Strål
säkerhets
myndigheten

Swedish Radiation Safety Authority

Author: Martin Castor, GEISTT, Stockholm
Jonathan Borgvall, GEISTT, Stockholm

Research

2018:24

Structural equation modelling for analysis of human and organizational factors in nuclear power plant control rooms

Report number: 2018:24 ISSN: 2000-0456
Available at www.stralsakerhetsmyndigheten.se

SSM perspective

Background

During the supervisory work of the Swedish Radiation Safety Authority (SSM), a need was identified to develop the methods used when evaluating control room work in the central control rooms of nuclear power plants. Benchmarking is commonly used today, with reference values from earlier Integrated System Validation (ISV), when ISV is available. Often, ISV performs well but has some weaknesses. Some of the elements of knowledge that are currently missing include how to establish strict clarity concerning the aspects that have individual importance and which aspects are important collectively, as well as how to match different measurable aspects. Improved knowledge in this area, in addition to an advanced method, can give a credible outcome and provide guidance when formulating specifications of requirements for requisite skills, provide input for education and training programmes which may need sharper focus, and achieve a higher level of knowledge in-house at SSM in relation to supervision in the field.

The assignment to investigate the methods used when evaluating control room work in the central control rooms was given to GEISTT, which as part of a research project, placed a focus on how the methods of evaluation might be improved. This was done by means of an in-depth study of how data can be analysed and presented using a static method for modelling called Structural Equation Modelling (SEM). Examples of useful output from/benefits of SEM include the possibility to integrate the analysis of several different data collection methods and scales, as well as the possibility to present the outcomes in a way defining the factors of greatest significance, e.g. in order to illustrate acceptance values for different criteria. SEM can also be used for plant modifications both large and small.

Objective

The method has not yet been applied by the Swedish nuclear power industry. On the other hand, forms of cooperation for development of evaluation methods have been established not only with other agencies that regulate the nuclear power industry, but also with IFE/Halden, which have shown great interest. Based on the stringency and the outcomes produced within the project, SSM expects that further research will be carried out using quantities of data formulated/designed/adapted?? to better suit SEM, to which other sources of funding will also contribute.

Results

The results indicate that SEM is a statistical modelling method that can meet needs and increase the level of knowledge to possibly benefit individual facilities, educational institutions and the Authority. The results also indicate that when conducting evaluations, it is essential from the outset, prior to the evaluations, to conscientiously look into how measures and variables are formulated and to set the parameters for the quantity of data while considering how the outcomes should be collected and analysed.

Need for further research

A current evaluation method such as ISV is designed to make detailed information available regarding Human Error Discrepancies, which is specific to and appropriate for the nuclear power plant in question, and limited to benchmarking only at this facility. However, since there is a need to have capability to perform comparisons on a more general level in order to develop this area of competence, SSM has established that there is a need for further research. One need that has been identified is investigating whether it is valuable to study the outcomes from previous evaluations of integrated systems and to raise them to a higher level of abstraction for the purpose of achieving comparability and reinforcing the reference values.

Project information

Contact person at SSM: Yvonne Johansson
Reference: SSM2016-690 / 7030057-00

Strål
säkerhets
myndigheten
Swedish Radiation Safety Authority

Author: Martin Castor, GEISTT, Stockholm
Jonathan Borgvall, GEISTT, Stockholm

2018:24
Structural equation modelling for analysis of human and organizational factors in nuclear power plant control rooms

Date: November 2018
Report number: 2018:24 **ISSN:** 2000-0456
Available at www.stralsakerhetsmyndigheten.se

Structural equation modelling for analysis of human and organizational factors in nuclear power plant control rooms

Authors: Martin Castor & Jonathan Borgvall, GEISTT AB

Content

Author Note

Dr. Martin Castor, GEISTT AB (Group for Effectiveness, Interaction, Simulation Technology, and Training), Stockholm, Sweden; Mr. Jonathan Borgvall, GEISTT AB, Stockholm, Sweden.

This work was funded by the Swedish Radiation Safety Authority (SSM) and commissioned by SSM to GEISTT AB. The authors express their gratitude to professor emeritus Erland Svensson for his review of a draft version of the report.

Correspondence concerning this report should be sent to Dr. Martin Castor, GEISTT AB, Isafjordsgatan 39B, 164 40 Kista, Stockholm, Sweden. E-mail: martin.castor@geistt.com.

Figures

Tables

1. Executive Summary

Major changes to a nuclear power plant, e.g. modernization of the central control room, usually requires an Integrated System Validation (ISV) evaluation to be conducted to ensure that safe operations can be maintained after the changes have been implemented. For changes to a central control room, the impact on human and organizational factors such as interactions between humans and systems as well as work process, instructions etc. need to be carefully evaluated, proposedly supported by a range of heterogeneous datasets to be assessed before changes can be declared safe for implementation.

Structural Equation Modelling (SEM) is a quantitative, second generation multivariate statistical analysis method that combines the benefits of path analysis, factor analysis, and multiple regression analysis. Basically, SEM tests if a modellers theory, expressed in the model, fits the data and SEM is thus primarily a so called confirmatory method. Several alternative models can usually be specified, but SEM provides several goodness of fit - values of how well a model explains the variance of the dataset, which can then be used to compare alternative models.

SEM requires a level of statistical understanding that is often beyond laymen’s understanding. The report provides description of SEM at several different levels, from high level descriptions of the potential of SEM application to step by step descriptions of how a model is developed.

A dummy data set was developed, inspired by a recent ISV process of a Swedish nuclear power plant central control room. SEM models from this dataset are presented to show the potential of SEM for ISV.

The fact that SEM can be used to express relations between variables collected during an ISV is not surprising, given that SEM is a general statistical method designed to describe relations between many variables. The conclusion of the report is that SEM represents a powerful statistical analysis method which is useful for analysis of large and heterogeneous datasets, which is often the case with datasets from operational settings and processes. Accordingly, SEM-analyses make it possible to draw scientifically valid conclusions in operational settings of high operational realism and complexity.

2. Sammanfattning

För att säkerställa att säker drift kan upprätthållas efter det att större anläggningsändringarna har genomförts i ett kärnkraftverk, t.ex. modernisering av det centrala kontrollrummet, krävs vanligen att Integrerad System Validering (ISV) genomförs. För anläggningsändringar i ett centralt kontrollrum ska påverkan på mänskliga och organisatoriska faktorer som t.ex. samspelet mellan människor och system samt arbetsprocesser, instruktioner m.m. utvärderas noggrant, förslagsvis med stöd av en rad heterogena datamängder som skall analyseras innan förändringar kan förklaras säkra att införa.

Strukturella ekvationsmodeller (SEM) är en kvantitativ, andra generationens multivariat statistisk analysmetod som kombinerar fördelarna med ”path analysis”, faktoranalys, och multipel regressionsanalys. I grund och botten svarar en SEM analys på hur väl en modellerares teorier, som uttryckts i modellen, passar gentemot en databas. SEM är alltså i första hand en hypotesprövande så kallad konfirmatic statistisk metod. Flera alternativa modeller prövas ofta, och SEM ger värden på hur väl en modell passar eller kan förklara en viss datamängd, vilket således kan användas för att jämföra alternativa modeller.

SEM kräver en nivå av statistisk förståelse som ofta är bortom lekmäns gängse kunskap. Denna rapport beskriver SEM i flera olika nivåer, från att övergripande beskriva SEM och dess potential till att förklara hur SEM bör genomföras steg för steg.

För att svara på frågan avseende lämpligheten i att använda SEM för att analysera ISV-data avseende utvärderingen av operatörers arbete i centralt kontrollrum utvecklades en s.k. ”dummy databas”. Det är alltså en fiktiv databas, men som utvecklats för att i hög utsträckning återspegla hur en databas från en ISV-process avseende centralt kontrollrum skulle kunna se ut. SEM modeller baserat på denna fiktiva databas presenteras för att studera lämplighet och potential i användandet SEM för ISV.

Det faktum att SEM kan användas för att förklara relationer mellan variabler som samlas in under en ISV är inte förvånande med tanke på att SEM är en generell statistisk metod som syftar till att beskriva relationerna mellan många variabler. Slutsatsen i rapporten är att SEM representerar en kraftfull statistisk analysmetod som är användbar för analys av stora och heterogena datamängder, vilket ofta är fallet med datamängder från operativa miljöer och processer.

3. Introduction

3.1. Background

The reported study was commissioned to GEISTT AB by the Swedish Radiation Safety Authority (SSM, *Sw.* Strålsäkerhetsmyndigheten) to inform them on the applicability and value of utilizing the statistical method Structural Equation Modelling (SEM) for analysis and modelling of Human and Organizational Factors (HOF)-data collected during Integrated System Validation (ISV) processes.

3.2. Rationale for the study

The primary purpose of the research was to investigate benefits of SEM as a component of the ISV toolbox. Hence, the focus of the research is method development by evaluating the applicability and value of SEM as a component of ISV.

Compared to many ISV processes, where the purpose is to identify detailed HEP or HED (Human Error Probabilities or Human Error Discrepancies) associated with plant or procedures changes, the purpose of this type of modelling effort is to create comparability across ISV results, and build scientific and operational understanding across studies.

SEM has been used as a statistical analysis and modelling method in many behavioural research efforts. SEM can be used to develop data-driven models that explain how selected behavioural constructs relate to each other. Through the modelling process it is also often the case that the measurement tools are refined and the understanding of how different measures relate to each other is further developed.

Data collection and data analysis is a constantly recurring challenge when describing and analysing a complex teamwork situation. Typically, numerous, heterogeneous data sources are used to describe the work process and the interaction between human operators, technical systems, and organizational factors. This may generate a dataset which is challenging to compile and present effectively, especially for some types of stakeholders, such as an reviewer at a regulatory agency. SEM provides the statistical capability to quantify the relations between directly measured variables and the not directly measurable variables believed to cause the variations in data. SEM can also express the results in a visual format which facilitates human interpretation.

For situations like control room environments, classical experimental designs are often less appropriate. The complexity and dynamics of the situation is obvious, and it is often not possible to maintain experimental control without losing realism and dynamics. As a complement to classical experimentation, SEM-analyses make it possible to draw scientific conclusions hard to achieve from classical experimental designs (cf. Svensson & Nählinder, 2014).

3.3. SEM at a glance

As noted, SEM is a quantitative, second generation statistical analysis method that combines the benefits of path analysis, factor analysis, and multiple regression analysis. LISREL and AMOS are two of the more commonly known software packages offering the computational capability of SEM.

SEM is based on correlational statistics, which means that the linear relationships between variables and the common variance between these variables form the basis for the analyses. Like all statistical methods, SEM has several statistical requirements on the dataset (e.g. normal distribution and independent measures) and assuming those requirements are fulfilled SEM offers powerful capabilities for analysing datasets with diverse variables, e.g. different types of measures (e.g. system-generated measures, self-observations, observer measures etc.) and different scales (e.g. ordinal and interval scales).

SEM presents the degree of relationship between variables in terms of explained variance by statistically testing a hypothesized model in a simultaneous analysis of the entire system of variables, to determine the extent to which the covariance or correlation matrix stipulated by the model is consistent with the matrix based on the empirical data. If the statistical goodness of fit between the two compared matrices is adequate, the model is a plausible representation of the relations between variables that the model developer has specified. It is worth noting that a stipulated model shall reduce the complexity of the manifest or measured variables in terms of a few latent variables or factors of high explanatory power.

Basically, SEM tests if the theory, expressed by the model, fits the data. However, it is important to realize that a SEM never can be accepted, it can only avoid being rejected and several alternative models can usually be specified. SEM provides several goodness of fit values, i.e. values of how well a model explains the variance of the dataset, which can then be used to compare alternative models. Basically, the fit-indices announce the proportion of the common variance between the measured variables that can be explained by the stipulated model. In this way SEM as a tool can advance understanding of the measures along with their relations, allowing an empirically supported model of current best fit to be proposed.

SEM requires a level of statistical understanding that is often beyond laymen's understanding. SEM is not a "silver bullet" for analysis of human and organizational factors during an ISV, but represents a powerful statistical analysis method which is useful for analysis of large and heterogeneous datasets, which is often the case with datasets from operational settings and processes.

3.4. Caveats and delimitations

Due to a lack of access to a larger data-set from ISV-evaluations of operator performance in the central control room (CKR), analyses and modelling of relevant concepts have been conducted based on dummy data generated for the purposes of showing the steps of SEM. A specific purpose of the report was also to show examples of how results can be reported. This exemplifying SEM-analysis is found in Appendix C.

The design of the dummy dataset was inspired by how data was collected during a recent ISV process conducted in Sweden, i.e. it contains similar variables and constructs, but is constructed by the report's authors. The authors of the report had no influence on the measures used during the real ISV process and there are some notable deviations between the dataset regarding the individual measures and the distribution of the data. The dummy data is far from a random dataset, and data points have been created with careful attention, but it is dummy data and any analysis of the data, e.g. the estimates (values) in the SEM-models presented in the report, must be seen in this light.

Late in the project, SEM was tried on a dataset from a real ISV process. However, this modelling effort did not result in models with satisfactory fit values, which is summarised in Appendix E.

3.5. Structure of the report

The study and this report are intended to describe SEM on several levels, from summaries of the method's potential to detailed descriptions.

The report initially presents an overview of the purpose and process of an ISV-process (Section 4). The report then describes relevant modelling constructs based on the literature on Human Reliability Analysis (HRA) and specifically Performance Shaping Functions, PSFs (Section 5 and Appendix B). The report then elaborates on the statistical method SEM (Section 6 and Appendix A). The report concludes with a discussion of the value of statistical modelling through SEM (Section 0). More in-depth information, e.g. concerning the SEM-process and PSFs are provided in appendices. A summary of the modelling effort with data from a live ISV process in provided in Appendix E.

To further the understanding of SEM-application, the development of a model, based on a dummy data set, is provided in Appendix C. This appendix contains a dummy example of the method and result sections that typically are reported in experimental studies. The method section describes the measures that the dummy data set was designed to replicate. Rather than just providing a final SEM-model and assuming SEM knowledge of the behalf of the reader, the report provides a worked example of a model development process, based on the dummy-data.

