to start with investigations is the "Haywood case" of the CFA; "Reliability of Management (Own Unit)" and "Trust Regarding Management." It is perhaps worth mentioning that an alternative name for the reliability of management scale, the name given to it by its original authors (Näswall et al., 2010, p. 11) is actually "*Trust.*" Furthermore, the conceptual definition of "Reliability of Management" include the adjective "trustworthy." As the stated preferred interpretation of "Trust Regarding Management" refers to "perceived trust in management," it becomes difficult not to reinterpret these observations as anything but convergent evidence of validity for the claim that both tests converge on a single entity. Thus, the interpretation that the tests measure distinct phenomena appears invalidated.

A more puzzling observation however is the extremely strong correlation between "Recognition" and the above-mentioned factors (r = .97). The stated interpretation refers to feeling recognized and appreciated for ones efforts. At the level of factor labels and construct conceptual definitions, there does not appear to be any good reason for this factor to exhibit such extreme correlations with the other two factors. Reviewing item content does not appear to shed much light on the matter, and it seems for this reason that "Recognition" and "Trust" constructs are both conceptually and operationally distinct, which lends credence to the notion that the measurement interpretation is valid despite statistical evidence suggesting otherwise. Unless a source of construct-irrelevant variance is demonstrated to be at play inflating the observed covariation, the observations seem to imply that either these distinct entities share a common cause for their variation, or variation in one of the latent variables entails variation in the other latent variable.

However, the "Recognition" construct includes an item in its operational definition (v06_005) that potentially transgress conceptually on the proverbial lawn of the "Fairness of the Supervisor" construct, which is represented by the factor that fails to discriminate with all of the other factors in the cluster. The conceptually offending operation refers to being treated fairly by ones unit management. It is conceivable that being recognized for ones efforts can constitute an instance of feeling that one is being treated fairly (a logical analysis placing "Recognition" higher up in the causal hierarchy), and the factor loading of the apparently offending item appears to lend credence to the notion.

The last two factors in the cluster, "Empowering Leadership" and " Social Support from the Supervisor" provides a more clear cut case, as it provides an obvious instance of operational failure of discrimination. While empowerment is defined as transferring power

and enablement, the operations refers to behaviors of leaders that can easily be described as "socially supporting," as the questions posed by means of the items refers to concepts such as "encouragement" and "contributions." Thus, it appears that socially supportive behaviors are empowering behaviors. Based on this analysis, it would seem that empowering behaviors can be considered a subset of socially supportive behaviors, which in turn makes one expect fairly strong relationships between the factors. However, it remains a floating issue whether this conceptual connection warrants a relationship as close to perfect as the one observed here.

To summarize and conclude; conceptual, operational, and logical analyses ultimately suggest that one would expect these latent variables to exhibit strong correlations with each other. A causal explanation can perhaps be articulated as follows: Being recognized for ones efforts constitutes behavior that is perceived as "fair," and when management is fair, they are perceived as trustworthy. Being recognized for ones efforts furthermore constitutes an example of socially supportive behavior that can double as empowering behavior. Also, if management is socially supportive they can be perceived as being fair.

Another plausible explanation is that the items in unison simply capture a general (nonspecific) sense of (dis)content (or simply valence) with the unit management (Reliability of Management – Next Level, after all, does successfully discriminate from this cluster), as specifying a single common factor accounts for the vast majority of factor covariation. In contrast with the above explanation, this constitutes a reversal of the proposed ordering of the causal relationships between the latent variables, as it posits that general experiences causally influence specific experiences (which at the face of it does not seem unreasonable).

These suggested ad-hoc explanations are not necessarily in competition, as they are not mutually exclusive. It is perfectly conceivable that specific experiences causally influence general experiences, subsequently biasing perceptions of specific particulars, which in turn reinforce general perceptions and experiences (e.g., confirmation bias). This line of reasoning is consistent with a "network" view of psychosocial phenomena (Cramer, 2012; Schmittmann et al., 2013) as systems of causally coupled phenomena forming self-reinforcing gestalts.

***Cluster 2: Experiences with climate, colleagues, community, and unit.***

Cluster 2 (Factor2 in table 5.1) points to four factors being strongly related to each other – Cohesion in Work Teams, Interpersonal Conflicts, Social Climate, and Social Community at Work. As the conceptual definitions consistently refer to experiences directed at particulars such as climate, colleagues, community and unit, it appears reasonable to label this factor cluster "experiences with climate, colleagues, community, and unit." The factor-cluster seems

Table 5.3.
*Correlation Matrix for the Factors Loading Strongly on Factor-Cluster 2.*

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Cohesion in Work Teams | 1 | | | |
| 2. Interpersonal Conflicts | 0,77 | 1 | | |
| 3. Social Climate | **0,89** | **0,94** | 1 | |
| 4. Social Community at Work | **0,86** | **0,81** | **0,95** | 1 |

Note: Inter-factor correlations ≥ .8 marked by boldface.

to indicate that the "Cohesion in Work Teams" factor is something of an odd factor out, as it exhibits a relatively low factor loading (≈ .6) compared to the remaining factors (all ≥ .9).

Inspecting the factor correlation matrix (table 5.3) does however reveal that it exhibits strong correlations with the remaining factors, discriminating only (on the liberal criterion) with the "Interpersonal Conflicts" factor (see figure 5.2 for an illustration of factor-clustering on the liberal criterion for discrimination). Comparing the constructs' conceptual definitions suggests that "Cohesion in Work Teams" and "Interpersonal Conflicts" are more specific than "Social Climate" and "Social Community at Work" concerning what aspect of the working environment that is being assessed, as they refer to specific ways in which a working climate can be "good" or "bad." This specificity might in turn explain why they are the factors in the cluster that discriminate the most strongly from each other. In contrast, the "Social Climate" construct appears the least specific as it simply refers to "a good social climate," the meaning of which must apparently be derived from its operations.

Inspecting the content of the items constituting the "Social Climate" scale reveals that the items refer to such properties of climates as "distrustfulness" and "suspiciousness" (item v01_008), "encouragement" and "supportiveness" (item v01_016), and "relaxedness" and "comfortableness" (item v01_018). The built in operational-conceptual overlap with the remaining, more specific constructs appears obvious. Thus, the factor appears to capture a

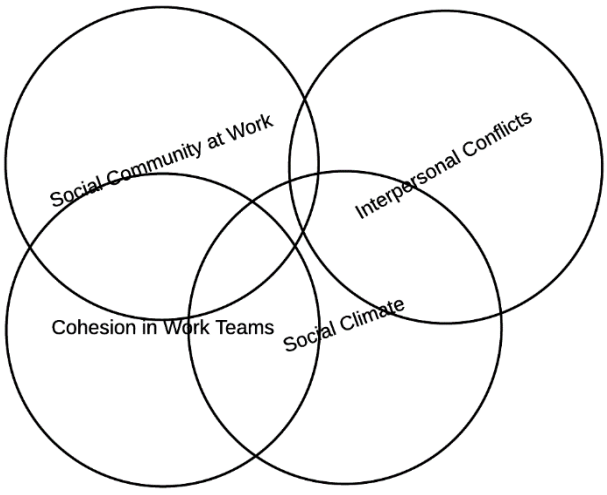


*Figure 5.2.* Diagram illustrating factor clustering on the liberal criterion for discrimination for factor-cluster 2. The circles represents individual factors. Overlapping circles represents factor correlations ≥ .8 while non-overlapping circles represent factor correlations < .8. The figure illustrates how all factors except "Cohesion in Work Teams" and "Interpersonal Conflict" fail to discriminate on the liberal criterion.

general state of the working environment at a higher level of abstraction than the remaining constructs, placing it further down in the causal hierarchy. Furthermore, properties such as "cohesion" and "interpersonal conflicts" can conceivably constitute sources of (and as such antecedents to) a sense of social community. Thus, it appears that the clustering and lack of discrimination between the factors can reasonably be attributed to conceptual, operational, and theoretical overlap.

