

## Mapping of Time-Consumption During Input Data Management Activities

Anders Skoogh, Björn Johansson, Chalmers University of Technology, Sweden

{Anders.Skoogh, Bjorn.Johansson}@chalmers.se

SNE Simulation Notes Europe SNE 19(2), 2009, 39-46, doi: 10.11128/sne.19.tn.09937

The success of a discrete event simulation project relies heavily on input data quality. In order to achieve high quality data a significant amount of time needs to be spent, either due to absence of data or problems with defining and extracting existing data from databases. This paper presents a distribution of the time-consumption for the activities in the input data phase during discrete event simulation projects. The results show where efforts need to be focused to reduce time-consumption and improve quality of input data management.

### Introduction

The competition between companies in all markets has increased considerably during the recent decades and it is getting more and more important to optimise the efficiency in production [1]. To improve productivity, some organisations use analysis tools like Discrete Event Simulation (DES) in major change projects as well as for continuous improvements. However, the input data needed to analyse the production is often not available, or at least, it takes plenty of time to collect and prepare the data for further analysis.

DES is a powerful tool for productivity analysis and it is argued that input data management is the most crucial and time-consuming step in DES projects [2] [3]. The time spent on input data management is typically as much as 10-40% of the total time of a DES project [4]. This set-back sometimes tempts organisations to choose less complex analyses with lighter requirements on input data quality. As a result, these analyses yield results of poor, or at least, inferior quality.

Few previous studies have closely mapped the input data phase in order to find the reasons for the heavy time-consumption [5]. Even fewer studies focus on identifying the input data activities which are most favourable to improve. The aim of this work is to identify the most time-consuming activities in the input data phase of DES projects. The results will show where to put important efforts in future research, in order to reduce time-consumption and increase quality of input data management. Not only in simulation projects, but also for projects using other production analysis methods.

### 1 Input Data Management in Discrete Event Simulation

One always present step in DES projects is the input data phase, usually called “Data Collection”; see for example the widely applied methodologies described in Banks *et al.* [3], Law and Kelton [6], and Rabe *et al.* [7] (Figure 1).

These methodologies merely show the input data management step as a black box. However, in practice input data management includes several activities such as collection of raw data from various sources, transformation of data to information and documentation. Here, *data* is referred to as “a set of discrete, objective facts about events” [8] (e.g. 1000 repair times for a machine). *Information* on the other hand, is slightly simplified defined as “data with meaning” [9]. In this case, information can be exemplified by a statistical representation of Mean Time To Repair (MTTR), which contains both relevance and purpose for the receiver (the simulation model).

In this paper, the input data phase is described in more detail than on the black box level. We have divided the internal time-consumption within the input data phase into separate activities and measured the time-consumption for each activity.

The focus on input data is surprisingly low in previous scientific contributions within the field of DES. Perera and Liyanage [5] is one of few contributions that really address the difficulties related to the input data management in DES projects. They rank the major pitfalls in input data collection as follows:

1. Poor data availability
2. High level of model details

3. Difficulty in identifying available data sources
4. Complexity of the system under investigation
5. Lack of clear objectives
6. Limited facilities in simulation software to organise and manipulate input data
7. Wrong problem definitions

There is also lack of publications on systematic guidelines to overcome these issues and to reduce time-consumption in input data management (one is Bernhard and Wenzel [10]). Instead, earlier research performed on reduction of time-consumption in input data management has primarily focused on the level of human involvement in the process. A study made by Robertson and Perera [2] describes four alternative solutions for managing data for DES models:

#### 1. Tailor-made solution

- Data primarily derived from the project team
- Data manually supplied to the model by the model builder
- Data resides in the simulation tool

#### 2. Spreadsheet solution

- Data primarily derived from the project team
- Data manually supplied to the computer application (e.g. MS Excel spreadsheet)
- Data automatically read by the model via a computer application
- Data Resides in the computer application