4. Integrated System Validation (ISV)

This section contains a brief description of the ISV-process to provide context for the further analysis. The literature offers numerous alternatives for more detailed descriptions of ISV processes, including recommendations and potential for development (e.g. Boring & Lau, 2017; NRC, 2012a; Simonsen, 2016a; Simonsen, 2016b; Rollenhagen, Bladh, Borg, & Evénius, 1998).

NUREG-0711 (NRC, 2012a) state that the objective of validation is to provide evidence, that the integrated system supports plants personnel in safe operation of the plants, i.e., that the integrated design remains within acceptable performance envelopes. ISV is intended to be a validation of the composed functionality of the control room together with all human operators from a human factors perspective. ISV of the central control room (CKR) is further described as performance-based tests, which attempt to determine whether an integrated system's design (i.e. hardware, software, and personnel elements) meets performance requirements and supports the plant's safe operation. HEDs (Human Error Discrepancies) are identified if performance criteria are not met. Baseline comparisons are often used, with the requirement that the new control room must function at least as well as the old control room.

NUREG-0711 further describes that ISV employs a hierarchical set of performance measures including measures of plant performance, personnel task performance, situation awareness, cognitive workload, and anthropometric/physiological factors. Errors of omission and commission also are identified. The hierarchal set of measures provides sufficient information to validate the integrated system design and affords a basis to evaluate deficiencies in performance and thereby identify needed improvements. Pass/fail measures are those used to determine whether the design is considered to have successfully passed the validation or not. Diagnostic measures are used to better understand personnel performance and to facilitate the analyses of errors and HEDs.

Changes conducted during for example the upgrade of a CKR of a nuclear power plant (NPP) warrants an ISV process to be conducted. Most commonly these are technical system changes, e.g. modernization of a CKR, which results in an ISV process, but potentially changes triggering an ISV process may be of other natures, e.g. organizational changes. Halbert, Park, Boring, and Jung (2016) describes a long range of common human performance issues identified in scientific and operational literature regarding the development of digital control rooms that could be triggered or be detected during an ISV. They summarize their findings in a list of fifteen categories:

- Change in the role/function of human operators
- Cognitive workload
- Confirmation/trust on a digital system
- Crew performance
- Dealing with different information available across different sources
- Decrease of the range of vision (visual momentum)
- Digital environment
- Digital fatigue

- HMI complexity
- Novel human error in a digital system
- Opacity in a digital system
- Physical workload
- Recovery of human error in a digital system
- Situation assessment
- Training

To identify issues such as the ones listed above, the NPP-community often used methods labelled Human Reliability Analysis (HRA) which are briefly described in the next Section.

5. Modelling constructs from Human Reliability Analysis (HRA)

The phenomena or concepts of interest to human factors researchers are often not directly measurable. In statistics, these abstract phenomena have been called latent variables, factors or constructs. Examples of latent variables in psychology are, e.g. different types of intelligence or motivation. The same measurement problem is true for many of the constructs relevant during an ISV. The label construct or latent variable is used from here on in the report.

A clear example and analogy of a latent variable from the physical sciences is provided by Wilson et al. (2004). The temperature can be measured with several different scales such as the Kelvin (K), Réaumur (R), Fahrenheit (F), and Celsius (C) scales. However, the manifest and measurable variation in the scales is a consequence of the amount of excitation of nuclear particles, and it is not the movement of the particles that is observed directly. Thus, temperature can be considered as the hypothetical phenomenon affecting and explaining the variation in the scales, and is thus considered a latent variable which finds manifest expression on the different temperature scales, as illustrated in Figure 1.

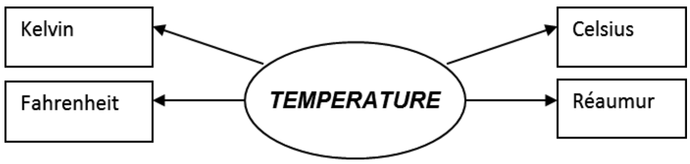


Figure 1. The latent construct Temperature and some manifest measures.

A critical question in any SEM effort is of course which latent variables that should be included in the modeling. Which and how many latent variables that are important for the modelling effort naturally depends of the purpose of the model. For application of SEM to an ISV-process, the authors consulted the scientific HRA-literature to identify latent variables that would be important to assess during an ISV in a nuclear power plant.

Human Reliability Analysis (HRA) should be a part of an ISV, and several methods exist, see e.g. Bell and Holroyd (2009) for a concise review. Gertman, Blackman, Marble, Byers, Haney, and Smith (2005) describe six HRA methods common in the nuclear domain:

- THERP
- HEART
- CREAM
- ASEP
- SHARP
- SPAR-H

These six methods and several other methods found in the HRA literature use the concept of Performance Shaping Functions (PSFs) to describe constructs that are important to consider when assessing the performance of human operators. The differences, pros and cons of these different HRA methods are beyond the scope of the present report, but NUREG/IA-0216 (IAR, 2009) describe twelve HRA methods and compare experiences of their application on a set of common scenarios.

Groth (2009) lists 53 PSF's often used in THERP HRA (i.e. originally from Swain & Guttman, 1983) analyses. Based on these, Groth (2009) has refined and rearranged the 53 PSFs into a 9-bubble model (Figure 2) while Groth and Mosleh (2010a, 2010b) describes a 6-bubble model (Figure 3) to provide a tiered classification of PSFs, as different levels of resolution are suitable for different purposes. Further detailed description of these PSF taxonomies are provided in Appendix B.

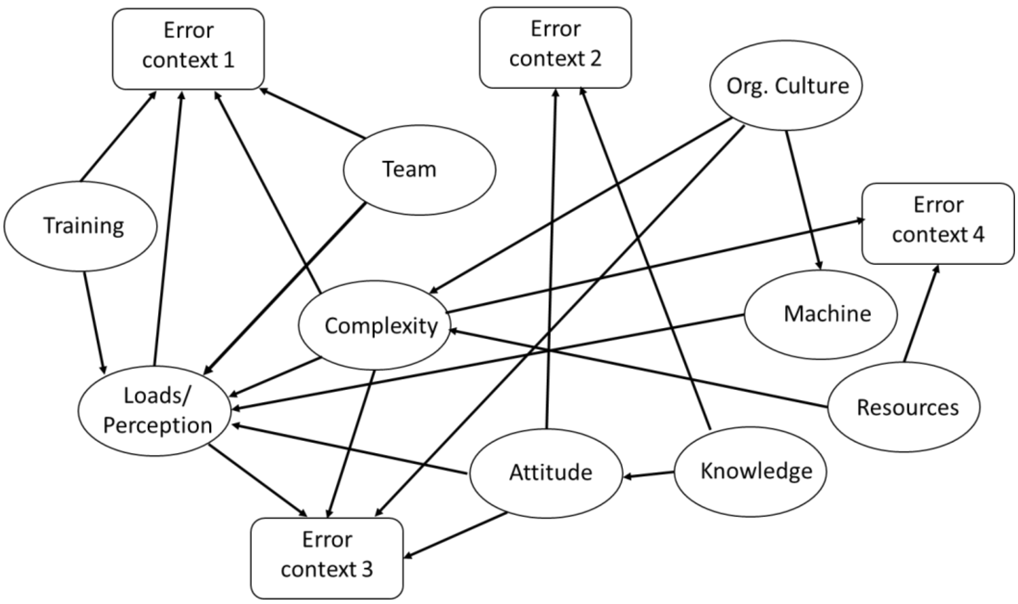


Figure 2. 9-bubble model from Groth (2009).

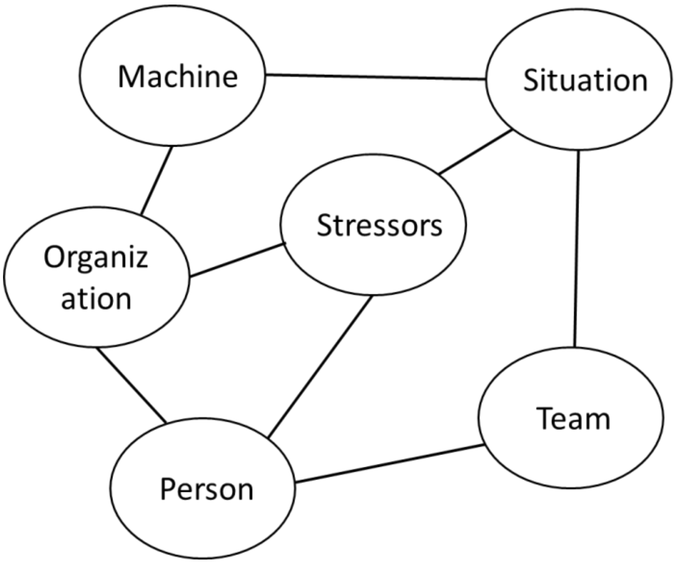


Figure 3. 6-bubble model from Groth and Mosleh (2010a).

Gertman, Blackman, Marble, Byers, Haney, and Smith (2005) and Blackman, Gertman, and Boring (2008) describe the Standardized Plant Analysis of Risk-Human Reliability (SPAR-H) method. SPAR-H uses a number of PSFs derived from the THERP method, while SPAR-H is designed to be a less resource demanding method than THERP. SPAR-H suggests a set of nine PSFs which are presented in Figure 4. The SPAR-H PSFs all have their individual scales described in Blackman et al. (2008). The path diagram in Figure 4 shows the predicted relationships between the PSFs in the model of Blackman et al., where solid lines indicate high degree of relationship and dashed lines indicate medium degree of relationship.

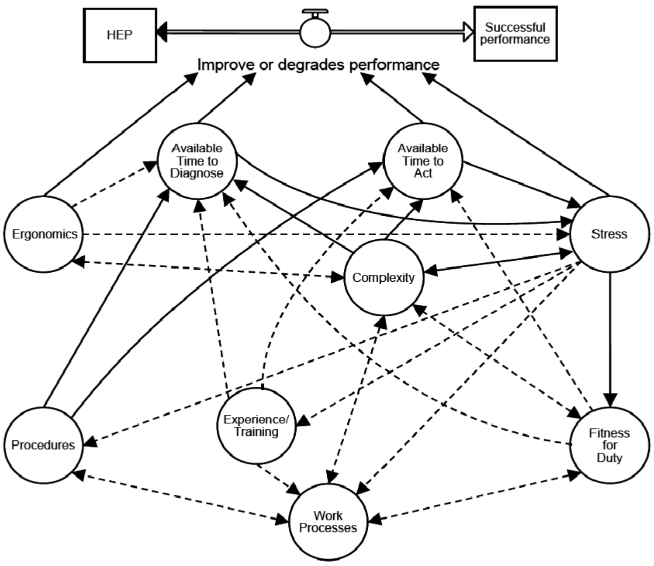


Figure 4. Path diagram showing the relative relationship among SPAR-H PSFs.

Boring (2010) provides a crosswalk, see Table 1, of PSFs used in four HRA methods which have explicit PSF models:

- HRA good practices guide (Kolaczkowski, Forester, Lois, & Cooper, 2005)
- SPAR-H (Gertman, Blackman, Marble, Byers, Haney, & Smith, 2005)
- CREAM (Hollnagel, 2005)
- 9-bubble PSF model (Groth, 2009)

Table 1. *Crosswalk of PSFs by Boring (2010).*

Good practices (Kolaczkowski et al., 2005)	**SPAR-H (Gertman et al., 2005)**	**CREAM (Hollnagel, 2005)**	**9-bubble model (Groth, 2009)**
Training and Experience	Experience/Training	Adequacy of Training and Preparation	Training Knowledge
Procedures and Administrative Controls	Procedures	Availability of Procedures/Plans	Resources
Instrumentation	Ergonomics/HMI	Adequacy of HMI and Operational Support	Machine
Time Available	Available Time	Available Time	Loads/Perceptions
Complexity Workload/Time Pressure/Stress	Complexity Stress/Stressors	Number of Simultaneous Goals	Complexity Loads/Perceptions
Team/Crew Dynamics	Work Processes	Crew Collaboration Quality	Team
Available Staffing	Work Processes	Adequacy of Organization	Resources
Human-System Interface	Ergonomics/HMI	Adequacy of HMI and Operational Support	Machine
Environment	Stress/Stressors	Working Conditions	Complexity
Accessibility/Operability of Equipment	Ergonomics/HMI	Adequacy of HMI and Operational Support	Machine
Need for Special Tools	Ergonomics/HMI	Adequacy of HMI and Operational Support	Resources
Communication	Work Processes	Crew Collaboration Quality	Team
Special (Equipment) Fitness Needs	Ergonomics/HMI	Adequacy of HMI and Operational Support	Resources
Considerations of 'Realistic' Accident Sequence Diversions and Deviations	--	--	--
--	Fitness for Duty	Time of day	--
--	Work Processes	Adequacy of Organization	Organizational Culture
--	--	--	Attitude

NUREG-0711 (NRC, 2012a), in the section on performance measurement, recommends evaluation of the following performance related factors as a part of ISV processes:

- Plant performance
- Personnel task performance
- Situation awareness
- Cognitive workload
- Anthropometric/physiological factors

During recent and current ISVs conducted at Swedish NPPs, a data collection setup developed by the Institute for Energy Technology (IFE, *Nw.* Institutt for Energiteknikk) has been used. The measurement model used in Appendix C is inspired by the measurement setup used in the OKG O2 ISV (Braarud, Eitrheim, & Svengren, 2015), which were developed by IFE, based on NUREG-0711 (NRC, 2012a) and earlier ISV experiences, where the following constructs is assessed:

- Task performance
- Situation Awareness
- Workload
- Teamwork
- Usability

As hopefully evident from the above descriptions of PSFs in different HRA methods, it would be very useful from both operational and scientific perspectives to quantify the relations between identified PSFs. SEM provides a method for quantifying these relations and to present the results in a graphical form.

6. Structural Equation Modelling (SEM)

6.1. Background to SEM

To advance scientific theories and compare results from different studies, many researchers in different fields have had the need of statistical methods that enable them to quantify the relations between different types of not directly observable constructs or latent variables. The researchers often need to: a) to be able to estimate latent variables of interest through multiple directly measurable and manifest variables to get better measurement, b) to be able to accommodate for measurement error, and c) to be able to statistically compare alternative models.