To summarize and conclude; much like the previous cluster it appears that a causal explanation between the construct-entities can account for the strong correlations. However, with the exception of "Cohesion in Work Teams" and "Interpersonal Conflicts," the causal link appears to be one of levels of abstraction and not of intrinsic properties of latent variable entities that belong at the same level of analysis. Thus, it seems that the failure of statistical discrimination can be attributed to lack of both conceptual and operational distinctiveness. As such, the two (or three, depending on how one looks at it) general explanations for failures of discrimination proposed for the previous cluster apply for this cluster as well. That is, either the specific causes the general, the general causes the specific, or the general and the specific mutually cause and reinforce each other.

***Cluster 3: Experiences with and perceptions of job properties.***

Cluster 3 (Factor3 in table 5.1) points to four factors as clustering around a single dimension; "Illegitimate Tasks," Role Conflicts," "Work-SoC – Comprehensibility," and "Work-SoC Manageability." In contrast to the two preceding clusters, pinning down the "identity" of the factor is not immediately obvious beyond the fact that each of the common denominator of the constructs is the particular job of the employee. The uncommon denominators of the constructs are the specific properties attached to that particular. At face value, it appears plausible that the properties should correlate strongly (i.e., that there are causal connections between the properties make intuitive sense), though does not seem obvious that they should covary to the point of discriminatory failure.

Inspecting the inter-factor correlation matrix (table 5.4), it appears that the clusters form two relatively distinct sub-clusters if one employ the liberal criterion for discrimination

Table 5.4.
*Correlation Matrix for the Factors Loading Strongly on Factor-Cluster 3.*

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1. Illegitimate Tasks | 1 | | | |
| 2. Role Conflict | **0,97** | 1 | | |
| 3. Work SoC - Comprehensibility | -0,66 | -0,73 | 1 | |
| 4. Work SoC - Manageability | -0,67 | -0,74 | **0,98** | 1 |

Note: Inter-factor correlations ≥ .8 marked by boldface.

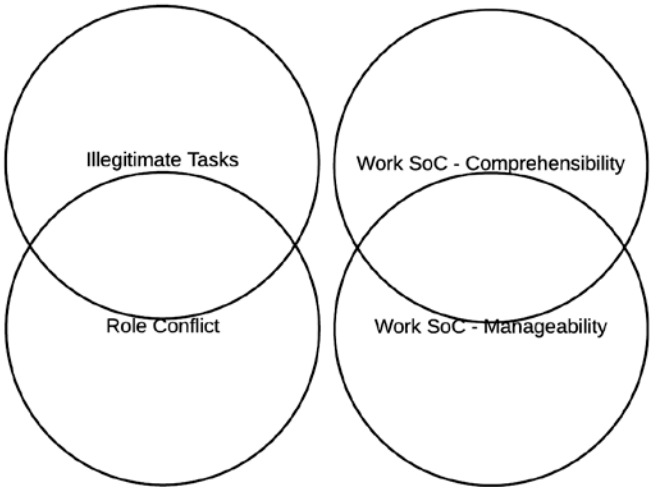


*Figure 5.3.* Diagram illustrating factor clustering on the liberal criterion for discrimination for factor-cluster 3. The circles represents individual factors, and overlapping circles represents factor correlations ≥ .8 while non-overlapping circles represent factor correlations < .8. The figure illustrates how the factors form two distinct sub-clusters, as "Illegitimate Tasks" and "Role Conflicts" overlap and discriminate from the second sub-cluster composed of "Work SoC – Comprehensibility" and "Work SoC – Manageability."

(see figure 5.3 for an illustration). Specifically, the factors "Illegitimate Tasks" and "Role Conflicts" form one sub-cluster, and "Work SoC – Comprehensibility" and "Work SoC – Manageability" form the other. Furthermore, the "internal cohesion" of these sub-clusters can only be characterized as extreme, as the correlation coefficients between the factor-pairs forming each sub-cluster both approximate perfection.

The constructs represented by the factors constituting the first cluster share their theoretical foundations in role theory. Conceptually, it seems the constructs should be distinct, though the correlation between the factors representing them suggest that they are virtually equivalent. This indicates that the factors intended to represent distinct latent variables in actuality converge on a single one. Recalling that both factors exhibited convergence failure (Role Conflict more severely and Illegitimate Tasks less so), and what was being measured was deemed unclear on the grounds of inspecting content-related evidence. It seemed that at least one item pertaining to each construct appeared as if it better fit the conceptual definition of the other, indicating lack of operational distinctiveness. Considering the severe failure of discrimination in light of the strong failure of convergence, it would appear that the factors converge on some unidentified distal rather than proximal common cause of variation.

As for the second sub-cluster, the failure of discrimination is expected as the factors are hypothesized manifestations of a "sense of coherence" (Antonovsky, 1996). As such, from a construct validity-theoretical perspective, the extreme correlation between the factors might be interpreted as evidence of successful convergence rather than of discrimination failure. The fact that the factor representing "Work SoC – Meaningfulness" does not belong to this cluster indicates that the principal common cause of variation in the factors "comprehensibility" and

"manageability" (for which the observations seem to suggest that there is one) is not the principal common cause of variation in the "meaningfulness" factor. As for the overarching "Work-SoC" score interpretation – that the respondents experience the workplace as health promoting – it is not obvious why it is necessary to access the particular in such a roundabout manner, rather than simply posing the question directly.

Regarding the strong connection between the clusters, it makes sense to compare the operations of the factors whose identities are difficult to interpret (i.e., those of sub-cluster 1) to those of the factors whose identities are easier to interpret (i.e., those of sub-cluster 2). Coding item content of cluster 1 for overlap indicates failure of operational distinctiveness with the "Work-SoC" constructs. For example, item v03_002 of "Role Conflict" transgress directly on the conceptual domain of "Manageability" by echoing its default interpretation nearly verbatim. This item was previously outed as a substantially unreliable manifestation of role conflict, as it does not seem apparent that inadequate resources must be a consequence of conflicting roles. It seems plausible that inadequate resources could cause one to perceive one's work as less manageable (make note of the formative causal implication). Thus, this item appears as if it might artificially drag the "Role Conflict" factor towards the "Work-SoC Manageability" construct by means of operational conflation.

To summarize and conclude; this cluster proved challenging due to the confounding nature of the first sub-cluster (i.e., it is difficult to pin down an identity interpretation due to their failures of convergence). The second cluster provides a more clear cut case as the factors constituting the cluster are supposed to serve as indicators of a higher-order construct (i.e., a sense of coherence at work). As such, the second sub-cluster can be considered evidence of successful convergence rather than as failure of discrimination. As for the strong correlation between the two sub-clusters, examining item content indicate clear conceptual-operational overlap between some of the items of sub-cluster 1 with the operational and conceptual definitions of the constructs represented by the factors constituting sub-cluster 2, sowing doubt on the legitimacy of their proposed distinctiveness.

**General Suggestions for Addressing Failures of Convergence and Discrimination**

Failures of discrimination and convergence can be addressed by two principal routes; altering the questionnaire (a distal fix) or altering the interpretation of the questionnaire (a proximal remedy). In absence of additional redeeming evidence, such failures must be addressed as they sow doubt on- and thus invalidate the standing default interpretations. The previous sections addressed specific cases of failures of factor convergence by individual items and

discrimination between individual factors, suggesting reinterpretations of what entities specific items might measure, or what latent variables the factors represent.