#### 3. Off-line database solution

- Data primarily derived from a Corporate Business System (CBS)
- Data automatically supplied to an off-line database from the CBS
- Data automatically read by the model
- Data resides in an intermediary simulation database

#### 4. On-line database solution

- Data primarily derived from the CBS
- Data automatically supplied to the model from the CBS
- Data resides in the CBS

The same publication states that solution 1 and 2 were most frequently used in industry, which is most likely still a valid statement. However, some research work and industrial applications have strived towards less human involvement in the input data management process. For example, some years ago the tendency shifted towards integration of systems, in which DES is one component that share data and information with many other applications within the same package. DELMIA from Dassault Systèmes [11] and SIEMENS Teamcenter [12] are two examples of such Product Lifecycle Management (PLM) software packages. Moreover, simulations driven by an off-line simulation database using input data from Enterprise Resource Planning (ERP) systems have also been performed [13]. This is one example of the contributions towards solution 3 and 4, described above.

However, the situation remains; Robertson and Perera [2] state that: "It is strongly argued that data collection is the most crucial and time consuming stage in the model building process". Therefore, this paper evaluates if this statement is still valid and shows where future efforts should be concentrated. This is done by summarising the time-consumption within DES projects in general, in the input data phase in specific and even more important in the activities of the input data phase.

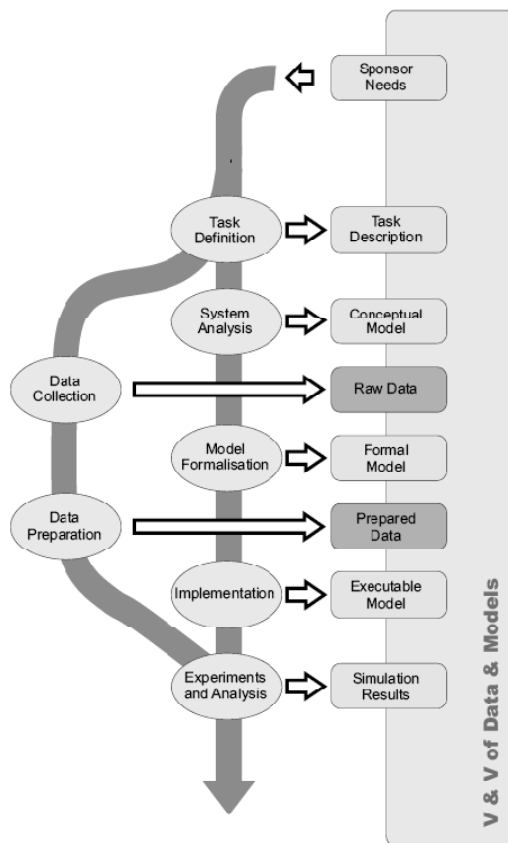


Figure 1. Steps in a simulation study [7].

## 2 Material and Methods

The study embraces the analysis of 15 DES projects performed between 2000 and 2007. The projects have been performed in a wide range of companies with regard to line of business, size of organisation and previous experience in DES. The plants in which the projects were performed are all located in the Nordic countries, mainly in Sweden. Both pure industrial cases and simulation projects performed in cooperation between industry and academia are included among the 15 projects.

Semi-structured interviews [14] were conducted with members from each project in order to define the work procedure and activities in the input data phase of the projects. The agenda of the interviews was focused on the kind of problems, related to input data, which arose during the project. Furthermore, an additional aim with the interviews was to identify key factors for rapid and precise input data management, from a practitioner's viewpoint.

The respondents were also asked to fill in a questionnaire where the time-consumption for the whole project as well as for each specific activity in the input data phase was specified. Moreover, information about availability and sources of input data in the projects were gathered in order to detect reasons of extensive time-consumption as well as factors for successful input data management. All times responded in the study are given in the unit "man-days". One man-day equals to one eight-hour working-day for one person. For example, if two persons have spent two days to carry out a task together, the amount of time reported to this study is four man-days. The respondents were asked to write the time given with a resolution by minimum one man-day but if they were able to recall in greater detail they were allowed to answer in fractions of man-days.