For many human factors related studies it is also a fact that experimental design, data collection and analysis of human work, at least from realistic settings, are resource demanding activities. Different methodological approaches and measurement setups in different studies also make it hard to integrate experimental results from different studies with each other.

SEM is a quantitative statistical method that was developed to manage these types of methodological needs. SEM combines the benefits of path analysis, factor analysis, and multiple regression analysis (Jöreskog & Sörbom, 1984, 1993; Tabachnick & Fidell, 1996). SEM is based on correlational statistics, i.e. the linear relationships between variables, and the common variance between the variables forms the basis for the analyses. SEM identifies the degree of relationship between variables in terms of explained variance. One or more hypothesized models are tested statistically in a simultaneous analysis of the entire system of variables, to determine the extent to which the covariance or correlation matrix stipulated by the model, is consistent with the matrix based on the empirical data. If the statistical goodness of fit between the two compared matrices is adequate, the model developer can argue that the model is a plausible representation of the relations between variables that the model developer has specified.

Due to these methodological possibilities, SEM has been used for many years and is a popular methodology for non-experimental research, where methods for testing theories are not well developed, and ethical or practical considerations make traditional experimental designs unfeasible. Diamantopoulos and Siguaw (2000) provide an excellent introduction to the SEM process and recent introductory reviews to SEM can be found in Byrne (2016) and Blunch (2013).

While most other multivariate procedures essentially are descriptive by nature (e.g. exploratory factor analysis), SEM takes a confirmatory (i.e. hypothesis-testing) approach to data analysis, even though exploratory research questions can be addressed. Whereas traditional multivariate procedures are incapable of assessing measurement errors, SEM provides explicit estimates of these parameters.

Hoyle (1995) describes three main differences between SEM and other approaches. First, SEM requires formal specification of a model to be estimated and tested. It forces the model developer to think carefully about their data and to formulate hypotheses regarding each variable. Second, SEM has the capacity to estimate and test relationships between latent variables. Third, SEM is a more comprehensive and flexible approach to research design and data analysis than any other single statistical model in standard use by social and behavioural scientists. Hoyle also describes SEM as similar to correlation analysis and multiple regression analysis in four specific ways. First, SEM is based on linear statistical models. Second, there are similar requirements, such as independence of observations and multivariate normality. Third, SEM promises no test of causality - it merely tests relations among different variables. Finally, like any other quantitative analysis, post-hoc adjustments to a SEM model require cross-validations.

Causality is a natural and important aspect of modelling, and in a model of explanatory power you can predict the effects of some factors on others. Even if SEM-analyses do not promise a test of causality, the fit or explanatory power of a SEM-model gives support for a causal model. That some manifest measures or variables are measured before others in time can be used to strengthen conclusions of causality in SEM-analyses. Background-variables, measures of information complexity, mental workload and situation awareness are often registered before different aspects of performance. And, most important, good scientific judgement and experience of the domain must be applied (in Swedish called ‘saklogik’ by the LISREL-developer K. G. Jöreskog).

The development of a structural equation model is supported by special software packages. The first software package developed was LISREL (Jöreskog & Sörbom, 1984; 1993) which is an acronym for LInear Structural RELations (*Sw.* Linjära Strukturella Relationer). LISREL was originally developed by the two Swedish professors Karl Gustaf Jöreskog and Dag Sörbom, and one of the earliest references to LISREL methodology is Jöreskog (1973). Since then, SEM-methodology and software has been developed by many researchers and companies. AMOS (SPSS, 2016) and EQS (MVSOFT, 2016) are probably the most widely spread, apart from LISREL. The strongly evolving R community has also developed several SEM packages[1].

6.2. Basic SEM concepts

A structural equation model has several components. One component that is present in all structural equation models is the measurement model, which defines the latent variables through manifest variables. Another important component is the structural model. The structural model tests relationships between several latent variables.

Measurement model

The measurement model is the part of a SEM model which defines relations between the latent variables or constructs and their manifest variables. The manifest variables are often the items/questions of a questionnaire, but can be any type of measured data. To provide a well-rounded measurement of the construct the manifest variables should be chosen or designed so that they assess different aspects of the construct, i.e. the manifest variables should not be too similar. A pure measurement model represents a confirmatory factor analysis (CFA) model in which there is undetermined covariance

[1] R, the platform of open-source statistics packages contains packages named e.g. SEM, LAVA and LAVAAN. Bayesian SEM estimation is implemented in the blavaan package.

between each possible pair of latent variables. The pure measurement model is frequently used as the “null model”, where all covariances in the covariance matrix for the latent variables are all assumed to be zero, i.e. the constructs are totally unrelated to each other. For the proposed structural model, i.e. the part where relations between the constructs are hypothesized, to be investigated further, differences from the null model must be significant.

Structural model
The structural model describes how the researcher has defined the relationships between the latent variables. It consists of a set of latent variables in the model, together with the direct effects connecting them, and the error variance for these variables. Diamantopoulos and Siguaw (2000) state that models with five to six latent variables, each measured by three to four manifest variables can be considered an appropriate upper level of complexity. Many models found in the literature are not as complex and consist of two or three latent variables. Increases in model size typically results in increasing difficulty to meet the recommended thresholds for model fit.

Residual error terms
For the majority of variables that are of interest within the HOF field it is typically challenging to design measures that capture a phenomenon perfectly. Thus, error in measurement is assumed, and with SEM such are addressed by the inclusion of error terms for each variable. Residual error terms reflect the unexplained variance in latent variables.

6.3. Example model

To exemplify, a structural equation model from the military aviation domain Castor (2009) is shown in Figure 5, where ovals are latent variables forming the structural model, while squares are manifest variables forming the measurement model. Error residuals are not shown in this figure.

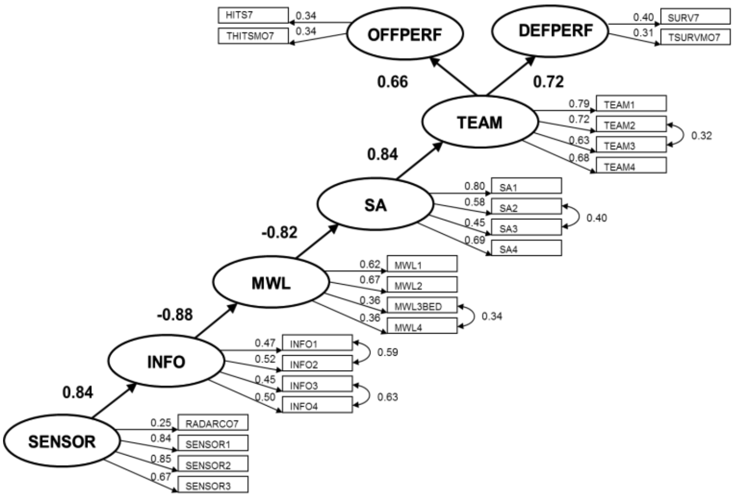


Figure 5. Example model from aviation research (Castor, 2009).

In Castor (2009) the data from 308 simulated fighter aircraft engagements with four pilots and 24 variables measured during each engagement were analysed and compiled into the statistical model shown in Figure 5. The database thus consisted of 1232 cases with 24 variables, generated by 37 pilots. The resulting model summarizes more than 700 hours of experienced pilots' complex behaviour in an operationally valid environment.

The model describes the relations between 24 manifest variables in the measurement model, which are used to describe relations among the seven latent variables. The model demonstrates how changes in Sensor management (SENSOR), explains changes in Usability of Information (INFO), Mental workload (MWL), Situation Awareness (SA), Teamwork (TEAM), Offensive Performance (OFFPERF) and Defensive Performance (DEFPERF). In other words, if SENSOR is high, then INFO is high, and then MWL is low (which is desirable). If MWL is low then SA is high, which relates to high TEAM, which in turn is used to describe changes in OFFPERF and DEFPERF. Instead of showing a correlation matrix of the 24 manifest variables against each other, a SEM analysis was conducted resulting in the model in Figure 5 which is an adequate and more succinct way to show important relations between the variables of the database.

As demonstrated by Figure 5, SEM forces an analyst or researcher to be very explicit concerning the relations between all chosen latent and manifest variables, e.g. the omission of an arrow in the model is a quite strong theoretical statement. For example, in Figure 5 there is no direct effect between SENSOR and OFFPERF/DEFPERF. These effects are instead mediated by the other latent variables. This implies that changes in SENSOR, e.g. through increased technical performance of an aircraft radar, are mediated by the pilot and the team, (INFO-TEAM) before there are effects in the tactical outcome (number of enemies shot down and the own team's survival, i.e. the manifest variables of OFFPERF/DEFPERF). The estimates between the latent variables warrant them all as being useful as separate concepts and show that there is explanatory power between them. Too high estimates would imply that they are the same thing and too low estimates would show that concepts not really, at least not directly, are statistically related.

Experiences from SEM-modelling in the aviation domain, such as Castor (2009) and Nählinder, Berggren, and Svensson (2004), were used as starting point for the analysis reported in Appendix C. One resulting model from this modelling effort, based on the dummy data replicating an ISV is presented in Figure 6. This should thus be seen as an example of potential outcome of SEM modelling during an ISV, i.e. a model that can be analysed to further understand the work processes it was designed to assess. For further explanation of this model, see Appendix C.

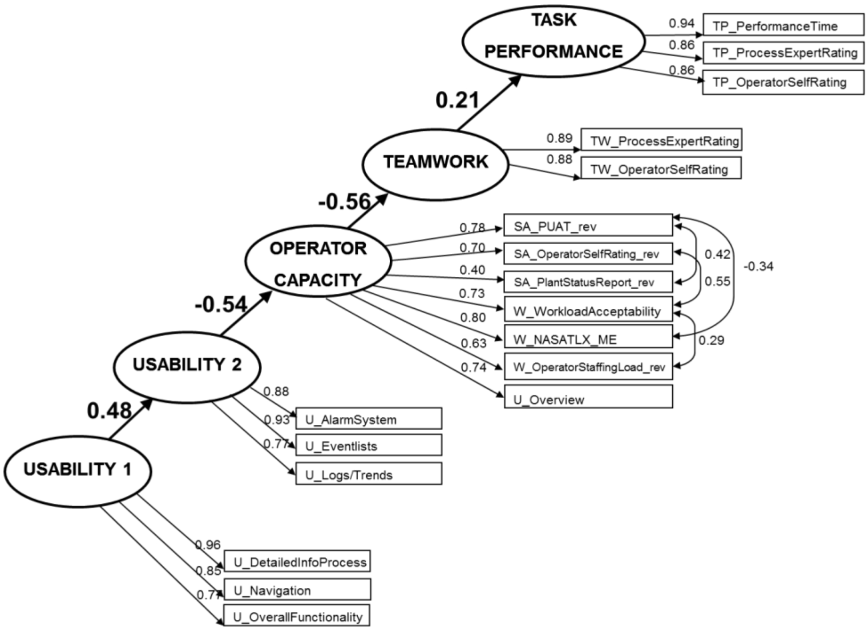


Figure 6. Example model based on ISV dummy data.

6.4. SEM development process

The SEM development process is quite complicated, with many steps and requirements concerning e.g. the data distribution, model complexity versus amount of data, model fit indices, and so forth. This process and its requirements are further elaborated in Appendix A.

7. Discussion

For many studies of human behaviour in applied settings, SEM should not be seen as a replacement to more classical statistical analyses, e.g. variance analysis through ANOVA or presentation of descriptive data. Rather, SEM provides a powerful complementary analysis capability, which provides advanced understanding of the process where data was collected. For some studies and research questions, SEM can provide one of the few or the only statistical analysis method(s) that can be applied.

Repeated testing and analysis of the manifest variables and their properties are the foundation for useful and valid SEM models. The finally proposed model from one data collection campaign, e.g. a model from one ISV process, can provide important understanding of how variables, in a data-driven view, relate to each other. If the same patterns are seen in other studies and a similar structural model repeatedly can be found, e.g. as reported by Nählinder, Berggren and Svensson (2004), the theoretical importance of the SEM model grows over time.

A recent empirical study using SEM-analyses concerns performance of a command and control centre of the Swedish Armed Forces. The study analysed models of mental workload, individual and team performance, and relates highly to performance and function of nuclear power plant central control rooms. Among other things of practical value, it was found that mental workload has a predictive power on individual and team performance up to 30 minutes ahead (Svensson, Rencrantz, Marklund, & Berggren, 2014).

SEM typically supports the analyst in his or her understanding of selected manifest and latent variables. Examples of conclusions/findings from the analyses presented in Appendix C, i.e. analysis of Figure 7, are presented below to show example results. However, remember that the database was designed to demonstrate these findings.

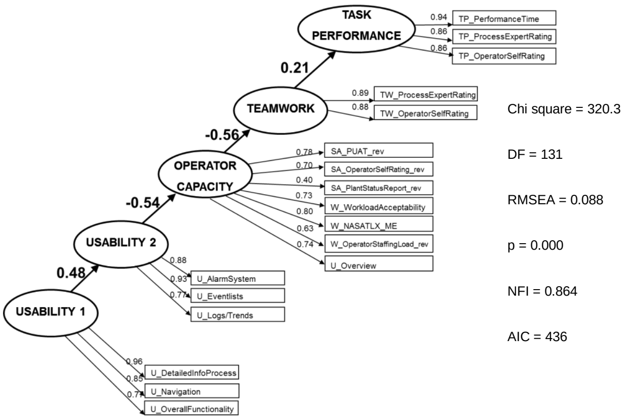


Figure 7. Model based on dummy data before specification search.

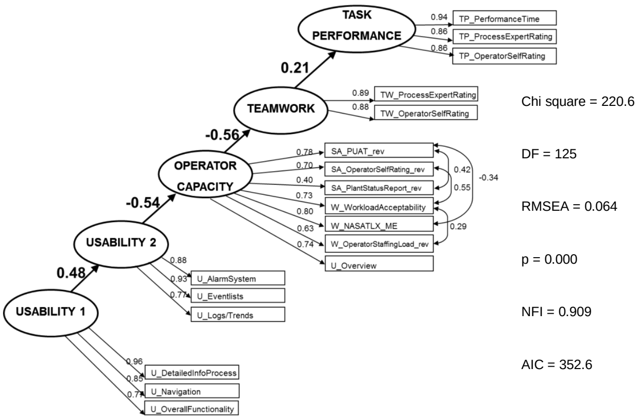


Figure 8. Model based on dummy data after specification search.