Those suggestions work with what is available, and constitutes nested rearrangement of items, the hope being that it will result in valid measurement interpretations by facilitating convergence and discrimination (i.e., potential proximal remedies). Here, suggestions for adjustments that can be made to the questionnaire itself (i.e., distal fixes) are proposed which might facilitate the validity of its interpretation(s) by examining, addressing, and controlling for detrimental effects associated with response styles (Podsakoff et al., 2003; Podsakoff et al., 2012; Weijters, Cabooter, & Schillewaert, 2010; Weijters, Schillewaert, & Geuens, 2008).

The suggestions offered basically reduce to two proposals; that is, (1) capitalizing on well-functioning items to control for systematic error variance associated with "acquiescence" and (2) employing a "planned missingness" design in data collection (Enders, 2010; Graham, Taylor, Olchowski, & Cumsille, 2006) to counteract construct-irrelevant variance associated with "respondent reactivity" (T. D. Little, Jorgensen, Lang, & Moore, 2014). Employing the terminology of Podsakoff et al. (2012), the first suggestion would enable a statistical remedy, while the second suggestion constitutes a procedural remedy. In terms of the LVIV model (p. 17, appendix A), this would allow for examining and controlling for context effects.

**Capitalizing on well-functioning items to control for acquiescence.**

The first suggestion involves building measures of response bias into KIWEST that contribute to its primary goal of assessing working-environmental variables; this by including negated copies of "well-functioning" items in the questionnaire (e.g., by adding the "not" connective to a statement), basically asking the same question twice with the only difference being its polarity. These items could replace those that demonstrate poor functioning on statistical and substantive grounds. Including such items would allow for filtering out of a potential source of systematic error variance; acquiescence (Hinz, Michalski, Schwarz, & Herzberg, 2007).

"Well-functioning" items would be those that perform well in terms of providing strong evidence pertaining to the content- and internal structure categories of evidence (i.e., they appear to contribute to the justification of both measurement and identity interpretations). That is, even in absence of response-process related evidence, indirect circumstantial evidence makes it appear highly probable that variation in the item is in fact principally caused by the targeted latent variable. Such an item would allow one to compute a difference score between the original and negated item that could serve the purpose of measuring acquiescence (Podsakoff et al., 2003; Podsakoff et al., 2012; Weijters et al., 2008).

These items could perhaps be construed as “super-usefully redundant” (contrasted to the “usefull reduncancy" that one normally aim at in measurement; DeVellis, 2017; Edwards, 2010) by doing double work. That is, we would be reasonably confident that they represent (i.e., that their variance is generally caused by) the desired latent variable with high fidelity, and we would control for a potential source of systematic construct irrelevant variance, effectively filtering it out. If enough items of this kind are included in the questionnaire (e.g., at least three or four negated items), it would allow for modelling acquiescence as a latent variable in its own right (see figure 5.4), as any non-zero mirrored difference across pairs of negated and non-negated items would be indicative of a systematic (dis)acquiescent slant.

In theory, the consequence of controlling for acquiescence would be that whatever lack of agreement between items is due to difference in polarity would be controlled for. As such, if acquiescence is at play, the correlation between items with opposite polarities should increase, and controlling for acquiescence should thus contribute to item-factor convergence. However, controlling for acquiescence would conversely cause the correlation between items sharing the same polarity to take a hit proportional to the extent to which their correlations can be attributed to the respondents’ levels of acquiescence, detracting from convergence.

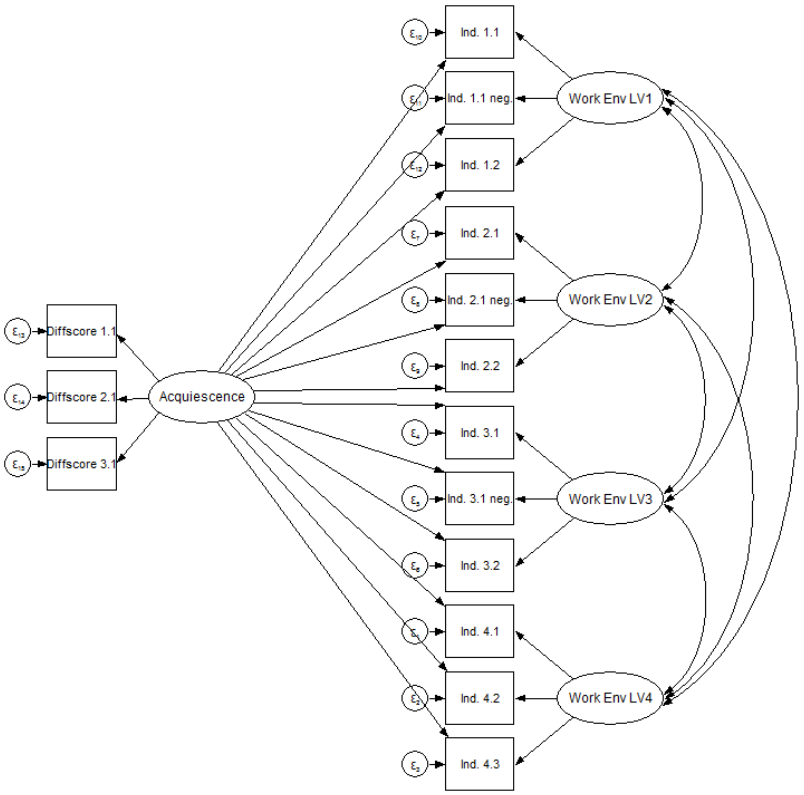


*Figure 5.4*. Modelling of- and controlling for Acquiescence as a systematic source of indicator error (co)variance. Makes use of difference scores computed from on negated and non-negated versions of well-functioning indicators as indicators of Acquiescence, and regressing all indicators belonging to working-environmental latent variables of interest on the Acquiescence latent variable. This approach to statistically remedy method bias is recommended by Podsakoff et al. (2003, 2012), preferably in conjunction with additional modelled sources of potential bias (e.g., positive and negative affectivity; see Podsakoff et al., 2003, p. 896). Such techniques should allow one to partial out construct irrelevant variance in indicators from specific sources. Abbreviations: Ind. = Indicator, Neg. = Negated, Work Env LV = Working Environment Latent Variable, Diffscore = Difference score computed from negated and non-negated version of an item.

While this might seem undesirable, by deciding to not control for conceivable sources of systematic error variance would be committing the "begging the question" fallacy (Kane, 2013a). By choosing not to do so,[5] one is decreasing ones confidence in (and thus the validity of) the interpretations of test scores as measuring a single latent variable. However, there is a carrot to go along with the stick. Filtering out the effects of acquiescence might contribute to successful discrimination between factors, as whatever covariance between them attributable to "yes-saying" can be filtered out. Thus, controlling for construct-irrelevant covariance can allow one to disentangle (statistically) strongly related but substantially distinct entities.

**Planned missingness design to counteract respondent reactivity to surveying.** Behind the scenes, there are murmurs of a concern that the KIWEST questionnaire contains too many items, and there is talk about making it leaner. The outcomes of this thesis suggest that there is indeed room for trimming, as a number of factors exhibit failure of statistical discrimination. By demonstrating near (if not substantive, at least functional) equivalence, the usefulness of the distinctions are questionable, as the inclusion of measures for different entities that prove near functional equivalence reduce the efficiency of the questionnaire by encumbering it with unnecessary items. Particularly in assessment settings perceived by test-takers as "low stakes" (Wise & DeMars, 2005), it is conceivable that unnecessarily detailed surveys can introduce motivational method effects (Podsakoff et al., 2003; Podsakoff et al., 2012) such as, for example, nonresponding (Heggestad, Rogelberg, Goh, & Oswald, 2015) and respondent fatigue (Bradley & Daly, 1994; Hess, Hensher, & Daly, 2012).

