UHASSELT
KNOWLEDGE IN ACTION
www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Validation of a short survey concerning workability

Jimmy Komalceh

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Steven ABRAMS
Mevrouw Lieve VAN DYCK

2024
2025

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Validation of a short survey concerning workability

Jimmy Komalceh

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Steven ABRAMS

Mevrouw Lieve VAN DYCK

Acknowledgements

My mountain-sized thanks go to my supervisors, Prof. Dr. Steven Abrams (Hasselt University) and Mrs. Lieve Van Dyck (Mensura), whose expertise and careful guidance have brought invaluable insights to this research. This final report of my work is far better for their effort. The remaining errors and lack of clarity are the result of not heeding their advice.

I am grateful to Hasselt University, both the teaching and non-teaching staff, as well as to my colleagues, for creating such a stimulating and supportive learning environment.

I am forever indebted to my wife, Florence, for her unwavering love and support, and to my two sons, Phinehas Komael and Kemuel Caden, for their joy and patience during this journey. I acknowledge the strong educational foundation laid by my mother, Mrs. Dorine Ocen, who never had this opportunity herself, and I express deep gratitude to my sister, Apokowat Winnie, for her financial support.

Finally, to the One who founded and holds the universe, with whom time and space belong, and who is able to make all things—both challenges and opportunities—work together for my good: my opportunity to undertake this Master's thesis and complete my studies was not by chance, but purely by His grace and perfect timing.

Abstract

Performing a reliable assessment of workplace well-being in diverse populations remains a central challenge in occupational health research. Instruments must demonstrate robust measurement equivalence to ensure fair and meaningful comparisons. This study presents a comprehensive psychometric validation of the OHS Barometer e-survey, designed to assess multiple dimensions of workplace well-being by evaluating factor structure, internal consistency, and measurement invariance across key demographic and methodological subgroups.

Data were collected from 699 employees across four Belgian companies in the retail and wholesale sectors. Participants completed Dutch (88.3%) or French (10.4%) survey versions, with 25% using mobile devices. The evaluation done in this master's thesis work included confirmatory factor analysis, reliability assessments using multiple measures (McDonald's omega, polychoric alpha, composite reliability) suited to ordinal data, and measurement invariance testing using both traditional null hypothesis significance testing and modern equivalence-based approaches. Missing data (5–6%) were handled using multiple imputation with bootstrap-based uncertainty estimation.

CFA results supported a four-factor structure comprising Psychosocial, Ergonomics, Safety, and Hygiene domains, with excellent model fit indices (CFI = 0.987, TLI = 0.984, SRMR = 0.075). All domains demonstrated strong internal consistency, with polychoric alpha values ranging from 0.817 to 0.935. Full scalar measurement invariance was established across language groups, data collection methods, and gender. Age-based analyses revealed developmental differences in response patterns, particularly greater sensitivity to physical workplace conditions among older employees, rather than measurement bias.

Two critical findings emerged. First, the strong inter-factor correlation between Ergonomics and Hygiene domains ($r = 0.777$) violated the Fornell-Larcker criterion for discriminant validity, suggesting empirical inseparability despite theoretical distinctions. Second, traditional measurement invariance testing for the Environmental domain failed to converge. Subsequent differential item functioning analysis indicated that observed issues stemmed from questionnaire version differences rather than instrument problems. These findings highlight the need to account for survey version effects when assessing measurement equivalence, challenging traditional assumptions about measurement invariance in organizational research.

Although limited to cross-sectional design within Belgian retail and wholesale sectors, the OHS Barometer emerges as a psychometrically sound and cross-platform compatible tool, offering practitioners a validated instrument for workplace well-being assessment while providing researchers with important insights into measurement dynamics in organizational contexts.

Keywords: workplace well-being, psychometric validation, measurement invariance, confirmatory factor analysis, occupational health assessment, organizational context

Contents

List of Figures

List of Tables

List of Acronyms

AVE	Average Variance Extracted
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
CR	Composite Reliability
DIF	Differential Item Functioning
ET	Equivalence Testing
EU-OSHA	European Union information Agency for Occupational Safety and Health
FMI	Fraction of Missing Information
HR	Human Resource
IFI	Imputation Fit Index
IRT	Item Response Theory
KMO	Kaiser-Meyer-Olkin
MCAR	Missing Completely At Random
MGCFA	Multi-Group Confirmatory Factor Analysis
MICE	Multiple Imputation with Chained Equations
MNAR	Missing Not At Random
NHT	Null Hypothesis Testing
OHS/OSH	Occupational Health and Safety
R&D	Research and Development
RMSEA	Root Mean Square Error of Approximation
SRMR	Standardized Root Mean Square Residual
TLI/NNFI	Tucker-Lewis Index / Non-Normed Fit Index

VIP	Variance Inflation Factor
WAI	Work Ability Index
WHO	World Health Organization
WLSMV	Weighted Least Squares Mean and Variance adjusted

Chapter 1

Introduction

1.1 Background

Workability is a pivotal concept in occupational health, representing an employee's capacity to meet job demands while maintaining physical, mental, and social well-being. This multifaceted construct is shaped by individual characteristics, workplace conditions, and broader societal influences (Gould et al., 2008). Accurately assessing workability is critical for employers to implement targeted interventions that support employee health and productivity throughout their careers.

Occupational health and safety (OHS/OSH) remains a top priority within the European Union (EU), with policies aimed at fostering safe and healthy working environments. The OSH Barometer, an EU-wide public information system, offers valuable insights into workplace safety, health, and well-being (EU-OSHA, 2025). According to a recent report by EU-OSHA, *Safety and Health at Work in Europe: Status and Trends in 2023*, fatal workplace accidents have decreased by 57% in recent decades. However, these improvements have plateaued in recent years, signaling the need for renewed focus and innovative approaches (Eurogip, 2024).

In Belgium, the responsibility for OHS lies with the Minister of Employment and the Federal Public Service for Employment, Labour, and Social Dialogue. The Act of 4 August 1996 serves as the cornerstone of Belgium's legislative framework, mandating that employers prioritize worker well-being through prevention, safety training, and risk mitigation (SPF Emploi, Travail et Concertation sociale, 2025). This legislation underscores the importance of proactive workplace policies in cultivating a culture of safety and well-being.

1.2 Concept of Workability

The concept of workability was first introduced by Ilmarinen and Tuomi (1993) as a multidimensional construct that encompasses an individual's ability to perform work effectively in relation to job demands, health status, and mental resources. Traditionally, workability has been measured using the Work Ability Index (WAI) (Ilmarinen & Tuomi, 1993), which assumes a unidimensional structure focused primarily on a medical perspective. However, recent studies suggest that workability is better understood through a multidimensional lens (Martus et al., 2010; Radkiewicz & Widerszal-Bazyl, 2005), leading to an evolution toward a more comprehensive model that balances various components, including job demands, environmental factors, and individual resources. This shift highlights the complexity and dynamic nature of workability, emphasizing the need for a holistic approach to its assessment that goes beyond traditional occupational health perspectives.

In this study, we adopt a multidimensional approach to workability, grounded in theoretical frameworks, developed by experts at Mensura, a leading Belgian occupational health service. The survey employed in this study assesses subjective workability through self-assessment rather than relying solely on evaluations by occupational physicians or other healthcare professionals. According to Tuomi (1991), subjective estimates are strong predictors of future work ability and disability (Tuomi et al., 1991). In addition, subjective assessment methods, such as the demand-specific workability approach (Nabe-Nielsen et al., 2014), offer nuanced insights into employees' perceived workability across diverse job contexts.

In human resource management, extending workers' careers is unfeasible if employees reach a point where they are unable to continue working. Therefore, it is crucial for employers to understand workability comprehensively to develop effective interventions, foster employee engagement, and ensure organizational sustainability (Pak et al., 2021).

1.3 Mensura's OHS Barometer

Mensura is developing the OHS Barometer, a data-driven tool designed to provide employers with actionable insights into employee well-being through a comprehensive assessment of workability dimensions.

This barometer focuses on occupational well-being, a multidimensional construct encompassing physical, mental, and social health dimensions that enable optimal workplace functioning. It combines subjective experiences of satisfaction and engagement with objective health measures (Dodge et al., 2012). The World Health Organization (WHO) defines it as a state of complete physical, mental, and social well-being, not merely the absence of disease in relation to work (World Health Organization, 2018). As highlighted in Section 1.2, this concept aligns with modern workability perspectives, reflecting the dynamic balance between individual capabilities, job demands, and environmental factors that support sustainable employment (Ilmarinen, 2019; Schulte & Vainio, 2010).

1.3.1 OHS Barometer Domains

The *OHS Barometer* evaluates four core domains of workplace well-being:

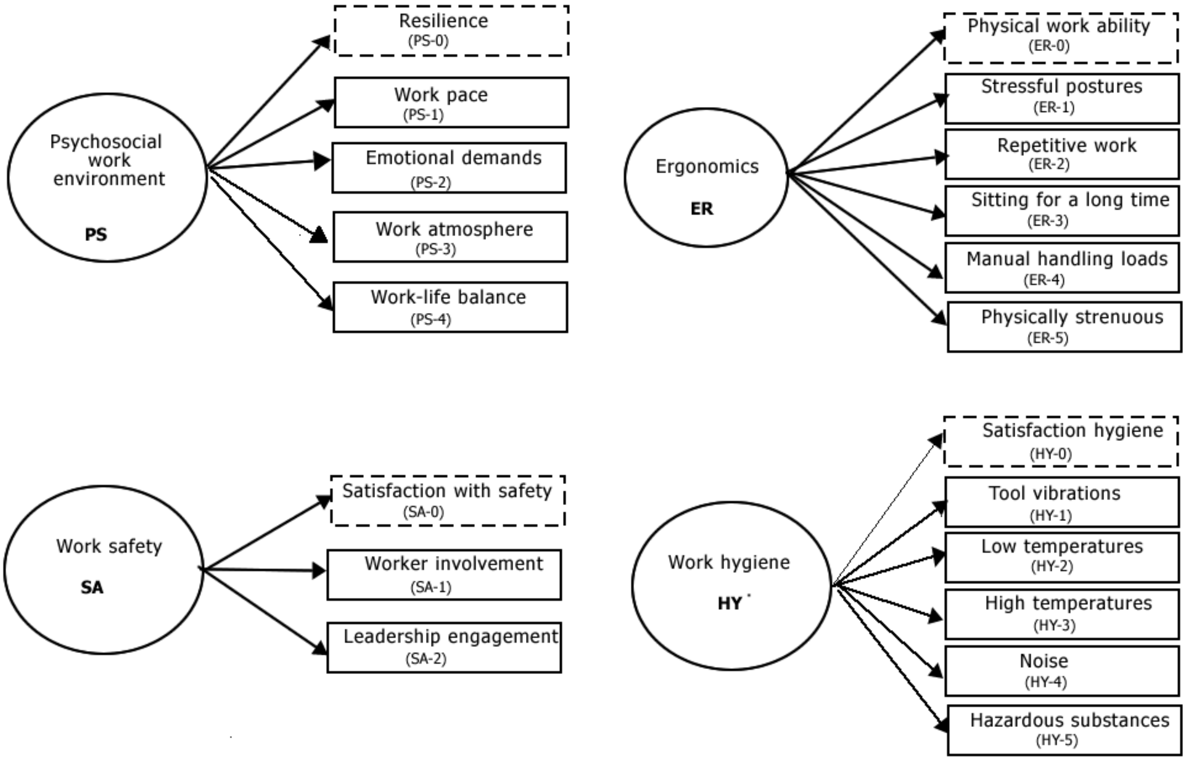


Figure 1.1: Model framework including four well-being domains as integrated in the OHS Barometer developed by Mensura. More specifically, the model diagram includes ER, HY, SA, and PS work environments. The link between well-being domains and specific items in the barometer are indicated by arrows. Solid boxes refer to reflective indicators (i.e., reflective of the work environment), dashed boxes represent formative indicators (i.e., defining the work environment itself).

1. **Psychosocial work environment (PS):** Measured through indicators such as work pace, emotional demands, work atmosphere, and work-life balance, with resilience being a formative indicator[1] for the PS domain.
2. **Ergonomics (ER):** Assessed through indicators such as stressful postures, repetitive work, sitting for a long time, manual handling of loads, and physically strenuous activities, with physical work ability being a formative indicator for the ER domain.
3. **Work safety (SA):** Evaluated through indicators such as worker involvement and leadership engagement, with satisfaction with safety being a formative indicator for the SA domain.
4. **Work hygiene (HY):** Measured through indicators such as exposure to tool vibrations, low temperatures, high temperatures, noise, and hazardous substances, with satisfaction with hygiene being a formative indicator for the HY domain.

In the extended version of the OHS Barometer, a fifth domain—**Environment (EN)**—was included to assess environmental aspects of workplace well-being, allowing for a comprehensive measurement across the physical, social, and environmental dimensions of occupational health.

[1] Formative indicators are measurement variables that collectively define and form a latent construct. Unlike reflective indicators (which are manifestations of an underlying construct), formative indicators are considered to cause or determine the construct they measure.

The OHS barometer aims to provide actionable data-driven recommendations for targeted interventions to enhance workplace well-being. Currently, benchmarking is based on data from pilot companies, although the development of meaningful industry-specific benchmarks against established standards remains an area for future development and refinement. Mensura's R&D department has developed a streamlined, user-friendly online assessment tool (hereafter referred to as the e-survey) designed for cross-sectoral application and longitudinal monitoring. This efficient instrument accommodates diverse work environments while facilitating consistent data collection over time. However, prior to full-scale implementation, comprehensive psychometric validation is essential to establish the reliability, validity, and measurement precision of e-survey across different organizational contexts.

1.4 Problem Statement

For the OHS Barometer to serve as an effective tool in workplace well-being assessment, rigorous validation is essential to ensure that the survey instrument reliably and accurately measures workability across diverse occupational contexts. This validation requires addressing two critical methodological aspects:

1. **Internal consistency and reliability:** Evaluating of whether the survey provides consistent measurements of the intended constructs.
2. **Measurement invariance:** Ensuring that the survey functions equivalently across different conditions and populations.

1.5 Study Objectives

More specifically, this study aims to:

1. Assess the reliability and internal consistency of the OHS Barometer e-survey using data collected from 699 employees across four companies in the retail and wholesale sectors.
2. Evaluate measurement invariance across different conditions:
   - Survey versions (with and without the EN)
   - Language groups (Dutch and French)
   - Data collection methods (mobile vs. desktop)
   - Gender groups
   - Age categories
3. Provide evidence-based recommendations for refining the survey instrument based on comprehensive validation findings, enhancing its utility as an occupational well-being assessment tool.
4. Establish the psychometric properties of the EN as a potential addition to the core OHS Barometer framework, determining whether its inclusion affects the measurement properties of existing domains.

1.6 Research Hypotheses and Assessment Criteria

This study employs Confirmatory Factor Analysis (CFA) as the primary analytical framework to validate the reliability and measurement invariance of the OHS Barometer e-survey. CFA allows for the specification and testing of hypothesized latent factor structures, enabling the assessment of how well observed survey items reflect their intended underlying constructs. Within this framework, both formal statistical hypothesis testing and exploratory assessment criteria are used to evaluate the psychometric properties of the instrument.

All statistical metrics and methods referenced in this section are formally introduced and detailed in the methods section (Chapter 2) of this master thesis.

1.6.1 Formal Statistical Hypotheses

Factor Loading Significance

For each item-factor relationship in the CFA:

- Null hypothesis (H_0): $\lambda_{ik} \leq 0.5$ (The standardized factor loading is inadequate for practical significance)
- Alternative hypothesis (H_1): $\lambda_{ik} > 0.5$ (The standardized factor loading demonstrates practical significance)
- **Test statistic:** $t = \dfrac{\lambda_{ik} - 0.5}{\mathrm{SE}(\lambda_{ik})}$
- **Decision rule:** Reject H_0 if $t > t_{\alpha,df}$ (one-tailed test, $\alpha = 0.05$)

Note: λ_{ik} refers to the factor loading of item i on factor k.

Measurement Invariance Testing

For each group comparison (Survey version, language groups, data collection methods, gender, age categories):

Configural vs. Metric Invariance

$$H_0 : \text{Factor loadings are equal across groups (metric invariance holds)}$$
$$H_1 : \text{Factor loadings differ significantly across groups}$$

Metric vs. Scalar Invariance

$$H_0 : \text{Item thresholds are equal across groups (scalar invariance holds)}$$
$$H_1 : \text{Item thresholds differ significantly across groups}$$

Test statistic: Chi-square difference test ($\Delta\chi^2$)
Decision rule: Reject H_0 if $p < 0.05$ for $\Delta\chi^2$ test

Differential Item Functioning (DIF) Analysis

For items showing potential bias across groups:

$$H_0 : \text{Item functions equivalently across groups (no DIF present)}$$
$$H_1 : \text{Item functions differently across groups (DIF present)}$$

Test statistics: Wald test, likelihood ratio test
Decision rule: Reject H_0 if $p < 0.05$

1.6.2 Exploratory Assessment Criteria

The following assessments use established thresholds from psychometric literature but do not constitute formal statistical tests:

Reliability Assessment

Within the CFA framework, latent factors are hypothesized to represent underlying constructs, and their internal consistency is evaluated using established thresholds:

- **Internal consistency:** Cronbach's Alpha ≥ 0.70 (acceptable), ≥ 0.80 (good)
- **Composite reliability (CR):** CR ≥ 0.70 (acceptable)
- **Construct reliability:** Average Variance Extracted (AVE) ≥ 0.50
- **Item reliability:** Factor loadings $\lambda \geq 0.50$ (acceptable), ≥ 0.70 (preferred)

1.6.3 Survey Version Comparison (with/without EN)

Due to substantial sample size imbalance ($n = 44$ vs. $n = 655$), traditional multi-group confirmatory factor analysis may not be feasible. Alternative approaches will be explored:

- Differential Item Functioning (DIF) analysis to assess item-level equivalence
- Descriptive comparisons of reliability metrics across versions
- Qualitative assessment of factor structure consistency

This approach acknowledges the practical limitations while maximizing the analytical value of available data.

1.7 Significance of the Study

This validation study addresses a critical gap in occupational health measurement by providing a psychometrically sound, efficient workplace well-being assessment tool. The research significance spans methodological innovation, practical application, and broader impact on measurement science and practice.

1.7.1 Methodological Significance

This study advances measurement science by introducing the first systematic integration of multiple imputation with bootstrap procedures for workplace assessment validation. The research provides empirical evidence for contextual influences on measurement invariance, particularly the role of organizational characteristics in shaping item response behavior. These findings contribute to the theoretical understanding of how workplace-specific factors affect the validity of employee well-being assessments across heterogeneous settings.

1.7.2 Practical Significance

The validated e-survey addresses key limitations of existing workplace assessment tools by demonstrating: full measurement invariance across survey version, device types, cross-linguistic functionality, gender, and age categories; and robust psychometric properties, enabling real-time automated scoring. These empirically validated features allow for seamless integration into routine HR and OHS processes across diverse workforce demographics, providing reliable data for evidence-based decision making.

1.7.3 Broader Significance

These contributions advance both measurement science and occupational health practice by bridging the gap between rigorous psychometric validation and practical workplace implementation needs. By establishing psychometric equivalence across Belgium's multilingual workforce, this research promotes equitable workplace health monitoring and supports alignment with the European Union's Strategic Framework on Health and Safety at Work. The validated instrument provides a foundation for targeted, evidence-based workplace well-being initiatives that, according to existing research, can reduce workplace-related healthcare costs by up to 30% while improving employee retention and productivity (Baicker et al., 2010). This creates a pathway for healthier, more sustainable work environments that benefit individual workers, organizational performance, and broader public health outcomes across diverse organizational contexts.

1.8 Scope of the Study

The validation process incorporates e-survey data from four companies in the retail and wholesale sectors (N = 699), as detailed in Chapter 3, providing a diverse sample for psychometric evaluation of the OHS Barometer.

