Safety and Reliability – Safe Societies in a Changing World – Haugen et al. (Eds)
© 2018 Taylor & Francis Group, London, ISBN 978-0-8153-8682-7

Challenges with data for human reliability analysis

K. Laumann
Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

H. Blackman
Division of Research and Economic Development, Boise State University, Boise, Idaho, USA

M. Rasmussen
Department of Psychology, Norwegian University of Science and Technology, Trondheim, Norway

ABSTRACT: A lack of empirical data is often been presented as a large challenge for HRA, which begs the question: why is this so difficult? HRA methods were not developed as objective quantitative test methods, but more as qualitative evaluation methods because objective data did not exist. Since HRA methods include substantial qualitative evaluation of the meaning of the elements in HRA methods, such as definitions of the performance shaping factors as well as their strength, these elements cannot be objective measured. This paper also discusses other challenges with collection data from event reports, literature reviews, experiments and databases. The conclusion in this paper is that a decision should be made about how we should look at HRA methods: as qualitative evaluation methods or objective quantitative test methods. Quantitative and qualitative methods have different approaches to evaluate the quality of the methods making it difficult to be something in between.

1 INTRODUCTION

Most of the Human Reliability Analysis (HRA) methods and techniques have been developed to estimate human reliability for tasks within Probabilistic Risk Analysis (PRA) in nuclear power plants. Human reliability analysis (HRA) was developed because of the lack of empirical data on human error probability. If data on the likelihood of human error on specific tasks existed, we would not need an HRA method. Williams (1992, p.20) said: "Therefore in cases where an assessor may have access to more specific and accurate task failure data, these should be used in preference to the HEART generic data-set." So if we had the data, the data should be used and not a HRA method.

HRA methods like; THERP (Swain & Guttmann, 1983), HEART (Williams, 1992), SPAR-H (Gertman,et al. 2005; Whaley, et al. 2011) and ATHEANA (Forester, et al. 2007), were developed as methods to support more qualitative expert judgements. The expert judgement provides gross estimates of failure probabilities for tasks defined by the PRA, when better data was missing. In the HEART manual (Williams, 1992, p. 4) states: "When considering system safety and reliability, engineers are generally concerned with gross changes in the probability of failure within system e.g. factors of 10, the proverbial order of magnitude. To be of value, therefore human reliability assessment techniques should be concerned with those factors, which are likely to produce probability of failure modification in excess of a factor of 3, and which, when cumulated, could produce significant changes in performance, and possible threaten system safety, operability and reliability."

Also HRA methods are often said to be simplified methods since it is often not enough resources in performing a comprehensive analysis which mean that the analyst is analyzing the most important influences on human reliability with limited time to read and understand guidelines and to perform the analysis.

Even though HRA analysis results in a quantitative likelihood for failure or success, HRA methods were not developed as positivistic quantitative test methods. HRA methods seem to be closer to a post positivistic research view. In positivism one assume that an objective reality exists, that this reality can be objective measured by scientific methods and that it is possible to develop scientific laws that can be generalized across settings (Guba & Lincoln, 1994). Within this approach, only quantitative research methods are used. In post positivism one assumes that an objective reality exists. However, this reality is complex and that

it can only be imperfectly apprehended and that we can never be sure that a true reality has actually been found (Guba & Lincoln, 1994). Within this approach, both qualitative and quantitative methods are used.

In spite of these characteristics of HRA methods described above, many authors claim that the biggest challenge in HRA is the lack of objective empirical data (for example; Boring et al. 2012; Hallbert et al. 2004; Kim, et al. 2015; Swain,1990; Williams, 1985). A question raised here is a question that was first presented in Laumann (in review) about what an HRA method actually is: Is an HRA method a qualitative evaluation method that leaves a lot up to the analyst qualitative evaluation and where mainly expert judgement was used to develop HRA? Or is an HRA method a quantitative test method where empirical tests where used to develop and test the methods? HRA today, is much more a qualitative evaluation method than a quantitative test method. Probably the words used about HRA methods such as "analysis" and "techniques" reflect the qualitative basis of HRA, rather than a quantitative.