Respondent fatigue (ibid.) refers to a type of rank-order effect, where the test-takers engagement in responding drops as the questionnaire progresses. This arguably represents construct-irrelevant variance, as it is plausible that a respondent will turn towards engaging superficial heuristics when retrieving information necessary to respond to items with high construct fidelity. Thus, item response might be unduly influenced by fallible modes of reasoning (Evans & Stanovich, 2013; Tversky & Kahneman, 1973) that might cause the respondent to access more general information when responding to questions pertaining to

[5] It actually happens that even highly esteemed scholars refuse to control for systematic sources of error (co)variance *because* it detracts from reliability. As an example; Morgeson and Humphrey (2006, p. 1324), the creators of the "Work Design Questionnaire" managed to get the following sentence past peer review: "*because negatively worded items have been shown to produce factor structure problems in other work design measures […], items were positively worded such that greater levels of agreement indicated the presence of more of the work characteristic.*" This appears to be exactly the wrong way to think about indicators of convergence, providing a quintessential and prototypical example of putting the cart in front of the horse. To say that scholars are simply confusing means and ends by not controlling for systematic error variance would be giving them the benefit of the doubt. Not giving them the benefit of the doubt would be accusing them of intellectual dishonesty.

matters of specificity. In the context of the current validation effort; if this phenomenon is at play, it might help explain the failures of discrimination between factors intended to represent distinct properties (e.g., supportiveness and trustworthiness) of one particular (e.g., superiors).

Planned missingness designs (Graham et al., 2006) are strategies for efficiency in data collection that capitalize on the MCAR and MAR mechanisms of missing data, as well as the potency modern of missing data treatment techniques such as MI and FIML (Enders, 2010). What a planned missingness design entails is straightforward; a given survey is split into a number of forms that each contain a subset of the total number of survey items. Graham et al. (2006) propose one such design labelled "matrix sampling," suggesting that the alternative design forms include all of the sets but one, along with an "X-set" of items which are included in all of the forms. The X-set should contain the "central" items of the survey, which for KIWEST would perhaps be those pertaining to constructs representing the entities singled out in §1-1 of the NWEA (i.e., health-promotion, meaningfulness, safety, inclusiveness, equality).

This data collection strategy has the potential of drastically improving the efficiency of the questionnaire, with little to no cost to effectiveness. In fact, by potentially counteracting construct-irrelevant variance due to suppressing the actualization of motivational factors such as nonresponding and respondent fatigue, it might serve to increase effectiveness by reducing respondent reactivity to surveying. For example, the three-set matrix sampling design offered by Graham et al. (2006) would reduce the number of items administered to a respondent by up to a third (minus the X-set), with no loss of statistical power owing to the sophistication of the modern missing data treatment techniques (Enders, 2010; Enders & Bandalos, 2001). Planned missingness designs in questionnaire based research basically constitutes "win-win" in terms of efficiency and effectiveness that ought to be maximally exploited (T. D. Little et al., 2014).

**Limitations and Suggestions for Further Research**

An obvious limitation of the thesis is the complete and total absence of evidence pertaining to the source category of "response processes," which is the reason for why the thesis principally has dealt with the weak claims of latent variable measurement interpretations. Researchers engaged in more substantive research will no doubt find the claims validated in this thesis wanting for their own purposes, as they would require valid latent variable identity claims in order to examine and do research on the relationships between psychosocial phenomena. Future validation research should for this reason focus on accumulating and evaluating evidence informing "construct representation" in order to validate identity claims of the latent variables for which the measurement claims were successfully validated in this thesis.

It is perhaps worth noting that the response process category of evidence in the Standards appears underdeveloped and generally shunned by validators; even those that champion the CTM in validation, and thus for whom it would appear to be the most relevant (e.g., Markus & Borsboom, 2013b). For example, the theory of validation offered by Markus and Borsboom (2013a, p. 239) explicitly "*does not attempt to provide a theory of the tacit cognitive operations that it models. Test use does not seem to require a theory of that nature and it may be some time before cognitive psychology advances to the point that it could reasonably provide one.*" It seems ironic (and honestly appears disingenuous) that when push comes to shove, even the vocal proponents of establishing "*a causal assertion that a given construct causes a given set of test scores*" (Markus & Borsboom, 2013a, p. 221) in validation treat the response process category as a black box. They do however point out that there is potentially a lot of epistemological and methodological work to be done when it comes to developing the theory and practice of validating response process interpretations.

A possible avenue of research would be to examine whether- and the extent to which the application of grounded theory procedures (e.g., Charmaz, 2014; Corbin & Strauss, 2015) can prove useful for this purpose. Another possibility is to bring intervention practice into validation (e.g., MIMIC modelling of latent variables; Brown, 2015; Joreskog & Goldberger, 1975; Kline, 2016), as the identities of entities often suggest how the states of the entities can be manipulated (Markus & Borsboom, 2013a). For example, if perceived job autonomy is caused by structural freedom at work, then it follows that perceived job autonomy should increase if the employees' degrees of structural freedom at work is increased. What is clear is that there is need for alternatives to the impoverished practice of examining content evidence.

Furthermore, there is an elephant in the room; the complete and total absence of – as well as reference to – evidence pertaining to the source category of "consequences." This thesis has not concerned itself with particular uses of test scores beyond the practice of doing measurement-related interpretations. How these interpretations are to be used, that is, whether the interpretations lead to decisions and uses that affect, for example, individuals, working environments or organizations, have not been considered. The consequences category has basically been treated as wholly irrelevant to the question of whether the variance in sets of items appear to be caused by a phenomena that is distinct from phenomena causing variation in other sets of items included in KIWEST. In order to conform to the edicts of the Standards, future endeavors that would make some specific use of the outcomes of this thesis should give this source category of evidence its due consideration, this in order to examine the extent to which decisions made on the basis of test scores are justifiable (Cizek, 2012).

## Conclusion

This thesis has investigated the extent to which- and whether the latent variable measurement interpretation of the observations collected by administering the KIWEST 2.0 questionnaire on its target population can be claimed "valid" based on CFA. The results seem to suggest that; while the interpretation is for the most part comprehensive (i.e., low frequency and magnitude of convergence failures), it appears to suffer from lack of parsimony (i.e., high frequency and magnitude of discrimination failure; particularly when it comes to constructs relating to perceived properties of superiors and colleagues).

More specifically, fourteen of the 33 modelled latent variables managed to satisfy both of the criteria for discrimination and convergence, which might lead one to conclude that the KIWEST latent variable measurement interpretation is "42.4% valid." This might seem unduly conservative in light of its apparent comprehensiveness; but it is what one is justified in concluding based on available evidence relative to the established criteria. Thus, future endeavors should focus on developing valid measurement interpretations for which valid identity interpretations in turn can be formulated. The investigations performed in this thesis has not controlled for potential systematic sources of error variance however, which might help items converge on factors and to disentangle factors that fail to discriminate.

The results do indicate that one can proceed with validating the latent variable identity interpretations of the factors that managed to satisfy the criteria for valid measurement interpretations. The thesis has offered some suggestions for how one might approach this issue, as well as minor alterations to the questionnaire itself that can aid future validation work by procedurally (planned missingness) and statistically (incorporating measures of response styles) controlling for systematic construct-irrelevant variance; thus building measures of- and countermeasures to construct irrelevant variance into KIWEST itself.

**References**

AERA, APA, & NCME. (1999). *Standards for Educational and Psychological Testing* (5th ed.): American Educational Research Association.

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing* (6th ed.): American Educational Research Association.

Allison, P. D. (2002). Missing Data. *Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136*.

Andreassen, T. W., Lorentzen, B. G., & Olsson, U. H. (2006). The Impact of Non-Normality and Estimation Methods in SEM on Satisfaction Research in Marketing. *Quality and Quantity, 40*(1), 39-58. doi:10.1007/s11135-005-4510-y

Antonovsky, A. (1979). *Health, Stress, and Coping*: Jossey-Bass.