1.8.1 Sample Characteristics

Companies 1–3 (N = 605): These companies represent diverse subsectors within retail and wholesale. The sample features varying technological engagement (19.0% mobile completion) and linguistic diversity (86.4% Dutch, 12.1% French, 1.5% English), allowing robust cross-group comparisons for the formal hypotheses outlined in Section 1.6.
Company 4 (N = 94): This supplementary sample extends the validation by incorporating an additional EN domain and collecting data on occupational classification (white/blue collar). To assess the impact of domain inclusion, two survey variants were administered:

- **Form A (N = 46):** Standard assessment excluding EN items
- **Form B (N = 48):** Enhanced assessment including EN items

1.8.2 Measurement Equivalence Assessment

A comprehensive multi-method approach to equivalence testing examines the instrument's stability across diverse conditions:

Traditional Measurement Invariance Testing: Hierarchical model comparisons (configural, metric, scalar) assess equivalence across:

- Language versions (Dutch vs. French)
- Data collection modalities (mobile vs. desktop)
- Gender groups
- Age categories

Differential Item Functioning (DIF) Analysis: Examines response patterns between survey versions (with/without EN items) through:

- Full-sample analysis to establish general patterns
- Company-specific analyses to control for questionnaire version consistency effects

1.8.3 Analytical Considerations

The validation employs sophisticated statistical methodologies to ensure robust findings:

- **Missing Data Management:** Multiple imputation techniques preserve sample representativeness and statistical power while accounting for uncertainty in the imputation process.
- **Robust Parameter Estimation:** Bootstrap procedures quantify uncertainty with respect to the fit indices used to assess the fit of the CFA model to the observed data, while appropriate transformations address potential non-normality in the distribution of the estimators.

1.9 Model Framework

This study is grounded in the Healthy Workplace Model developed by the WHO (World Health Organization, 2010), which offers a comprehensive framework for occupational health interventions. At its core, the model emphasizes collaboration between workers and management, built upon a foundation of ethics and organizational values. Surrounding this core are essential processes of leadership engagement and worker involvement, ensuring shared responsibility for workplace well-being.

These components align with the measurement domains of the OHS Barometer, as illustrated in Figure 1.1:

- **Physical work environment:** Captured by the ER, HY, and SA domains. It should be noted that leadership engagement and worker involvement, while currently measured within the SA domain as indicators of safety culture, represent broader organizational culture principles that hypothetically play key roles in well-being policy implementation across all domains.
- **Psychosocial work environment:** Assessed through the PS domain.

The Belgian well-being domains measured by the OHS Barometer can be considered nested within the WHO model's four avenues of influence, providing the content framework for targeted interventions. While leadership engagement and worker involvement are currently contextualized within work safety, future development may expand these as general process indicators reflecting broader organizational culture principles that support comprehensive well-being initiatives.

Chapter 2

Materials and Methods

2.1 Data Description

2.1.1 Overview

This study analyzes data from the OHS Barometer validation project collected between February and June 2024. Four companies in the retail and wholesale sectors were contacted and agreed to participate, with voluntary employee participation resulting in a dataset of 699 respondents. Table 2.1 provides an overview of the participating companies and their response rates. The e-survey was administered in three languages (Dutch, French, and English), and participants were able to complete the survey using either mobile devices or desktop computers. For Company 4, we used an experimental design in which participants were randomly assigned to a basic survey version or an extended version that included an additional Environment (EN) domain. The dataset includes 39 variables measuring four primary workplace well-being domains: Psychosocial (PS), Ergonomics (ER), Work Safety (SA), and Work Hygiene (HY) across all companies, with the Environment (EN) domain administered to a subset of Company 4 respondents. This structure enables measurement invariance testing across survey versions, languages,data collection methods, gender, and age categories.

Table 2.1: Survey distribution and response rates by company

Company	**Sector**	**Employees Invited**	**Respondents**	**Response Rate (%)**
Company 1	Construction & Metal	417	149	35.7
Company 2	Hardware Wholesale	673	298	44.3
Company 3	IT & Software	203	158	77.8
Company 4	Mixed Retail & Wholesale	217	94	43.3
Total	—	1510	699	46.3

Data collection targeted a minimum response rate of 40% across participating companies, based on established guidelines for organizational survey research (Rogelberg & Stanton, 2007). This threshold was selected to ensure adequate representation while accounting for typical response patterns in workplace electronic surveys, which generally achieve rates between 30–60%

(Baruch & Holtom, 2008).

2.1.2 Data Preprocessing and Missing Data Handling

Prior to the main analyses, several preprocessing and data handling steps were implemented to ensure data quality and address missing data:

- **Pre-processing by Data Provider:** Prior to analysis, Mensura performed data cleaning and manipulations to address structural missingness patterns, including cases where missing responses could be logically deduced from previous questions.
- **Missing Data Analysis:** Following initial data cleaning, examination revealed approximately 5–6% missingness across key model variables, with systematic missingness (93.7%) for EN domain variables by design. Little's MCAR (Missing Completely At Random) test showed no systematic relationship between missingness and observed variables for the core domains. Although MCAR findings would support unbiased complete case analysis, multiple imputation was implemented to gain statistical efficiency and retain the full sample for planned analyses, acknowledging that MCAR cannot rule out missing not at random (MNAR) mechanisms.
- **Data Structure Verification:** We confirmed that reverse-coded items were properly aligned in the provided dataset, with consistent scoring directionality throughout the instrument (higher item scores uniformly reflected better workplace conditions). This verification ensured valid measurement of the intended constructs.
- **Outlier Detection:** Multivariate outliers were identified using Mahalanobis distance calculated from factor scores representing the continuous latent dimensions (PS, ER, SA, HY, and EN where applicable) with a conservative threshold ($p < 0.001$). These cases were reviewed but retained, as they may reflect valid workplace response patterns.

Multiple Imputation Procedure

To handle missing data, Multiple Imputation with Chained Equations (MICE) was performed using the `mice` package in R (van Buuren & Groothuis-Oudshoorn, 2011), with the following specifications:

- **Number of datasets:** 10 imputed datasets were generated.
- **Iterations:** 20 iterations per imputation, with convergence assessed separately for each of the 10 imputed datasets through trace plots of means and standard deviations.
- **Imputation methods:**
  - Ordinal variables: Proportional odds logistic regression (polr)
  - Continuous variables: Predictive mean matching (pmm)
  - Binary variables: Logistic regression (logreg)
- **Special case handling:** For cases without EN domain data (based on the `has_environment_domain` variable), environmental variables were set to a placeholder value (-999) prior to imputation to prevent these structurally. Age was measured as an ordered categorical variable ("<25", "25-34", "35-44", "45-54", ">=55") and was converted to an ordered factor with explicit sequencing before imputation.

Imputation Quality Assessment

To evaluate imputation quality, the following diagnostics were performed:

- **Convergence diagnostics:** Trace plots of means and standard deviations across iterations were examined separately for each of the 10 imputed datasets to confirm algorithm convergence within each imputation.
- **Distributional checks:** Distributions of observed and imputed data were compared to ensure preservation of key characteristics.
- **Variance assessment:** Between- and within-imputation variances were examined to evaluate imputation stability.
- **Imputation Fit Index (IFI):** To evaluate imputation quality beyond standard diagnostics (Rubin, 1976), we employed the *Imputation Fit Index (IFI)*, which assesses the congruence between observed and imputed data distributions by comparing their standard errors. This metric addresses limitations of direct standard error comparisons and provides a standardized measure of imputation accuracy. The mathematical formulation of the IFI is detailed in Appendix A.1.

2.1.3 Sample Size Considerations

Sample size adequacy is crucial for obtaining stable parameter estimates in CFA and measurement invariance testing. For the main analyses of the combined dataset, the sample size substantially exceeds common recommendations of at least 200 participants (Kline, 2016) or a minimum ratio of 5–10 participants per parameter (Bentler & Chou, 1987).

For subgroup analyses in measurement invariance testing, sample size adequacy varies across comparison groups:

- **Language groups:** Substantial variation in group sizes necessitates focusing the primary language invariance analysis on the two largest groups, with the smallest language group excluded from standalone analysis due to insufficient sample size.
- **Device types:** Both desktop and mobile user groups exceed the minimum threshold of 100 observations recommended for group comparisons (Chen, 2007).
- **Survey versions:** The experimental comparison groups approach the minimum threshold of 50 observations often cited for factor analysis (Hair et al., 2019), though they remain somewhat limited for robust CFA. Cautious interpretation of results and robust estimation techniques will be employed to address these moderate sample sizes.
- **Company comparisons:** All organizational subsamples exceed the minimum recommendation of 50 observations for factor analysis, enabling meaningful between-company comparisons.
- **Gender groups:** Both groups substantially exceed minimum thresholds for robust group comparisons.
- **Age categories:** Sample sizes vary across the five age groups, with the youngest category falling below typical minimums and potentially requiring cautious interpretation or category consolidation for invariance testing.

Bootstrapping will be implemented primarily for interval estimation purposes, circumventing problems in deriving asymptotic variances for the quantities of interest. While bootstrap procedures can provide more stable variance estimates compared to asymptotic methods, the point estimates themselves are expected to remain close to those obtained from the original data. This approach is particularly valuable for smaller subgroups where asymptotic assumptions may be less reliable. *Complete sample size distributions for all planned measurement invariance comparisons are reported in the results (Section 3.1.1).*

2.2 Software and Computational Environment

All analyses were conducted in **R** (version 4.2.2) using specialized packages for confirmatory factor analysis (`lavaan`), measurement invariance testing (`semTools`), multiple imputation (`mice`), and related procedures. Code and outputs were managed using `R Markdown` to ensure full reproducibility. Complete software specifications and computational details are provided in Appendix A.2.

2.3 Confirmatory Factor Analysis (CFA)

Confirmatory Factor Analysis (CFA) is a statistical technique used to validate the factor structure of a measurement instrument by testing whether observed variables (survey items) reliably reflect their intended latent constructs (Kline, 2016). CFA is hypothesis-driven and confirms predefined theoretical models, making it an essential tool for psychometric research, particularly in validating survey instruments.

In this study, CFA is applied to assess the validity of the OHS Barometer e-survey. Our analysis examines whether the survey items appropriately measure their respective dimensions as proposed in our theoretical framework (see Figure 1.1).

The CFA confirms whether the items load onto their hypothesized factors with sufficient strength and whether the overall measurement model provides an acceptable fit to the data, thereby establishing the construct validity of the instrument for the assessment of workplace well-being.

2.3.1 Components of CFA in this Study

CFA relies on key components to define and assess the factor structure of the OHS Barometer e-survey:

- **Latent Variables**: These represent unobserved workplace well-being constructs inferred from observed responses. In this study, the latent factors correspond to the four primary domains (PS, ER, SA, HY) and the additional EN domain for the extended survey version.
- **Observed Variables (Reflective Indicators)**: These include the individual survey items that serve as reflective indicators of the latent variables. In reflective measurement models, the observed indicators are conceptualized as effects or manifestations of the underlying latent construct, meaning changes in the latent variable cause changes in the observed indicators. The strength of their association with the latent factor is represented by their factor loadings, with higher loadings indicating that the observed variable more strongly reflects the underlying construct.
- **Error Terms**: Each observed variable includes a measurement error component, repre-

senting the portion of variance not explained by the latent factor. This acknowledges that survey items are imperfect measures of underlying constructs.

- **Factor Correlations**: The model allows latent factors to correlate with each other, reflecting the interconnected nature of workplace well-being domains while maintaining their distinctiveness as separate constructs.
- **Path Diagram**: A structural representation of the CFA model, visually mapping the relationships between survey items and their corresponding latent factors (see Figure 2.1).
- **Model Fit Assessment**: Statistical measures evaluate how well the proposed measurement model fits the observed data, using multiple indices to assess different aspects of model adequacy. The specific fit indices and their interpretive criteria are detailed in Appendix A.4.

This comprehensive approach ensures rigorous evaluation of the measurement model's psychometric properties and its suitability for workplace well-being assessment across diverse organizational contexts.

2.3.2 Mathematical Model of CFA

CFA models the relationship between observed survey responses and latent workability constructs. While the equation can be expressed for individual responses, a matrix formulation provides a more comprehensive representation of the complete model:

$$\mathbf{Y} = \mu + \mathbf{\Lambda F} + \mathbf{e}$$

Where:

- Y is the $j \times i$ matrix of observed responses from j subjects on i items.
- μ is the $i \times 1$ vector of intercepts for each item.
- Λ is the $i \times m$ matrix of factor loadings, where m is the number of latent factors.
- F is the $j \times m$ matrix of latent factor scores.
- e is the $j \times i$ matrix of error terms.

This matrix formulation elegantly captures the multivariate nature of the CFA model, accommodating multiple latent factors and their relationships to observed variables. It also provides the basis for estimating the model's parameters through maximum likelihood or other estimation methods.

For an individual observation, the confirmatory factor analysis (CFA) model can be expressed as:

$$Y_{ji} = \mu_i + \sum_{k=1}^{m} \lambda_{ik} F_{jk} + e_{ji}$$

where:

- λ_{ik} represents the loading of item i on factor k,
- F_{jk} is the score of subject j on factor k,
- Y_{ji} is the response of subject j to item i, and
- e_{ji} is the error term for subject j on item i.

In this study, $m = 4$ for the standard survey version (covering the domains *PS, ER, SA, HY)*, or $m = 5$ for the extended version that includes the additional *EN* domain.

2.3.3 Model Diagnostics for CFA

Prior to conducting CFA, comprehensive model diagnostics were performed to ensure data suitability and validity of statistical assumptions. Key assessments included:

- **Multivariate Normality:** Evaluated using univariate skewness/kurtosis indices and Mardia's multivariate tests. Robust estimation methods (WLSMV) were employed to accommodate ordinal data and potential violations of normality assumptions.
- **Multicollinearity:** Assessed through Variance Inflation Factors (VIF), correlation matrices, and determinant analysis to detect problematic redundancy among observed variables.
- **Factorability:** Evaluated using Bartlett's Test of Sphericity and Kaiser-Meyer-Olkin (KMO) measures to confirm that inter-item correlations were sufficient for meaningful factor extraction.

Comprehensive diagnostic procedures, mathematical specifications, and an interpretation guide are provided in the Appendix A.3.

2.4 CFA for OHS Barometer e-Survey

2.4.1 Model Specification

Based on theoretical frameworks and prior research, we specified CFA models representing workplace well-being domains:

- **Four-factor model (Standard version)**: Psychosocial Work Environment (PS), Ergonomics (ER), Work Safety (SA), and Work Hygiene (HY)
- **Five-factor model (Extended version)**: The four-factor model plus Environment (EN) domain

The models assume:

- Each observed variable loads on only one latent factor (simple structure)
- Latent factors are allowed to correlate, reflecting the interconnected nature of workplace well-being domains
- Factor variances are fixed to 1 for identification purposes
- Error terms are uncorrelated unless modification indices suggest otherwise

2.5 Parameter Interpretation and Model Evaluation

2.5.1 Item-Level Properties

Factor Loadings (λ_{ik}): Standardized loadings indicate the strength of relationship between observed variables and latent factors.

- $\lambda > 0.50$: Acceptable discrimination
- $\lambda > 0.70$: Strong discrimination (Hair et al., 2010)

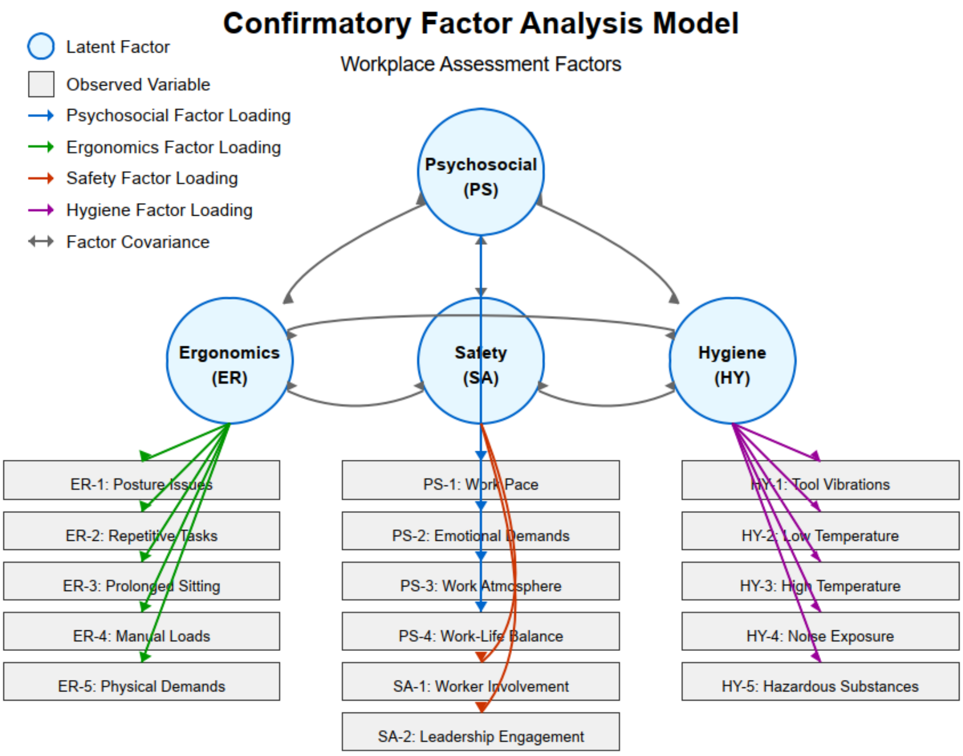


Figure 2.1: Illustrates the four-factor CFA model structure, showing factor loadings from each latent factor to its corresponding observed variables and covariances between latent factors.

2.5.2 Threshold Parameters for Ordinal Indicators

For ordinal survey responses analyzed with WLSMV estimation, threshold parameters (τ) represent the cut-points on an underlying continuous latent response variable that determine transitions between observed response categories. Each ordinal item reflects an underlying continuous latent response propensity, with thresholds defining the boundaries between adjacent response categories. Threshold estimates inform interpretations of item difficulty and reveal where items are most informative along the latent trait continuum (see Appendix A.5 for detailed mathematical specification and interpretation guidelines).

For ordinal survey responses analyzed with *WLSMV* estimation, **threshold parameters** (τ) represent the cut-points on an underlying continuous latent response variable that determine transitions between observed response categories.

2.5.3 Model Fit Assessment

Overall model adequacy was evaluated using multiple fit indices, each capturing different aspects of model adequacy and compensating for the limitations of individual measures. Complete definitions of all fit indices and their interpretive cut-offs are provided in Appendix A.4.

2.5.4 Robust Estimation Methods

Primary Estimation

Given the ordinal nature of Likert-scale survey data and potential violations of multivariate normality (assessed in Section 2.3.3), we employed robust estimation method.

WLSMV (Weighted Least Squares Mean and Variance Adjusted):

- Designed specifically for ordinal data using polychoric correlations
- Does not assume multivariate normality
- Provides robust standard errors and fit indices

Bootstrap Implementation

To enhance robustness and provide additional uncertainty quantification, bootstrap procedures with pooled results were implemented according to Rubin's rules in order to:

- Generate empirical standard errors for parameters where analytical standard errors may be unreliable
- Assess stability of fit indices across resampled datasets
- Provide robust uncertainty quantification independent of distributional assumptions

Technical Specifications:

- Number of bootstrap samples: 1,000 replications
- Estimation method: WLSMV
- Random seed set for reproducibility

2.5.5 Combining Multiple Imputation with Bootstrap

To address the uncertainty of missing data and sampling variability simultaneously, we implemented a comprehensive approach combining multiple imputation with bootstrap resampling (Schomaker & Heumann, 2018).