In this paper, we will present challenges for HRA methods to be quantitative test methods and challenges with collecting quantitative HRA data. First, we will define challenges that exist for obtaining quantitative data for HRA that are the same challenges found with all kinds of quantitative methods. Then we will present challenges to obtain HRA data that exist with more specific methods such as; literature reviews, experiments designed to collect HRA data, event reports and databases. For simplicity, we have chosen to present example from SPAR-H (Gertman, et al. 2005; Whaley, et al. 2011) and HEART (Williams, 1992). However, the challenges for collecting empirical data presented exist for all HRA methods. SPAR-H and HEART were also chosen because these are methods where quantitative data collection have been much discussed.

2 DISCUSSION

2.1 *Challenges that exists for all kinds of data collections in HRA*

There are some demands for all kinds of quantitative test methods; they should be valid and reliable. HRA methods have a strong similarity to psychological test methods where human behavior is predicted from different "psychological" construct. HRA is also about predicting human behavior from constructs, since the elements in the methods such as Performance Shaping Factors (PSFs) or Error Producing Conditions (EPCs) are constructs, which are assumed to affect human behavior. In psychological test methods, validity is divided into different facets; content validity, construct validity, concurrent and predictive validity (Murphy & Davidshoffer, 2014). These demands will not be described here since they are well known and can be found in many textbooks as for example Murphy and Davidshoffer (2014).

A large challenge for HRA is content validity. Content validity is achieved by a) judgments and descriptions of the constructs and of the structure within these concepts and b) definitions and development of measurement scales to measure the constructs and their structure. If the content of the constructs, their structure, and how they should be measured (measurement scales) have not been clearly defined, then it is not possible to test the other aspects of validity such as construct validity, concurrent validity and predictive validity or reliability. Next, we will give some examples of content validity issues in two HRA methods SPAR-H and HEART. Laumann and Rasmussen have also presented challenges with content validity in SPAR-H in several papers (Laumann & Rasmussen, 2016; Rasmussen, Sandal & Laumann, 2015; Rasmussen & Laumann, in review).

The elements in SPAR-H (Gertman, et al. 2005; Whaley, et al. 2011) are: two nominal tasks with two nominal failure rates and eight performance shaping factors. The two nominal tasks in SPAR-H (Gertman,et al. 2005; Whaley, et al. 2011) are diagnosis and action. A task should be classified as diagnosis if it involves cognitive processing. An action involves limited cognitions. The separation between cognition and action within SPAR-H is very peculiar since probably nothing an operator does within an accident scenario is purely action without cognition. The diagnosis/action separation gives room for a much interpretations of what this actually means.

In the SPAR-H manual (Gertman et al. 2005), there are no specifications about what is meant by a task or at what task level the analysis should be performed. How a task is defined have a large effect on the result of the analysis or the probability for errors (for a discussion see Rasmussen & Laumann, 2017). SPAR-H leaves it up to the analyst to define at what task level SPAR-H should be applied and this gives much room for analyst choice and interpretation.

The elements in the HEART (Williams, 1992) method described in the user manual are 14 generic task types with a nominal human unreliability value and their suggested uncertainty bounds and 38 EPCs. For the EPCs the analyst should also assess the proportion of affects. Williams and Bell (2017) have recently reviewed HEART with a large literature review. Based on this review 32 out of the 38 original EPCs were kept, six of the EPCs were revised slightly and two new ones was incorporated in HEART.

It is difficult to find a definition of what a generic task actually is in HEART. It is said about generic tasks in HEART (Williams, 1992, p.8): "The first is the assumption that basic human reliability is dependent upon the generic nature of the task to be performed, i.e. for each task in life there is a basic probability of failure." So a generic task has something to do with the generic nature of the task which is not a very specific definition. Error producing conditions in HEART are defined as (Williams, 1992, p.1): 'Error producing conditions are factors that can affect human performance, making it less reliable than it would otherwise be.' The separation between what is a GTT and what is an EPC is not obvious in HEART, and some of the GTTs include elements that are very similar to the EPCs. This gives room for different interpretation by different analysts.

HEART defines that a task should be analyzed at the level that fits the GTT. How to analyze the proportion of affects or the strengths of the EPCs are not well defined in HEART, which gives much room for interpretation by the analyst.