Antonovsky, A. (1996). The Salutogenic Model as a Theory to Guide Health Promotion. *Health Promotion International, 11*(1), 11-18. doi:10.1093/heapro/11.1.11

APA. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

ARK. (2016). Arbeidsmiljø- og klimaundersøkelser: KIWEST Version 2 (ARKKIWEST#2). Retrieved from https://hunt-db.medisin.ntnu.no/ark/#studyp369

ASD. (2005, 01.10.2015). Arbeidsmiljøloven. Retrieved from http://lovdata.no/dokument/NL/lov/2005-06-17-62

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly, 2*(1), 1-34. doi:10.1207/s15434311laq0201_1

Bakker, A. B., Demerouti, E., & Sanz-Vergel, A. I. (2014). Burnout and work engagement: The JD–R approach. *The Annual Review of Organizational Psychology and Organizational Behavior., 1*(1), 389-411. doi:10.1146/annurev-orgpsych-031413-091235

Bauer, G. F., & Jenny, G. J. (2007). Development, Implementation and Dissemination of Occupational Health Management (OHM): Putting Salutogenesis into Practice. In S. McIntyre & J. Houdmont (Eds.), *Occupational Health Psychology: European Perspectives on Research, Education and Practice.* (pp. 219-250): ISMAI.

Beehr, T. A., Walsh, J. T., & Taber, T. D. (1976). Relationships of stress to individually and organizationally valued states: Higher order needs as a moderator. *Journal of Applied Psychology, 61*(1), 41-47. doi:10.1037/0021-9010.61.1.41

Bollen, K. A. (2002). Latent Variables in Psychology and the Social Sciences. *Annual Review of Psychology, 53*(1), 605-634. doi:10.1146/annurev.psych.53.100901.135239

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological bulletin, 110*(2), 305-314. doi:10.1037/0033-2909.110.2.305

Bollen, K. A., & Long, J. S. (Eds.). (1993). *Testing Structural Equation Models*: Sage Publications, Inc.

Boring, E. G. (1945). The use of operational definitions in science. *Psychological Review., 52*(5), 243-245. doi:10.1037/h0054934

Borsboom, D. (2005). *Measuring the Mind: Conceptul Issues in Contemporary Psychometrics*: Cambridge University Press.

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives, 6*(1-2), 25-53. doi:10.1080/15366360802035497

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franić, S. (2009). The End of Construct Validity. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications* (pp. 135-170). Charlotte, NC, US: Information Age Publishing.

Borsboom, D., & Markus, K. A. (2013). Truth and Evidence in Validity Theory. *Journal of Educational Measurement, 50*(1), 110-114. doi:10.1111/jedm.12006

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*(2), 203-219. doi:10.1037/0033-295X.110.2.203

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review, 111*(4), 1061-1071. doi:10.1037/0033-295X.111.4.1061

Bradley, M., & Daly, A. (1994). Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. *Transportation, 21*(2), 167-184. doi:10.1007/bf01098791

Bridgman, P. W. (1927). *The Logic of Modern Physics*. New York: The Macmillan Company.

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.): Guilford Press.

Brown, T. A., & Moore, M. T. (2012). Confirmatory Factor Analysis. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling.* (pp. 361-379): Guilford Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin, 56*(2), 81-105. doi:10.1037/h0046016

Caplan, R. D. (1971). *Organizational Stress and Individual Strain: A Social-Psychological Study of Risk Factors in Coronar Heart Diseases among Administrators, Engineers, and Scientists.*: Institute for Social Research, University of Michigan.

Carless, S. A., & De Paola, C. (2000). The Measurement of Cohesion in Work Teams. *Small Group Research, 31*(1), 71-88. doi:10.1177/104649640003100104

Carnap, R. (1959). Psychology in Physical Language. In A. J. Ayer (Ed.), *Logical positivism* (pp. 165-198): Free Press.

Charmaz, K. (2014). *Constructing Grounded Theory* (2 ed.): SAGE Publications Ltd.

Christensen, M., Aronsson, G., Borg, V., Clausen, T., Gutenberg, J., Hakanen, J., . . . Straume, L. V. (2012). *Building Engagement and Healthy Organisations: Validation of the Nordic Questionnaire on Positive Organisational Psychology (N-POP). The Third Report from the Nordic Project.* Copenhagen: Nordic Council of Ministers.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43. doi:10.1037/a0026975

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of Validity Evidence for Educational and Psychological Tests: A Follow-Up Study. *Educational and Psychological Measurement, 70*(5), 732-743. doi:10.1177/0013164410379323

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2007). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement, 68*(3), 397-412. doi:10.1177/0013164407310130

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology, 94*, S95-S120. doi:10.2307/2780243

Collie, R. J., & Zumbo, B. D. (2014). Validity Evidence in the Journal of Educational Psychology: Documenting Current Practice and a Comparison with Earlier Practice. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (Vol. 54, pp. 113-135): Springer International Publishing.

Cook, J., Hepworth, S. J., Wall, T. D., & Warr, P. B. (1981). *The Experience of Work: A Compendium and Review of 249 Measures and their Use.*: Academic Press.

Cook, J., & Wall, T. (1980). New Work Attitude Measures of Trust, Organizational Commitment and Personal Need Non-Fulfilment. *Journal of Occupational Psychology, 53*, 35-52.

Corbin, J., & Strauss, A. (2015). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (4th ed.): Sage Publications Inc.

Cramer, A. O. J. (2012). Why the Item “23 +1” Is Not in a Depression Questionnaire: Validity From a Network Perspective. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 50-54. doi:10.1080/15366367.2012.681973

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281-302. doi:10.1037/h0040957

D'Agostino, R. B., & Belanger, A. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. *The American Statistician, 44*(4), 316-321. doi:10.2307/2684359

Dallner, M., Elo, A.-L., Gamerale, F., Hottinen, V., Knardahl, S., Lindström, K., . . . Ørhede, E. (2000). *Validation of the General Nordic Questionaire (QPSNordic) for Psychological and Social Factors at Work.* Copenhagen: Nordic Council of Ministers.

DeVellis, R. F. (2017). *Scale Development: Theory and Applications* (4th ed.): SAGE Publications, Inc.

Doornik, J. A., & Hansen, H. (2008). An Omnibus Test for Univariate and Multivariate Normality*. *Oxford Bulletin of Economics and Statistics, 70*, 927-939. doi:10.1111/j.1468-0084.2008.00537.x

Edwards, J. R. (2010). The Fallacy of Formative Measurement. *Organizational Research Methods, 14*(2), 370-388. doi:10.1177/1094428110378369

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5*(2), 155-174. doi:10.1037/1082-989X.5.2.155

Eignor, D. R. (2013). The standards for educational and psychological testing. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 245-250). Washington, DC, US: American Psychological Association.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380-396. doi:10.1037/1082-989X.3.3.380

Embretson, S. E. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? *Educational Researcher, 36*(8), 449-455. doi:10.3102/0013189x07311600

Enders, C. K. (2010). *Applied Missing Data Analysis*: Guilford Publications.

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models.

*Structural Equation Modeling: A Multidisciplinary Journal, 8*(3), 430-457. doi:10.1207/S15328007SEM0803_5

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition. *Perspectives on Psychological Science, 8*(3), 223-241. doi:10.1177/1745691612460685

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research, 18*(1), 39-50. doi:10.2307/3151312

Frisbie, D. A. (2005). Measurement 101: Some Fundamentals Revisited. *Educational Measurement: Issues and Practice, 24*(3), 21-28. doi:10.1111/j.1745-3992.2005.00016.x

Frone, M. R. (2003). Work-Family Balance. In J. C. Quick & L. E. Tetrick (Eds.), *Occupational Healt Psychology* (pp. 143-162): American Psychological Association.