The authors compiled all collected information in a data-sheet and analysed it in order to map the time-consumption for all activities and to find patterns in prerequisites and work procedures, which can reduce time-consumption in data management. The findings from the questionnaires were then combined with the information from the interviews. The results are presented in Section 4.

## 3 Input Data Management Activities

In the presented analysis of time-consumption, the input data process in DES projects is divided into nine separate activities. Each activity consists of

several tasks. The number of tasks and the way to execute each task can differ slightly between simulation projects because of differences in prerequisites and objectives. However, the work procedures are structurally very similar among simulation input data phases, and the activities defined below cover the process of all studied projects.

Below, each input data management activity is briefly described to enable measurements of the time-consumption. However, a more thorough description, including supportive guidelines, is provided in Skoogh and Johansson [15].

### 3.1 Identification of Input Data Parameters

The identification of required input data parameters has earlier been addressed as one of the key activities for successful input data management. The process is often performed in cooperation with people having expert knowledge of the modelled manufacturing process. The parameters to include are often dependent on project objectives, on model complexity and on level of model detail. Therefore, there is an ongoing interaction between construction of the conceptual model and identification of input data parameters [3].

### 3.2 Accuracy Requirement Specification

It is of great advantage if the project team can forecast each input parameter's impact on model behaviour. If accuracy requirements can be specified for each factor, the effort spent on information collection can be optimised. Accordingly, more resources and time can be assigned to important parameters instead of less central ones. As a result of this activity, the required number of unique data-points for each parameter is decided.

### 3.3 Mapping of Available Data

Once the relevant parameters are selected, the project team needs to search for and map the input data already available, without need for manual gathering in the real world production system. Such available data can generally be found in simple manual systems (e.g. spreadsheets with previously performed time studies) or in more complex computer based systems such as ERP-systems, Manufacturing Execution Systems (MES) or other databases holding process information (e.g. time-stamps logged by Programmable Logic Controllers (PLCs)).

However, it is hazardous to instantly rely on the applicability of information from this kind of systems,

without further investigations. Despite that database specifications and people with extensive practical experience say that data is available, simulation engineers frequently find the data in a crude form or measured in a manner that makes it useless for simulation. Consequently, the activity of mapping available data includes identifying sources, understanding the sources and making sure that it will be possible to extract required data from the systems.

### 3.4 Choice of Gathering Methods

When the available data has been mapped, a gap between required data and available data will be detected in most simulation projects. Hence, some additions will be necessary. In this activity the project team decides which methods to use in order to gather missing data from the modelled system. The choice will mainly depend on possibilities to measure and on the expected accuracy of each parameter according to earlier specifications (section 3.2). Examples of gathering methods are time studies, frequency studies and interviews.

### 3.5 Document Creation

In order to store all data that will be collected from available sources or from real world measurements, a document needs to be created. A well-designed document helps to structure the data collection procedure. It also gives greater possibilities to reuse data in future studies and to make small adjustments if errors occur, or if the modelled system changes during the project time.

### 3.6 Data Collection

The data collection activity can be divided into two parts. One is the extraction and compilation of available data from the identified sources. The other is to gather the missing data according to the previously specified methods (Section 3.4).

**Extract and Compile Available Data** Despite the availability of data, some efforts are almost always needed to extract relevant information from the data sources. As mentioned before, more complex databases often contain data in forms that require some transformation before it can be used for further analysis in a simulation project. One example is breakdown data that is often logged in a crude form where start and stop times of all stops are stored. In this case, efforts are needed both to sort out the stops of interest for the analysis, and to calculate the absolute length of breakdowns.