To show an example of the difficulties with content validity, SPAR-H and HEARTs definitions of the PSF available time (SPAR-H) and time shortage (HEART) will be presented.

SPAR-H [Gertman et al. 2005, page 20] defines one of its PSFs available time as: "Available time refers to the amount of time that an operator or a crew has to diagnose and act upon an abnormal event. A shortage of time can affect the operator's ability to think clearly and consider alternatives. It may also affect the operator's ability to perform. Multipliers differ somewhat, depending on whether the activity is a diagnosis activity or an action."

The SPAR-H Step-by-Step (Whaley et al. 2014, page 2–4) give the following definitions of the levels for available time in SPAR-H:

Inadequate time—the time margin is negative because less time is available than is required.

Barely Adequate Time—the time margin is zero because the time available equals the time required

Nominal Time—there is a small time margin because the time available is slightly greater than the time required.

Extra Time—the time margin is greater than zero but less than the time required; the time available is greater than the time required

Expansive Time—the time margin exceeds the time required; the time available is much greater than the time required.

With these definitions, subjective evaluation depending on the characteristics of the tasks and the contexts are necessary to decide on a level for available time for a particular task. There is no way to objective define the level since there is no objective description of how much time one should assume to define the different levels. In addition, it is a question, what is the unit of analysis in SPAR-H. It has been claimed by the authors of SPAR-H that it is an analysis of the average operator. However, if there is barely adequate time for the average operator one might expect that there is too little time for the slower than average operators. How much failure one expects becomes circular with the definition of the unit and the unit of analysis is not well defined in SPAR-H.

HEART has a similar EPC, which is described as: "A shortage of time available for error detection and correction". HEART gives no advice on, how the analyst should go about analyzing this EPC. As shown under the discussion of SPAR-H, a lot of information needs to be clarified to analyze this EPC and since it is not available in the method, it is up to each analyst subjective judgement. Something that is peculiar with HEART is that the analyst is not instructed on how he/she should go about collecting information about the GTTs and the EPCs. The EPCs are defined by one sentence and then it is up to the analyst to interpret how this sentence fit their contexts/tasks, or what the sentence actually means, as for example—how much shortage of time should exist before the maximum predicted nominal amount should be chosen.

These characteristics with HRA methods (as SPAR-H and HEART) show that they are more qualitative evaluation methods than objective quantitative test methods and that they are far from being objective quantitative test methods. It is a question if we should expect methods that include so little definitions and descriptions (content validity) as SPAR-H and HEART to show interrater reliability. To obtain interrater reliability the concepts need to be precisely, defined. However, with a qualitative evaluation method view of the method, we will focus more on the analyst ability to predict correct error rates based on qualitative evaluations with use of a HRA method. With this approach, we might not expect high interrater reliability, but rather look at the quality of the data and the evaluations that the prediction is based on.

We claim that if HRA is going to be tested with quantitative methods they need to be improved and that the place to start is to develop good definitions of the content of the concepts included into the method. If good definitions exist, it might be possible to develop measurement scales for the PSFs/EPCs. If we have these measurement scales of the PSFs/EFCs it might be possible to predict how different levels of PSFs, such as for example complexity, affects performance.

However, it is a question, if this is possible. It could be that with the different elements included into HRA, it is impossible to be so well defined that quantitative measurements can be developed. For example, for time available, it could also be that the

tasks and contexts that are evaluated in HRA are so different that it is not possible to develop the exact meaning of the PSFs and the PSFs levels/strengths that counts for all kinds of contexts and tasks. If this is the case, the qualitative evaluation view of HRA methods might be better, but then also the qualitative part of the analysis need to be further developed.

2.2 *Challenges with data from psychological and human factors studies (literature reviews)*

HEART was developed based on human factor literature (Williams, 1992). Williams (1992) and Williams & Bell (2017) has done literature reviews to investigate studies that include EPCs and their maximum multipliers and nominal error rates on GTTs in experimental designs. Also, Laumann, Sandal & Rasmussen (Laumann & Rasmussen, 2016; Rasmussen, Sandal & Laumann, 2015; Rasmussen & Laumann, in review) have done literature reviews to investigate the meanings of the PSFs and how large effect the PSFs have on affecting human errors on tasks.