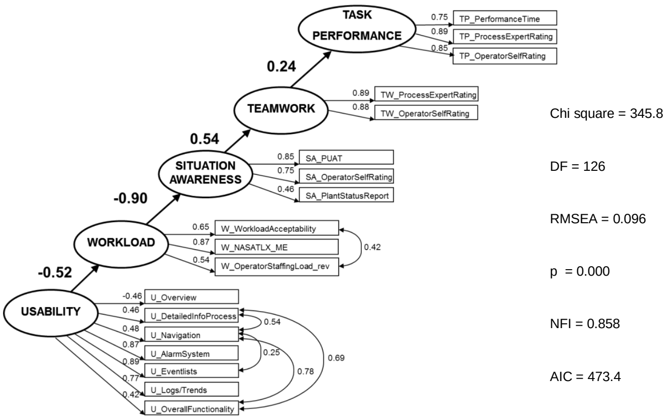


Figure 9. Model from dummy database, testing the origina

Some changes (or potential for changes) that the model development in Appendix C led to are summarized below:

- Changes to the measurement model, e.g. the EFA showed that U_Overview manifest variable should load on the SA/Workload latent variable.
- Changes to the structural model, e.g. the new construct of SA/Workload, called Operator. The original hypothesis for the analysis, see Section X, considered Situation Awareness and Workload to be two separate constructs, based on earlier research. However, in this analysis (based on dummy data), data suggested that the two constructs should be regarded as the same construct.
- Management of items with low loadings. This was not shown in examples in Appendix C, but could be applied to e.g. SA_PlantStatusReport in the analysis of the dummy database. If the variable does not load strongly anywhere, the variable either measures something else, which not is covered by the selected latent variables, or has other measurement issues.
- Management of variables with low variance and non-normal distributions. The modelling and the preparations for modelling put "an analytical

spotlight” on variables that have low variance or non-normal distributions. The usefulness of a variable with low variance must be discussed, but the potential decision to exclude it must be made on a case by case basis, depending on what the variable measures.

- Manage co-variances between the manifest variables, as exemplified by comparing Figure 7 and Figure 8. The arrows between the error variables estimate the co-variance between manifest variables, and ca to some extent be tolerated as the manifest variables can be assumed to co-vary in this study. Through inclusion of these co-variances the model fit of the proposed model barely reached common acceptable criteria, c.f. recommendations on model fit indices in Appendix A.
- Comparison of models, e.g. if the models in Figure 8 and Figure 9 are compared with regard to fit indices, the model in Figure 8 is at better model of the data than the model of Figure 9. Theoretical justification might lead a modeller to retain worse fitting models, but the goal is to find the model which both has theoretical justification, explanatory value and acceptable fit indices.

The AMOS software that was used in the reported analysis provides a number of additional analysis capabilities that not are described in the report. Further details of these capabilities are provided by Byrne (2016).

- Bayesian SEM analyses, which is useful when the analysts, in general or for the specific analysis, cannot not accept the assumption that ordinal data, generated by Likert-type questions, e.g. rating on a 1-5 scale concerning some relevant aspect, can be treated as data with continuous scale properties data.
- Multigroup analyses, which are useful when different groups of data samples need to be compared, e.g. do the SEM look different for data from different ISV’s or for different control room operator roles.
- Bootstrapping, useful when analysing data with non-normal distributions.

As stated above, SEM is a powerful statistical method, but there are several issues that must be understood and accepted by the reader to appreciate SEM.

Multidimensional constructs

The latent variables are often multidimensional and hard to exactly define, which is one of the reasons to analyse them through SEM. Some of the latent variables described in Appendix C have been debated in the scientific literature, (c.f. the discussion concerning the validity of the situation awareness construct). However, they all represent very useful constructs that instructors and designers use, and the terms are also used in the regulatory documents.

Depending on the reason for the ISV, it might be necessary to study other latent variables than the ones used in the example modelling of this report. The latent variables chosen for the OKG O2 ISV (Braruud et al., 2015) would often be important, but for a ISV triggered by procedural changes or training, some new latent and manifest variables might need to be identified. The PSF taxonomies described in Section 5 and Appendix B provide a good guide to latent variables other researchers have found to be important. Boring's (2010) crosswalk of PSF build on a large amount of research from different perspectives and could be used as a starting point.

Subjective ratings

Subjective ratings or answers to surveys, questionnaires and interview questions often provide very valuable data concerning psychological phenomena. However, they must be carefully designed and tested to ensure validity and reliability.

Sources of variance

For the modelling effort presented in this report several assumptions concerning the source of variance in the data have been made.

The number of domain experts, e.g. nuclear power plant operators, that practically can be assessed is generally too low for multivariate statistical analysis, even if the whole national population would be part of an evaluation, which, on the other hand, reduces the statistical problem of generalizability from sample to population.

Consider the case of the dummy database that was developed for the purposes of this report. The database was inspired by typical ISV data (c.f. Braruud et al., 2015) from control room operators, but to be statistically strict the data of the 189 cases (i.e. rows) in the database should be independent from the other cases. Given the available number of operators and the fact that the real work tasks largely are team tasks, this strict independence requirement will be practically unachievable. The database, from which the model in Appendix C was based, was generated by 21 operators distributed over seven team across nine scenarios. So, from a strict statistical point of view, the 189 cases do not represent independent measurements.

For assessments of operational performance in realistic domains, the variance that exists in the database can be assumed to be a result of interindividual (difference in rating patterns for each participant), intra-individual (difference in rating patterns between participants), and situational variance (changes in rating due to events in scenarios). If the assumption that the events in the scenarios contributes with more variance than the participants rating patterns can be accepted, and by combining inter-intra-, and situational variance sources in repeated measurement studies, databases and models of practical and theoretical importance can be developed, especially if similar structural models repeatedly can be demonstrated.

The differences between how variables vary across the different scenarios and operator types should be analysed before they are combined into one database. Castor (2009) provides some examples of how differences can be analysed to justify whether they can be combined into the same database (which briefly relate to comparing the correlation matrices between different subsets of the data, e.g. the three types of operators or scenarios). Byrne (2016) also demonstrates different ways to compare different datasets, e.g. through latent mean analysis.

Causality in the model

As stated earlier, SEM does not in itself provide clues concerning causality between variables. From a SEM perspective, the effect (e.g. the arrow from Usability to Workload) could just as well be in the other direction. Hypotheses about causality and the direction of arrows representing regression weights in a model are typically based on earlier observations and understanding of the domain. More insights concerning defending the causal interpretation of structural equations can be found for example in Bollen and Pearl (2013), and Pearl (2009). It is also important to realise that the presence or absence of specific arrows in a model are rather strong statements. For example, if the is no direct arrow between two latent variables, the modeller explicitly states that there is no direct relation between these variables, and that any effect are mediated by other latent variables.

It also important to realize that SEM models are not process models, but models of how the relations between variables in the collected dataset can be described.

Properties of data

Proper application of SEM requires the data to display multi-variate normality. If severe deviations from non-normality are observed in the dataset the analysis is at risk, but there are also means to manage the non-normality, e.g. through the bootstrapping functions of AMOS (e.g. Byrne, 2016).

One of many models

The fact that SEM can be used to express relations between variables collected during an ISV is not surprising, given that SEM is a multivariate statistical method designed to describe relations between many variables.

Appendix C provides an example model, based on the dummy data set that was created by the authors. Any conclusions regarding the estimates between the latent variables in Figure 8 or Figure 9 must be avoided as it is based on dummy data, but the figures exemplify results that can be expected after a SEM modelling process.

The potential to draw operationally relevant conclusions from a SEM model is based on the design of the measurement model. If the measurement model does not contain any criteria variables, e.g. measures of production or safety, the model can of course not be used to draw conclusion concerning this, however it may still be a very useful descriptive model of how included variables relate to each other.

SEM provides a powerful analysis tool that enables theories that are more specific than “everything is connected” as it quantifies relations and shows abstracted relations. It is also very important to realize that the essence of SEM is modelling, implying a simplification of reality.

Models, as almost anything else, can be described on different levels of abstraction, and to exemplify, Figure 10 show three models of a snowflake on different levels of abstraction. All three of the models in the figure, to different degrees, capture essential properties of a snowflake, even though every real snowflake is said to be unique. There are patterns which clearly identify a snowflake and these patterns are found in every snowflake, i.e. it is possible to design models of what snowflakes look like.

Figure 10. Three models of a snowflake, described on three levels of abstraction.

Expanding on the visual analogy shown in Figure 10, and considering the ISV process and the need to model work process in a central control room, with all technical systems, operators capabilities and processes, Figure 11 is another visual analogy of the challenges of statistical modelling, i.e. what resolution is needed in order to recognize the important characteristics of the process.

Figure 11. Visual analogy of the challenge of selecting the appropriate level of resolution in a model.

Useful models capture the essential properties of a system or process and facilitate insights of their nature. Thereby models can be used as predictive tools, providing a foundation for important decisions. Regardless of simplicity, the model still needs to contain the essential information to be useful. As conceptualized in Figure 10 and Figure 11, the search for the "one and only" model or level of representation is a dead end, and the abstraction level of choice instead depends upon the purpose of the model. A model can, as shown, be described on different levels of abstraction, and any model will face challenges regardless of level of abstraction. The model can be challenged because it fails to provide an idealization about the structure of the system, which approximates the actual behaviour of the system good enough, or that it buries the important processes in a mass of "irrelevant" detail.

For a modeller, it is a trade-off between maximizing explanatory power without oversimplifying. It is the modeller who defines the frame of the model and chooses which variables to include, based on experience, previous scientific findings, theory, and model purpose.

8. References

Bell, J., & Holroyd, J. (2009). Review of human reliability assessment methods. HSE RR679. Buxton, Health and Safety Laboratory.

Bollen, K.A., & Pearl, J. (2013). Eight Myths About Causality and Structural Equation Models. In S.L. Morgan (Ed.), Handbook of Causal Analysis for Social Research, Chapter 15, 301-328, Springer. Available at: http://ftp.cs.ucla.edu/pub/stat_ser/r393-reprint.pdf

Boring, R. (2010). How Many Performance Shaping Factors are Necessary for Human Reliability Analysis. Proceedings of 10th International Conference on Probabilistic Safety Assessment and Management PSAM 2010.

Boring, R., & Lau, N. (2017). Measurement Sufficiency Versus Completeness: Integrating Safety Cases into Verification and Validation in Nuclear Control Room Modernization. In S. Cetiner, P. Fechtelkotter, & M. Legatt, (Eds.) Proceedings of AHFE 2016 - Advances in Human Factors in Energy: Oil, Gas, Nuclear and Electric Power Industries, pp.79-90.

Blunch, N.J. (2013). Introduction to Structural Equation Modelling using IBM SPSS Statistics and AMOS. Los Angeles, CA: SAGE.

Braarud, P.Ø., & Berntsson, O. (2016). Assessment of Situation Understanding Mission, Control and Teamwork in the Control Room: The Development and Initial Testing of the SCORE Measure. EHPG MTO-6.7 HWR-1125. Proceedings of EHPG 2016.

Braarud, P.Ø., Eitrheim, M., & Svengren, H. (2015). Results from the Integrated System Validation of the Modernized OKG Oskarshamn 2 Control Room. IFE/HRF/F-20/1632. Halden, Institute for Energy Technology.

Byrne, B. (2016). Structural Equation Modeling with AMOS 3rd Edition. New York, NY; Routledge.

Collier, S., & Follesø, K. (1995). SACRI: A measure of situation awareness for nuclear power plant control rooms", Proceedings of an International Conference: Experimental Analysis and Measurement of Situation Awareness, pp. 115-122, 1995-November-1-3.

Diamantopoulos, A., & Siguaw, J. (2000). Introducing LISREL. London, Sage Publications.

Garson, G. D. (2015). Structural Equation Modeling. Asheboro, NC: Statistical Associates Publishers.

Gertman, D., Blackman, H., Marble, J., Byers, J., & Smith, C. (2005). The SPAR-H Human Reliability Analysis method. NUREG/CR-6883. Washington, DC: United States Nuclear Regulatory Commission.

Groth, K.M. (2009). A Data-Informed Model of Performance Shaping Factors for use in Human Reliability Analysis. PhD thesis. College Park, MD: University of Maryland. Available at: http://drum.lib.umd.edu/bitstream/handle/ 1903/9975/ Groth_umd_0117E_10944.pdf;sequence=1

Groth, K.M., & Mosleh, A. (2010a). A data-informed model of performance shaping factors and their interdependencies for use in human reliability analysis. In Briš, Guedes Soares & Martorell (Eds.) Reliability, Risk and Safety: Theory and Applications. London, Taylor & Francis Group. Available at: http://www.sandia.gov/~kgroth/publicationfiles/Groth_ESREL2009.pdf

Groth, K.M., & Mosleh, A. (2010b). A Performance Shaping Factors Causal Model for Nuclear Power Plant Human Reliability Analysis. Proceedings of 10th International Conference on Probabilistic Safety Assessment and Management PSAM 2010.

Halbert, B., Park, J., Boring, R, & Jung, W. (2016). Methods and measures for characterizing nuclear power plant operator performance to support control room modernization. Proceedings of 13th International Conference on Probabilistic Safety Assessment and Management PSAM 2016.

Hart, S.G., & Staveland, L. (1988). Development of the NASA task load index (TLX): Results of empirical and theoretical research. In P.A. Hancock and N. Meshkati (Eds.), Human Mental Workload, pp 139-183. Amsterdam, Elsevier.

Hollnagel, E. (1998). Cognitive reliability and error analysis method. Oxford, Elsevier.

IAR (2009). International HRA Empirical Study – Phase 1 Report. NUREG/IA-0216 Vol 1.

Jöreskog, K.G., & Sörbom, D. (1984). Analysis of linear structural relationships by maximum likelyhood, instrumental variables, and least squares methods. Uppsala, University of Uppsala.

Jöreskog, K.G., & Sörbom, D. (1993). LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language. Chicago, IL: Scientific Software International.

Kline, R. B. (2011). Principles and practice of structural equation modeling. New York, NY: Guilford.