Gadamer, H.-G. (1974/2004). *Truth and Method* (J. Weinsheimer & D. G. Marshall, Trans. 2nd. ed.): Continuum.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned Missing Data Designs in Psychological Research. *Psychological Methods, 11*(4). doi:10.1037/1082-989X.11.4.323

Guion, R. M. (1977). Content Validity—The Source of My Discontent. *Applied Psychological Measurement, 1*(1), 1-10. doi:10.1177/014662167700100103

Guion, R. M. (1980). On Trinitarian doctrines of validity. *Professional Psychology, 11*(3), 385-398. doi:10.1037/0735-7028.11.3.385

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*: Cambridge University Press.

Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology, 60*(2), 159-170. doi:10.1037/h0076546

Haig, B. D. (2012). From Construct Validity to Theory Validation. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 59-62. doi:10.1080/15366367.2012.681975

Hair, J. F. J., Black, W. C., Babin, B. J., & Anderson, R. E. (2013). *Multivariate Data Analysis* (7th ed.): Pearson Education Limited.

Hanson, G. C., Hammer, L. B., & Colton, C. L. (2006). Development and validation of a multidimensional scale of perceived work-family positive spillover. *Journal of Occupational Health Psychology, 11*(3), 249-265. doi:10.1037/1076-8998.11.3.249

Heggestad, E. D., Rogelberg, S., Goh, A., & Oswald, F. L. (2015). Considering the effects of nonresponse on correlations between surveyed variables: A simulation study to provide context to evaluate survey results. *Journal of Personnel Psychology, 14*(2), 91-103. doi:10.1027/1866-5888/a000129

Hellgren, J., Sverke, M., & Näswall, K. (2008). Changing Work Roles: New Demands and Challanges. In K. Näswall, J. Hellgren, & M. Sverke (Eds.), *The Individual in the Changing Working Life* (pp. 46-66): Cambridge University Press.

Hess, S., Hensher, D. A., & Daly, A. (2012). Not bored yet – Revisiting respondent fatigue in stated choice experiments. *Transportation Research Part A: Policy and Practice, 46*(3), 626-644. doi:10.1016/j.tra.2011.11.008

Hinz, A., Michalski, D., Schwarz, R., & Herzberg, P. Y. (2007). The acquiescence effect in responding to a questionnaire. *GMS Psycho-Social Medicine, 4*, Doc07.

Hood, S. B. (2012). In Defense of an Instrument-Based Approach to Validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 63-65. doi:10.1080/15366367.2012.681976

Hovmark, S., & Thomsson, H. (1995). *ASK - ett frågeformulär för at mäte arbetsbelastning, socialt stöd, kontroll och kompetens i arbetslivet (Reports from the Department of Psychology 86).* Stockholm, Sweden: Institute of Psychology, The University of Stockholm.

Hoyle, R. H. (Ed.) (2012). *Handbook of Structural Equation Modeling*: Guilford Press.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi:10.1080/10705519909540118

Ilgen, D. R., & Hollenbeck, J. R. (1991). The Structure of Work: Job Design and Roles. In D. M. D & L. M. Hough (Eds.), *Handbook of Industrial and Organizational Psychology* (2nd ed., Vol. 2, pp. 165-207): Consulting Psychologists Press.

Innstrand, S. T., Christensen, M., Undebakke, K. G., & Svarva, K. (2015). The Presentation and Preliminary Validation of KIWEST Using a Large Sample of Norwegian University Staff. *Scandinavian Journal of Public Health*, 1-12. doi:10.1177/1403494815600562

Innstrand, S. T., Langballe, E. M., Falkum, E., Espnes, G. A., & Aasland, O. G. (2009). Gender-specific perceptions of four dimensions of the work/family interaction. *Journal of Career Assessment, 17*(4), 402-416. doi:10.1177/1069072709334238

James, W. (1890/2015). *The Principles of Psychology* (Vol. 1): Forgotten Books.

Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician), 47*(1), 183-189. doi:10.1111/1467-9884.00122

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*(4), 411-414. doi:10.1037/1082-989X.5.4.411

Joreskog, K. G., & Goldberger, A. S. (1975). Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable. *Journal of the American Statistical Association, 70*(351), 631-639. doi:10.2307/2285946

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika, 34*(2), 183-202. doi:10.1007/BF02289343

Jöreskog, K. G. (1993). Testing Structural Equation Models. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 294-316): Sage Publications, Inc.

Kagan, J. (2005). A Time for Specificity. *Journal of Personality Assessment, 85*(2), 125-127. doi:10.1207/s15327752jpa8502_03

Kahn, R. L., Wolfe, D. M., Quinn, R. P., Snoek, J. D., & Rosenthal, R. (1964). *Organizational Stress: Studies in Role Conflict and Ambiguity.*: Wiley.

Kane, M. T. (2012). All Validity Is Construct Validity. Or Is It? *Measurement: Interdisciplinary Research and Perspectives*, *10*(1-2), 66-70. doi:10.1080/15366367.2012.681977

Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi:10.1111/jedm.12000

Kane, M. T. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50*(1), 115-122. doi:10.1111/jedm.12007

Katz, D., & Kahn, R. L. (1978). *The Social Psychology of Organizations.* (2 ed.): Wiley.

Kelloway, E. K., & Barling, J. (1990). Item content versus item wording: Disentangling role conflict and role ambiguity. *Journal of Applied Psychology, 75*(6), 738-742. doi:10.1037/0021-9010.75.6.738

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling* (4th ed.): Guilford.

Kristensen, T. S., Hannerz, H., Høgh, A., & Borg, V. (2005). The Copenhagen Psychosocial Questionnaire-a tool for the assessment and improvement of the psychosocial work environment. *Scandinavian Journal of Work, Environment & Health, 31*(6), 438-449. doi:10.2307/40967527

Kuhn, T. S. (1962/2012). *The Structure of Scientific Revolutions. 50th Anniversary Edition.*: Chicago University Press.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria. *Organizational Research Methods, 9*(2), 202-220. doi:10.1177/1094428105284919

Li, C. (2013). Little's test of missing completely at random. *Stata Journal, 13*(4), 795-809.

Lissitz, R. W. (Ed.) (2009). *The Concept of Validity: Revisions, New Directions, and Applications*: Information Age Publishing.

Lissitz, R. W., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher, 36*(8), 437-448. doi:10.3102/0013189x07311286

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association, 83*(404), 1198-1202. doi:10.1080/01621459.1988.10478722

Little, T. D., Jorgensen, T. D., Lang, K. M., & Moore, E. W. G. (2014). On the Joys of Missing Data. *Journal of Pediatric Psychology, 39*(2), 151-162. doi:10.1093/jpepsy/jst048

Loevinger, J. (1957). Objective Tests As Instruments Of Psychological Theory: Monograph Supplement 9. *Psychological reports, 3*(3), 635-694. doi:10.2466/pr0.1957.3.3.635

Lovasz, N., & Slaney, K. L. (2013). What makes a hypothetical construct “hypothetical”? Tracing the origins and uses of the ‘hypothetical construct’ concept in psychological science. *New Ideas in Psychology, 31*(1), 22-31. doi:10.1016/j.newideapsych.2011.02.005

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201-226. doi:10.1146/annurev.psych.51.1.201

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological bulletin, 111*(3), 490-504. doi:10.1037/0033-2909.111.3.490

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55*(2), 95-107. doi:10.1037/h0056029

Maraun, M. D., & Gabriel, S. M. (2013). Illegitimate concept equating in the partial fusion of construct validation theory and latent variable modeling. *New Ideas in Psychology, 31*(1), 32-42. doi:10.1016/j.newideapsych.2011.02.006

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530. doi:10.1093/biomet/57.3.519

Markus, K. A. (2008). Constructs, Concepts and the Worlds of Possibility: Connecting the Measurement, Manipulation, and Meaning of Variables. *Measurement: Interdisciplinary Research and Perspectives, 6*(1-2), 54-77. doi:10.1080/15366360802035513

Markus, K. A., & Borsboom, D. (2013a). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*: Routledge.