One challenge with using literature reviews on psychological and human factors studies to collect information for HRA is that these studies were not designed with the purpose of testing HRA methods and therefore it is difficult to transform the data to fit the HRA method. For example, literature studies usually only included one negative level for a PSF and this level is difficult to match with the level description for example in SPAR-H. It is also difficult to match the description of PSFs to the one manipulated PSFs in this kind of studies.

It is not obvious how the literature review to collect information on EPCs in HEART was done. Williams and Bell (2017) say that they have looked for the maximum multipliers of the EPCs. However, the human factors studies do usually not intent to manipulate the maximum multipliers of the EPCs. They usually just manipulate one level and it is often not described or discussed what this level actually is. In addition, the experiments usually intend to study why the PSFs/EPCs have an influence on performance rather than how much it affects performance. In the human factors literature there is also not developed measurement scales for the PSFs/EFCs. We have looked at some of the studies that are referred to by Williams & Bell (2017) and it is difficult to see that the maximum multiplier, in the meaning of 'the highest possible negative multipliers' for the PSFs/EFCs,' were the manipulations in these studies.

We are hopeful that Williams & Bell (2017) will present more from their literature review and discuss the evidence for the EPCs' maximum multipliers and nominal failure rates on the GTTs. In this way, other researchers can better understand the authors' arguments, and perhaps add to and relate this evidence to other methods. We do not think one should look at this data as objective evidence for an EPC/PSF but rather an evaluation done on the available evidence. Since it is an evaluation, it is important to understand the authors' arguments on for example, including or excluding experiments and the authors' argument about how similar the experimental manipulations found in these experiments are to the concepts in HEART.

2.3 *Challenges with performing new experiments to collect data that is relevant for HRA*

The unspecific definitions of the concepts in the HRA methods are also a large challenge for developing experiments since it is difficult to develop manipulations and measurements that fit with the HRA methods. An example of this is an experiment performed by Liu and Li (2014) where experimental data were compared to the multipliers in SPAR-H. In this experiment, one can see the difficulties the authors have in matching the definitions of the PSFs and the levels in their experiment to SPAR-H definitions and levels. For example, experience and training were defined as the 20 first trials as the negative level and the later 20 trials as the nominal level. This manipulation does not fit with the negative level description of experience and training given in SPAR-H, which is less than 6 months of relevant experience and training. It was difficult to develop an experiment manipulation that fits with this level description in SPAR-H. In this experiment, also complexity was manipulated, but the measurement of complexity measured the complexity of the procedures, and then it is a question if this should have been looked at as complexity or procedures. There were also questions on, how to match the manipulated levels to SPAR-H levels also for complexity and available time.

For the HEART EPCs, one should in experiments only manipulate the maximum strength of the EPCs, since these are the elements included into the method. However, usually, the maximum strength of an EPC, does not seem to be a meaningful experimental manipulation, if the "maximum multiplier" is interpret literary. For example, in an experiment on EPC 1 (Williams, 1992, p.22): 'Unfamiliarity with a situation which is potentially important but which only occurs infrequently or which is novel' one would give the participants no training on a completely new task. In this situation, one would have expected a human error probability for failure close to 1, and we might not need to test this because the result is too obvious.

For some of the GTTs in HEART and the nominal tasks in SPAR-H an obvious challenge for new experiments is the number of subjects needed, when errors is expected to occur in 1 of 100, 2 of 100 or 1 of 1000 subject.

There are many challenges with performing experiments that are relevant for HRA. One frequently mentioned challenge is that if these experiments should be done with actual operators in a simulated control room, the cost of the experiments are high and little data would be collected (Boring, 2012).

Another challenge that exists when performing experiments on PSFs (with both operators and other participants such as students) is that PSFs are difficult to completely control and without that control, it is difficult to measure the independent effect of the different nominal tasks and/or the PSFs/EPCs. We have seen in experiments that other PSFs than those manipulated often affect the results. For example, poor teamwork is a variable that is difficult to control, since this might exist within the crews before they come to the experiment, or develop during the experimental run. Other examples of PSFs that are difficult to control are; operators that are ill, hungry, stressed, fatigued or demotivated. The crews themselves might also increase the complexity of a scenario by some erroneous actions or by forgetting a procedural step. Then the manipulation for some crews might be different from the one the experimenter planned.