Procedure Steps

1. **Bootstrap Sampling**: For each of the $M = 10$ imputed datasets, $B = 1,000$ replacement bootstrap samples were drawn
2. **Model Fitting**: The CFA model was fitted to each bootstrap sample using WLSMV estimation
3. **Parameter Extraction**: Fit indices (CFI, TLI, RMSEA, SRMR, χ^2) and parameter estimates (factor loadings, factor correlations, item thresholds) were extracted from each fitted model
4. **Bootstrap Standard Errors Estimates**: For each parameter θ in imputed dataset j, bootstrap-based standard errors estimates were calculated as:
   where $\hat{\theta}j^{(b)}$ is the parameter estimate from bootstrap sample b in imputed dataset j, and $\bar{\theta}j = \frac{1}{B}\sum b = 1^B\hat{\theta}j^{(b)}$ is the bootstrap mean.
5. **Variance-Stabilizing Transformations**: Applied to bounded indices to improve distributional properties:
   - Fisher transformation: $z = 0.5 \times \ln\left(\frac{1+x}{1-x}\right)$ for CFI and TLI
   - Log transformation: $z = \ln(x)$ for RMSEA

   Pooling via Rubin's Rules:
6. **Pooling via Rubin's Rules**: For each parameter θ, results from the M imputed datasets are combined using Rubin's rules to obtain pooled point estimates, within-imputation variance, between-imputation variance, and total variance (see Appendix A.6 for detailed formulas).

7. **Degrees of Freedom Adjustment**: Calculated using Rubin's formula (Rubin, 1987) :

$$df = (M-1)\left[1+\frac{W}{(1+\frac{1}{M})B}\right]^2$$

8. **Back-Transformation**: Results transformed back to original scale for interpretation purposes.
9. **Missing Information Assessment**: The fraction of missing information (FMI) was computed as:

$$\text{FMI} = \frac{\left(1+\frac{1}{M}\right)B}{T}$$

This represents the proportion of the total variance attributable to missing data uncertainty.

The bootstrap approach, combined with multiple imputation, enhances parameter estimation and hypothesis testing by mitigating bias from missing data, addressing non-normality through robust estimation, quantifying uncertainty from both missing data and sampling variability, adjusting degrees of freedom for valid inference, and ensuring conservative testing when missing data impact is substantial, thus providing a reliable framework for workplace well-being measurement across diverse organizational contexts.

2.6 Assessing Internal Consistency and Reliability

Following confirmatory factor analysis model estimation, a comprehensive assessment of internal consistency and reliability is essential for establishing the psychometric quality of the OHS Barometer e-survey. This study employs multiple complementary reliability indices to address the limitations of any single coefficient and account for the diverse characteristics of our workplace well-being measurement domains (Brown, 2015).

Reliability Assessment Framework

Given the ordinal nature of our Likert-scale survey data and the potential violation of tau-equivalence assumptions across items, we implemented a multi-faceted reliability approach incorporating McDonald's Omega, Cronbach's Alpha, Polychoric Alpha, Composite Reliability, Average Variance Extracted, and Spearman-Brown Reliability. Additionally, discriminant validity is assessed using the Fornell-Larcker Criterion.

Interpretive guidelines and mathematical specifications for all reliability measures are provided in Appendix A.7.

2.6.1 Cronbach's Alpha (α)

Cronbach's Alpha represents the traditional approach to reliability assessment, providing a measure of internal consistency when the assumption of tau-equivalence holds. Tau-equivalence assumes that all items measuring a construct have equal factor loadings, meaning they contribute equally to the measurement of the latent variable (Edwards et al., 2021).

Formula:

$$\alpha = \frac{I}{I-1}\left(1-\frac{\sum_{i=1}^{I}\psi_i}{\sigma^2_{\text{total}}}\right)$$

Where I is the number of items, ψ_i represents the error variance of each item $i = 1, ..., I$, and σ^2_{total} represents the total variance of the composite scores.

Tau-equivalence assessment: This assumption can be evaluated by comparing constrained CFA models (where factor loadings are set equal) with unconstrained models (where factor loadings are freely estimated) through chi-square difference testing. Additionally, examining the magnitude of standardized factor loadings in the unconstrained model provides insight into whether loadings are approximately equal.

While widely used, Cronbach's Alpha may underestimate or overestimate reliability when tau-equivalence is violated, necessitating alternative approaches for congeneric measurement models.

2.6.2 McDonald's Omega (ω)

When tau-equivalence cannot be assumed, McDonald's Omega serves as a more robust alternative to Cronbach's Alpha for estimating internal consistency reliability. Unlike Alpha, which presumes equal factor loadings, Omega accommodates congeneric measurement by incorporating the actual factor loadings and error variances obtained from CFA. This allows ω to quantify the proportion of variance in the composite score attributable to the latent construct rather than to measurement error, thereby providing a more accurate and theoretically grounded reliability estimate (McDonald, 1999; Raykov, 2001; Zinbarg et al., 2005).

Formula for a single construct:

$$\omega = \frac{(\sum_{i=1}^{I} \lambda_i)^2}{(\sum_{i=1}^{I} \lambda_i)^2 + \sum_{i=1}^{I} \psi_i}$$

Where λ_i represents standardized factor loadings for items $i = 1, ..., I$, and ψ_i the unique variances for each item.

Interpretation: The numerator represents the total explained variance, while the denominator incorporates both explained variance and measurement error, providing the proportion of reliable variance in the observed composite.

2.6.3 Composite Reliability (CR)

Composite Reliability assesses the overall reliability of indicators measuring a latent construct within the CFA framework, incorporating differential factor loadings without assuming tau-equivalence.

Formula:

$$CR = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum (1 - \lambda_i^2)}$$

Where λ_i represents standardized factor loadings.

2.6.4 Polychoric Alpha

Given the ordinal nature of our Likert-scale responses, Polychoric Alpha provides reliability estimates that account for the categorical structure of the data by utilizing polychoric correlations.

Formula:

$$\alpha_p = \frac{I}{I-1} \left(1 - \frac{\sum_{i=1}^{I} \sigma_i^2}{\sigma^2_{\text{sum}}} \right)$$

Where σ_i^2 represents the variance of item i estimated from the polychoric covariance matrix, and σ_{sum}^2 represents the variance of the sum score computed from that matrix.

2.6.5 Supplementary Reliability Measures

Average Variance Extracted (AVE)

AVE measures the proportion of variance captured by a construct relative to measurement error, serving as an indicator of convergent validity (Fornell & Larcker, 1981).

Formula:

$$AVE = \frac{\sum_{i=1}^{I} \lambda_i^2}{I}$$

Interpretation: AVE values ≥ 0.50 indicate that the construct explains more variance in its indicators than is attributable to error. It is conceptually related to both Omega and Composite Reliability, though it emphasizes average explained variance per item.

Spearman-Brown Reliability

Used for domains with only two items.

Formula:

$$r_{SB} = \frac{2r_{hh}}{1 + r_{hh}}$$

where r_{hh} is the correlation between the two items of the scale.

2.6.6 Discriminant Validity Assessment

Fornell-Larcker Criterion

This criterion assesses whether the constructs are empirically distinct (Fornell & Larcker, 1981).

Criterion:

$$\sqrt{AVE_a} > r_{ab} \quad \text{and} \quad \sqrt{AVE_b} > r_{ab}$$

Where r_{ab} is the correlation between the constructs a and b.

2.6.7 Integration with CFA Results

All reliability and validity assessments are based on parameter estimates from the CFA models described in Section 2.3, ensuring coherence between measurement model specification and psychometric evaluation. This integrated approach provides a comprehensive assessment of the OHS Barometer's measurement quality.

2.7 Measurement Invariance and Equivalence Testing

This study employed complementary methodological approaches to assess whether the OHS Barometer functions equivalently across different conditions, supporting valid cross-group comparisons and ensuring the instrument's utility across diverse workplace contexts.

2.7.1 Multi-Group Confirmatory Factor Analysis (MGCFA)

Analytical Framework

MGCFA was conducted using two complementary approaches to overcome limitations of traditional invariance testing:

- **Traditional Null Hypothesis Testing (NHT)**: Compares nested models with increasing constraints to detect statistically significant fit deterioration, following the sequential testing paradigm established by (Meredith, 1993).
- **Equivalence Testing (ET)**: Evaluates whether parameter differences are practically negligible rather than statistically significant, addressing the limitation that traditional approaches may reject invariance due to trivial differences in large samples (Counsell et al., 2020). Equivalence thresholds were set at 0.2 for factor loadings and 0.3 for item thresholds, representing small to medium effect sizes that are unlikely to affect substantive interpretations.

Sequential Testing Hierarchy

Invariance was assessed sequentially using WLSMV estimation, consistent with the approach described in Section 2.3:

- **Configural Invariance**: Tests whether the same factor structure holds across groups (same pattern of loadings).
- **Metric Invariance**: Tests whether factor loadings are equivalent across groups (same measurement units).
- **Scalar Invariance**: Tests whether item intercepts are equivalent across groups (enables mean comparisons).

Each level serves as a prerequisite for the next, with failure at any stage precluding interpretation of subsequent levels.

Evaluation Criteria

Model Fit Assessment: Following the criteria detailed in Appendix A.4, acceptable fit was defined as CFI > 0.95, RMSEA < 0.06, and SRMR < 0.08.

Invariance Support: Following (Chen, 2007):

- Metric invariance: ΔCFI > -0.01, ΔRMSEA < 0.015, ΔSRMR < 0.03[1]
- Scalar invariance: ΔCFI > -0.01, ΔRMSEA < 0.015, ΔSRMR < 0.01

Equivalence Testing: Invariance was supported when 90% confidence intervals for parameter differences fell entirely within equivalence bounds, indicating practical equivalence rather than mere non-significance (Lakens, 2017).

2.7.2 Differential Item Functioning (DIF) Analysis

Rationale and Application

DIF analysis was employed as a complementary approach to MGCFA, particularly valuable when sample size constraints prevent reliable multi-group modeling. DIF evaluates whether individual items function differently across groups while controlling for the underlying trait level (Raykov et al., 2018).

Detection Methods

Three complementary DIF detection methods were implemented:

[1] Δ (Delta) indicates the change in fit indices between increasingly constrained models in measurement invariance testing.

- **Mantel-Haenszel Procedure**: Examines odds ratios for item responses across groups, stratified by total score levels. DIF magnitude was classified using Educational Testing Service criteria:
  - Negligible (A): $|\log(\alpha_{MH})| < 1.0$
  - Moderate (B): $1.0 \leq |\log(\alpha_{MH})| < 1.5$
  - Large (C): $|\log(\alpha_{MH})| \geq 1.5$
- **Logistic Regression Approach**:
  - Uniform DIF: Different item difficulty across groups (main effect)
  - Non-uniform DIF: Different item discrimination across groups (interaction effect)
- **IRT-based Detection**: Compares item parameters (difficulty and discrimination) across groups using likelihood ratio tests to identify statistically significant differences.

Methodological Integration

Both MGCFA and DIF analyses were integrated with the multiple imputation and bootstrap framework described in Section 2.4. For MGCFA, invariance tests were conducted across all imputed datasets, with the results pooled using appropriate combining rules for nested model comparisons. For DIF, detection procedures were applied to each imputed dataset separately, with the final classification of DIF based on consistency across imputations and pooled effect size estimates.

2.7.3 Group Comparisons and Sample Considerations

Primary Comparisons

The measurement invariance testing plan encompassed:

- **Language Groups**: Dutch versus French survey versions (linguistic equivalence)
- **Data Collection Methods**: Mobile versus desktop completion (method effects)
- **Survey Versions**: Standard versus extended versions with EN (structural equivalence)
- **Gender Groups**: Men versus women (demographic equivalence)
- **Age Categories**: Younger (<35), middle-aged (35–54), and older (≥55) workers (generational equivalence)

Analytical Strategy by Group Type

- **Adequate Sample Sizes**: Full MGCFA approach with both NHT and ET evaluation for groups with $n \geq 100$.
- **Imbalanced Groups**: Primary reliance on DIF analysis, supplemented by descriptive MGCFA where feasible—particularly for the EN comparison.
- **Company-Specific Analysis**: Focused analysis within Company 4 for the EN comparison to ensure the consistency of the questionnaire version and eliminate confounding of the survey design.

Chapter 3

Results

3.1 Data Description

The analysis dataset consisted of 699 observations across 39 variables, including 20 primary model variables, 4 EN variables, 10 auxiliary variables, and 5 grouping variables. See Table 3.1 for model variable details and Appendix Table F.6 for auxiliary variable specifications. Company participation details are provided in Table 2.1.

3.1.1 Sample Characteristics

Language Distribution: Dutch speakers constituted the majority at 88.3% ($N = 617$), with French speakers representing 10.4% ($N = 73$) and English speakers 1.3% ($N = 9$). Due to the small English subsample, measurement invariance testing focused on Dutch-French comparisons as planned in the methodology.

Data Collection Method: Approximately 25% of respondents ($N = 175$) completed the survey using mobile devices, with the remainder ($N = 524$) using desktop computers, enabling robust cross-platform invariance testing.

Gender and Age Distribution: The sample included 63.4% men ($N = 443$) and 35.2% women ($N = 246$), with 1.4% missing gender information. Age distribution was as follows: <25 years (4.0%, $N = 28$), 25–34 years (18.9%, $N = 132$), 35–44 years (27.5%, $N = 192$), 45–54 years (32.6%, $N = 228$), and ≥55 years (16.2%, $N = 113$), with 0.9% missing age data.

Environment Domain: The extended survey version including the EN domain was administered to Company 4 participants, with 48 receiving the extended version and 46 receiving the standard version, implementing the planned experimental design.

3.1.2 Missing Data Patterns

Prior to multiple imputation, missing data analysis revealed approximately 5–6% missingness across core model variables, with systematic missingness (93.7%) for EN variables by design. Little's MCAR test confirmed no systematic relationship between missingness and observed variables for the core domains ($\chi^2 = 1847.3$, df = 1839, $p = 0.564$), supporting the appropriateness of multiple imputation procedures.

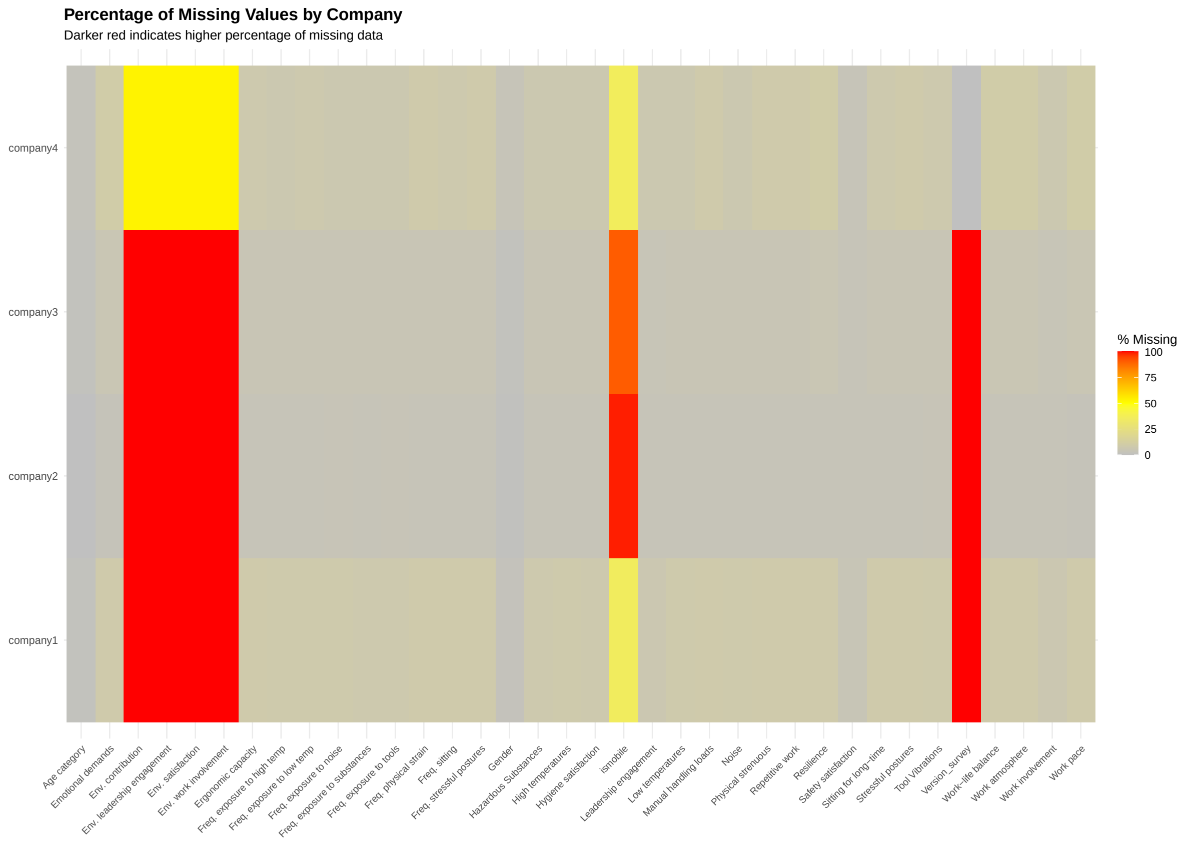


Figure 3.1: Heat map showing the percentage of missing values by company and variable

3.1.3 Descriptive Statistics and Distributional Properties

Table 3.1 presents comprehensive descriptive statistics for all model variables. Notable distributional characteristics include:

- **Psychosocial Domain:** Variables measured on 3-point scales (1 = not good, 2 = reasonable, 3 = good) showed means ranging from 2.42 to 2.58, indicating generally positive workplace perceptions. Moderate negative skewness values (-0.81 to -1.09) reflected the tendency toward higher response categories.
- **Ergonomics Domain:** Variables measured on 4- or 5-point scales exhibited means between 3.05 and 3.76, with substantial distributional asymmetry. *Physical Strenuous Work* showed pronounced ceiling effects (skewness = -2.83, kurtosis = 6.94), indicating most respondents reported low physical demands.
- **Safety Domain:** Variables demonstrated consistent means around 2.6 on 3-point scales, with moderate distributional characteristics supporting reliable measurement.
- **Hygiene Domain:** Variables displayed the most pronounced ceiling effects, with *Hazardous Substances Exposure* showing a mean of 3.83 on a 4-point scale and substantial distributional asymmetry (skewness = -3.41, kurtosis = 11.58), reflecting the low prevalence of such exposures in the sampled organizations.
- **Environment Domain:** Available only for the Company 4 subsample ($N = 48$), these variables showed more balanced distributions with means between 2.84 and 3.07 and mod-

erate skewness values (-0.34 to -0.67), suggesting less pronounced ceiling effects than other domains.

Table 3.1: Descriptive statistics for model variables

Domain	Variable	N	Mean	SD	Median	Min	Max	Skewness	Kurtosis
Psychosocial									
	Emotional demands	661	2.49	0.67	3	1	3	-0.94	-0.31
	Work atmosphere	659	2.46	0.68	3	1	3	-0.88	-0.43
	Work pace	661	2.42	0.71	3	1	3	-0.81	-0.62
	Work-life balance	659	2.58	0.60	3	1	3	-1.09	0.17
Ergonomics									
	Manual handling loads	661	3.55	0.87	4	1	4	-1.70	1.50
	Physically strenuous	661	3.76	0.68	4	1	4	-2.83	6.94
	Repetitive work	661	3.67	0.74	4	1	4	-2.07	2.98
	Sitting for long periods	662	3.05	0.96	3	1	4	-0.47	-1.05
	Stressful postures	661	3.42	0.94	4	1	4	-1.26	0.08
Safety									
	Leadership engagement	668	2.62	0.59	3	1	3	-1.27	0.58
	Worker involvement	668	2.62	0.59	3	1	3	-1.32	0.69
Hygiene									
	Hazardous substances	665	3.83	0.55	4	1	4	-3.41	11.58
	High temperatures	664	3.63	0.80	4	1	4	-1.99	2.63
	Low temperatures	664	3.51	0.91	4	1	4	-1.55	0.90
	Noise	664	3.40	0.94	4	1	4	-1.23	0.07
	Tool vibrations	662	3.71	0.68	4	1	4	-2.36	4.85
Environment									
	Environmental leadership	44	2.84	0.75	3	1	4	-0.41	-0.01
	Environmental satisfaction	44	3.07	0.70	3	1	4	-0.49	0.34
	Environmental contribution	44	2.91	0.47	3	1	4	-1.57	5.53
	Environmental involvement	44	3.00	0.75	3	1	4	-0.33	-0.37

Implications for Analysis

The data's distributional characteristics, including ceiling effects and asymmetry, justified the use of WLSMV estimation and polychoric correlations, suitable for ordinal and nonnormal data. Sample size distributions confirmed adequate power for measurement invariance testing, with the English subgroup excluded due to its small sample size (N $<$ 50), aligning with best practices. These features support the validity of the measurement model findings.