Markus, K. A., & Borsboom, D. (2013b). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology, 31*(1), 54-64. doi:10.1016/j.newideapsych.2011.02.008

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320-341. doi:10.1207/s15328007sem1103_2

Maslach, C., Schaufeli, W. B., & Leiter, M. P. (2001). Job Burnout. *Annual Review of Psychology, 52*(1), 397-422. doi:doi:10.1146/annurev.psych.52.1.397

McGrath, R. E. (2005a). Conceptual Complexity and Construct Validity. *Journal of Personality Assessment, 85*(2), 112-124. doi:10.1207/s15327752jpa8502_02

McGrath, R. E. (2005b). Rethinking Psychosocial Constructs: Reply to Comments by Barrett, Kagan, and Maraun and Peters. *Journal of Personality Assessment, 85*(2), 141-145. doi:10.1207/s15327752jpa8502_06

Mehmetoglu, M., & Jakobsen, T. G. (2017). *Applied Statistics Using Stata: A Guide for the Social Sciences*: SAGE Publications Ltd.

Mellor, S., Mathieu, J. E., & Swim, J. K. (1994). Cross-level analysis of the influence of local union structure on women's and men's union commitment. *Journal of Applied Psychology, 79*(2), 203-210. doi:10.1037/0021-9010.79.2.203

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*(10), 955-966. doi:10.1037/0003-066X.30.10.955

Messick, S. (1987). Validity. *ETS Research Report Series, 1987*(2), i-208. doi:10.1002/j.2330-8516.1987.tb00244.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi:10.1037/0003-066X.50.9.741

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology, 31*(1), 13-21. doi:10.1016/j.newideapsych.2011.02.004

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*(3), 248-260. doi:10.1037/1082-989X.2.3.248

Millsap, R. E., & Everson, H. T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement, 17*(4), 297-334. doi:10.1177/014662169301700401

Mislevy, R. J. (2009). Validity from the Perspective of Model-Based Reasoning. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications* (pp. 83-108): Information Age Publishing.

Moran, E. T., & Volkwein, J. F. (1992). The Cultural Approach to the Formation of Organizational Climate. *Human Relations, 45*(1), 19-47. doi:10.1177/001872679204500102

Morgeson, F. P., & Humphrey, S. E. (2006). The Work Design Questionnaire (WDQ): Developing and validating a comprehensive measure for assessing job design and the nature of work. *Journal of Applied Psychology, 91*(6), 1321-1339. doi:10.1037/0021-9010.91.6.1321

Newton, P. E. (2012a). Clarifying the Consensus Definition of Validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 1-29. doi:10.1080/15366367.2012.669666

Newton, P. E. (2012b). Questioning the Consensus Definition of Validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 110-122. doi:10.1080/15366367.2012.688456

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301-319. doi:10.1037/a0032969

Newton, P. E., & Shaw, S. D. (2014). *Validity in Educational and Psychological Assessment*: SAGE Publications Ltd.

Nooteboom, B. (Ed.) (2003). *The Trust Process in Organizations: Empirical Studies of the Determinants and the Process of Trust Development*: Edward Elgar Publishing.

Näswall, K., Låstad, L., Vetting, T.-S., Larsson, R., Richter, A., & Sverke, M. (2010). *Job insecurity from a gender perspective: Data collection and psychometric properties*. Stockholm: Stockholm University Department of Psychology.

Pejtersen, J. H., Kristensen, T. S., Borg, V., & Bjorner, J. B. (2010). The second version of the Copenhagen Psychosocial Questionnaire. *Scandinavian Journal of Public Health, 38*(Suppl 3), 8-24. doi:10.1177/1403494809349858

Penfield, R. D. (2013). Item analysis. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 121-138). Washington, DC, US: American Psychological Association.

Perrewé, P. L., Hochwarter, W. A., & Kiewitz, C. (1999). Value attainment: An explanation for the negative effects of work–family conflict on job and life satisfaction. *Journal of Occupational Health Psychology, 4*(4), 318-326. doi:10.1037/1076-8998.4.4.318

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879-903. doi:10.1037/0021-9010.88.5.879

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of Method Bias in Social Science Research and Recommendations on How to Control It. *Annual Review of Psychology, 63*(1), 539-569. doi:10.1146/annurev-psych-120710-100452

Polanyi, M. (1966/2009). *The Tacit Dimension*: The University of Chicago Press.

Popper, K. R. (1963/2002). *Conjectures and Refutations: The Growth of Scientific Knowledge*: Routledge.

Raykov, T. (1997). Estimation of Composite Reliability for Congeneric Measures. *Applied Psychological Measurement, 21*(2), 173-184. doi:10.1177/01466216970212006

Rizzo, J. R., House, R. J., & Lirtzman, S. I. (1970). Role Conflict and Ambiguity in Complex Organizations. *Administrative Science Quarterly, 15*(2), 150-163. doi:10.2307/2391486

Robinson, S. L. (1996). Trust and Breach of the Psychological Contract. *Administrative Science Quarterly, 41*(4), 574-599. doi:10.2307/2393868

Rogers, H. J., & Swaminathan, H. (2016). Concepts and Methods in Research on Differential Functioning of Test Items: Past, Present, and Future. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational Measurement: From Foundations to Future* (pp. 126-142): Guilford Press.

Rousseau, D. M. (2011). The individual–organization relationship: The psychological contract *APA handbook of industrial and organizational psychology, Vol 3: Maintaining, expanding, and contracting the organization* (pp. 191-220). Washington, DC, US: American Psychological Association.

Schaufeli, W. B., & Bakker, A. B. (2004). Job Demands, Job Resources, and their Relationship with Burnout and Engagement: A Multi-Sample Study. *Journal of organizational Behavior., 25*(3), 293-315. doi:10.1002/job.248

Schaufeli, W. B., Bakker, A. B., & Salanova, M. (2006). The Measurement of Work Engagement With a Short Questionnaire: A Cross-National Study. *Educational and Psychological Measurement, 66*(4), 701-716. doi:10.1177/0013164405282471

Schaufeli, W. B., Salanova, M., González-romá, V., & Bakker, A. B. (2002). The Measurement of Engagement and Burnout: A Two Sample Confirmatory Factor Analytic Approach. *Journal of Happiness Studies, 3*(1), 71-92. doi:10.1023/A:1015630930326

Schaufeli, W. B., Shimazu, A., & Taris, T. W. (2009a). Being Driven to Work Excessively Hard: The Evaluation of a Two-Factor Measure of Workaholism in The Netherlands and Japan. *Cross-Cultural Research*. doi:10.1177/1069397109337239

Schaufeli, W. B., Shimazu, A., & Taris, T. W. (2009b). Being driven to work excessively hard: The evaluation of a two-factor measure of workaholism in the Netherlands and Japan. *Cross-Cultural Research: The Journal of Comparative Social Science, 43*(4), 320-348. doi:10.1177/1069397109337239

Scherermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online, 8*(2), 23-74.

Schmittmann, V. D., Cramer, A. O. J., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology, 31*(1), 43-53. doi:10.1016/j.newideapsych.2011.02.007

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93-105. doi:10.1037/0003-066X.54.2.93

Semmer, N. K., Amstad, F., & Elfering, A. (2006). *Dysfunctional Support.* Paper presented at the The 6th International Conference on Occupational Stress and Health., Miami, Florida.