PSFs have a tendency to exist from before the experiments or occur during the experiments, they cannot be completely controlled, and some you might observe (e.g. poor teamwork) and other might be more hidden for the experimenter (e.g. fatigue, stress and illness).

As an example of this, the first author experienced that in an experiment at the Halden Reactor Project where we intended to manipulate available time and information load, but for some of the crews we also observed that poor teamwork also occurred. The poor teamwork and short available time combined had a very negative effect on performance for some of the crews (Laumann, Braarud & Svengren, 2005).

In addition, another challenge with simulator experiments with operators is that these simulators such as the simulator at the Halden Reactor Project, often are computer based simulators rather than analog simulators that the operators use at their own plant, making it difficult to know how much "the new interface for the operator" PSF interacts with the manipulated PSFs and how this affect error rates.

2.4 *Challenges with event report as a basis for HRA data*

One possible source of data in HRA is event reports. One problem with using event reports as a basis for HRA data is that the event reports only investigate and report when an error occurred and then we do not know if the PSFs are usually present when this task is done or if there were some particular PSFs that were present when the error occurred (Kim et al. 2015).

Another issue is that event reports are often written by operators that probably do not have much knowledge about PSFs/EFCs and how to investigate the presence of PSFs/EFCs. The event reports that we have seen have not been very specific about how they collected the data on the PSFs and how the data were interpreted. In addition, the strengths or levels of the PSFs are not defined in the event reports and much interpretation have to be done to decide on a specific level or strength.

Another problem with event reports might also be that the events occur so infrequently that they do not give much data for HRA (Boring et al. 2012).

It is also a problem for HRA that usually in the event report more than one PSF has occurred in the event, which makes it difficult to estimate the effects on orthogonal PSFs/EFCs, which is included in the HRA methods.

It is also a problem with event report that many organizations prefer to not be open about such matter as human errors and why they occur since this is regarded as sensitive information.

2.5 *Challenges with databases*

There have been several attempts to develop databases for HRA data, which included data from event reports, literature reviews, and/or from experiments or simulations. Examples of such databases are NUCLARR (Gertman et al, 1990). HERA (Halbert et al, and COREDATA (Gibson, Basra & Kirwan, 1999).

The challenges described for event reports, literature reviews and event reports are also challenges for databases because this is the information that is entered into the databases.

The general reason for a database is to organize data in some predefined ways. Also for databases, the unclear definitions in HRA are a large issue because when a data bases structure is developed the definitions, for example of PSFs/EPCs and PSFs/EPCs levels/strengths from one or more method, have to be used as template and the data then has to be interpreted based on these definitions in the database. Data from databases are never going to be better than the data included in the first place. To include quality data into a database, a good and precise structure and definitions, which were also used during the data-collection from either event reports or experiments is required.

Since the HRA methods, include so diverse definitions of the elements in the methods one structure in the databases for each method is necessary, which is very resource demanding.

In HRA, the structure and purpose of the databases are often not clearly specified and the

argument for them seems to be that one day, some analyst (a very smart analyst) could find a good way to analyze this data in a way that fits HRA. However if the developers of the database do not know more exactly how the database should be used and what is the purpose of it, the work invested in it might be useless. One might wonder if all the resources to develop HRA databases have been a good investment based on the amount of data relevant for HRA that has been provided so far.

3 CONCLUSION

HRA methods have been criticized for the lack of predictive data and validation of their results. However, is seems like HRA methods are criticized for not being something they never intended to be: quantitative test methods. It is not enough to say that HRA methods, are methods to estimate human reliability on tasks. Within HRA one should make a choice, whether we should look at HRA methods as a qualitative evaluation methods that gives gross and crude differences based on expert judgement or if HRA methods should be developed to be quantitative test methods. The inventors of SPAR-H and HEART seems sometimes to present their method as much more objective quantitative test methods than they actually are. Of course, this likely because they were developed to support probabilistic risk assessment where a quantitative result is required. ATHEANA went in another direction and defined an HRA method that is mainly a qualitative method, with an expert based quantification technique added at the back end. We think that also SPAR-H and HEART are mainly qualitative methods, requiring substantial analysts' judgement in order to produce the quantitative result.