Kolaczkowski, A., Forester, J., Lois, E., & Cooper, S. (2005). Good practices for implementing human reliability analysis. NUREG-1792. Washington, DC; US Nuclear Regulatory Commission.

NRC (2012a). Human Factors Engineering Program Review Model. NUREG-0711, Rev. 3. Washington, DC: US Nuclear Regulatory Commission. Available at: http://www.nrc.gov/docs/ML1232/ML12324A013.pdf

NRC (2012b). The International HRA Empirical Study - Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. NUREG-2127. Washington, DC: US Nuclear Regulatory Commission.

Nählinder, S., Berggren, P., & Svensson, E. (2004). Re-occurring LISREL patterns describing mental workload, situation awareness and performance. Proceedings of the HFES 48th Annual meeting. New Orleans, LA: Human Factors and Ergonomics Society.

Pearl, J. (2009). Causality – Models, reasoning, inference 2nd Ed. Cambridge, Cambridge University Press.

Rollenhagen, C., Bladh, K., Borg, A. & Evénius, P. (1998). Handbok – Validering av kontrollrumsändringar.

Simonsen, E. (2016a). Measures and method characteristics for early evaluation of safe operation in nuclear power plant control room systems. Licentiate thesis no 99-2016. Göteborg, Chalmers University of Technology. Available at: http://publications.lib.chalmers.se/records/fulltext/232267/232267.pdf

Simonsen, E. (2016b). Measures and method characteristics for early evaluation of safe operation in nuclear power plant control room systems. SSM report 2016:31. Stockholm, Swedish Radiation Safety Authority. Available at: http://www.stralsakerhetsmyndigheten.se/Global/Publikationer/Rapport/Sakerhet-vid-karnkraftverken/2016/SSM-Rapport-2016-31.pdf

Strand, S., & Svengren, H. (in progress). The Process Understanding Assessment Technique (PUAT). HWR-907. Halden, OECD Halden Reactor Project.

Swain, A.D., & Guttman, H.E. (1983). Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications. NUREG/CR- 1278. Albuquerque, NM: Sandia National Laboratories.

Svensson, E., & Nählinder, S. (2014). Dynamic measures of effectiveness in command and control. In P. Berggren, S. Nählinder & E. Svensson (Eds.), Assessing Command and Control Effectiveness: Dealing with a changing world, pp 49-70. Farnham: Ashgate Publishing Limited.

Svensson, E., Rencrantz, C., Marklund, J., & Berggren, P. (2014). Empirical studies of command and control centres at the Swedish Airforce. In P. Berggren, S. Nählinder & E. Svensson (Eds.), Assessing Command and Control Effectiveness: Dealing with a changing world, pp 103-125. Farnham: Ashgate Publishing Limited.

Wilson, G.F., Frazer, W., Beamont, M., Grandt, M., Gundel, A., Varoneckas, G., Veltman, H., Svensson, E., Burov, A., Hockey, B., Edgar, G., Stone, H., Balkin, T., Gilliland, K., Schlegel, R.E. & van Orden, K. (2004). Operator Functional State Assessment (RTO-TR-HFM-104). Paris: NATO Research and Technology Organisation. Available at: handle.dtic.mil/100.2/ADA422195

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models an evaluation of power, bias, and solution propriety. Educational and Psychological Measurement, 73(6), 913-934.

9. Appendices

9.1. Appendix A: Structural equation modelling process

Structural equation modelling (SEM) is almost a research field of its own, and therefore only a brief introduction to the model development process is provided here. Introductory texts concerning the SEM development process accessible for non-experts, are, for example, provided in Diamantopoulos and Siguaw (2000) and Byrne (2016).

Jöreskog (1993) distinguishes between three use scenarios of SEM: Strictly Confirmatory, Alternative Models, and Model Generating. In the *Strictly Confirmatory scenario* the researcher formulates a single model based on theory, collects the appropriate data, and then test the fit of the model to the collected data. The researcher does no modifications to the model and either accepts or rejects the model. However, as other unexamined or nested models may fit the data as well or better, an accepted model is only a model that has not been rejected.

In the *Alternative Models scenario*, the researcher proposes several alternative competing theory-driven models. Based on the analysis of the collected data, the most appropriate model is chosen. Although this approach is desirable in principle, a problem is that in many specific research topic areas, the researcher does not find two or more well-developed alternative models to test.

In the *Model Generating scenario*, the researcher proceeds in a more exploratory fashion, often after first having had to reject an initial model after assessment of its poor fit. Jöreskog (1993) notes that although re-specification may be either driven by theory or data, the goal is to find a model that is both theoretically meaningful and with strong statistical fit. The problem with the model development approach is that models developed in this way are post-hoc models, which may not be stable and may not fit new datasets. Using a cross-validation strategy, where the initial model is developed using one data sample and then tested statistically for another independent sample, some of this concern can be addressed. The models presented in this report most closely matches the Model Generating scenario.

Regardless of which of these three approaches that have been chosen, SEM does not in itself provide clues concerning causality in a model, i.e. in what directions the effects go (and specifically in the modelling software, in which directions the arrows point). The causality must be justified by theory and the scientific judgment by the modeler.

In a description of the SEM development process, Jöreskog and Sörbom (1993) describe the validation of the measurement model and the fitting of the structural model as the two main steps. The validation of the measurement model is accomplished primarily through confirmatory factor analysis, while the fitting of the structural model is accomplished primarily through path analysis with latent variables. The model that is being developed is specified based on available theory. Constructs are chosen and operationalized by multiple manifest variables and tested through confirmatory factor analysis to establish that indicators seem to measure the corresponding constructs. The researcher proceeds to development of the structural model only when the measurement model has been validated. Two or more alternative models (one of which may be the null model) are then compared in terms of model fit, which measures the extent to which the covariance predicted by the model correspond to the observed covariance in the data. Modification indexes, suggested by the analysis software, may be used by the researcher to alter one or more model specifications to improve fit, but only if supported by theory.

Hence, a solid theoretical foundation is thus needed before a structural equation model is developed, as theory warns us of potential problems such as, for example, excluded variables. Theoretical support is also necessary to distinguish between statistically equivalent models. Good definitions are also helpful when identifying appropriate manifest variables/measures.

In another step by step description of the SEM development process Diamantopoulos and Siguaw (2000) describes eight relatively distinct but related steps that a researcher goes through when developing a structural equation model:

1. Model conceptualization
2. Path diagram construction
3. Model specification
4. Model identification
5. Parameter estimation
6. Assessment of model fit
7. Model modification
8. Model cross validation

Brief descriptions of the basic outline and considerations of each of Diamantopoulos and Siguaw's steps will be provided below.

Model conceptualization
In this initial step, the researcher defines his or her conceptual model, which translates theoretical assumptions into a conceptual framework. This conceptual model needs to be identified based on existing literature and theory. In this step, the researcher decides which latent variables or constructs that will need to be included, and how they are to be operationalized through manifest variables. During this stage, it is crucial to make every effort to include any important factors that can affect the variables that are included in the model. An omission of important factors represents a specification error and the result can be that the proposed model in the end does not represent the "whole" truth. The structural model consists of a set of exogenous and endogenous latent variables in the model, together with the direct effects connecting them, and the error variance for these variables. The error variance reflects the effects of unmeasured variables and error in measurement. The exogenous latent variables are those that are conceptualized as to cause variance in the values of other latent variables in the model. Changes in the values of exogenous variables are not explained by the model and they are considered to be influenced by factors external to the model. Endogenous latent variables, those that are influenced by the exogenous variables in the model, either directly, or indirectly affect each other. Variance in the values of

endogenous variables is explained by the model because all latent variables that influence them are included in the model specification.

Successful development of a structural equation model is to a large extent based on a sound model conceptualization. It is rare that a SEM development process that does not start from well-established theoretical concepts and tested manifest measures result in a useful model.

Path diagram construction
In this second step of the modelling process the model developer can describe his or her model graphically as a path diagram. This is not a mandatory step, but it is helpful to make the model more explicit for the model developer.

Model specification
The third step is model specification, where the researcher specifies which effects that are null, which are fixed to some constant and which ones that vary through the specification of a syntax file for the analysis software. The researcher now needs to be very explicit on which variables that will be included and how they shall relate. The specification of a syntax file can be either through a text or a graphical format.

Effects are represented by an arrow in a path diagram, while null effects result in the absence of an arrow. Note that the existence or absence of an arrow represents a rather strong theoretical assumption. A model where no effect is constrained to zero will always fit the data, and the closer one is to this most complex model, the better the fit of the model to the data. Thus, for a model where many effects are included in the specification, the fit indices reported are better, but the model is also more complex and harder to grasp for the researcher.

Model identification
The fourth step in the process is model identification, which is performed by the analysis program, e.g. AMOS. In this step, the empirical data is investigated to see whether there is enough information in the data to do the parameter estimation that is performed in the next step, i.e. that a unique value can be identified for each parameter in the model. If there is a lack of information, i.e. the number of parameters estimated is less than the number of variances and covariances, the model becomes under-identified and the analysis is cancelled. The model can also become just-identified or over-identified. If the number of parameters estimated are greater than the number of variances and covariances, the model is over-identified.

To exemplify what is done during the model identification the following simple example can be used: Is there enough information to uniquely identify the values of A and B in the equation A * B = 100? The answer is no, as there are several different possible solutions and this would equal to when a model is unidentified. However, if A is fixed to 10 you know that B must be 10, and the equation can be identified.

Parameter estimation
If the model can be identified, the parameter estimation step can be executed. During the parameter estimation, the analysis software creates a covariance matrix based on the specified model. If there is no relation between two variables specified during the model specification the covariance is set to zero. The covariance matrix that is proposed by the model is then compared to the matrix produced by the data.

The selection of method of estimation is also an important component of the model specification. Several methods of estimation can be used and ordinarily one will get similar estimates by any of the methods (Garson, 2015). Maximum Likelihood estimation is by far the most common method and Garson (2015) recommends that it is used, unless the researcher has good reason or counterarguments. Unlike some of the other estimation methods, Maximum Likelihood does not assume uncorrelated error terms. Key assumptions are large samples, manifest variables with multivariate normal distribution, valid specification of the model, and manifest variables on an interval or ratio scale, although ordinal variables are widely used in practice. If ordinal data are used, they should have at least five categories and not be strongly skewed.

Assessment of model fit

Once a model converges and parameter estimates are presented, the question is to what extent the empirical data fit the proposed model. In other words, how well the correlation or covariance matrix produced by the data matches the matrix that is implied by the model. Assessment of model fit is one of the more complex tasks of a SEM analysis. Model fit is related to data, model, and estimation methodology and a plethora of fit indices has been developed over the years.

Jaccard and Wan (1996) describe three classes of fit indices (absolute, parsimonious, and relative) that should be considered when evaluating the fit of a structural equation model. Absolute fit compares the predicted and observed covariance matrices. The chi-square ($\chi 2$), goodness of fit index (GFI), and standardized root mean square residual (Standardized RMR) are indicators of absolute fit.

Large values of chi-square reflect a discrepancy between the observed and predicted matrices. The chi-square is reported with the number of degrees of freedom associated with the model, and a significance test. The degrees of freedom are a function of the number of covariances provided and the number of paths specified and a statistically significant model suggests that the specified paths do not provide a perfect fit to the data. Hence a non-significant value ($p > 0.05$) is desired, but Hair et al. (1995) note that the chi-square is sensitive to sample size and that it is rare to find a non-significant value when sample size is over 500 cases.

The GFI is a function of the absolute discrepancies between the observed and predicted covariance matrices. The recommended threshold for the GFI is 0.90. GFI is sensitive to sample size.

The Root Mean Square Residuals (RMR) are the coefficients which result from taking the square root of the mean of the squared residuals, which are the amounts by which the sample variances and covariances differ from the corresponding estimated variances and covariances. The Standardized RMR (S RMR) is the average difference between the predicted and observed variances and covariances in the model, based on standardized residuals. The recommended threshold for the standardized RMR is 0.05.

The second category also considers absolute fit, but penalizes model complexity. The more paths specified, the lower the models' parsimony. The Root Mean Square Error of Approximation (RMSEA) is the common choice for measure of parsimony. The RMSEA fit index values approaching zero are desired. Many recommendations state that it should be less than 0.05 to represent a good model fit, but for example Bollen

(1989), and Browne and Cudeck (1993), state that a value of 0.08 or less could be considered acceptable. RMSEA is sensitive to sample size.

The third category of fit scales compares the absolute fit to an alternative model. The relative goodness of fit measures compares the evaluated model to the fit of another model. When none is specified, the analysis software packages usually default to comparing the model with the independence model, or even allow this as the only option. The Comparative Fit Index (CFI) is a commonly used fit index and Byrne suggest that the CFI should be a fit statistic of choice. The value for the CFI indicates the fit of the model compared to the null model and the recommended threshold is 0.90.

Numerous measures based on information theory have also been developed. These measures are appropriate when comparing models which have been estimated using maximum likelihood estimation. They do not have thresholds, like 0.90, and rather they are used when comparing models, with a lower value representing a better fit. AIC is the Akaike Information Criterion and is a goodness-of-fit measure which, adjusts model chi-square to penalize for model complexity. CAIC is the Consistent AIC, which penalizes for sample size as well as model complexity.

Most important when considering different fit indices, and expressed by Byrne (2016) is that model adequacy should be based on theoretical, statistical as well as practical considerations. Thus, the causal logic and good judgment of the model developer can never be underestimated. This has also been emphasized from the beginning by Jöreskog and Sörbom (1984; 1993).

Model modification

When a model has been evaluated with respect to its fit, the modeler can decide whether the model is acceptable or that it needs to be modified to better fit the empirical data. The SEM software packages presents suggestions for model improvement, so called modification indices. These modifications are entirely data driven and careful deliberation and theoretical support must substantiate any changes to the model based of the modification indices.

Model cross-validation

The last step of the modelling process is to conduct cross-validation of the proposed model against a new dataset, or a part of the dataset that have been kept aside for cross-validation purposes. This step is extra important if major changes have been made to the model as a result of the model modification phase. This has not been conducted for the current “dummy data” model reported in this report.

Guidelines for model development

There are many issues to consider when developing a model, which hopefully is evident from the above description of the development process. Thompson (2000) has suggested the following ten guidelines when developing and reporting structural equation models:

1. Do not conclude that a model is the only model to fit the data.
2. Test re-specified models with split-halves data or new data.
3. Test multiple rival models.
4. Use a two-step approach of testing the measurement model first, then the structural model.
5. Evaluate models by theory as well as statistical fit.