3.2 Multiple Imputation Results

3.2.1 Imputation Implementation

In line with the procedures described in Section 2.1.2, 10 imputed datasets were constructed to address missingness in 5–6% of the core variable observations. These imputed datasets were subsequently used for all CFA and measurement invariance analyses as detailed in Section 2.5.5.

3.2.2 Imputation Quality Assessment

Distribution Preservation

Distribution statistics comparing original and imputed data are provided in Appendix Table F.8. The imputation procedure generally preserved the distribution characteristics of the original data, with imputed means consistently slightly lower than original means and imputed standard deviations slightly higher, reflecting modest regression toward the mean typical in multiple imputation procedures.

Imputation Variance Components

Rubin's variance components analysis (detailed in Appendix Table F.7) demonstrated extremely low between-imputation variances relative to within-imputation variances, resulting in minimal Fraction of Missing Information (FMI) values across all variables. This pattern reflects the low percentage of missingness (5–6%) and indicates that the uncertainty introduced by missing data is negligible compared to the sampling variance. FMI values ranged from effectively zero to 2.93×10^{-4}, well below the 0.30 threshold indicating problematic missing data impact.

Imputation Diagnostics

Detailed imputation diagnostics, including Imputation Fit Index (IFI) analysis comparing standard errors between observed and imputed data, are provided in Appendix Table F.9. While some variables showed larger discrepancies than others (particularly those with ceiling effects), the low overall missingness rate (5–6%) suggests minimal practical impact on subsequent analyses.

3.2.3 Convergence and Implementation

Convergence diagnostics confirmed that imputation algorithms converged within the specified 20 iterations across all imputation runs, with no substantial differences in variable distributions observed after imputation. While imputed distributions may not exactly match observed data distributions (depending on missingness patterns and variable relationships in the imputation model), the overall distributional characteristics were well-preserved. Detailed convergence diagnostics and trace plots are provided in Appendix A.9.

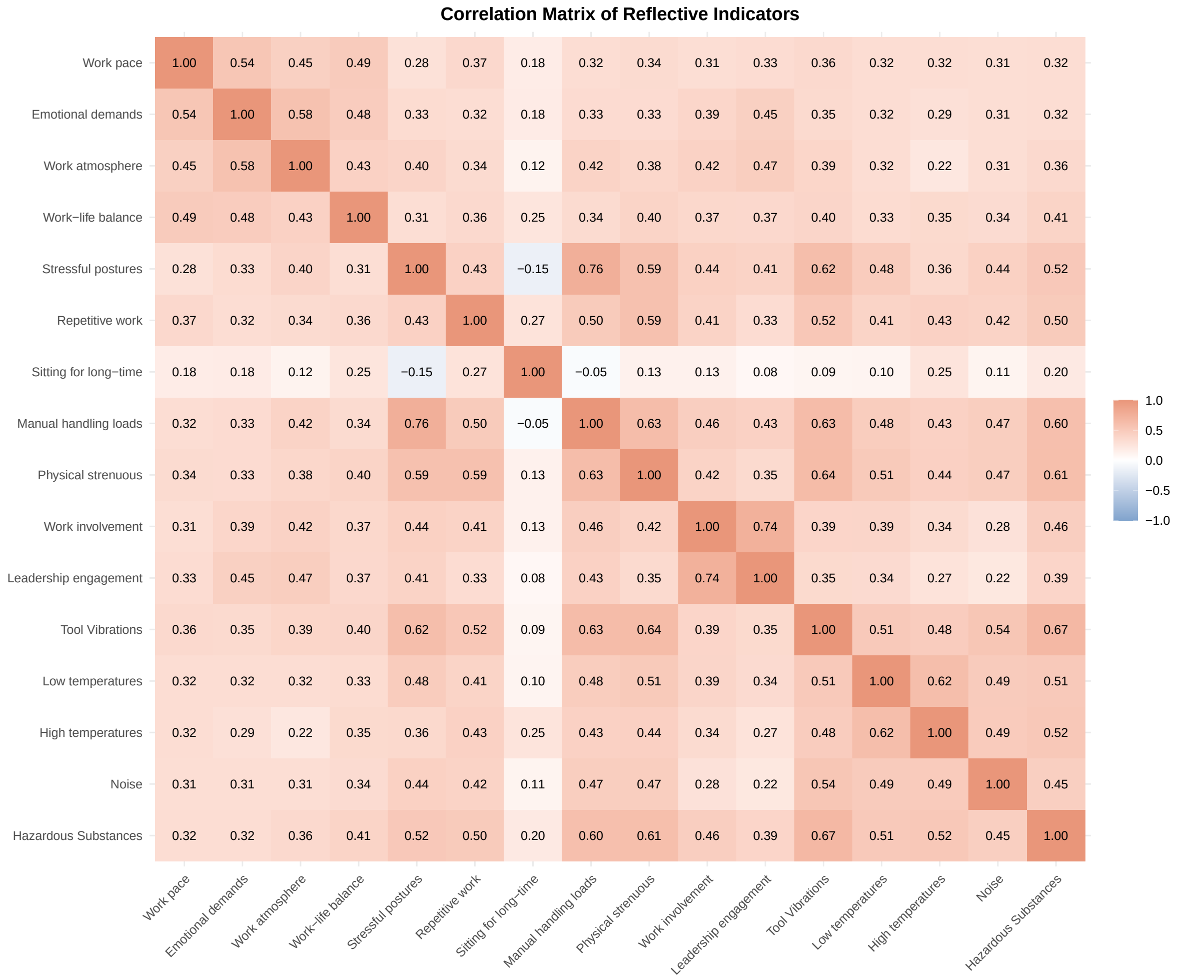
The 10 multiply-imputed datasets were subsequently used for all CFA and measurement invariance testing, with results pooled using Rubin's rules as detailed in Section 2.5.5.

3.3 Confirmatory Factor Analysis Results

3.3.1 CFA Assumptions and Diagnostics

Prior to model estimation, comprehensive diagnostics confirmed the appropriateness of the analytical approach. Assessment of distributional characteristics revealed substantial departures from multivariate normality assumptions, with Mardia's multivariate skewness (32909.28, $p < 0.001$) and kurtosis (194.05, $p < 0.001$) statistics indicating significant distributional asymmetry. All variables failed univariate normality tests (Anderson-Darling, $p < 0.001$). These findings confirmed the appropriateness of WLSMV estimation for the ordinal survey data, as specified in the methodology.

The multicollinearity assessment through the correlation matrix (Figure 3.2) revealed interpretable patterns supporting the theoretical factor structure. Within-domain correlations were generally stronger than cross-domain correlations, with the strongest relationships observed between *Manual handling loads* and *Stressful postures* ($r = 0.76$), *Leadership engagement* and *Work involvement* ($r = 0.74$), and *Tool vibrations* and *Hazardous substances* ($r = 0.67$). No correlations exceeded the 0.85 threshold indicating problematic multicollinearity. However, *Sitting for long-time* showed weak correlations overall, including a concerning negative correlation with *Stressful postures* ($r = -0.15$).



	Work pace	Emotional demands	Work atmosphere	Work−life balance	Stressful postures	Repetitive work	Sitting for long−time	Manual handling loads	Physical strenuous	Work involvement	Leadership engagement	Tool Vibrations	Low temperatures	High temperatures	Noise	Hazardous Substances
Work pace	1.00	0.54	0.45	0.49	0.28	0.37	0.18	0.32	0.34	0.31	0.33	0.36	0.32	0.32	0.31	0.32
Emotional demands	0.54	1.00	0.58	0.48	0.33	0.32	0.18	0.33	0.33	0.39	0.45	0.35	0.32	0.29	0.31	0.32
Work atmosphere	0.45	0.58	1.00	0.43	0.40	0.34	0.12	0.42	0.38	0.42	0.47	0.39	0.32	0.22	0.31	0.36
Work−life balance	0.49	0.48	0.43	1.00	0.31	0.36	0.25	0.34	0.40	0.37	0.37	0.40	0.33	0.35	0.34	0.41
Stressful postures	0.28	0.33	0.40	0.31	1.00	0.43	−0.15	0.76	0.59	0.44	0.41	0.62	0.48	0.36	0.44	0.52
Repetitive work	0.37	0.32	0.34	0.36	0.43	1.00	0.27	0.50	0.59	0.41	0.33	0.52	0.41	0.43	0.42	0.50
Sitting for long−time	0.18	0.18	0.12	0.25	−0.15	0.27	1.00	−0.05	0.13	0.13	0.08	0.09	0.10	0.25	0.11	0.20
Manual handling loads	0.32	0.33	0.42	0.34	0.76	0.50	−0.05	1.00	0.63	0.46	0.43	0.63	0.48	0.43	0.47	0.60
Physical strenuous	0.34	0.33	0.38	0.40	0.59	0.59	0.13	0.63	1.00	0.42	0.35	0.64	0.51	0.44	0.47	0.61
Work involvement	0.31	0.39	0.42	0.37	0.44	0.41	0.13	0.46	0.42	1.00	0.74	0.39	0.39	0.34	0.28	0.46
Leadership engagement	0.33	0.45	0.47	0.37	0.41	0.33	0.08	0.43	0.35	0.74	1.00	0.35	0.34	0.27	0.22	0.39
Tool Vibrations	0.36	0.35	0.39	0.40	0.62	0.52	0.09	0.63	0.64	0.39	0.35	1.00	0.51	0.48	0.54	0.67
Low temperatures	0.32	0.32	0.32	0.33	0.48	0.41	0.10	0.48	0.51	0.39	0.34	0.51	1.00	0.62	0.49	0.51
High temperatures	0.32	0.29	0.22	0.35	0.36	0.43	0.25	0.43	0.44	0.34	0.27	0.48	0.62	1.00	0.49	0.52
Noise	0.31	0.31	0.31	0.34	0.44	0.42	0.11	0.47	0.47	0.28	0.22	0.54	0.49	0.49	1.00	0.45
Hazardous Substances	0.32	0.32	0.36	0.41	0.52	0.50	0.20	0.60	0.61	0.46	0.39	0.67	0.51	0.52	0.45	1.00

Figure 3.2: Correlation matrix of reflective indicators

Note. Reflective indicators are observed variables that are manifestations of an underlying latent construct; the latent construct is assumed to cause variation in these indicators.

Factorability diagnostics strongly supported the appropriateness of factor analysis for most variables. The Kaiser-Meyer-Olkin (KMO) measure indicated meritorious sampling adequacy (KMO $= 0.89$), with individual KMO values > 0.80 for most variables. Bartlett's test of sphericity ($\chi^2 = 26145.97$, df $= 171$, $p < 0.001$) strongly rejected the identity matrix hypothesis. However, the *Sitting for long-time* variable failed to meet factorability requirements with KMO $=$ 0.68 (below the 0.70 threshold) and inadequate correlations with other indicators. Despite this, the variable was retained in the initial CFA model to assess its empirical performance, with the understanding that poor factor loading performance would support its removal in model refinement.

3.3.2 Model Fit and Parameter Estimates

The four-factor CFA model demonstrated good overall fit to the data. Pooled fit indices across multiply-imputed datasets showed excellent performance for CFI (0.987, 95% CI [0.978, 0.992]), TLI (0.984, 95% CI [0.972, 0.991]), and SRMR (0.075, 95% CI [0.064, 0.086]). RMSEA was slightly above the preferred threshold (0.082, 95% CI [0.070, 0.096]) but remained within acceptable limits. FMI values ranged from 0.298–0.323, indicating moderate impact of missing data on parameter estimation.

Factor loadings revealed generally strong relationships between indicators and their respective constructs (Table 3.2). Most standardized loadings exceeded 0.70, with particularly strong performance for *Manual handling loads* ($\lambda = 0.942$) and *Work involvement* ($\lambda = 0.944$). The critical exception was *Sitting for long-time* ($\lambda = 0.124$), which fell well below acceptable thresholds and confirmed concerns identified in assumption testing.

Table 3.2: Standardized and unstandardized factor loadings (discrimination) and intercepts (difficulty)

Factor	Description	Unstd Loading (SE)	Std Loading (SE)	Intercept Estimate (SE)
Ergonomics (ER)				
ER	Manual handling loads	1.059 (0.022)	0.942 (0.010)	1.187 (0.055)
ER	Physical strenuous	1.044 (0.024)	0.929 (0.016)	1.324 (0.059)
ER	Stressful postures	1.000* (0.000)	0.890 (0.014)	0.986 (0.051)
ER	Repetitive work	0.893 (0.033)	0.794 (0.026)	1.278 (0.057)
ER	Sitting for long-time	0.139 (0.034)	0.124 (0.030)	0.754 (0.048)
Hygiene (HY)				
HY	High temperatures	0.852 (0.026)	0.793 (0.025)	1.254 (0.057)
HY	Low temperatures	0.910 (0.021)	0.846 (0.020)	1.181 (0.055)
HY	Noise	0.740 (0.032)	0.689 (0.029)	0.974 (0.051)
HY	Hazardous substances	0.989 (0.024)	0.920 (0.020)	1.415 (0.062)
HY	Tool vibrations	1.000* (0.000)	0.930 (0.014)	1.342 (0.060)
Psychosocial Work Environment (PS)				
PS	Emotional demands	1.046 (0.052)	0.790 (0.028)	0.562 (0.044)
PS	Work pace	1.000* (0.000)	0.755 (0.027)	0.478 (0.042)
PS	Work atmosphere	1.086 (0.044)	0.820 (0.023)	0.531 (0.043)
PS	Work-life balance	1.034 (0.050)	0.781 (0.029)	0.744 (0.046)
Safety (SA)				
SA	Leadership engagement	0.987 (0.040)	0.931 (0.021)	0.807 (0.047)
SA	Work involvement	1.000* (0.000)	0.944 (0.021)	0.795 (0.047)

*Fixed for identification

Factor correlations (Appendix Table F.10) showed varying degrees of association between constructs. The strongest correlation emerged between Psychosocial and Safety factors ($r = 0.681$), while the weakest was between Psychosocial and Hygiene factors ($r = 0.295$). The moderate correlation between Ergonomics and Hygiene ($r = 0.586$) suggested potential discriminant validity concerns that required formal assessment using the Fornell-Larcker criterion.

3.3.3 Reliability and Validity Assessment

Reliability analysis demonstrated good to excellent internal consistency across all factors. Bootstrap confidence intervals across multiply-imputed datasets (10 imputations) showed (Table 3.3):

Table 3.3: Reliability coefficients for CFA by factor (N = 10 per factor)

Factor	N Valid	Cronbach's α	Polychoric α_{poly}	McDonald's ω	Spearman-Brown	Polychoric SB
PS	10	0.792 [0.790, 0.795]	0.860 [0.858, 0.862]	0.796 [0.794, 0.799]	–	–
ER	10	0.746 [0.742, 0.751]	0.817 [0.813, 0.821]	0.795 [0.788, 0.801]	–	–
SA	10	0.859 [0.854, 0.864]	0.935 [0.932, 0.939]	–	0.859 [0.854, 0.864]	0.935 [0.932, 0.939]
HY	10	0.828 [0.816, 0.839]	0.915 [0.908, 0.922]	0.856 [0.848, 0.864]	–	–

All reliability estimates substantially exceeded the 0.70 acceptability threshold, with polychoric measures showing consistently higher values when accounting for ordinal data structure. Confidence intervals were narrow, indicating stable reliability estimation. The Safety factor demonstrated the highest reliability across measures, while Ergonomics showed the lowest but still acceptable performance.

Convergent validity assessment (Table 3.4) showed all factors met established criteria, with Composite Reliability values ranging from 0.752–0.922 and Average Variance Extracted (AVE) values from 0.511–0.856. The Safety factor demonstrated the strongest convergent validity (AVE = 0.856), while Psychosocial showed the lowest but acceptable (AVE = 0.511).

Table 3.4: Convergent validity assessment

Factor	**Composite Reliability**	**AVE**
Psychosocial (PS)	0.804	0.511
Ergonomics (ER)	0.752	0.571
Safety (SA)	0.922	0.856
Hygiene (HY)	0.848	0.531

Note. CR values ≥ 0.70 indicate acceptable reliability. AVE values ≥ 0.50 indicate acceptable convergent validity.

Discriminant validity assessment using the Fornell-Larcker criterion revealed mixed results (Table 3.5). Good discriminant validity was found for most factor pairs, except Ergonomics and Hygiene. The correlation between these (0.777) exceeded the square root of AVE for Hygiene (0.729), indicating problematic overlap that challenges their conceptual distinctiveness.

Table 3.5: Fornell-Larcker matrix

Factor	**PS**	**ER**	**SA**	**HY**
PS	**0.715**	0.329	0.562	0.277
ER	0.329	**0.755**	0.466	0.777
SA	0.562	0.466	**0.925**	0.313
HY	0.277	0.777	0.313	**0.729**

Note: Diagonal elements (in bold) represent the square root of AVE. Off-diagonal elements represent correlations between factors.

3.3.4 Threshold and Intercept Parameters

Threshold estimates for ordinal indicators (Appendix Table F.11) showed consistent patterns, with values concentrated in the negative range of the latent continuum. This suggests the instrument is most sensitive to moderate-to-low workplace concern levels. Item intercepts varied across domains, with Hygiene items generally showing higher values, reflecting greater difficulty in endorsing negative conditions. The *Sitting for long-time* indicator again showed unique characteristics, with a positive t_3 threshold (0.235), further distinguishing it from other model indicators.

3.4 3.6 Measurement Invariance Results

Measurement invariance was tested across five key conditions using both traditional Null Hypothesis Testing (NHT) and modern Equivalence Testing (ET) approaches as detailed in Section 2.7.

3.4.1 Overview of Invariance Testing Results

Summary of Invariance Decisions

Table 3.6: Measurement invariance results by approach and group comparison

Comparison	**Sample Sizes**	**Metric Invariance**	**Scalar Invariance**	**NHT**	**ET**
Dutch vs French	$n_1 = 617, n_2 = 73$	Supported	Supported	Supported	Supported
Mobile vs Desktop	$n_1 = 175, n_2 = 524$	Supported	Supported	Supported	Supported
Men vs Women	$n_1 = 443, n_2 = 246$	Supported	Supported	Supported	Supported
Age Categories	5 groups (28-228)	**Not Supported**	Supported	Supported	Supported
Environment Domain	$n_1 = 44, n_2 = 655$	Not Testable	Not Testable	Not Testable	Not Testable

Note: NHT = Null Hypothesis Testing (ΔCFI > -0.01, ΔRMSEA < 0.015, ΔSRMR < 0.03 for metric; < 0.01 for scalar); ET = Equivalence Testing (0.2 threshold for loadings, 0.3 for thresholds). Bold indicates divergent results between approaches.

Model Fit Indices Across All Tests

Table 3.7: Fit indices for measurement invariance models

Comparison	Model	χ^2	df	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR
Dutch vs French	Configural	319.18	168	0.995	0.051	0.068	—	—	—
	Metric	346.56	179	0.994	0.052	0.071	-0.001	+0.001	+0.003
	Scalar	341.84	193	0.995	0.047	0.068	+0.001	-0.005	-0.002
Mobile vs Desktop	Configural	290.62	168	0.997	0.046	0.059	—	—	—
	Metric	318.23	179	0.996	0.047	0.061	-0.000	+0.001	+0.002
	Scalar	379.98	199	0.995	0.051	0.059	-0.001	+0.004	-0.002
Men vs Women	Configural	240.93	142	0.996	0.045	0.069	—	—	—
	Metric	279.34	152	0.995	0.049	0.075	-0.001	+0.004	+0.005
	Scalar	260.12	164	0.996	0.041	0.070	+0.001	-0.008	-0.004
Age Categories	Configural	299.00	284	0.999	0.016	0.074	—	—	—
	Metric	376.92	314	0.998	0.034	0.081	-0.002	+0.018*	+0.008
	Scalar	342.36	350	1.000	0.002	0.074	+0.002	-0.031	-0.007

Note: Asterisk indicates threshold violation (ΔRMSEA = 0.018 > 0.015 threshold).