Semmer, N. K., Tschan, F., Meier, L. L., Facchin, S., & Jacobshagen, N. (2010). Illegitimate tasks and counterproductive work behavior. *Applied Psychology: An International Review, 59*(1), 70-96. doi:10.1111/j.1464-0597.2009.00416.x

Shear, B. R., & Zumbo, B. D. (2014). What Counts as Evidence: A Review of Validity Studies in Educational and Psychological Measurement. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and Validation in Social, Behavioral, and Health Sciences* (Vol. 54, pp. 91-111): Springer International Publishing.

Sheldon, S., & Burke, P. J. (2000). The Past, Present, and Future of an Identity Theory. *Social Psychology Quarterly, 63*(4), 284-297. doi:10.2307/2695840

Sieber, S. D. (1974). Toward a Theory of Role Accumulation. *American Sociological Review, 39*(4), 567-578.

Siegrist, J. (1996). Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology, 1*(1), 27-41. doi:10.1037/1076-8998.1.1.27

Sireci, S. G. (2007). On Validity Theory and Test Validation. *Educational Researcher, 36*(8), 477-481. doi:10.3102/0013189x07311609

Sireci, S. G. (2009). Packing and Unpacking Sources of Validity Evidence: History Repeats Itself Again. In R. Lissitz (Ed.), *The Concept of Validity: Revisions, New Directions, and Applications* (pp. 19-37): Information Age Publishing.

Sireci, S. G., & Faulkner-Bond, M. (2016). The Times They Are A-Changing, but the Song Remains the Same: Future Issues and Practices in Test Validation. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational Measurement: From Foundations to Future* (pp. 435-448): Guilford Press.

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology* (pp. 61-84). Washington, DC, US: American Psychological Association.

Slaney, K. L., & Racine, T. P. (2013a). Constructing an understanding of constructs. *New Ideas in Psychology, 31*(1), 1-3. doi:10.1016/j.newideapsych.2011.02.010

Slaney, K. L., & Racine, T. P. (2013b). What's in a name? Psychology's ever evasive construct. *New Ideas in Psychology, 31*(1), 4-12. doi:10.1016/j.newideapsych.2011.02.003

Spearman, C. (1961). "General Intelligence" Objectively Determined and Measured. In J. J. J. D. G. Paterson (Ed.), *Studies in individual differences: The search for intelligence* (pp. 59-73). East Norwalk, CT, US: Appleton-Century-Crofts.

Spector, P. E., & Fox, S. (2002). An emotion-centered model of voluntary work behavior: Some parallels between counterproductive work behavior and organizational citizenship behavior. *Human Resource Management Review, 12*(2), 269-292. doi:10.1016/S1053-4822(02)00049-9

Stang, I. (2003). Bemyndigelse: en Innføring i Begrepet "Empowermenttenkningens" Relevans for Ansatte i Velferdsstaten. In H. A. Hauge & M. B. Mittelmark (Eds.), *Helsefremmende Arbeid i en Brytningstid: fra Monolog til Dialog?* (pp. 141-161): Fagbokforlaget.

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research, 25*(2), 173-180. doi:10.1207/s15327906mbr2502_4

Stevens, S. S. (1935). The operational definition of psychological concepts. *Psychological Review., 42*(6), 517-527. doi:10.1037/h0056973

Sverke, M., & Sjoberg, A. (1994). Dual Commitment to Company and Union in Sweden: An Examination of Predictors and Taxonomic Split Methods. *Economic and Industrial Democracy, 15*(4), 531-564. doi:10.1177/0143831x94154003

Tadić, M., Bakker, A. B., & Oerlemans, W. G. M. (2015). Challenge versus hindrance job demands and well-being: A diary study on the moderating role of job resources. *Journal of Occupational and Organizational Psychology, 88*(4), 702-725. doi:10.1111/joop.12094

Tanaka, J. S. (1993). Multifaceted Conceptions of Fit in Structural Equation Models. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 10-39): Sage Publications, Inc.

Thoits, P. A. (1991). On Merging Identity Theory and Stress Research. *Social Psychology Quarterly, 54*(2), 101-112. doi:10.2307/2786929

Thomas, K. W., & Velthouse, B. A. (1990). Cognitive Elements of Empowerment: An "Interpretive" Model of Intrinsic Task Motivation. *Academy of Management Review, 15*(4), 666-681. doi:10.5465/amr.1990.4310926

Tolman, E. C. (1936). An Operational Analysis of "Demands". *Erkenntnis, 6*, 383-392.

Tomassi, P. (1999). *Logic* (1st ed.): Routledge.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist, 24*(2), 83-91. doi:10.1037/h0027108

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207-232. doi:10.1016/0010-0285(73)90033-9

Tyler, T. R. (1989). The psychology of procedural justice: A test of the group-value model. *Journal of personality and social psychology, 57*(5), 830-838. doi:10.1037/0022-3514.57.5.830

Undebakke, K. G., Innstrand, S. T., Anthun, K. S., & Christensen, M. (2014). *ARK: The ARK Intervention Programme, Who - What - How*. Retrieved from http://www.ntnu.no/documents/34221120/0/2015_01_ARK_web.pdf/65dd9de7-b61f-4f0b-9829-bf87dd58cfda

van der Vliet, C., & Hellgren, J. (2002). *The Modern Working Life: Its Impact on Employee Attitutes, Performance, and Health (SALTSA report 4).* Sweden: National Institute for Working Life & SALTSA.

Vogt, K., Jenny, G. J., & Bauer, G. F. (2013). Comprehensibility, manageability and meaningfulness at work: Construct validity of a scale measuring work-related sense of coherence. *SA Journal of Industrial Psychology, 39*(1), 1-8. doi:10.4102/sajip.v39i1.1111

Voydanoff, P. (2002). Linkages Between the Work-family Interface and Work, Family, and Individual Outcomes: An Integrative Model. *Journal of Family Issues, 23*(1), 138-164. doi:10.1177/0192513x02023001007

Wacker, J. G. (1998). A Definition of Theory: Research Guidelines for Different Theory-Building Research Methods in Operations Management. *Journal of Operations Management., 16*(4), 361-385. doi:10.1016/S0272-6963(98)00019-9

Wacker, J. G. (2004). A Theory of Formal Conceptual Definitions: Developing Theory-Building Measurement Instruments. *Journal of Operations Management., 22*(6), 629-650. doi:10.1016/j.jom.2004.08.002

Walsh, J. T., Taber, T. D., & Beehr, T. A. (1980). An integrated model of perceived job characteristics. *Organizational Behavior and Human Performance, 25*(2), 252-267. doi:10.1016/0030-5073(80)90066-5

Wayne, J. H., Musisca, N., & Fleeson, W. (2004). Considering the role of personality in the work–family experience: Relationships of the big five to work–family conflict and

facilitation. *Journal of Vocational Behavior, 64*(1), 108-130. doi:10.1016/S0001-8791(03)00035-6

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236-247. doi:10.1016/j.ijresmar.2010.02.004

Weijters, B., Schillewaert, N., & Geuens, M. (2008). Assessing response styles across modes of data collection. *Journal of the Academy of Marketing Science, 36*(3), 409-422. doi:10.1007/s11747-007-0077-6

West, S. G., Taylor, A. B., & Wu, W. (2012). Model Fit and Model Selection in Structural Equation Modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling.* (pp. 209-231): Guilford Press.

Wise, S. L., & DeMars, C. E. (2005). Low Examinee Effort in Low-Stakes Assessment: Problems and Potential Solutions. *Educational Assessment, 10*(1), 1-17. doi:10.1207/s15326977ea1001_1

Wothke, W. (1993). Nonpositive Definite Matrices in Structural Modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 256-293): Sage Publications, Inc.

Zumbo, B. D., & Chan, E. K. H. (Eds.). (2014). *Validity and Validation in Social, Behavioral, and Health Sciences*: Springer International Publishing.

**Appendix 1: The LVIV Model of Latent Variable Interpretation Validation**