6. Report multiple fit indices.
7. Show that you meet the assumption of multivariate normality.
8. Seek parsimonious models.
9. Consider the level of measurement and distribution of variables in the model.
10. Do not use small samples.

Mulaik and Millsap (2000) suggested a stringent four-step approach to modelling:

1. Exploratory factor analysis to estimate the number of latent variables or factors.
2. Confirmatory factor analysis to confirm the measurement model. As a further refinement, factor loadings can be constrained to 0 for any measured variable's cross loadings on other latent variables, so every measured variable loads only on one latent variable.
3. Test the structural model.
4. Test nested models to get the most parsimonious one. Alternatively, test other researchers' findings or theory by constraining parameters as they suggest should be the case.

9.2. Appendix B: Performance shaping functions (PSF) taxonomies

Table 2. *THERP PSFs (Swain & Guttman, 1983).*

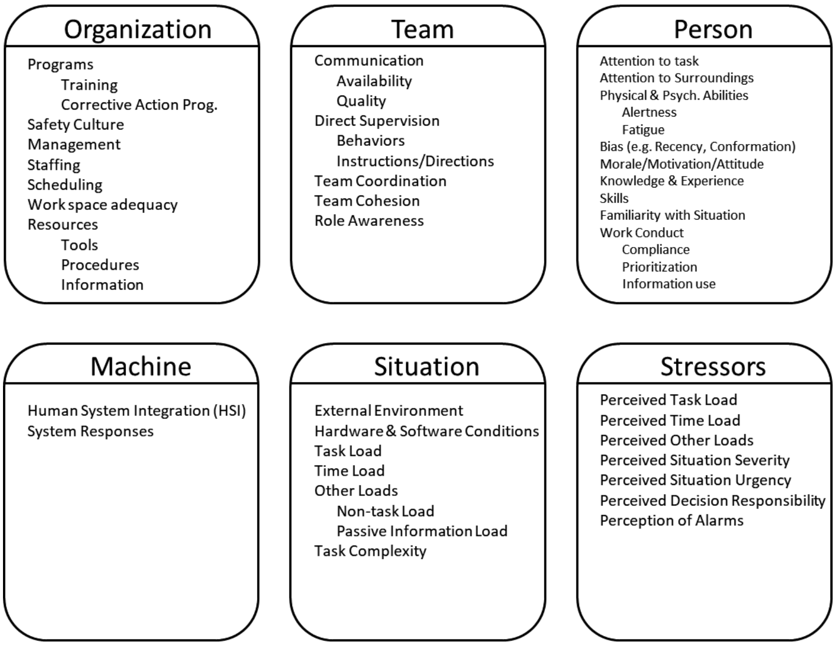
External PSFs		
	Situational Characteristics	a) Control room architectural features b) Quality of the working environment c) Work hours and work breaks d) Shift rotation and night work e) Availability/Adequacy of special equipment/tools and supplies f) Manning parameters g) Organizational structure and actions by others h) Rewards, recognition and benefits
	Task and equipment characteristics	a) Perceptual requirements b) Motor requirements c) Control-Display requirements d) Anticipatory requirements e) Interpretation f) Decision-making g) Complexity/Information load h) Frequency and repetitiveness i) Task criticality j) Long- and short-term memory k) Calculation requirements l) Feedback m) Dynamic versus Step by step activities n) Team structure o) Man-machine interface factors
	Job and task instructions	a) Operating procedures b) Oral instructions
Internal PSFs		
	Psychological Stressors	a) Suddenness of onset b) Duration of stress c) Task speed d) Task load e) High jeopardy risk f) Threat of failure, loss of Job g) Monotonous, degrading or meaningless Work h) Long, uneventful vigilance periods i) Conflicts of motives about Job performance j) Reinforcement absent or negative k) Sensory deprivation l) Distraction (Noise, Glare, Movement,

		Flicker, Colour) m) Inconsistent cueing
	Physiological Stressors	a) Duration of stress b) Fatigue c) Pain or discomfort d) Hunger or thirst e) Temperature extremes f) Radiation g) G-Force extremes h) Atmospheric insufficiency i) Vibration j) Movement constriction k) Lack of physical exercise l) Disruption of circadian rhythms
	Organizational Factors	a) Previous training/Experience b) State of current practice or skill c) Personality and attitudes d) Motivation and attitudes e) Knowledge of required performance standards f) Sex differences g) Physical condition h) Attitudes based on influence of family and other outside persons or agencies i) Group identification

Table 3. Groth's (2009) 9-bubble model of PSFs.

Model Node	**Included PSFs**
Training	Training
Organizational Culture	Safety Culture, Management Activities, Corrective Action Program
Resources	Procedures, Tools, Necessary Information
Team	Communication, Team Coordination, Team Cohesion, Direct Supervision, Role Awareness
Attitude	Morale/Motivation/Attitude, Bias, Attention
Knowledge	Skills, Knowledge and Experience, Familiarity with Situation, Physical and Psychological Abilities
Machine	Human-System Interfaces, System Responses
Loads/Perceptions	Task Load, Time Load, Other Loads, Perceived Situation Severity, Perceived Situation Urgency, Perceived Decision Responsibility
Complexity	Task Complexity, Hardware and Software Conditions

Table 4. Groth and Mosleh's (2010a) 6-bubble model of PSFs.



9.3. Appendix C: Dummy data modelling

The underlying research question for the project reported was if and how SEM can contribute with important analytical support during an ISV process.

To exemplify how results would be reported, this appendix contains a modelling effort based on the dummy database that was developed in the project. This appendix entails two sections, Method and Results. The structure and format of these two sections quite strictly follows the academic publication guidelines of American Psychological Association (APA) 6th Edition Publication Manual (2010).

The goal for this dummy analysis is to present a SEM model of the performance of teams of shift operators in a nuclear power plant's (NPP) central control room (*Sw.* Centralt kontrollrum, CKR), as expressed by the latent variables of usability, workload, situation awareness, teamwork, and task performance. These latent variables were chosen as they were evaluated during the PLEX ISV of the OKG O2 power plant in 2015 (c.f. Braruud et al., 2015). The model can thus be said to be a model that quantifies the relation between the latent variables and expresses the relation between properties of technical systems, individual's cognition and interaction within a team, and the performance of the total human-machine system. Theoretically, and almost philosophically, it is interesting whether the complex processes studied can be expressed in something as abstracted and simplified as a simplex structure. The main hypothesis is presented graphically in Figure 12. For the model described in this appendix, the goal was to describe the relation between the selected variables in one of the conceptually simplest, yet modelling wise most challenging way, i.e. in a so called simplex structure.

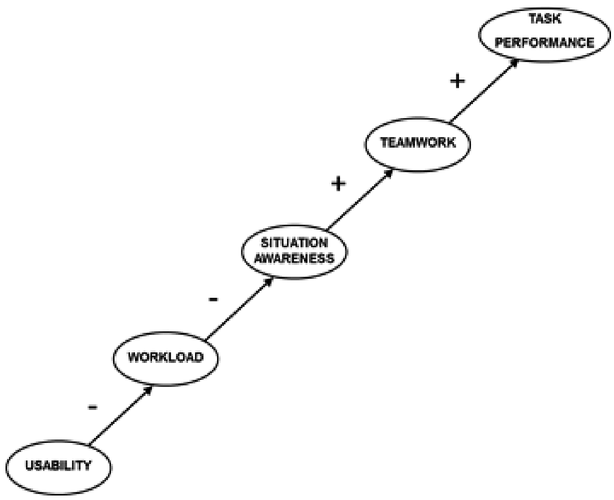


Figure 12. Main hypothesis for example model based on ISV dummy data.

The rationale behind this graphical representation of this hypothesis was Castor's (2009) conceptualization that lower level processes (to the left in the figure) form the foundation of higher level processes that eventually builds up to performance, c.f. Figure 5. It also relates to the modelling efforts reported by Nählinder, Berggren, and Svensson (2004), where the constructs of workload, situation awareness and performance show reoccurring patterns.

This main hypothesis could also be described through several sub-hypotheses that are presented in the bullet list below:

- H1: Usability is negatively correlated with Workload.
- H2: Workload is negatively correlated with Situation Awareness.
- H3: Situation Awareness is positively correlated with Teamwork.
- H4: Teamwork is positively correlated with Task Performance.
- H5: All latent variables are useful and appropriate (valid and reliable) in a SEM model describing the collected data.
- H6: It is possible to describe the constructs in a simplex structure while retaining reasonable model fit.

9.3.1. Method

Participants

Participant characteristics

To be eligible to participate in the experimental part of the study, participants had to be regular employees at the nuclear power plant and be member of a shift team, be between 25 and 65 years of age and have been employed as a control room operator for at least 3 years.

Sampling procedures

The participants were all regular control rooms operators and members of the regular shift team at the power plant. No participants were in any type of dependency toward the research team, or any other people or organizations involved in this research. Thirty persons were approached and a total of 21 were available, willing, and eligible to participate. Participation in the ISV and the data collection used for the reported modelling was conducted as a part of the operators' regular work duties and no extra benefits were received for participation.

Sample size, power, and precision

The required sample size in SEM depends on model complexity, but several other factors, e.g., normality of the data, affect the sample size decision. Many recommendations suggest between 5 to 15 cases per parameter that will be estimated or at least 200 cases, for an overview see e.g. Kline (2011). Recent simulations by Wolf, Harrington, Clark, and Miller (2013) indicate the sample size requirements are lower. For the current study, the ISV sampled the whole population of shift teams at the power plant, and the databases thus consisted of 189 cases, generated by seven shift teams of three operators, i.e. 21 actual operators.

Measures and covariates

The measures used in this report to a large, but not complete, extent replicate the ISV evaluation used at OKG O2 in late 2015. In the current modelling effort, and the dummy database, 18 manifest variables/measures were used to replicate the OKG O2 ISV which are listed below and briefly described at their respective subheadings. Full descriptions of the measures are available in Braarud, Eitrheim, and Svengren (2015).

- TP_Performance time
- TP_ProcessExpertRating
- TP_OperatorSelfRating
- TW_ProcessExpertRating
- TW_OperatorSelfRating
- SA_PUAT
- SA_OperatorSelfRating
- SA_PlantStatusReport
- W_WorkloadAcceptability
- W_NASATLX_ME
- W_OperatorRatingStaffingLoad
- U_Overview
- U_DetailedInfoProcess
- U_Navigation
- U_AlarmSystems
- U_EventLists
- U_LogsTrends
- U_OverallFunctionality

The Institute for Energy Technology in Halden (*Nw.* Institutt for Energiteknikk, IFE) is currently developing a measurement tool/method called SCORE (Supervisory COntrol and Resilience Evaluation) (e.g. Braarud, Eitrheim, & Fernandez, 2015; Fernandez, & Braarud, 2015; Braarud, & Berntsson, 2016; Braarud, Eitrheim, Holmgren, & McDonald, 2016). A replication of the SCORE expert ratings was not included in the dummy database.

Task Performance

Task Performance was operationalized through the measures:

- TP_Performance time – in this report and the dummy database this is represented as a deviation from "goodness" concerning the time to manage incidents during the scenarios.
- TP_ProcessExpertRating – Process expert rating of task performance on a 1-6 scale which represented from Not Acceptable to Acceptable.
- TP_OperatorSelfRating – Operator self-rating of task performance on a 1-6 scale which represented from Not Acceptable to Acceptable.

Teamwork

Teamwork was operationalised through the measures:

- TW_ProcessExpertRating – Process expert rating of teamwork on a 1-6 scale which represented from Not Acceptable to Acceptable.
- TW_OperatorSelfRating – Operator self-rating of teamwork on a 1-6 scale which represented from Not Acceptable to Acceptable.

Situation Awareness

SA was operationalised through the measures:

- SA_PUAT – The PUAT (Process Understanding Assessment Technique) by Strand and Svengren (in progress) is a variant of the SACRI measure (Collier

& Follesø, 1995) where questions concerning process understanding are answered by the operator and converted in a scale from 0=worst to 1=best. In this report one PUAT variable is used.

- SA_OperatorSelfRating – Operator self-rating of situation awareness on a 1-6 scale which represented from Not Acceptable to Acceptable.
- SA_PlantStatusReport – Process expert rating of how a plant status report by the shift team adhered to guidelines for such a report. Rated on a scale from 0=worst to 1=best.

Workload

Workload was operationalised through the measures:

- W_WorkloadAcceptability – Workload acceptability rated by the operators on a 1-6 scale which represented from Not Acceptable to Acceptable.
- W_NASATLX_ME – the mental effort dimension from the NASA TLX (Hart & Staveland, 1988) measure. NASA TLX contains six dimensions but for this ISV process the mental effort rating was deemed necessary. The NASA TLX value was in this report transformed into a 1-6 scale.
- W_OperatorRatingStaffingLevel – Operator ratings of frequency of staffing inadequacy. The Operator rating of staffing level value was for this report transformed into a 1-6 scale.

Usability

During the OKG O2 PLEX ISV Usability was operationalised through 78 ratings on 7 dimensions. In this report and the dummy database Usability was measured through mean ratings on the seven manifest variables below:

- U_Overview – 16 ratings on a 1-6 scale concerning different aspects of the usability of Overview information.
- U_DetailedInfoProcess – 15 ratings on a 1-6 scale concerning different aspects of the usability of detailed information about processes, parameters, components and other objects.
- U_Navigation – 6 ratings on a 1-6 scale concerning different aspects of usability of the navigation in the technical systems.
- U_AlarmSystem – 21 ratings on a 1-6 scale concerning different aspects of the alarm system.
- U_EventLists – 4 ratings on a 1-6 scale concerning different aspects of the usability of the event lists.
- U_LogsTrends – 8 ratings on a 1-6 scale concerning different aspects of usability of the logs and trend displays.
- U_OverallFunctionality – 8 ratings on a 1-6 scale concerning different aspects of usability of the overall functionality and interaction of the technical systems.

OKG 02 central control room simulator

The experiment was conducted in a CKR simulator, with technical capabilities to simulate both the previous and the modernized OKG O2 CKR.