Differential Item Functioning Results

Table 3.8: DIF analysis results for environment domain comparison

Analysis Type	**Sample**	**Total Items**	**Items with DIF**	**% with DIF**
Full Sample	All companies	16	**16**	**100%**
Company 4 Only	Single organization	16	**0**	**0%**

Note: DIF = Differential Item Functioning using Mantel-Haenszel procedure.

3.4.2 Language Invariance (Dutch vs. French)

Sample Characteristics: Dutch speakers ($N = 617$) vs. French speakers ($N = 73$). Three variables with sparse French responses were recoded to binary format to ensure adequate cell frequencies.

Results: Strong invariance support across all levels under both NHT and ET approaches, with all change indices well within established thresholds and maximum parameter differences of 0.01 (well below ET equivalence bounds). Model fit remained excellent across all invariance levels (see Tables 3.6 and 3.7).

3.4.3 Device Method Invariance (Mobile vs. Desktop)

Sample Characteristics: Mobile users ($N = 175$) vs. Desktop users ($N = 524$).

Results: Consistent invariance support under both approaches, with minimal fit deterioration and parameter differences well below equivalence thresholds. Visual evidence in Appendix Figure A.6 demonstrates stable CFI values (> 0.99) across all invariance levels.

3.4.4 Gender Invariance (Men vs. Women)

Sample Characteristics: Men ($N = 443$) vs. Women ($N = 246$). Modified model excluding problematic `Low temperatures` variable.

Results: Strong invariance support across all levels under both approaches, with excellent model fit maintained and minimal parameter differences observed.

3.4.5 Age Category Invariance

Sample Characteristics: Five age groups: <25 ($N = 28$), 25-34 ($N = 132$), 35-44 ($N = 192$), 45-54 ($N = 228$), ≥55 ($N = 113$).

Divergent Findings: While scalar invariance was supported by both approaches, metric invariance showed divergent results—NHT approach indicated non-support (ΔRMSEA $= 0.018 > 0.015$ threshold), while ET approach supported equivalence (max loading difference $= 0 < 0.2$ threshold).

Source of Divergence: Three items showed substantial loading differences across age groups, with older workers (≥ 55) consistently showing stronger item-factor relationships for physical workplace conditions:

- **High temperature:** 0.238 loading difference (0.859 vs. 0.620)
- **Noise:** 0.213 loading difference (0.737 vs. 0.524)
- **Repetitive work:** 0.165 loading difference (0.923 vs. 0.758)

This pattern suggests age-related differences in sensitivity to physical workplace factors, representing a theoretically meaningful finding rather than measurement error.

3.4.6 Environment Domain Invariance

Multi-Group CFA Challenges

Traditional MGCFA could not be implemented for the EN comparison ($N = 44$) vs.($N = 655$) due to insufficient sample size and sparse cell frequencies. Multiple methodological approaches were attempted, including binary recoding, simplified models, and company-specific analysis, but these resulted in model non-convergence/inadmissible solutions.

Differential Item Functioning Analysis

Full Sample Results: DIF analysis revealed universal differential functioning—all 16 items (100%) across all four domains showed significant DIF between groups with and without the EN (Table 3.8). All items received a "C" classification, indicating large effect sizes.

Company 4 Results: In stark contrast, when the analysis was restricted to Company 4, no items showed significant DIF (Table 3.8). This dramatic difference (100% → 0% DIF) highlights the critical role of questionnaire version consistency in measurement equivalence, as Company 4 was the only organization where both comparison groups completed the same survey version including the environmental extension.

Chapter 4

Discussion

4.1 Discussion and Interpretation of the Results

This comprehensive psychometric validation of the OHS Barometer e-survey provides essential evidence for its utility as a workplace well-being assessment tool while revealing important insights about measurement invariance, questionnaire version effects, and age-related differences in workplace assessment. The findings establish a solid foundation for evidence-based workplace well-being measurement while identifying key areas requiring attention for optimal instrument performance and application.

The confirmatory factor analysis revealed a generally well-functioning four-factor measurement model with excellent overall fit (CFI = 0.987, TLI = 0.984, SRMR = 0.075) and robust reliability across all domains. The reliability analysis demonstrated strong internal consistency, with particularly noteworthy findings regarding the consistent superiority of polychoric measures over traditional reliability coefficients when accounting for the ordinal nature of the data. Polychoric alpha values reached as high as 0.935 for the Safety factor, underscoring the importance of using measurement approaches specifically designed for ordinal data in workplace assessment contexts.

However, several critical findings emerged that require careful theoretical consideration. The discriminant validity analysis revealed problematic overlap between the Ergonomics and Hygiene factors ($r = 0.777 > \sqrt{\text{AVE}_{\text{HY}}} = 0.729$), challenging their conceptual distinctiveness and suggesting these domains may not be empirically separable in workplace assessments despite their theoretical differentiation. This finding aligns with contemporary perspectives emphasizing the interconnected nature of physical work environment factors and suggests that workplace well-being may be better conceptualized through integrative rather than compartmentalized models (Health and Safety Executive, 2019).

The strong correlation between Psychosocial and Safety factors ($r = 0.681$) further supports this integrative perspective, suggesting that psychological safety and broader safety culture are closely intertwined.

The identification of the *Sitting for long-time* variable as consistently problematic across multiple assessments—demonstrating inadequate factorability (KMO = 0.68), extremely weak factor loading ($\lambda = 0.124$), and poor correlations with other ergonomic indicators—suggests fundamental misalignment with the intended factor structure. This is further evidenced in the threshold estimates, where *Sitting for long-time* is the only variable across all domains to exhibit a positive threshold (t3 = 0.235), indicating that endorsement of the highest category requires exceeding the distribution mean, contrasting sharply with all other ergonomic variables that show consistently negative thresholds. This unique threshold pattern reinforces the variable's conceptual and statistical divergence from other ergonomic factors, supporting its problematic fit within the ergonomic domain. This finding indicates that not all theoretically relevant workplace factors necessarily translate into psychometrically sound measurement indicators, highlighting the critical importance of empirical validation in instrument development.

The measurement invariance testing yielded the most theoretically and methodologically significant findings. Strong measurement invariance was established across language groups (Dutch–French), data collection methods (mobile–desktop), and gender, supporting valid cross-group comparisons for these conditions. The consistency between traditional null hypothesis testing and modern equivalence testing approaches strengthened confidence in these conclusions and demonstrated the value of employing multiple analytical perspectives in invariance research.

The age-related analysis yielded particularly compelling insights that advance our understanding of developmental perspectives in workplace assessment. Both analytical approaches consistently supported scalar invariance, indicating equivalent item thresholds across age groups. However, metric invariance results diverged markedly: traditional null hypothesis testing suggested non-invariance due to ΔRMSEA exceeding the 0.015 threshold, while equivalence testing demonstrated practical equivalence, with maximum factor loading differences remaining well below the 0.2 criterion.

This methodological divergence reveals critical distinctions between statistical and practical significance in measurement invariance research. The findings suggest that age-related differences in factor loadings may reflect legitimate developmental variations in workplace priorities and experiences rather than measurement bias. Older workers' differential weighting of workplace factors could represent meaningful life-stage perspectives rather than psychometric limitations, thereby contributing valuable insights to developmental theories of occupational health assessment.

These results highlight fundamental tensions in current measurement invariance practice (Putnick & Bornstein, 2016) and contribute to evolving debates about appropriate equivalence standards in organizational research (Marsh et al., 2004; Vandenberg & Lance, 2000). The contrast between traditional threshold-based approaches (Chen, 2007) and emerging equivalence testing frameworks (Lakens, 2017) underscores the need for context-sensitive decision criteria rather than universal cutoff values (Nye & Drasgow, 2011).

The specific pattern of age-related loading differences proved particularly revealing, with older workers (≥ 55) consistently demonstrating stronger item-factor loadings for physical workplace conditions (high temperature, noise, repetitive work) compared to younger age groups. Rather than representing measurement error, these differences appear to capture genuine developmental processes whereby older workers become more sensitive to physical workplace conditions due to accumulated experience and changing physiological capacities. This pattern reflects theoretically meaningful developmental changes, aligning with research on age-related changes in physical tolerance and workplace priorities (Kenny et al., 2010).

Perhaps the most striking methodological insight emerged from the differential item functioning (DIF) analysis, which successfully addressed the convergence challenges encountered with traditional invariance testing approaches. In the full sample spanning all participating companies, universal differential functioning was observed across all four domains (100% of items exhibited significant DIF).

In stark contrast, when the analysis was limited to Company 4—the only organization that implemented the environmental questionnaire extension—no DIF was detected (0% of items), representing a complete reversal of the full sample pattern. This dramatic shift demonstrates that apparent measurement non-equivalence may stem from differences in questionnaire versions rather than from true measurement bias or organizational context effects.

This pattern reveals that survey extensions can create systematic artifacts in DIF analyses. What appears as measurement non-invariance may actually reflect methodological confounding introduced by comparing responses across companies with differing instrument versions. These findings demonstrate that traditional approaches to measurement validation may conflate true measurement bias with effects driven by survey version inconsistencies, potentially leading to erroneous conclusions about instrument reliability or fairness.

The results underscore a critical methodological requirement for organizational survey research: measurement equivalence testing must control for survey version consistency. The universal DIF observed in the full sample likely reflects methodological confounding due to Company 4's questionnaire extension rather than actual psychometric shortcomings. Consequently, cross-organizational DIF analyses must ensure identical instrument versions across groups to avoid artifactual findings that may misrepresent the measurement properties of workplace assessment tools.

From a practical implementation perspective, these findings provide clear guidance for instrument refinement and application. The immediate need to remove or substantially revise the *Sitting for long-time* indicator is evident from its consistently poor performance. The poor discriminant validity between Ergonomics and Hygiene domains suggests these should be considered for merger into a unified Physical Work Environment factor, which would better reflect the empirical relationships while maintaining theoretical coherence. The age-related findings necessitate implementing partial metric invariance approaches when conducting age-based comparisons, specifically freeing constraints on the three problematic items while maintaining scalar invariance for valid mean comparisons.

For organizations implementing the OHS Barometer, the established measurement invariance across language, device, and gender groups enables confident cross-group comparisons for these conditions. Organizations can validly compare workplace well-being scores between Dutch and French speakers, mobile and desktop users, and men and women without concern for measurement artifacts. Age-related differences in factor loadings may reflect meaningful developmental variations in workplace priorities rather than measurement limitations, suggesting that observed differences between age groups represent genuine life-stage perspectives on workplace factors. Additionally, organizations should ensure questionnaire version consistency when making cross-organizational comparisons, as survey extensions or modifications can create apparent measurement differences that reflect methodological artifacts rather than genuine workplace differences.

4.2 Study Limitations

Several methodological limitations must be acknowledged that may influence the interpretation and generalizability of these findings. The sample characteristics present the most significant constraint, with data collection restricted to retail and wholesale sectors within Belgium. This sectoral limitation potentially restricts generalizability to other industries where workplace well-being domain relationships may differ substantially. Manufacturing environments with heavy physical demands, healthcare settings with unique psychosocial stressors, or knowledge work contexts with predominantly cognitive demands may exhibit different factor structures and measurement properties than those observed in retail and wholesale settings.

The geographic restriction to Belgian organizations similarly limits international generalizability, as cultural differences in workplace expectations, regulatory frameworks, and assessment patterns may influence instrument performance in other national contexts. The sample size distribution across demographic groups presented analytical challenges, particularly for age-based analyses, with the smallest age group (<25 years, n = 28) falling substantially below optimal sample sizes for robust measurement invariance testing. Similarly, the English-speaking subsample (n = 9) was too small to include in cross-linguistic measurement invariance testing, limiting the analysis to Dutch-French comparisons and restricting conclusions about measurement equivalence for English-speaking workers.

The environment domain analysis suffered from severe methodological constraints due to the substantial sample size imbalance (n = 44 vs. n = 655), preventing traditional measurement invariance testing and limiting conclusions about the domain's measurement properties. The cross-sectional design represents another significant limitation, preventing conclusions about temporal stability, test-retest reliability, and responsiveness to workplace changes—critical applications for workplace assessment tools.

The reliance exclusively on self-report survey data without objective workplace assessments or external validation criteria represents a methodological constraint that limits understanding of how subjective assessments relate to objective workplace conditions. The convenience sampling approach may have introduced selection bias through the voluntary participation of organizations, potentially restricting the representativeness of findings to the broader population of Belgian workplaces.

While methodologically sophisticated, the multiple imputation approach relied on missing-at-random assumptions that, although statistically supported, cannot be definitively verified. Additionally, measurement invariance testing was limited to the specific grouping variables examined, potentially missing other important sources of measurement non-equivalence such as job level, tenure, or organizational size.

4.3 Ethical Thinking, Societal Relevance, and Stakeholder Awareness

The development and validation of workplace well-being assessment tools carries significant ethical responsibilities and societal implications that extend beyond technical psychometric considerations. This research contributes to fundamental ethical imperatives in occupational health by providing organizations with evidence-based tools to assess and improve conditions affecting worker health, safety, and well-being, ensuring that decisions affecting workers' lives are based on reliable, valid measurements rather than subjective impressions.

The cross-linguistic validation particularly addresses ethical concerns about equity and inclusion in workplace assessment. By establishing measurement equivalence across Dutch and French language groups, this research helps ensure that linguistic minorities are not disadvantaged by measurement bias in workplace evaluations, supporting broader social justice goals and reducing the potential for systematic bias against French-speaking workers in multilingual organizational contexts.

The identification of age-related measurement differences raises important ethical considerations about age discrimination and inclusive workplace practices. Rather than representing problematic bias, the finding that older workers show stronger sensitivity to physical workplace conditions provides valuable information for developing age-inclusive workplace policies that consider developmental differences in workplace design and assessment.

The questionnaire version effects revealed through the environment domain analysis highlight methodological responsibilities in workplace research and assessment. The finding that apparent measurement problems disappeared when survey version consistency was controlled demonstrates the importance of considering methodological factors rather than attributing assessment challenges to individual or instrumental failures. This supports more rigorous approaches to workplace evaluation that ensure questionnaire consistency across organizational comparisons and avoid misinterpreting methodological artifacts as genuine measurement problems or organizational differences.

For various stakeholders, this research provides tools that support ethical practice and decision-making. Occupational health professionals benefit from validated instruments that enable more accurate assessment of workplace conditions affecting worker health while providing efficient approaches that respect workers' time and organizational resources. Employers benefit from evidence-based assessment tools that support ethical decision-making about workplace conditions and enable proactive identification of workplace well-being concerns. Workers themselves benefit from validated assessment tools that accurately capture their workplace experiences and provide reliable foundations for improvement efforts.

The broader societal relevance extends to public health and economic implications of workplace well-being. Reliable assessment tools support the development of healthier workplaces that reduce healthcare costs, improve productivity, and enhance quality of life for workers and their families. The comprehensive assessment approach recognizes workers as whole persons with interconnected physical, psychological, and social needs rather than reducing them to isolated dimensions.

4.4 Future Research Directions

The findings and limitations of this validation study point toward several important directions for future research that could significantly advance workplace well-being assessment and measurement science. Longitudinal validation represents the most critical immediate need, as the cross-sectional design prevents conclusions about temporal stability, test-retest reliability, and responsiveness to workplace changes. Future research should examine the instrument's performance over various time intervals and assess whether the factor structure remains stable over time and whether measurement invariance holds across temporal contexts.

Expanding the validation scope to diverse organizational and cultural contexts represents another critical research priority. The current restriction to retail and wholesale sectors in Belgium limits understanding of how workplace well-being factor structures and measurement properties may vary across industries with different characteristics. Manufacturing environments, healthcare settings, educational institutions, and knowledge work contexts may exhibit different measurement properties that would inform both theoretical understanding and practical application.

The environment domain requires focused development attention given the methodological constraints that prevented adequate evaluation in this study. Future research should prioritize collection of larger, more balanced samples specifically designed to enable robust psychometric evaluation of environmental workplace factors and clarify the conceptual boundaries of environmental workplace well-being.

Criterion validation research represents a significant gap that limits conclusions about practical utility and predictive validity. Future studies should incorporate objective workplace measurements, health outcomes, performance indicators, and organizational metrics to establish how subjective well-being assessments relate to concrete workplace conditions and outcomes, examining relationships with absenteeism, turnover, productivity measures, safety incidents, and healthcare utilization.

The age-related measurement differences warrant deeper investigation through mixed-methods approaches that combine quantitative analysis with qualitative exploration of how different age groups understand and interpret workplace assessment items. Advanced analytical approaches could provide new insights into workplace well-being structure and relationships, including network analysis to explore complex patterns of relationships and multilevel modeling to examine organizational-level variance in measurement properties.

The questionnaire version effects revealed in this study suggest important new research directions examining how survey design consistency and methodological controls influence measurement properties in organizational assessment processes. Research could investigate how different questionnaire versions or extensions affect measurement equivalence across organizations and develop standardized protocols to ensure valid cross-organizational comparisons while avoiding methodological artifacts that can be misinterpreted as genuine organizational or measurement differences.

Chapter 5

Conclusion

This comprehensive psychometric validation establishes the OHS Barometer e-survey as a valuable tool for workplace well-being assessment while providing important insights for both measurement science and practical application. The research demonstrates that rigorous psychometric validation can be successfully conducted in organizational contexts using sophisticated analytical approaches that address missing data, measurement invariance, and multiple sources of uncertainty.

The generally strong psychometric properties, including robust reliability across domains and good overall model fit, support the instrument's utility for workplace assessment applications. The established measurement invariance across language groups, data collection methods, and gender provides confidence that the instrument functions equivalently across these critical conditions, enabling valid cross-group comparisons and supporting inclusive assessment practices.

The methodological contributions extend beyond instrument validation to advance measurement science through innovative approaches to combining multiple imputation with bootstrap procedures and systematic comparison of traditional and modern measurement invariance testing approaches. The identification of questionnaire version effects in measurement equivalence represents a significant contribution to measurement methodology, demonstrating that apparent measurement problems may reflect survey design inconsistencies rather than instrumental factors.

However, several important limitations require acknowledgment and attention. The problematic *Sitting for long-time* indicator requires removal or substantial revision, while the poor discriminant validity between Ergonomics and Hygiene domains suggests the need for structural refinement. The age-related measurement differences, while theoretically meaningful, necessitate careful interpretation and may require age-specific implementation guidelines.

Despite these limitations, the research provides a solid foundation for evidence-based workplace well-being assessment while identifying clear pathways for continued development and refinement. The comprehensive approach to validation, honest acknowledgment of limitations, and detailed implementation guidance provide a model for rigorous psychometric research in workplace contexts. The instrument represents an important step forward in efficient, multidimensional workplace assessment that respects diverse linguistic and technological preferences while providing reliable assessment of workplace well-being factors.

References

Baicker, K., Cutler, D., & Song, Z. (2010). Workplace wellness programs can generate savings. *Health Affairs*, *29*(2), 304–311. https://doi.org/10.1377/hlthaff.2009.0626

Baruch, Y., & Holtom, B. C. (2008). Survey response rate levels and trends in organizational research. *Human Relations*, *61*(8), 1139–1160.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*(1), 78–117. https://doi.org/10.1177/0049124187016001004

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Chen, F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Counsell, A., Cribbie, R. A., & Flora, D. B. (2020). Equivalence testing for measurement invariance: A user-friendly primer. *The Quantitative Methods for Psychology*, *16*(4), 348–361.