To be a quantitative test method we need content validity and very good and specific definitions of the concepts the method includes, definitions of measurement scales for the concepts, definitions of who is the unit of the analysis, and definitions of how should a task be defined for that method. An important question is: Are these definitions possible within HRA? Maybe concepts such as PSFs and EPCs might be too difficult to be precisely defined, because the concepts include too much, and because they vary too much from context to context or from task to task. It might not be possible to develop definitions and measurement scales that can apply for all of the contexts and tasks where the HRA methods are used.

However, if we define HRA methods as qualitative evaluation methods, criteria for a good qualitative analysis should be developed and discussed. Qualitative research methods have other methods to evaluate the quality of the research than quantitative research methods. One paper by Laumann (in review) presents criteria for good qualitative analysis and discusses how these could be applied for HRA.

It might be that the definitions when a qualitative method is used do not need to be that specific and are more allowed to vary between contexts. However, even with a qualitative evaluation method view, as good as possible definitions and advice about how to perform the analyses, should be available.

This question about how we should look at HRA methods should be answered based on what we think about our data. Is it possible to precisely define the different elements in a HRA method that can be used across different contexts and tasks? If this is not possible, we have to collect HRA data with a qualitative approach.

After working for many years with HRA methods, definitions of PSFs and their levels and performing experiments within HRA, we doubt that it is possible to define and specify the PSFs/EFCs and measurement scales of the PSFs/EFCs enough that a quantitative approach is possible. An alternative for HRA then is to more focus on developing good qualitative methods for evaluation of PSFs/EPCs, PSF levels/EPC strengths, and error rates.

As HRA methods are today it would be a best to just admit that they are qualitative expert judgement methods trying to predict crude differences in performance, and that they are far from being objective test methods that can be empirical validated with quantitative methods. However, HRA methods are not good and systematic qualitative methods either and improvement in descriptions of how the qualitative analyses should be performed, are also needed. A qualitative method approach might demand lesser specification than a quantitative test approach.

One could argue, if HRA methods are not objective test methods why should they predict performance? There have been some studies to test the validity of HRA methods such as The international empirical study (Forester, et al. 2014) and the U.S HRA empirical study (Forester et al. 2014). These studies do not give an overall conclusion on how well HRA methods predict human errors. They give many and varied answer depending on the task, the HRA method and the analyst. However, in these studies one might wonder, what is actually tested? Is it the analysts' ability to use the HRA method to predict the likelihood of errors or is it the HRA method in itself that is validated? If one assume that the method was tested, the researcher should assure that the HRA method guideline was reliable followed by the analysts. This is not possible since some of the methods like SPAR-H and

HEART do not include complete and prescriptive descriptions of the qualitative parts of the analysis. In these studies, it seem to be the analysts' qualitative evaluation with use/help of the HRA method that was tested and not the method in itself.

HRA methods should not continue to be something in between qualitative and quantitative research methods, since then they are based neither on good qualitative research methods nor on good quantitative research methods. Qualitative and quantitative research methods have different assumptions about quality and have different ways to investigate the quality of the method or the quality of the research. A choice should be made within each method and the choice has to be made by the authors of the methods.

REFERENCES

Boring, R. et al. 2012 Microworlds, simulators, and simulation: Framework for a benchmark of human reliability data sources. In *Joint Probabilistic Safety Assessment and Management and European Safety and Reliability Conference,* 16B-Tu5-5.

Forester, J. et al. 2014. The International HRA Empirical Study. Lessons Learned from Comparing HRA Methods Predictions to HAMMLAB Simulator Data. *NUREG-2127,* US Nuclear Regulatory Commission, Washington, DC.

Forester, J. et al. 2014. The US HRA Empirical Study – Assessment of HRA Method Predictions against Operating Crew Performance on a US Nuclear Power Plant Simulator. *NUREG-2156,* US Nuclear Regulatory Commission, Washington, DC.