Figure 13. OKG O2 control room simulator.

Research design

Figure 14 schematically describes the research design with seven shift teams performing in nine scenarios with rating for the data collection conducted after each scenario. The seven teams of three operators with Shift supervisor (SS), Reactor operator (RO) and Turbine operator (TO) all participated in all scenarios

Scenario name and duration	Briefing	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8	Scenario 9
	60 min	60 min	120 min	120 min	120 min	60 min	120 min	120 min	120 min	60 min
Day	Day 1	Day 1	Day 1	Day 1	Day 2	Day 2	Day 2	Day 3	Day 3	Day 3
Number of injects		5	10	6	9	4	8	10	9	5
Ratings	1 trial	2	3	4	5	6	7	8	9	10

Figure 14. A schematic representation of the research design.

9.3.2. Results

Participant flow

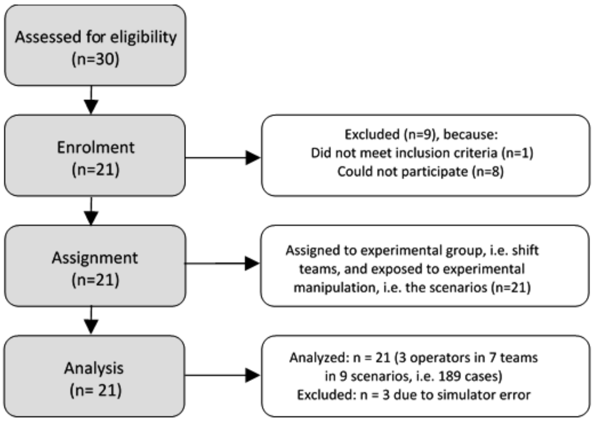


Figure 15. Participant flow through the stages of experiment preparation and scenario execution.

Recruitment

The data collection of the experimental phase of the study was conducted during a total of twenty-one days in May and June 2015. The recruitment of the participants was conducted during the weeks before each specific day of the experiment, with participant confirmation about a week in advance. The participants received no monetary compensation for their participation.

Data analysis and statistics

The statistical software packages SPSS 24 and AMOS 24 was used for all statistical analyses.

Normality of data

SEM requires the data to exhibit both single variable and multi-variate normality. Both statistical univariate and multivariate normality as well a set of assumptions concerning linearity, outliers, and multicollinearity of the collected data, which is a prerequisite for the planned statistical analyses, was initially assessed through inspection of different types of graphs and plots. Descriptive data are provided below, at the appropriate subsection. The Kolmogorov-Smirnov (K-S) values for each measure are presented in Table 5 - Table 9 below; a non-significant result (i.e. a sign. value >.05) indicates normality. For all the variables in the dummy database non-normality was detected. However, all variables were retained after visual inspection, accepting the risk of effects on the estimates in the models, given the observed distributions. Figure 16 shows an example of a histogram showing distribution of one of the variables.

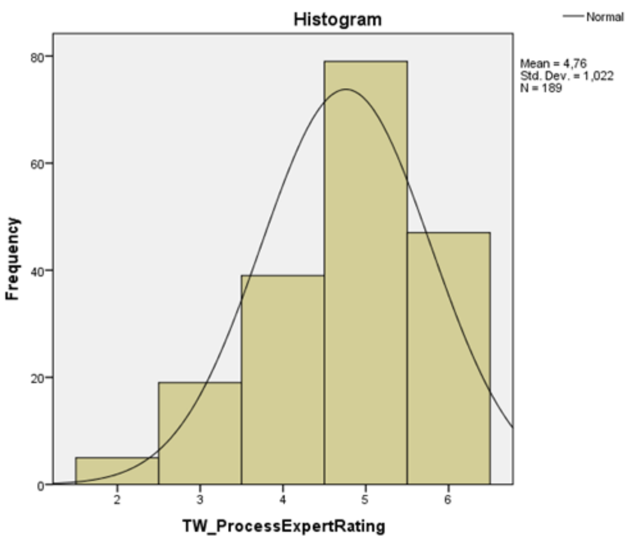


Figure 16. Data normality example.

Missing data

Due to the strict procedural discipline during the experiment no missing data exists. At three occasions, the simulator had to be restarted due to technical problems, but the individual participants could quickly complete the scenarios and data collection due to the simulator's capability to resume aborted scenarios at the exact place where problems occurred.

Demographics

The mean age of the participants in the dummy data set (4 female and 17 male) was 45.29 years with a standard deviation of 11.95, and a range of 25-64 years. As a comparison, the mean age of the operators in the nuclear power plant of FKA is 44.31 years.

Task Performance

Descriptive data concerning the Task Performance measures are provided in Table 5.

Table 5. *Descriptive statistics Task Performance measures.*

Measure	N	Mean	SD	Min	Max	Skewness	Kurtosis	K-S	K-S Sig.
TP_PerformanceTime	189	4.84	.886	2	6	-.727	-.307	.139	.000
TP_ProcessExpertRating	189	4.85	1.021	2	6	-.426	-.842	.203	.000
TP_OperatorSelfRating	189	5.11	.951	3	6	-.739	-.502	.260	.000

In Figure 17, the TP_PerformanceTime variable is presented as a bar graph to, as an example, show the variance in the data graphically.

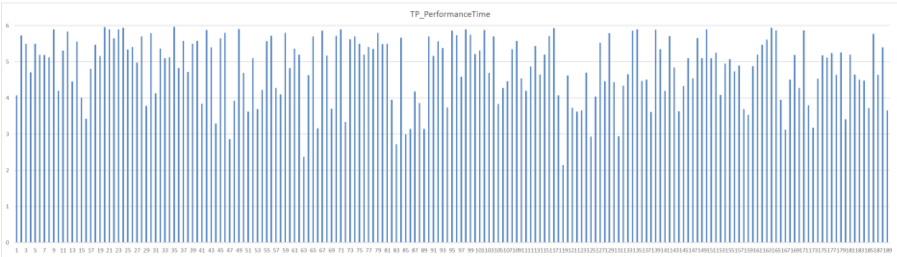


Figure 17. Bar graph of the TP_PerformanceTime variable.

Teamwork

Descriptive data concerning the Teamwork measures are provided in Table 6.

Table 6. *Descriptive statistics Teamwork measures.*

Measure	N	Mean	SD	Min	Max	Skewness	Kurtosis	K-S	K-S Sig.
TW_ProcessExpertRating	189	4.76	1.022	2	6	-.687	-.037	.259	.000
TW_OperatorSelfRating	189	4.66	1.012	2	6	-.604	-.165	.268	.000

Situation Awareness

Descriptive data concerning the Situation Awareness measures are provided in Table 7.

Table 7. *Descriptive statistics Situation Awareness measures.*

Measure	N	Mean	SD	Min	Max	Skewness	Kurtosis	K-S	K-S Sig.
SA_PUAT	189	.84	.110	0	1	-1.080	1.511	.093	.000
SA_OperatorSelfRating	189	4.75	.939	2	6	-.601	-.119	.284	.000
SA_PlantStatusReport	189	.83	.092	0	1	-.627	-.148	.119	.000

Workload

Descriptive data concerning the Workload measures are provided in Table 8.

Table 8. *Descriptive statistics Workload measures.*

Measure	N	Mean	SD	Min	Max	Skewness	Kurtosis	K-S	K-S Sig.
W_WorkloadAcceptability	189	3.84	1.227	1	6	.591	-.369	.212	.000
W_NASATLX_ME	189	3.38	.938	2	6	.889	.648	.307	.000
W_OperatorRatingStaffingLoad	189	4.74	.963	2	6	-.398	-.446	.225	.000

Usability

In Table 9 the means of ratings for the seven usability dimensions are presented.

Table 9. *Descriptive statistics Usability measures.*

Measure	N	Mean	SD	Min	Max	Skewness	Kurtosis	K-S	K-S Sig.
U_Overview	189	3.57	1.063	1	6	.053	-.734	.182	.000
U_DetailedInfoProcess	189	4.04	1.291	1	6	-.249	-.561	.193	.000
U_Navigation	189	4.35	1.182	1	6	-.257	-.525	.205	.000
U_AlarmSystems	189	5.15	.889	2	6	-.766	-.081	.263	.000
U_EventLists	189	5.17	.871	2	6	-.727	-.148	.269	.000
U_LogsTrends	189	4.29	1.243	1	6	-.289	-.527	.157	.000
U_OverallFunctionality	189	3.12	1.635	1	6	.212	-.571	.176	.000

Models

Exploratory factor analysis

After the initial statistical assumption and data quality testing that was reported above, the 18 variables in the database that was intended to be used as manifest variables were analysed with an Exploratory Factor Analysis, EFA. This was done to be able to do initial justification of the measurement model. The pattern of how the 18 manifest variables, in factor analysis terms called items, group themselves in the Pattern Matrix of Table 10, can be interpreted to see how the different items the variables load on the expected factors, which in turn can be used to specify and justify the measurement model.

The resulting pattern matrix presented in Table 10 show how the manifest variables load on five factors, with factors loadings less than .30 being suppressed for readability. The Kaiser-Meyer-Olkin value was .799, exceeding the recommend value of .6 and Bartlett's Test of Sphericity reached statistical significance, supporting the factorability of the correlation matrix. The maximum likelihood analysis revealed the presence of five factors with eigenvalues over 1, explaining 33.6%, 14.5%, 13.8%, 7.1% and 6.3% of the variance respectively. An inspection of the scree plot in Figure 18 suggests that a four or five factor solution should be considered. The Pattern Matrix also show that the individual items primarily load on one factor (all other loadings fall below the recommended threshold of .30 and are thus suppressed in Table 10 and not further considered in the analysis), i.e. a quite clear indication of how the manifest variables group themselves, from the perspective of the data, i.e. not considering theoretical differences between them and what they are assumed to measure.

Table 10. *Pattern Matrix for Exploratory Factor Analysis with Direct Oblimin rotation and Maximum Likelihood extraction.*

Items	1	2	3	4	5
TP_ProcessExpertRating			.831		
TP_OperatorSelfRating			.864		
TP_PerformanceTime			.961		
TW_ProcessExpertRating				.794	
TW_OperatorSelfRating				.982	
SA_PUAT_rev					.792
SA_OperatorSelfRating_rev					.437
SA_PlantStatusReport_rev					.474
CW_WorkloadAcceptability					.657
CW_NASATLX_ME					.812
CW_OperatorRatingStaffingLoad_rev					.582
U_Overview					.620
U_DetailedInfoProcess	.992				
U_Navigation	.805				
U_AlarmSystems		-.792			
U_EventLists		-.931			
U_LogsTrends		-.728			
U_OverallFunctionality	.730				

Values under .30 suppressed.

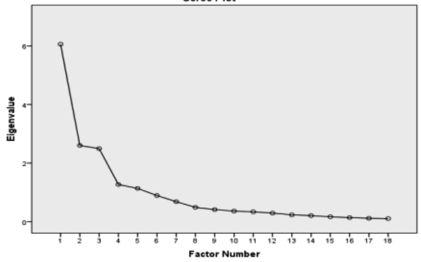


Figure 18. Scree plot from EFA.

The factor correlation matrix is presented in Table 11.

Table 11. *Factor correlation matrix.*

Factor	1	2	3	4	5
1	1,000	-,428	-,195	,185	-,173
2	-,428	1,000	,006	-,343	,368
3	-,195	,006	1,000	,198	,159
4	,185	-,343	,198	1,000	-,477
5	-,173	,368	,156	-,477	1,000

Extraction Method: Maximum Likelihood.

Rotation Method: Oblimin with Kaiser Normalization.

Some observations from the exploratory factor analysis are:

- The Task Performance and the Teamwork related items load on their own respective factors.
- The Situation Awareness and Workload related items load on the same factor and this can be interpreted, from a pure data perspective, that they should load on the same latent variable, tentatively called Operator Capacity. The raw data provide both negative and positively loadings on this factor, but some of the variables ratings have been reversed as indicated by the _rev endings in Table 10.
- The Usability related items do not load on the same factor. The Overview item load together with Situation Awareness and Workload, and the other six items load on two different factors, tentatively called Usability 1 and Usability 2.

Confirmatory factor analysis

Based on the result of the EFA, a confirmatory factor analysis, CFA, was conducted. See Figure 19. CFA results for results. One factor called OperatorCapacity and two Usability related factors, here called Usability 1 and Usability 2 were used. The U_Overview item here loads on Operator Capacity, rather than any of the Usability factors.

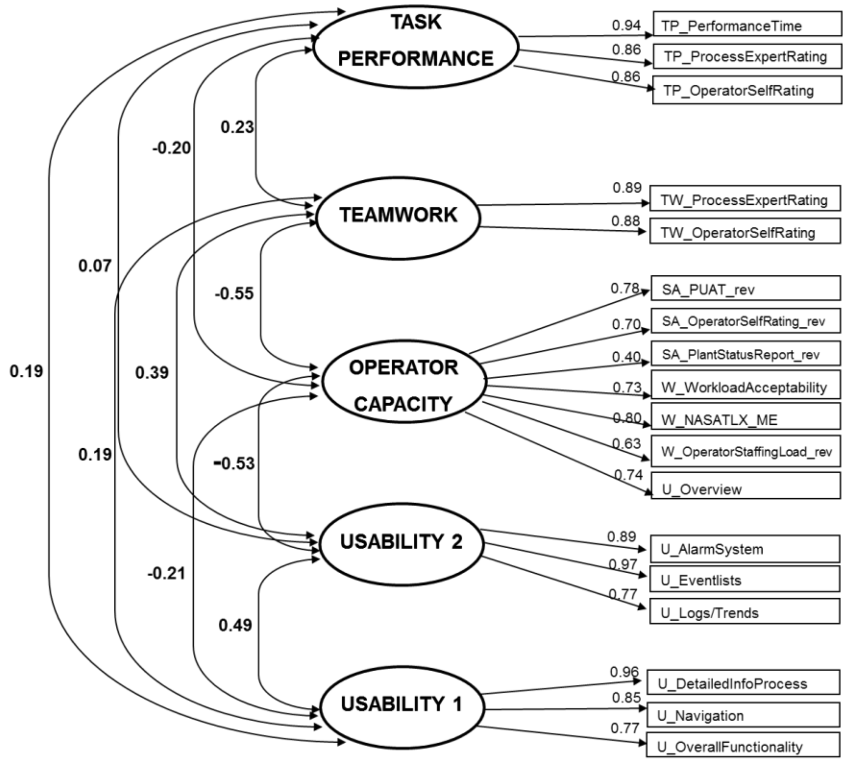


Figure 19. CFA results.