Dodge, R., Daly, A. P., Huyton, J., & Sanders, L. D. (2012). The challenge of defining wellbeing. *International Journal of Wellbeing*, *2*(3), 222–235. https://doi.org/10.5502/ijw.v2i3.4

Edwards, A. A., Joyner, K. J., & Schatschneider, C. (2021). A simulation study on the performance of different reliability estimation methods. *Educational and Psychological Measurement*, *81*(6), 1089–1117. https://doi.org/10.1177/0013164421994184

EU-OSHA. (2025, January). *OSH-barometer* [Accessed: 2025-02-27]. https://visualisation.osha.europa.eu/osh-barometer/

Eurogip. (2024, April). *Annual report: Key features of our activity in 2023*. Eurogip-193/E.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50.

Gould, R., Ilmarinen, J., Järvisalo, J., & Koskinen, S. (2008). *Dimensions of work ability*. Finnish Institute of Occupational Health (FIOH).

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th). Pearson Education.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th). Cengage Learning.

Health and Safety Executive. (2019). Ergonomic factors in the workplace [Accessed: 2025-06-13]. https://www.hse.gov.uk/pubns/indg90.htm

Ilmarinen, J., & Tuomi, K. (1993). Work ability index for aging workers. *Aging and Work*, 142–151.

Ilmarinen, J. (2019). From work ability research to implementation. *International Journal of Environmental Research and Public Health*, *16*(16), 2882. https://doi.org/10.3390/ijerph16162882

Kenny, G. P., Yardley, J. E., Brown, C., Sigal, R. J., & Jay, O. (2010). Heat stress in older individuals and patients with common chronic diseases. *CMAJ*, *182*(10), 1053–1060. https://doi.org/10.1503/cmaj.091062

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th). Guilford Press.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

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*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Martus, P., Jakob, O., Rose, U., Seibt, R., & Freude, G. (2010). A comparative analysis of the work ability index. *Occupational Medicine*, *60*(7), 517–524. https://doi.org/10.1093/occmed/kqq093

McDonald, R. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum Associates.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Nabe-Nielsen, K., Thielen, K., Nygaard, E., Thorsen, S. V., & Diderichsen, F. (2014). Demand-specific work ability, poor health and working conditions in middle-aged full-time employees. *Applied Ergonomics*, *45*(4), 1174–1180. https://doi.org/10.1016/j.apergo.2014.02.007

Nye, C. D., & Drasgow, F. (2011). Assessing goodness of fit: Simple rules of thumb simply do not work. *Organizational Research Methods*, *14*(3), 548–570. https://doi.org/10.1177/1094428110368562

Pak, K., Kooij, D. T. A. M., De Lange, A. H., van den Heuvel, S., & Van Veldhoven, M. J. P. M. (2021). The influence of human resource practices on perceived work ability and the preferred retirement age: A latent growth modelling approach. *Human Resource Management Journal*, *31*(2), 311–325. https://doi.org/10.1111/1748-8583.12304

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90.

Radkiewicz, P., & Widerszal-Bazyl, M. (2005). Psychometric properties of work ability index in the light of comparative survey study. *International Congress Series*, *1280*, 304–309. https://doi.org/10.1016/j.ics.2005.02.089

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, *54*(2), 315–323.

Raykov, T., Dimitrov, D. M., Marcoulides, G. A., Li, T., & Menold, N. (2018). Examining measurement invariance and differential item functioning with discrete latent construct indicators: A note on a multiple testing procedure [Epub 2016 Oct 25. PMID: 29795959; PMCID: PMC5965654]. *Educational and Psychological Measurement*, *78*(2), 343–352. https://doi.org/10.1177/0013164416670984

Rogelberg, S. G., & Stanton, J. M. (2007). Understanding response rates in organizational survey research: A meta-analytic and meta-narrative review. *Journal of Applied Psychology*, *92*(1), 121–133.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Schomaker, M., & Heumann, C. (2018). Bootstrap inference when using multiple imputation. *Statistics in Medicine*, *37*, 2252–2266. https://doi.org/10.1002/sim.7654

Schulte, P., & Vainio, H. (2010). Well-being at work – overview and perspective. *Scandinavian Journal of Work, Environment & Health*, *36*(5), 422–429. https://doi.org/10.5271/sjweh.3076

SPF Emploi, Travail et Concertation sociale. (2025, January). National focal points belgium [[Accessed: 2025-02-27]]. https://osha.europa.eu/en/about-eu-osha/national-focal-points/belgium

Tuomi, K., Eskelinen, L., Toikkanen, J., Järvinen, E., Ilmarinen, J., & Klockars, M. (1991). Work load and individual factors affecting work ability among aging municipal employees. *Scandinavian Journal of Work, Environment & Health*, *17*(Suppl 1), 128–134.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

World Health Organization. (2010). *WHO healthy workplace framework and model: Background and supporting literature and practices* [Accessed: 2025-05-29]. WHO Press. https://www.who.int/publications/i/item/who-healthy-workplace-framework-and-model

World Health Organization. (2018). Mental health: Strengthening our response [fact sheet].

Zinbarg, R., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, revelle's β, and mcdonald's ω: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.

Appendix A

Supplementary Material

This file contains additional figures, tables, and equations referenced in the main text.

A.1 Imputation Fit Index (IFI)

For each variable m and imputation k:

$$\mathrm{IFI}_{mk} = \left| SE(x_m^{(\mathrm{obs})}) - SE(x_m^{(k)}) \right|$$

Average IFI across all imputations:

$$\mathrm{IFI}_m = \frac{1}{K} \sum_{k=1}^{K} \mathrm{IFI}_{mk}$$

Standard deviation of IFI:

$$S_{\mathrm{IFI}_m}^2 = \frac{1}{K-1} \sum_{k=1}^{K} (\mathrm{IFI}_{mk} - \mathrm{IFI}_m)^2$$

Standardized IFI score:

$$Z_{\mathrm{IFI}_{mk}} = \frac{\mathrm{IFI}_{mk} - \mathrm{IFI}_m}{S_{\mathrm{IFI}_m}}$$

Where:

- $SE(x_m^{(\mathrm{obs})})$ = standard error of observed data for variable m
- $SE(x_m^{(k)})$ = standard error of imputed data for variable m in imputation k
- K = number of imputations (10 in this study)

A.2 Software and Computational Environment

A.2.1 Software Specifications

Analyses were conducted in **R (4.2.2)** using both core and specialized packages:

- **Core Packages**: `lavaan`, `semTools`, `mice`, `mitml`, `psych` — for factor analysis, invariance testing, and reliability estimation.
- **Specialized Packages**: `difR`, `MVN`, `boot` — for differential item functioning analysis, normality testing, and bootstrap procedures.

- **Data and Visualization**: `dplyr`, `ggplot2`, `knitr`, `rmarkdown` — for data manipulation, visualization, and dynamic reporting.

A.2.2 Computational Environment

- **Hardware**: 3.6 GHz processor, 32 GB RAM, running Windows 11 on SSD storage.
- **Optimization**: Utilized parallel processing, efficient memory management, and controlled random seeds (`set.seed(123)`).

A.2.3 Reproducibility Measures

- **Code Management**: Modular `.Rmd` scripts, version control systems (e.g., Git), and clearly separated code blocks for analyses.
- **Documentation**: Clear variable naming conventions, comprehensive commenting, and session information logging.
- **Output Handling**: Automated generation of tables and figures, consistent formatting styles, and saved intermediate outputs.

A.2.4 Code Availability

Analysis scripts are available upon request and are structured to support adaptation and future research applications.

A.3 Diagnostic Procedures and Statistical Assumptions

This appendix details the diagnostic procedures, mathematical formulations, and interpretation criteria referenced in Section 2.4.3.

A.3.1 Multivariate Normality

Multivariate normality was assessed using:

- **Univariate Skewness (γ_1):** Significant asymmetry indicated by $|\gamma_1| > 2$.
- **Univariate Kurtosis (γ_2):** Excess peakedness or flatness indicated by $|\gamma_2| > 7$.
- **Mardia's Tests:**
  - Multivariate skewness and kurtosis statistics.
  - $p < .05$ indicates deviation from multivariate normality.

A.3.2 Multicollinearity

Multicollinearity among observed variables was examined via:

- **Variance Inflation Factor (VIF):** Values below 10 indicate acceptable levels.
- **Inter-item correlations:** Correlation coefficients $|r| > .85$ may suggest redundancy.
- **Determinant of Correlation Matrix:** Values < 0.00001 indicate potential multicollinearity problems.

A.3.3 Factorability

To assess factorability:

- **Bartlett's Test of Sphericity:** $p < .05$ supports factorability.
- **Kaiser-Meyer-Olkin (KMO) Measure:**
  - KMO $> .60$ considered acceptable.

  - KMO $>$ .80 considered very good.
- **Correlation Matrix Inspection:** Majority of coefficients $|r| > .30$ indicates adequate inter-item correlation.

All diagnostics were reviewed before confirmatory analysis, with adjustments made (e.g., use of WLSMV estimator) to accommodate detected violations.

A.4 Model Fit Indices

A.4.1 Mathematical definitions of CFA model fit indices

Table A.1: Mathematical definitions of CFA model fit indices

Fit Index	**Formula**	**Interpretation**
Chi-Square (χ^2)	$(N-1) \times F_{\mathrm{ML}}$	Tests if model-implied and sample covariance matrices match.
CFI	$1 - \frac{\max[(\chi^2_{\mathrm{target}} - df_{\mathrm{target}}),0]}{\max[(\chi^2_{\mathrm{baseline}} - df_{\mathrm{baseline}}),(\chi^2_{\mathrm{target}} - df_{\mathrm{target}}),0]}$	Compares target model to baseline independence model.
TLI	$\frac{(\chi^2_{\mathrm{baseline}}/df_{\mathrm{baseline}})-(\chi^2_{\mathrm{target}}/df_{\mathrm{target}})}{(\chi^2_{\mathrm{baseline}}/df_{\mathrm{baseline}})-1}$	Adjusts for model complexity, less sensitive to sample size.
RMSEA	$\sqrt{\max\left(\frac{\chi^2 - df}{df(N-1)}, 0\right)}$	Evaluates discrepancy per degree of freedom, adjusted for sample size.
SRMR	$\sqrt{\frac{\sum_{ij}(s_{ij} - \hat{\sigma}_{ij})^2}{\sum_{ij} s_{ij}^2}}$	Measures average standardized residuals between observed and model-implied covariances.

A.4.2 Interpretation of Model Fit Indices

Table A.2: Interpretive guidelines for CFA model fit indices

Fit Index	**Excellent Fit**	**Good Fit**	**Acceptable Fit**	**Poor Fit**
Chi-Square (χ^2)	$p > 0.05$	$p > 0.01$	$p > 0.001$	$p \leq 0.001$
CFI	≥ 0.97	≥ 0.95	≥ 0.90	< 0.90
TLI	≥ 0.97	≥ 0.95	≥ 0.90	< 0.90
RMSEA	≤ 0.05	≤ 0.06	≤ 0.08	> 0.08
SRMR	≤ 0.05	≤ 0.06	≤ 0.08	> 0.08

A.4.3 Practical Considerations

Table A.3: Practical guidelines and limitations for model fit interpretation

Aspect	**Guidelines**
Sample Size	• Small samples (N $<$ 200): focus on descriptive fit indices. • Large samples (N $>$ 500): chi-square likely significant, emphasize practical fit indices. • Very large samples (N $>$ 1000): chi-square almost always significant, rely on CFI, TLI, RMSEA, and SRMR.
Model Complexity	• Simple models may show good fit but lack theoretical richness. • Complex models may exhibit lower fit indices but capture nuanced relationships. • Prefer simpler models with adequate fit over complex models with marginal improvement.
Combined Interpretation	• Good fit: CFI and TLI $\geq$ 0.95, RMSEA $\leq$ 0.06 with narrow CI, SRMR $\leq$ 0.06, non-significant χ^2 (if sample size allows). • Marginal fit: CFI and TLI $\geq$ 0.90, RMSEA $\leq$ 0.08, SRMR $\leq$ 0.08, strong theoretical justification.
Limitations	• Cut-off values are guidelines, not absolute thresholds. • Model type sensitivity affects interpretations. • Estimation method impacts fit indices. • Good fit doesn't guarantee theoretical or practical utility.

A.5 Threshold Parameters - Mathematical Specification

Conceptual Framework

Each ordinal item is assumed to reflect an underlying continuous latent response propensity. Respondents perceive this continuous trait level but report it using the limited set of discrete response categories (e.g., 1 = not good," 2 = reasonable," 3 = "good").

Mathematical Specification

For an item with C response categories, there are $C-1$ thresholds $(\tau_1, \tau_2, \ldots, \tau_{C-1})$ that define the boundaries between adjacent response categories:

- Response category 1 if latent response $\leq \tau_1$
- Response category 2 if $\tau_1 <$ latent response $\leq \tau_2$
- Response category 3 if $\tau_2 <$ latent response $\leq \tau_3$
- ...

Interpretation Guidelines

- Threshold location indicates the latent trait level required to transition between response categories
- Negative thresholds suggest transitions occur at below-average trait levels
- Positive thresholds indicate transitions require above-average trait levels
- Threshold spacing reflects item discrimination across different trait levels

Practical Implications

- Items with widely spaced thresholds discriminate well across a broad range of trait levels
- Items with closely spaced thresholds provide fine-grained measurement precision in a narrower trait region

This modeling approach is fundamental to ordinal confirmatory factor analysis (CFA) and provides essential insight into item functioning beyond factor loadings alone.

A.6 Rubin's Rules Formulas

For each parameter θ, the following formulas are used to combine results across M imputed datasets:

- **Pooled point estimate:**
$$\hat{\theta}_{\text{pool}} = \frac{1}{M}\sum_{j=1}^{M}\hat{\theta}_j$$
- **Within-imputation variance:**
$$W = \frac{1}{M}\sum_{j=1}^{M}\text{SE}_j^2(\hat{\theta}_j)$$
- **Between-imputation variance:**
$$B = \frac{1}{M-1}\sum_{j=1}^{M}(\hat{\theta}_j - \hat{\theta}_{\text{pool}})^2$$
- **Total variance:**
$$T = W + \left(1 + \frac{1}{M}\right)B$$

A.7 Reliability and Validity Measures

A.7.1 Mathematical Formulations

Table F.1: Mathematical formulas for reliability and validity measures

Measure	**Formula**	**Key Notes**
Cronbach's Alpha (α)	$\alpha = \frac{I}{I-1}\left(1 - \frac{\sum \psi_i}{\sigma^2_{\text{total}}}\right)$	I = items, ψ_i = error variance, assumes tau-equivalence
McDonald's Omega (ω)	$\omega = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum \psi_i}$	λ_i = standardized loadings, numerator = explained variance
Composite Reliability (CR)	$CR = \frac{(\sum \lambda_i)^2}{(\sum \lambda_i)^2 + \sum (1-\lambda_i^2)}$	Accounts for congeneric measures
Average Variance Extracted (AVE)	$AVE = \frac{\sum \lambda_i^2}{I}$	λ_i = standardized loadings, average variance explained
Polychoric Alpha (α_p)	$\alpha_p = \frac{I}{I-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma^2_{\text{sum}}}\right)$	For ordinal data; σ_i^2 from polychoric covariances
Spearman-Brown (r_{SB})	$r_{SB} = \frac{2r_{hh}}{1+r_{hh}}$	For two-item scales; r_{hh} = correlation between halves
Fornell-Larcker Criterion	$\sqrt{AVE_a} > r_{ab}$ and $\sqrt{AVE_b} > r_{ab}$	Discriminant validity test

A.7.2 Interpretive Thresholds

Table F.2: Interpretive thresholds for reliability and validity

Reliability Level	α, ω, CR	AVE	**Reference**
Excellent	≥ 0.90	≥ 0.70	Nunnally & Bernstein (1994)
Good	0.80–0.89	≥ 0.60	Nunnally & Bernstein (1994)
Acceptable	0.70–0.79	≥ 0.50	Nunnally & Bernstein (1994)
Questionable	0.60–0.69	0.40–0.49	Hair et al. (2010)
Poor	< 0.60	< 0.40	Hair et al. (2010)

A.7.3 Interpretation and Assessment Strategy

Table F.3: Interpretation for reliability and validity assessment

Situation	**Primary Measure**	**Additional Requirements / Action**
Tau-equivalence holds	Cronbach's Alpha	Confirm with constrained CFA model
Tau-equivalence violated	McDonald's Omega	Compare constrained vs. unconstrained CFA models
Ordinal/Likert data	Polychoric Alpha	Always report alongside primary measure
Two-item constructs	Spearman-Brown	Use split-half correlation
Convergent validity	AVE	AVE $\geq$ 0.50 required
Discriminant validity	Fornell-Larcker	$\sqrt{AVE}$ must exceed inter-construct correlations

A.7.4 Contextual Factors

Table F.4: Contextual factors in reliability evaluation

Factor	**Consideration**	**Implication**
Scale Length	2–3 items: Lower reliability expected; 4–7: Standard; 8+: Higher	Adjust interpretation based on item count
Data Type	Ordinal: Use Polychoric Alpha; Continuous: Standard; Binary: KR-20	Match method to data scale
Sample Size	N < 200: Less stable; N > 500: More stable; N > 1000: Overpowered	Interpret stability and significance accordingly

A.7.5 Troubleshooting Common Issues

Table F.5: Common issues and solutions in reliability analysis

Problem	**Potential Causes**	**Solutions**
Low Reliability (< 0.70)	Heterogeneous items, short scales, high error	Check item-total correlations, consider item removal
Failed Discriminant Validity	Conceptual overlap, method effects	Examine correlations, respecify constructs
Discrepant Estimates	$\alpha \ll \omega$: Tau-equivalence violated; Polychoric > Standard: Ordinal data	Report most appropriate measure, explain differences

A.7.6 Minimum Reporting Standards

- Report at least two reliability coefficients per construct (e.g., ω, CR)
- Include AVE values for all latent constructs
- Conduct and report discriminant validity assessment (e.g., Fornell-Larcker)
- Justify choice of primary reliability measure based on model and data characteristics

A.8 Data Description Appendix

A.8.1 Auxiliary Variables

Table F.6: Auxiliary variables for multiple imputation

Variable	Domain	Description	Missing (%)	Strongest Correlation
`age_cat`	Demographics	Age categories	0.9	N/A
`sex_cat`	Demographics	Gender categories	1.4	N/A
`hyg_freq_exp_substances`	Hygiene	Frequency of hazardous substance exposure	4.9	`Hazardous substances` (0.88)
`hyg_freq_exp_tools`	Hygiene	Frequency of vibrating tool exposure	5.0	`Tool vibrations` (0.86)
`hyg_freq_exp_high_temp`	Hygiene	Frequency of high temperature exposure	5.0	`High temperatures` (0.92)
`hyg_freq_exp_low_temp`	Hygiene	Frequency of low temperature exposure	5.2	`Low temperatures` (0.93)
`hyg_freq_exp_noise`	Hygiene	Frequency of noise exposure	5.2	`Noise` (0.89)
`erg_freq_exp_sitting`	Ergonomics	Frequency of prolonged sitting	5.3	`Sitting for long-time` (0.85)
`erg_freq_exp_posture`	Ergonomics	Frequency of stressful postures	5.4	`Stressful postures` (0.92)
`erg_freq_exp_physical`	Ergonomics	Frequency of physical demands	5.4	`Physical strenuous` (0.93)