Forester J. et al. 2007 ATHEANA user's guide. *NUREG-1880.* 2007. Nuclear Regulatory Commission, Washington, DC: U.S.

Gertman, D. et al. 2005. *The SPAR-H human reliability analysis method,* NUREG/CR-6883. U.S Nuclear Regulatory Commission, Washington, DC, USA.

Gertman, D.I. et al. 1990. Nuclear Computerized library for assessing reactor reliability (NUCLARR), *NUREG/CR-4639.* Nuclear Regulatory Commission, Washington, DC, US.

Gibson, H. et al. 1999. Development of the CORE-Data database. *Safety & Reliability Journal,* 19, 6–20. Guba E.G, Lincoln Y.S. 1994, Competing paradigms in qualitative research. In N.K. Denzin & Y.S. Lincoln, (eds), *Handbook of Qualitative Research.* Thousand Oaks, CA: Sage; p.105–117.

Hallbert, B. et al. 2004. The use of empirical data sources in HRA. *Reliability Engineering and System Safety,* 82, 139–143.

Hallbert, B. et al. 2006. Human Event Repository and Analysis (HERA) System Overview. NUREG/CR6903. Nuclear Regulatory Commission, Washington, DC, US.

Kim, Y. et al. 2015. A statistical approch to estimating effects of performance shaping factors on human error probabilities of soft control. *Reliability Engineering and System Safety,* 142, 378–387.

Laumann, K. et al. 2005. The task complexity experiment 2003/2004, *HWR-758.* Institute for Energy Technology, Halden, Norway.

Laumann, K. In review. Criteria for qualitative methods in Human Reliability Analysis. *Reliability Engineering and System Safety.*

Laumann, K., & Rasmussen, M. 2016. Suggested improvements to the definitions of Standardized Plant Analysis of Risk-Human Reliability Analysis (SPAR-H) performance shaping factors, their levels and multipliers and the nominal tasks. *Reliability Engineering & System Safety,* 145, 287–300.

Liu, P. & Li, Z. 2014. Human error data collection and comparison with prediction by SPAR-H. Risk Analysis, 34, 1706–1719. DOI: 10.111/risa.12199.

Murphy, K.R. & Davidshofer, C.O. 2014. *Psychological Testing Principles and Applications.* Sixth edition. Person Education Limited, Essex, UK.

Rasmussen, M. & Laumann, In review. The evaluation of fatigue as a performance shaping factor in the Petro-HRA method *Reliability Engineering and System Safety.*

Rasmussen, M. & Laumann. 2017. The impact of decomposetion level in human reliability analysis quantification. L. Walls, M. Revie & T. Bedford (Eds.), *Risk, Reliability and Safety: Innovating Theory and Practice.* Taylor & Francis group, London, ISBN 978-1-138-02997-2.

Rasmussen, M. et al. 2015. Task complexity as a performance shaping factor: a review and recommendations in standardized plant analysis risk-human reliability analysis (SPAR-H) adaption. *Safety Science,* 76, 228–238.

Swain, A.D. 1990. Human reliability analysis: Need, status, trends and limitations. *Reliability Engineering & System Safety,* 29, 301–313.

Swain, D.A, & Guttmann H.E. 1983, Handbook of human reliability analysis with emphasis on nuclear power plant application *NUREG/CR-1278,* Washington, D.C, USA.

Whaley A.M. et al. 2011. The SPAR-H step-by-step guidance. *INL/EXT-10-18533, Rev 2,* Idaho Falls, USA.

Williams, J.C. & Bell, J.L. 2016. Consolidation of the human error assessment and reduction technique. L. Walls, M. Revie & T. Bedford (Eds.), *Risk, Reliability and Safety: Innovating Theory and Practice.* Taylor & Francis group, London, ISBN 978-1-138-02997-2.

Williams, J.C. 1985. Validation of human reliability assessment techniques, *Reliability Engineering,* 11, 149–162.

Williams J.C. 1992. *A user manual for the HEART.* Stockport, UK: DNV Technica Ltd.