Structural Equation Model

In this section, the actual full SEM model results are described. In Figure 20 a SEM with a simplex structure is proposed, based on the observations from the EFA. In the figure, error terms and residuals not shown to maintain readability.

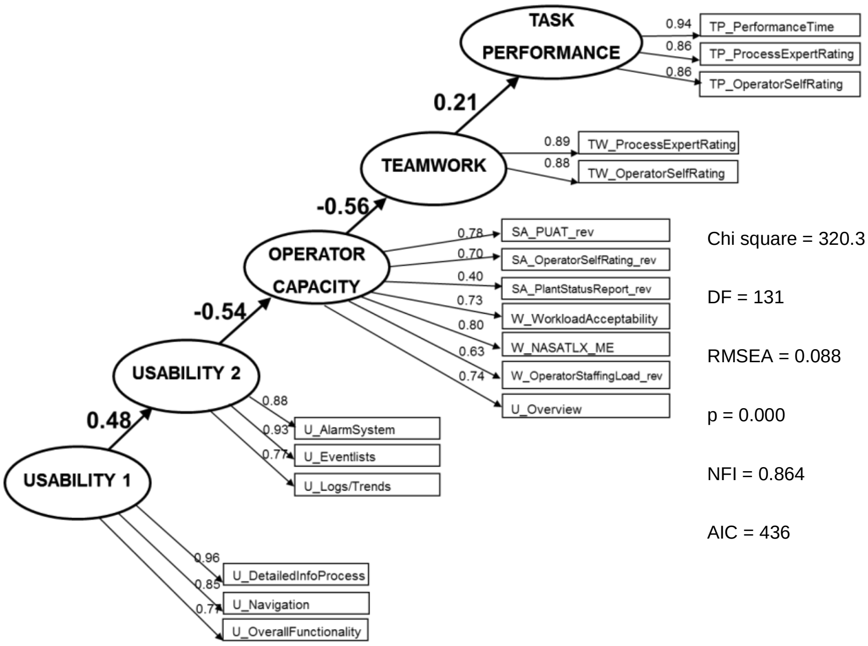


Figure 20. Full structural model, based on EFA observations.

The fit indices in Figure 20 indicate that the model to some extent capture the variance in the data, but would benefit from further refinement, since the RMSEA is above recommended thresholds levels and the NFI is below the threshold. See Appendix A for a further description of fit indices, their use and recommended thresholds.

The software packages that enable SEM calculations provide so called modification indices, that from a pure data driven perspective, provide recommendations on how to modify a model to achieve a better fitting models, called specification search. Figure 21 shows the model with some new arrows, as suggested by the modification indices. As a result, the RMSEA and the NFI indices have improved somewhat and just barely reach acceptable values.

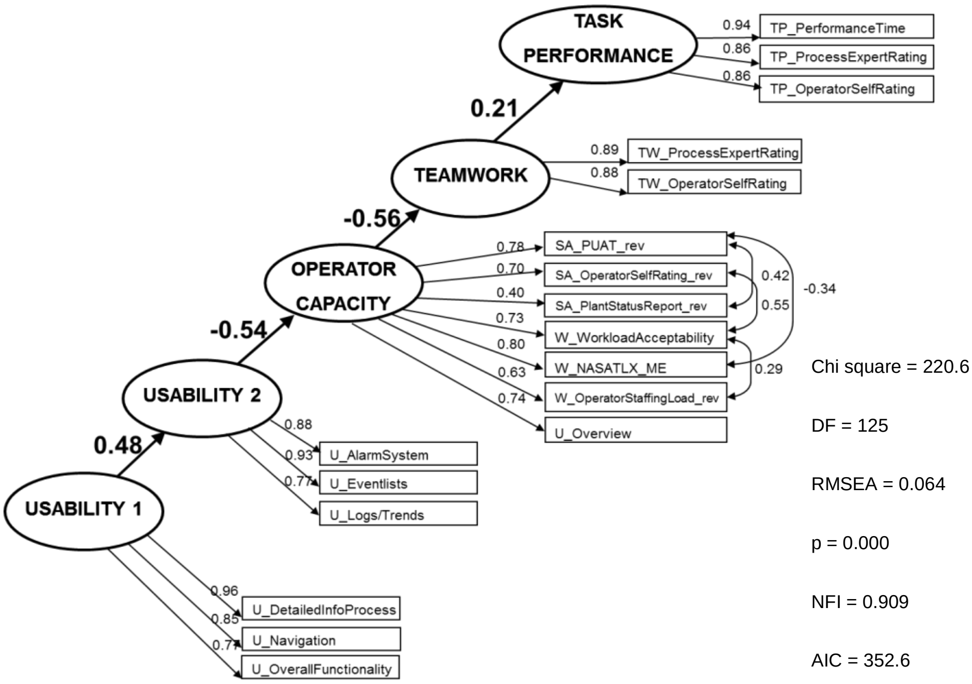


Figure 21. Full structural model, based on EFA observations and after consulting the modification indices.

Some observations concerning the full SEM models based on the dummy data set are:

- The model has a low regression weight, i.e. values on the arrow, between Teamwork and Task Performance, which should be interpreted as, the manifest variables that currently have been used to assess the latent variable named Teamwork, do no strongly explain the variation in the manifest variables used to measure the latent variable Task Performance. The variation in Task Performance probably depends on something else, variables which were not included in the data collection. In SEM nomenclature, this would be an example of misspecification, i.e. we do not assess all relevant factors, if we want to be able to explain changes in Task Performance. Another latent variable with associated manifest variables or other manifest variables for the current latent variables need to be considered. Note that the weak connection between Teamwork and Task Performance is only intended as a pedagogical example for the current report.

The model presented in Figure 22 disregards the EFA and test the main hypothesis, with manifest variables loading on the latent variables as expected by manifest variable names.

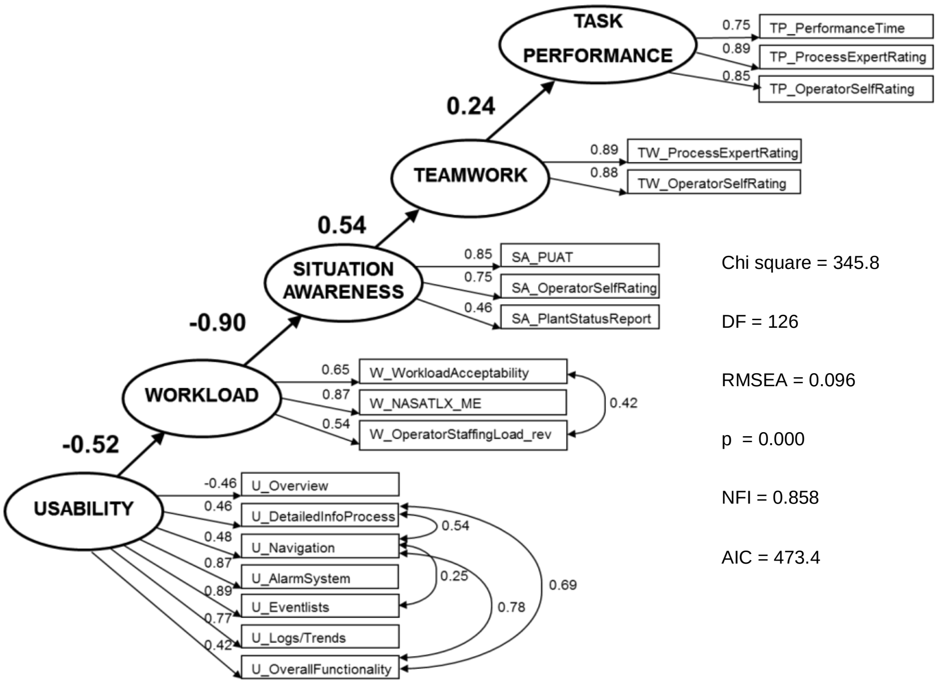


Figure 22. Full structural model testing the original main hypothesis.

The model presented in Figure 22, even after modification indices does not reach recommended thresholds for fit indices, and hence describe a model which not to the same degree as Figure 21 describe the variation in the data.

Thomsons (2000) guidelines, see Appendix A, have to the extent possible been followed in the development of these models of the dummy data. Recommendation #2 could not be followed due to the limited size of the database, and recommendation #10 is debatable for the same reason since the 189 cases in the database, generated by 21 operators cannot be a large sample.

9.4. Appendix D. Alternative and nested models

Numerous alternative models, e.g. nested models and sub-models, can typically be explored in a modelling process. Some examples are provided below, Figure 23 to Figure 28, however these are not elaborated in further detail since the report is intended to describe what SEM is and test its application for as part of CKR evaluation in ISV processes.

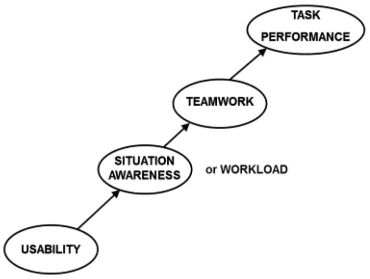


Figure 23. Model with a latent variable removed.

Figure 24. Model with new latent variables included.

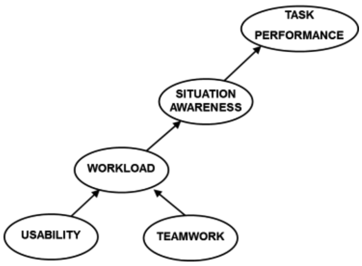


Figure 25. Model with more starting points, i.e. exogenous variables, e.g. if Teamwork and Usability are both used as starting points.

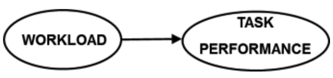


Figure 26. Sub-models, e.g. Workload and Task Performance.

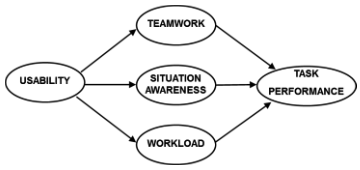


Figure 27. Model with alternative effect flow.

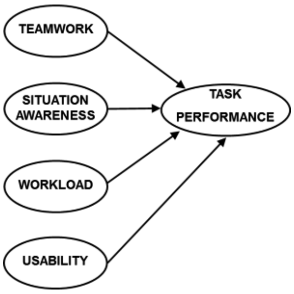


Figure 28. Model with another alternative effect flow.

9.5. Appendix E. Modelling test with data from a live ISV evaluation

During 2017 the authors of the report got access to a dataset collected during a live ISV evaluation. The dataset from the live evaluation had many similarities to the dummy data set described above, but also some important differences, as the battery of measures had been further refined. The results from more than 90 measures were included in the dataset.

A multitude of models was explored, where both the measurement model and the structural model were extensively varied, following the process described in the main report above. This was done to describe a model that would aggregate the ISV data into a SEM model that provided a first step towards a theoretical model of how the selected latent variables relate to each other. However, during the modelling of this live dataset, no theoretically interesting models, which also met SEM quality requirements, i.e. with adequate model fit, were identified by the authors of the report. Several reasons for this can be discussed:

- *Data distribution:* as described earlier, a scientifically publishable SEM model needs to meet certain requirements concerning the data both in terms of their distribution and amount of data in relation to model complexity. For this dataset, the non-normality requirement for the manifest variables required for SEM modelling was not met, i.e. many of the variables in the live dataset exhibited skew and kurtosis values that departed from statistical normality. SEM is reasonably robust to non-normality in the data, but the deviations were quite large for some variables. In retrospect, this might not be surprising, provided that the live dataset not had been intended to be used for SEM modelling, but this was not evident before the modelling effort was initiated. By design, from the researchers that designed the ISV data collection, many of the measures were expected to produce highly skewed results. For the actual ISV is was a feature, as HEDs then could be defined and identified by low means for the specific measure. The reader should note that this not reduces the validity of the measures as a part of an ISV evaluation, but it reduces the possibility to use SEM to model the data.
- *Correlations between measures:* The bivariate correlation analyses and factor analyses that were conducted before the actual SEM modelling exhibited low correlations between many of the individual measures and low factors intercorrelations. A comprehensive presentation of all these analyses between the 90 individual measures that were present in the live data will not be included in the present report to maintain report readability, but the interested reader can contact the report authors for further discussion.

Due to the reasons summarised above, no theoretically interesting models, which also met SEM quality requirements could be specified. However, it is important to point out that this current inability to specify theoretically useful and scientifically publishable models during this modelling effort not casts any doubt on the general applicability of SEM as a statistical analysis method nor on the methodological validity of the ISV evaluation methodology and ISV measures that was used during the live ISV evaluation.

2018:24

The Swedish Radiation Safety Authority has a comprehensive responsibility to ensure that society is safe from the effects of radiation. The Authority works to achieve radiation safety in a number of areas: nuclear power, medical care as well as commercial products and services. The Authority also works to achieve protection from natural radiation and to increase the level of radiation safety internationally.

The Swedish Radiation Safety Authority works proactively and preventively to protect people and the environment from the harmful effects of radiation, now and in the future. The Authority issues regulations and supervises compliance, while also supporting research, providing training and information, and issuing advice. Often, activities involving radiation require licences issued by the Authority. The Swedish Radiation Safety Authority maintains emergency preparedness around the clock with the aim of limiting the aftermath of radiation accidents and the unintentional spreading of radioactive substances. The Authority participates in international co-operation in order to promote radiation safety and finances projects aiming to raise the level of radiation safety in certain Eastern European countries.

The Authority reports to the Ministry of the Environment and has around 300 employees with competencies in the fields of engineering, natural and behavioural sciences, law, economics and communications. We have received quality, environmental and working environment certification.

Strålsäkerhetsmyndigheten
Swedish Radiation Safety Authority

SE-171 16 Stockholm
Solna strandväg 96

Tel: +46 8 799 40 00
Fax: +46 8 799 40 10

E-mail: registrator@ssm.se
Web: stralsakerhetsmyndigheten.se

Ale Tryckteam AB, Bohus, 2018