Between and Within Imputation Variance

Table F.7: Between and within imputation variance components by latent factor

Variable	Mean	Within Var	Between Var	Total Var	FMI
Psychosocial Work Environment (PS)					
Work pace	2.36	0.570	1.33×10^{-5}	0.570	2.56×10^{-5}
Emotional demands	2.43	0.506	8.05×10^{-5}	0.506	1.75×10^{-4}
Work atmosphere	2.39	0.539	4.91×10^{-6}	0.539	1.00×10^{-5}
Work-life balance	2.50	0.458	9.19×10^{-6}	0.458	2.21×10^{-5}
Ergonomics (ER)					
Stressful postures	3.31	1.051	1.21×10^{-4}	1.051	1.26×10^{-4}
Repetitive work	3.54	0.853	3.73×10^{-6}	0.853	4.81×10^{-6}
Sitting for long-time	2.95	1.065	7.75×10^{-6}	1.065	8.01×10^{-6}
Manual handling loads	3.42	1.023	6.57×10^{-6}	1.023	7.07×10^{-6}
Physical strenuous	3.62	0.796	1.24×10^{-5}	0.796	1.71×10^{-5}
Safety (SA)					
Work involvement	2.56	0.428	1.55×10^{-5}	0.428	3.98×10^{-5}
Leadership engagement	2.57	0.394	2.57×10^{-5}	0.394	7.17×10^{-5}
Hygiene (HY)					
Tool vibrations	3.57	0.786	1.84×10^{-6}	0.786	2.58×10^{-6}
Low temperatures	3.38	1.085	3.66×10^{-6}	1.085	3.71×10^{-6}
High temperatures	3.51	0.892	2.38×10^{-4}	0.892	2.93×10^{-4}
Noise	3.32	0.997	4.19×10^{-4}	0.997	4.62×10^{-4}
Hazardous substances	3.69	0.653	0.00000	0.653	0.00000

Distribution Comparison

Table F.8: Distribution statistics for original and imputed data

Variable	Original Distribution	Missing (%)	Original Mean (SD)	Imputed Mean / SD Range
Psychosocial (PS)				
Work pace	1: 85, 2: 211, 3: 365	5.4%	2.42 (0.71)	2.35–2.36 / 0.75–0.76
Emotional demands	1: 66, 2: 208, 3: 387	5.4%	2.49 (0.67)	2.42–2.45 / 0.69–0.72
Work atmosphere	1: 69, 2: 216, 3: 374	5.7%	2.46 (0.68)	2.38–2.39 / 0.73–0.74
Work-life balance	1: 37, 2: 204, 3: 418	5.7%	2.58 (0.60)	2.49–2.50 / 0.67–0.68
Ergonomics (ER)				
Stressful postures	1: 31, 2: 119, 3: 53, 4: 458	5.4%	3.42 (0.94)	3.30–3.33 / 1.00–1.05
Repetitive work	1: 12, 2: 71, 3: 40, 4: 538	5.4%	3.67 (0.74)	3.54–3.54 / 0.92–0.93
Sitting for long-time	1: 36, 2: 179, 3: 164, 4: 283	5.3%	3.05 (0.96)	2.94–2.95 / 1.03–1.03
Manual handling loads	1: 29, 2: 81, 3: 51, 4: 500	5.4%	3.55 (0.87)	3.42–3.42 / 1.01–1.01
Physical strenuous	1: 18, 2: 40, 3: 25, 4: 578	5.4%	3.76 (0.68)	3.62–3.63 / 0.88–0.90
Safety (SA)				
Work involvement	1: 39, 2: 174, 3: 455	4.4%	2.62 (0.59)	2.56–2.57 / 0.64–0.66
Leadership engagement	1: 37, 2: 182, 3: 449	4.4%	2.62 (0.59)	2.56–2.57 / 0.62–0.64
Hygiene (HY)				
Tool vibrations	1: 13, 2: 43, 3: 70, 4: 536	5.2%	3.71 (0.68)	3.57–3.57 / 0.88–0.89
Low temperatures	1: 32, 2: 97, 3: 38, 4: 497	5.0%	3.51 (0.91)	3.38–3.38 / 1.04–1.04
High temperatures	1: 22, 2: 70, 3: 42, 4: 530	5.0%	3.63 (0.80)	3.50–3.54 / 0.90–0.97
Noise	1: 33, 2: 114, 3: 72, 4: 445	5.0%	3.40 (0.94)	3.30–3.37 / 0.93–1.04
Hazardous substances	1: 9, 2: 24, 3: 41, 4: 591	4.9%	3.83 (0.55)	3.69–3.69 / 0.81–0.81

Imputation Fix Index

Table F.9: Imputation fit index (IFI) results by latent factor

Variable	**Orig. SE**	**Pooled SE**	$\mathbf{IFI_m}$	$\mathbf{S^2{}_{IFI}}$	$\mathbf{Z_{IFI}}$
Psychosocial Work Environment (PS)					
Work pace	0.0276	0.0286	0.0010	2.53×10^{-7}	-1.87
Emotional demands	0.0261	0.0269	0.0008	1.98×10^{-7}	-1.72
Work atmosphere	0.0264	0.0278	0.0014	3.94×10^{-7}	-1.68
Work-life balance	0.0233	0.0256	0.0023	6.13×10^{-7}	-1.46
Ergonomics (ER)					
Stressful postures	0.0366	0.0388	0.0022	8.47×10^{-7}	-1.52
Repetitive work	0.0287	0.0349	0.0062	3.89×10^{-6}	-0.93
Sitting for long-time	0.0372	0.0390	0.0018	4.12×10^{-7}	-1.66
Manual handling loads	0.0338	0.0382	0.0044	2.18×10^{-6}	-1.29
Physical strenuous	0.0266	0.0337	0.0071	5.04×10^{-6}	-0.85
Safety (SA)					
Work involvement	0.0230	0.0248	0.0018	3.27×10^{-7}	-1.70
Leadership engagement	0.0228	0.0237	0.0009	1.05×10^{-7}	-1.90
Hygiene (HY)					
Tool vibrations	0.0263	0.0335	0.0072	5.29×10^{-6}	-0.84
Low temperatures	0.0354	0.0394	0.0040	1.60×10^{-6}	-1.34
High temperatures	0.0312	0.0357	0.0045	3.17×10^{-6}	-1.18
Noise	0.0365	0.0378	0.0013	2.59×10^{-7}	-1.78
Hazardous substances	0.0212	0.0306	0.0094	8.84×10^{-6}	-0.77

Note. Z_{IFI} values indicate the standardized difference in standard errors between original and imputed data. Values ≤ -1.65 suggest minimal discrepancy (good imputation fit), while higher values indicate greater divergence and potential fit concerns.

A.8.2 Factor Correlations

Table F.10: Factor correlations

Factor 1	**Factor 2**	**Correlation (SE)**
PS	ER	0.372 (0.046)
PS	SA	0.681 (0.038)
PS	HY	0.295 (0.049)
ER	SA	0.450 (0.042)
ER	HY	0.586 (0.037)
SA	HY	0.408 (0.044)

A.8.3 Threshold Estimates

Table F.11: Threshold estimates from the confirmatory factor analysis

Domain	Variable	Threshold	Estimate (SE)
Psychosocial Work Environment (PS)			
	Work pace	t1	−0.956 (0.056)
	Work pace	t2	−0.067 (0.047)
	Emotional demands	t1	−1.113 (0.060)
	Emotional demands	t2	−0.148 (0.048)
	Work atmosphere	t1	−1.036 (0.058)
	Work atmosphere	t2	−0.097 (0.048)
	Work-life balance	t1	−1.261 (0.064)
	Work-life balance	t2	−0.259 (0.048)
Ergonomics (ER)			
	Stressful postures	t1	−1.474 (0.072)
	Stressful postures	t2	−0.625 (0.051)
	Stressful postures	t3	−0.403 (0.049)
	Repetitive work	t1	−1.501 (0.073)
	Repetitive work	t2	−0.959 (0.056)
	Repetitive work	t3	−0.752 (0.053)
	Sitting for long-time	t1	−1.281 (0.065)
	Sitting for long-time	t2	−0.365 (0.049)
	Sitting for long-time	t3	0.235 (0.048)
	Manual handling loads	t1	−1.336 (0.067)
	Manual handling loads	t2	−0.814 (0.054)
	Manual handling loads	t3	−0.579 (0.050)
	Physical strenuous	t1	−1.437 (0.070)
	Physical strenuous	t2	−1.112 (0.060)
	Physical strenuous	t3	−0.959 (0.056)
Safety (SA)			
	Work involvement	t1	−1.316 (0.066)
	Work involvement	t2	−0.389 (0.049)
	Leadership engagement	t1	−1.385 (0.068)
	Leadership engagement	t2	−0.369 (0.049)
Hygiene (HY)			
	Tool Vibrations	t1	−1.480 (0.072)
	Tool Vibrations	t2	−1.119 (0.060)
	Tool Vibrations	t3	−0.731 (0.052)
	Low temperatures	t1	−1.313 (0.066)
	Low temperatures	t2	−0.726 (0.052)
	Low temperatures	t3	−0.558 (0.050)
	High temperatures	t1	−1.453 (0.071)

Domain	Variable	Threshold	Estimate (SE)
	High temperatures	t2	−0.912 (0.055)
	High temperatures	t3	−0.701 (0.052)
	Noise	t1	−1.469 (0.072)
	Noise	t2	−0.718 (0.052)
	Noise	t3	−0.349 (0.049)
	Hazardous Substances	t1	−1.542 (0.075)
	Hazardous Substances	t2	−1.306 (0.065)
	Hazardous Substances	t3	−1.017 (0.058)

Convergence Diagnostics

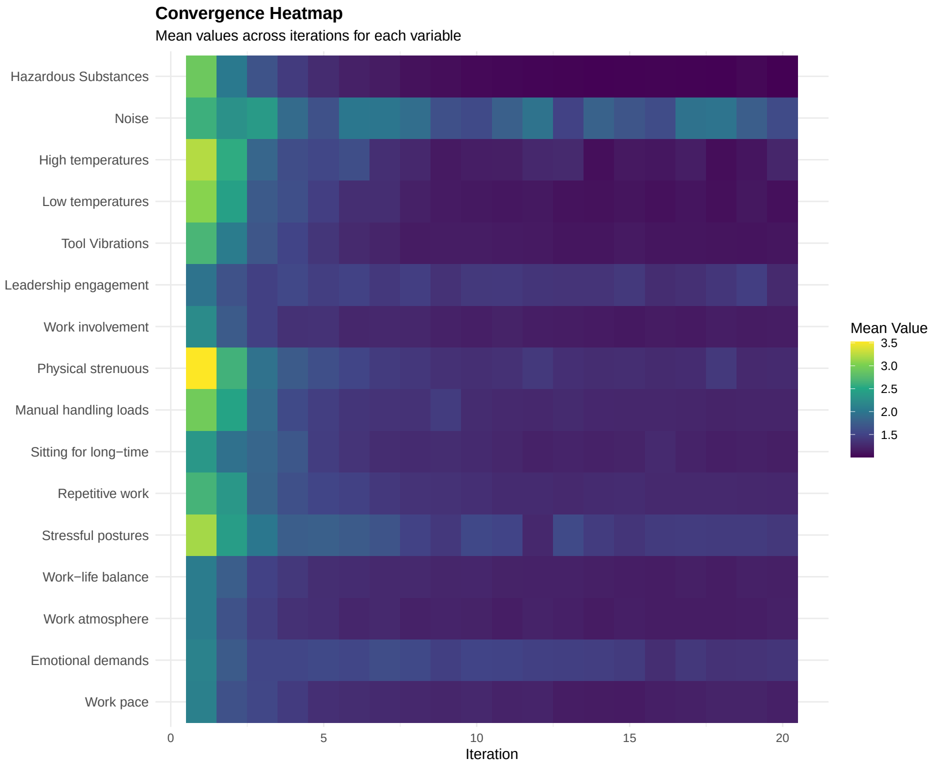


Figure A.1: Heat map of convergence mean value of imputation

A.8.4 Convergence Diagnostics

A.9 Figure showing convergence trace plot

Provide a detailed figure showing a convergence trace plot for the different domain factors for both mean and standard deviation.

A.9.1 Measurement invariance visual

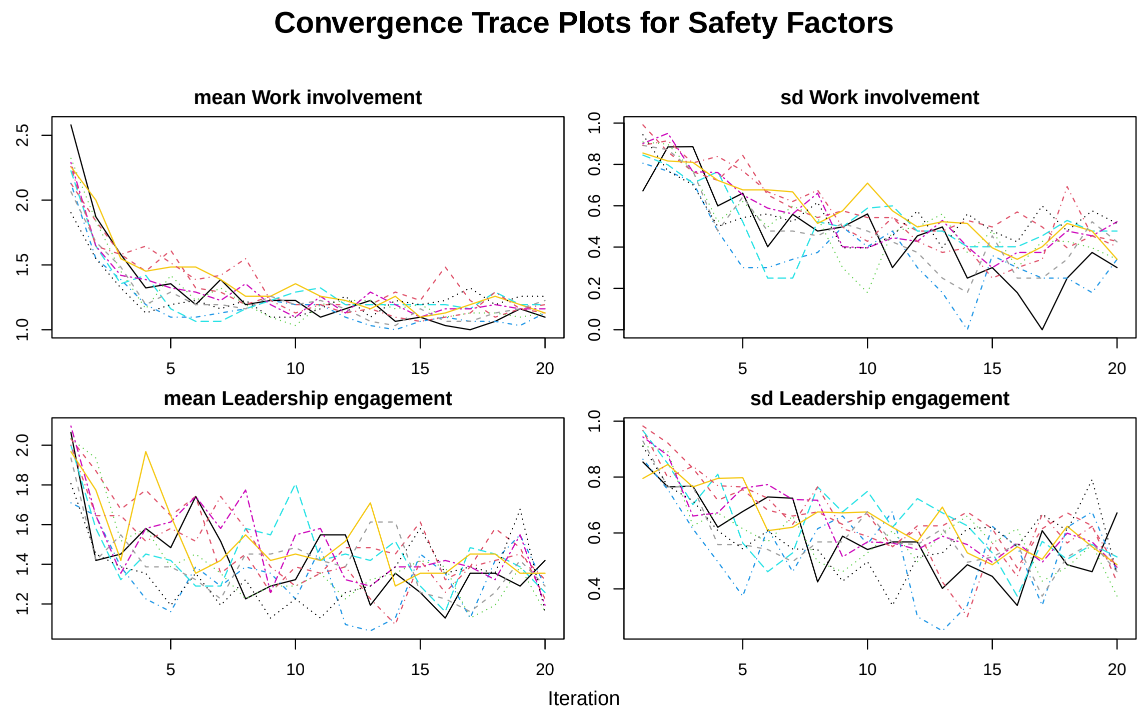


Figure A.2: Convergence trace plot for Safety

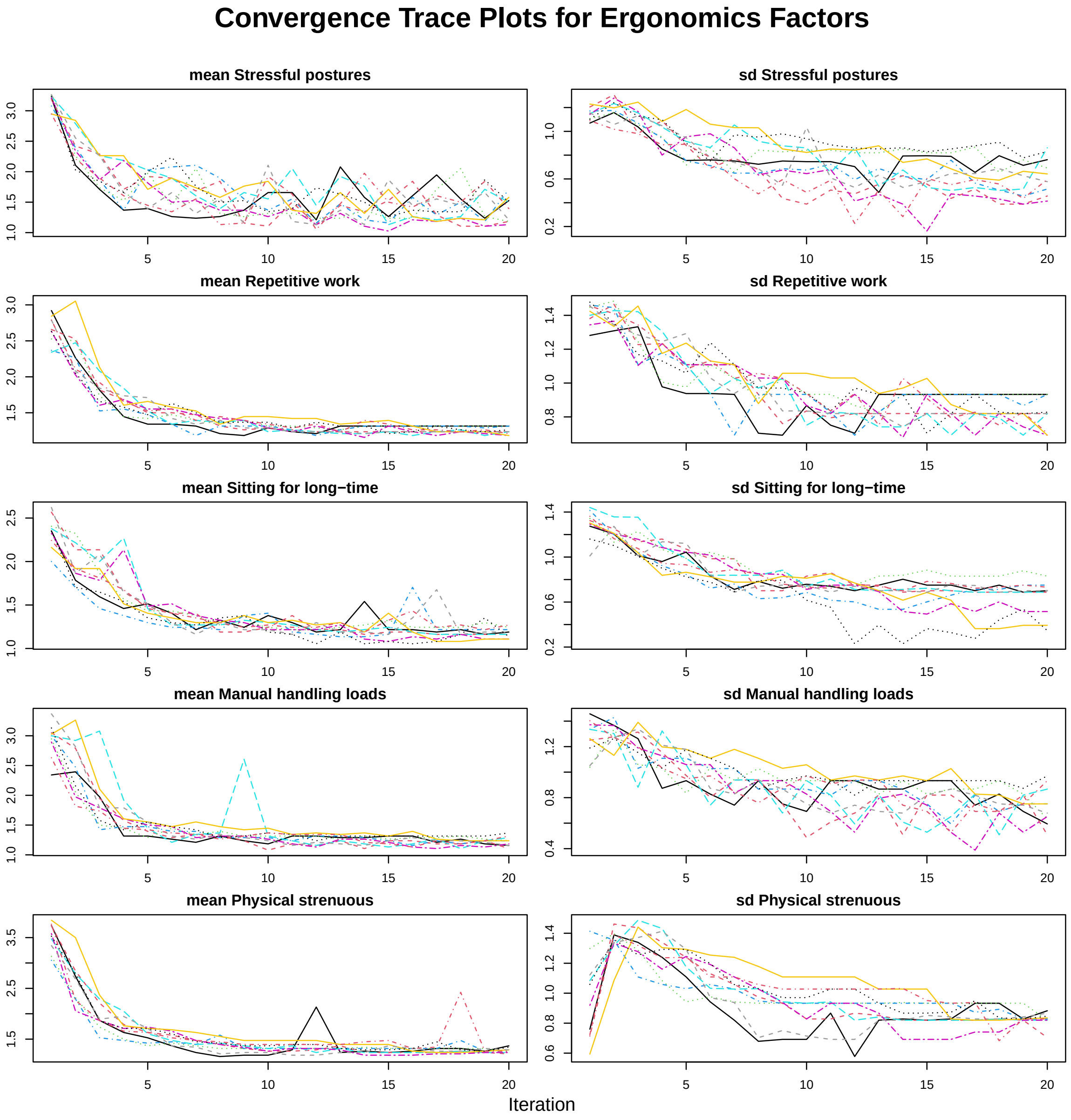


Figure A.3: Convergence trace plot for Ergonomics

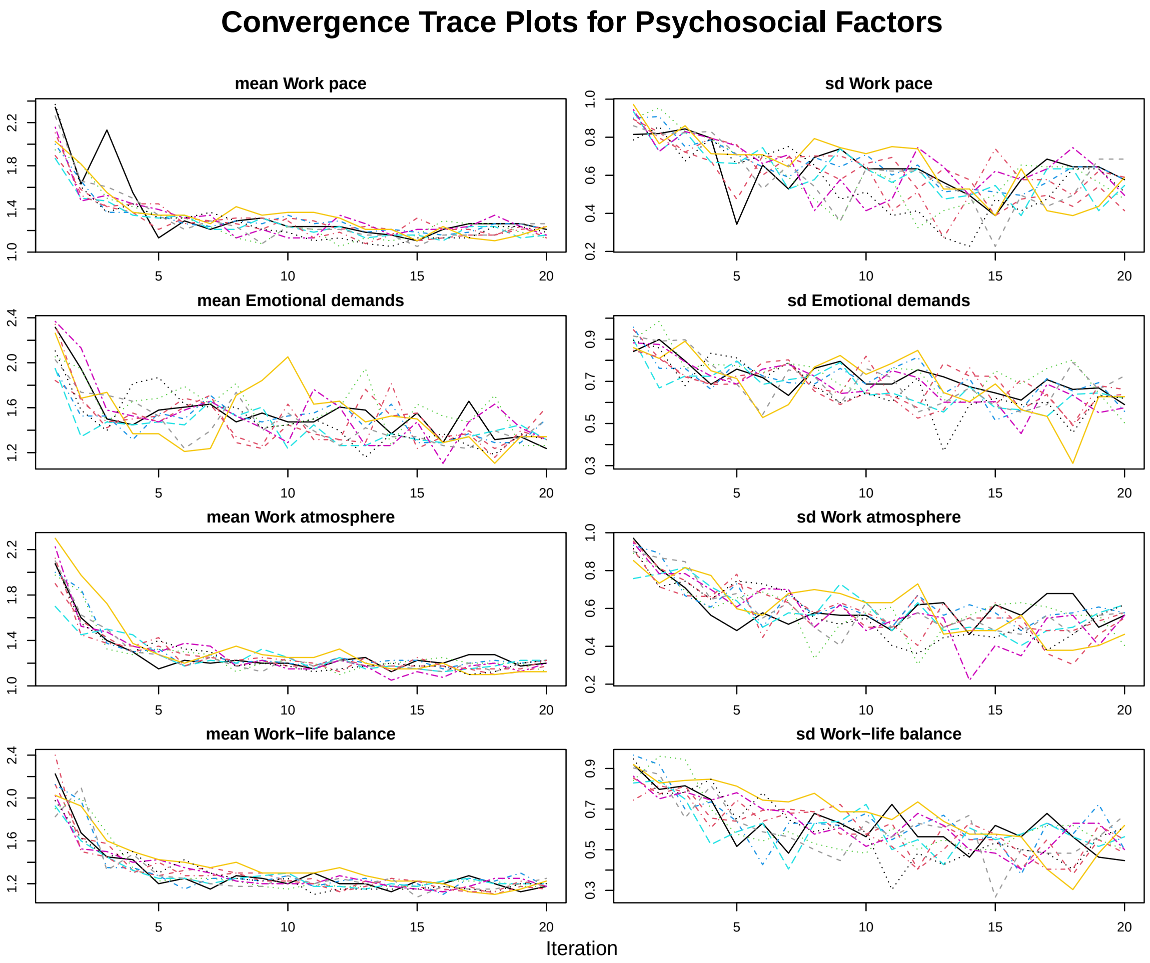


Figure A.4: Convergence trace plot for Psychosocial

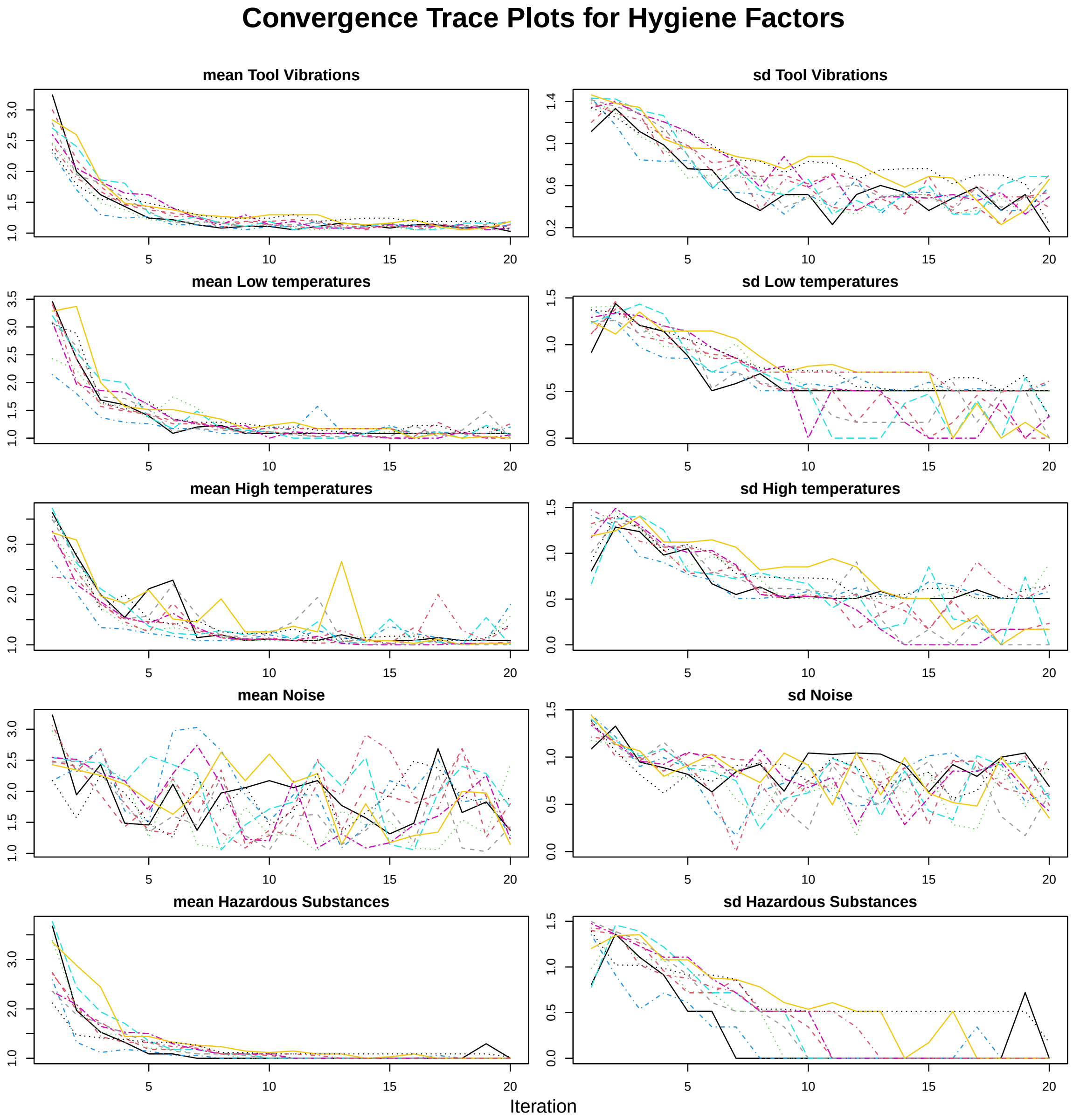


Figure A.5: Convergence trace plot for Hygiene

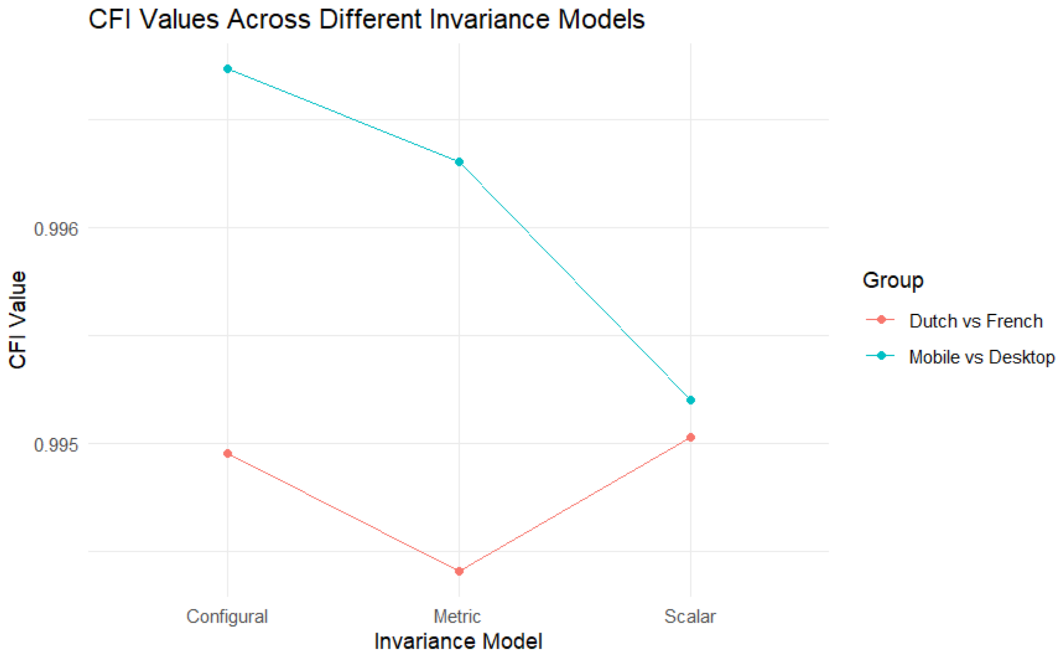


Figure A.6: CFI values across different invariance models for language groups (Dutch vs. French) and device types (Mobile vs. Desktop). Note the minimal changes in CFI values across models, with all values remaining above 0.99, indicating excellent fit at all levels of invariance testing.

Appendix B

Relevant R Code

Note: The complete code and the simulated data (not the company data used due to sensitivity) are available at: GitHub Repository.

B.1 Data Preparation and Variable Selection

Listing B.1: Define model variables for four-factor structure

```r
# Define model variables for four-factor structure
model_vars <- list(
  PS = c("veerkracht_calc", "psy_rating_pace", "psy_rating_emotional",
         "psy_rating_sphere", "psy_work_life"),
  ER = c("erg_capac", "erg_rating_posture", "erg_rating_repeat",
         "erg_rating_sitting", "erg_rating_loads", "erg_rating_physical"),
  SA = c("saf_satisfaction", "saf_rating_workinvolv", "saf_rating_leadengage"),
  HY = c("hyg_satisfaction", "hyg_rating_tools", "hyg_rating_low_temp",
         "hyg_rating_high_temp", "hyg_rating_noise", "hyg_rating_substances")
)

# Select auxiliary variables with <20% missing for imputation
auxiliary_vars <- c("age_cat", "sex_cat", "freq_exp_variables",
                   "health_indicators")
```

B.2 Missing Data Analysis

Listing B.2: Calculate missing data patterns by domain

```r
# Calculate missing data patterns by domain
domain_missingness <- lapply(names(domains), function(domain_name) {
  vars <- domains[[domain_name]]
  full_analysis_dataset %>%
    summarise(across(all_of(vars), ~mean(is.na(.))*100)) %>%
    pivot_longer(cols = everything(),
                 names_to = "variable",
                 values_to = "percent_missing") %>%
    mutate(domain = domain_name)
}) %>% bind_rows()
```

```r
# Little's MCAR test
mcar_test <- BaylorEdPsych::LittleMCAR(full_analysis_dataset[model_vars])
```

B.3 Multiple Imputation

Listing B.3: MICE imputation setup and execution

```r
# Set up MICE imputation with appropriate methods
imp_methods <- mice(imputation_dataset, maxit = 0)$method
imp_methods[ordinal_vars] <- "polr"       # Proportional odds for ordinal
imp_methods[continuous_vars] <- "pmm"      # Predictive mean matching
imp_methods[binary_vars] <- "logreg"       # Logistic regression

# Run multiple imputation
imputed_data <- mice(imputation_dataset,
                    method = imp_methods,
                    m = 10,          # 10 imputed datasets
                    maxit = 20,      # 20 iterations
                    seed = 12345)
```

B.4 Confirmatory Factor Analysis Model

Listing B.4: Four-factor CFA model specification

```r
# Define four-factor CFA model
cfa_model <- '
  # Psychosocial work environment factor
  PS =~ psy_rating_pace + psy_rating_emotional +
        psy_rating_sphere + psy_work_life

  # Ergonomics factor
  ER =~ erg_rating_posture + erg_rating_repeat + erg_rating_sitting +
        erg_rating_loads + erg_rating_physical

  # Safety factor
  SA =~ saf_rating_workinvolv + saf_rating_leadengage

  # Hygiene factor
  HY =~ hyg_rating_tools + hyg_rating_low_temp + hyg_rating_high_temp +
        hyg_rating_noise + hyg_rating_substances
'

# Fit model with WLSMV estimator for ordinal data
fit <- cfa(cfa_model,
          data = dataset,
          ordered = ordinal_vars,
          estimator = "WLSMV")
```

B.5 Reliability and Validity Assessment

Listing B.5: Composite Reliability and AVE calculation

```
# Function to calculate Composite Reliability and AVE
calculate_CR_AVE <- function(fit) {
  std_estimates <- standardizedSolution(fit)
  loadings <- subset(std_estimates, op == "=~")

  results <- data.frame()
  for(factor in unique(loadings$lhs)) {
    lambda <- subset(loadings, lhs == factor)$est.std
    lambda_squared <- lambda^2

    # Composite Reliability
    CR <- sum(lambda)^2 / (sum(lambda)^2 + sum(1 - lambda_squared))

    # Average Variance Extracted
    AVE <- sum(lambda_squared) / length(lambda)

    results <- rbind(results, data.frame(Factor = factor, CR = CR, AVE = AVE))
  }
  return(results)
}

# Discriminant validity: Fornell-Larcker criterion
# Square root of AVE should exceed inter-factor correlations
factor_cors <- lavInspect(fit, "cor.lv")
sqrt_AVE <- sqrt(CR_AVE_results$AVE)
```

B.6 Pooled Analysis Across Imputed Datasets

Listing B.6: Bootstrap and pooling across imputations

```
# Bootstrap and pool results across imputed datasets
for (m in 1:M) {
  # Bootstrap within each imputed dataset
  boot_fun <- function(data, indices) {
    boot_sample <- data[indices, ]
    fit <- cfa(cfa_model, data = boot_sample,
               ordered = ordinal_vars, estimator = "DWLS")
    return(fitMeasures(fit, c("cfi", "tli", "rmsea", "srmr")))
  }

  boot_results <- boot(data = completed_datasets[[m]],
                      statistic = boot_fun, R = 1000)
}

# Pool using Rubin's rules with transformations
# Fisher transformation for CFI/TLI, log transformation for RMSEA
pooled_estimate <- mean(transformed_estimates)
within_var <- mean(variances)
between_var <- var(estimates)
total_var <- within_var + (1 + 1/m) * between_var
```

B.7 Measurement Invariance Testing

Listing B.7: Measurement invariance testing with NHT and ET approaches

```
# Function for measurement invariance with NHT and ET approaches
test_invariance <- function(data, model, group_var, ordinal_vars) {
  # Configural invariance
  configural <- cfa(model, data = data, group = group_var,
                    ordered = ordinal_vars, estimator = "WLSMV")

  # Metric invariance (equal loadings)
  metric <- cfa(model, data = data, group = group_var,
                ordered = ordinal_vars, estimator = "WLSMV",
                group.equal = "loadings")

  # Scalar invariance (equal loadings and thresholds)
  scalar <- cfa(model, data = data, group = group_var,
                ordered = ordinal_vars, estimator = "WLSMV",
                group.equal = c("loadings", "thresholds"))

  # Evaluate using fit index changes
  # NHT: $\Delta$CFI > -0.01, $\Delta$RMSEA < 0.015, $\Delta$SRMR < 0.030 (
     metric) or 0.010 (scalar)


  # ET: Compare maximum parameter differences to equivalence bounds

  return(list(configural = configural, metric = metric, scalar = scalar))
}

# Test across multiple grouping variables
invariance_groups <- c("language_group", "device_type", "env_domain",
                       "sex_cat", "age_cat")
```

B.8 Differential Item Functioning Analysis

Listing B.8: Differential Item Functioning Analysis

```
r# Mantel-Haenszel DIF detection
perform_mh_dif <- function(data, group_var, items) {
  item_data <- data[, items]
  group_data <- data[[group_var]]
  total_scores <- rowSums(item_data, na.rm = TRUE)

  # Prepare data for difR
  dif_input <- cbind(item_data, group_data)
  colnames(dif_input) <- c(items, "group")

  # Run Mantel-Haenszel test
  mh_results <- difMH(Data = dif_input,
                      group = "group",
                      focal.name = levels(group_data)[2],
                     match = total_scores,
                     purify = TRUE)

  # Classify DIF magnitude (ETS criteria)
  alpha_mh <- mh_results$alphaMH
  dif_classification <- ifelse(abs(log(alpha_mh)) < 1.0, "A (Negligible)",
                              ifelse(abs(log(alpha_mh)) < 1.5, "B (Moderate)",
                                     "C (Large)"))

  return(list(
    statistics = mh_results$MH,
    alpha = alpha_mh,
    classification = dif_classification,
    significant = mh_results$MH > qchisq(0.95, 1)
  ))
}

# Logistic Regression DIF detection
perform_lr_dif <- function(data, group_var, items) {
  results <- list()

  for(item in items) {
    total_score <- rowSums(data[, items], na.rm = TRUE)
    item_median <- median(data[[item]], na.rm = TRUE)
    item_binary <- ifelse(data[[item]] > item_median, 1, 0)

    # Base model (ability only)
    model1 <- glm(item_binary ~ total_score, family = binomial)

    # Uniform DIF model (ability + group)
    model2 <- glm(item_binary ~ total_score + data[[group_var]], family =
       binomial)

    # Non-uniform DIF model (ability + group + interaction)
    model3 <- glm(item_binary ~ total_score + data[[group_var]] +
                 total_score:data[[group_var]], family = binomial)

    # Test for DIF
    uniform_test <- anova(model1, model2, test = "Chisq")
    nonuniform_test <- anova(model2, model3, test = "Chisq")

    results[[item]] <- list(
      uniform_p = uniform_test$`Pr(>Chi)`[2],
      nonuniform_p = nonuniform_test$`Pr(>Chi)`[2]
    )
  }

  return(results)
}
```

B.9 Key Functions for Pooled Reliability

Listing B.9: Reliability calculation across imputed datasets

```
# Calculate reliability across imputed datasets
calculate_pooled_reliability <- function(datasets, factor_vars) {
  for (factor_name in names(factor_vars)) {
    # Calculate alpha, omega, and polychoric alpha for each imputation
    alpha_values <- numeric(length(datasets))
    omega_values <- numeric(length(datasets))

    for (i in 1:length(datasets)) {
      factor_data <- datasets[[i]][, factor_vars[[factor_name]]]

      # Cronbach's alpha
      cor_matrix <- cor(factor_data, use = "complete.obs")
      k <- ncol(factor_data)
      avg_r <- (sum(cor_matrix) - k) / (k * (k - 1))
      alpha_values[i] <- (k * avg_r) / (1 + (k - 1) * avg_r)

      # McDonald's omega (for factors with >2 items)
      if (k > 2) {
        fit <- cfa(paste0(factor_name, " =~ ",
                         paste(factor_vars[[factor_name]], collapse = " + ")),
                  data = datasets[[i]], ordered = TRUE, estimator = "WLSMV")
        omega_values[i] <- reliability(fit)["omega", factor_name]
      }
    }

    # Pool estimates with confidence intervals
    pooled_alpha <- mean(alpha_values)
    se_alpha <- sd(alpha_values) / sqrt(length(datasets))
    ci_alpha <- pooled_alpha + c(-1.96, 1.96) * se_alpha
  }
}
```

B.10 Model Fit Summary Function

Listing B.10: Model fit indices extraction and interpretation

```
# Extract and format key fit indices
summarize_fit <- function(fit) {
  indices <- fitMeasures(fit, c("chisq", "df", "pvalue",
                               "cfi", "tli", "rmsea", "srmr"))

  fit_summary <- data.frame(
    Measure = c("Chi-square", "df", "p-value", "CFI", "TLI",
                "RMSEA", "SRMR"),
    Value = round(indices, 3),
    Interpretation = c(
      ifelse(indices["pvalue"] > 0.05, "Good fit", "Poor fit"),
      NA,
      NA,
      ifelse(indices["cfi"] > 0.95, "Excellent", "Acceptable"),
      ifelse(indices["tli"] > 0.95, "Excellent", "Acceptable"),
      ifelse(indices["rmsea"] < 0.06, "Good fit", "Acceptable"),
      ifelse(indices["srmr"] < 0.08, "Good fit", "Poor fit")
    )
  )
  return(fit_summary)
}
```

Note: The repository includes a simulated dataset that replicates the structure and characteristics of the original workplace assessment data, allowing full reproducibility of all analyses while maintaining participant privacy.

B.10.1 Repository Structure

```
workplace-wellbeing-assessment-thesis/
 CFA_rCode_JimmyKomalceh_Thesis.r    # Complete analysis code
 create_sample_data.r                 # Simulated data generator
 Simulated_data_SAMPLE.csv            # Simulated dataset
 README.md                            # Main documentation
 README_SampleGeneration.md           # Data simulation documentation
 requirements.txt                     # R package dependencies
 LICENSE                              # Repository license
 .gitignore                          # Git ignore file
```

B.10.2 Requirements

- R version 4.0+
- Required packages: `tidyverse`, `lavaan`, `mice`, `semTools`, `psych`, `MVN`
- See `requirements.txt` for complete package list

B.10.3 Citation

If using this code, please cite:

> [Jimmy Komalceh]. (2025). Workplace Well-being Assessment: Analysis Code. GitHub repository: https://github.com/jkomalceh/workplace-wellbeing-assessment_MasterThesis/blob/main/CFA_rCode_JimmyKomalceh_Thesis.r