**Gather Missing Data** Many times this activity is fairly straight-forward since the procedure is well outlined in previously presented activity (Section 3.4). However, depending on the chosen methods, type of modelled process and requirements of accuracy, it can be a time-consuming activity.

### 3.7 Data Analysis and Preparation

The outcome of the data collection activity is often a large set of data points, e.g. 100 measured cycle-times or 2000 repair times extracted from a maintenance database. In the data analysis and preparation activity, the way to represent the data in the simulation model is selected. Regardless of whether an empirical or statistical representation is chosen, some preparations are performed in this activity. For example, the statistical representation requires fitting the data set to a statistical distribution.

### 3.8 Data Validation

Before the data is used in the simulation model, a separate data validation activity helps to ensure accuracy in further analysis. An early control of the data representations' correctness usually saves iterations in later model validation, where more sources of error are involved. The data representation can be validated using production follow-ups or expert knowledge, e.g. Turing tests [16].

### 3.9 Final Documentation

It is important to document the results of the input data phase, since they are of vital importance for the model outputs and furthermore for the decisions taken with reference to the analysis. The final documentation is also necessary in order to make future simulation projects less time-consuming by enabling reuse of input data.

## 4 Survey and Interview Results

The results and analysis section is divided into two parts. The first part presents the analysis of time-consumption for input data activities and the second part shows the data availability in the studied DES projects.

### 4.1 Analysis of Time-Consumption

The respondents were asked to assess the time spent in each of the activities during the input data phase of the 15 DES projects included in the study. The percentages of time in each activity with regard to the duration of the entire input data phases are presented

Project	Input data parameter identification	Accuracy requirement specification	Mapping of available data	Choice of gathering methods	Document creation	Data collection	Data analysis and preparation	Data validation	Final documentation
# 1	12%	2%	2%	2%	0%	60%	12%	4%	5%
# 2	3%	0%	7%	7%	1%	51%	7%	22%	0%
# 3	5%	2%	12%	1%	2%	63%	1%	6%	6%
# 4	5%	2%	4%	5%	5%	61%	7%	5%	5%
# 5	3%	3%	12%	1%	6%	57%	12%	0%	6%
# 6	3%	0%	15%	3%	5%	58%	8%	5%	5%
# 7	1%	4%	2%	2%	1%	40%	25%	12%	12%
# 8	9%	0%	9%	4%	9%	52%	9%	4%	4%
# 9	5%	0%	9%	5%	5%	45%	23%	5%	5%
# 10	4%	4%	9%	4%	7%	50%	7%	9%	7%
# 11	33%	11%	11%	7%	0%	24%	2%	11%	2%
# 12	14%	7%	14%	11%	7%	21%	14%	4%	7%
# 13	5%	0%	10%	10%	5%	50%	10%	5%	5%
# 14	5%	3%	13%	5%	8%	56%	3%	5%	3%
# 15	10%	0%	21%	0%	0%	62%	8%	0%	0%
Average	8%	2%	10%	4%	4%	50%	10%	7%	5%

**Table 1.** Time-consumption for each input data activity with regard to the entire input data phase.

in Table 1. Around half of the input data management time is used for actual data collection, both from available sources and from manual gathering. *Mapping of available data* together with *data analysis and preparation* are the other two activities on the top-three ranking of time-consuming activities.

It is not surprising that the collection activity claims a significant amount of time. Some more detailed findings about the most time-consuming parameters, and how data availability influences time-consumption, will be further examined later in this section. However, the fact that mapping of available data is a top-three ranked activity is more conspicuous. But information from many of the respondents is very similar and claims that the major reasons are the complexity of the data sources and that the available data is not collected and stored in a way that is ready for use in simulation models. Hence, a lot of time is needed to understand the data sources and to ensure that the data is relevant in the specific case. Ensuring that it will be possible to extract and transform required data into a suitable representation for the simulation model, also adds to the extensive time-consumption.

Table 2 shows the ranking among input data parameter classes with regard to required collection time. Process times, breakdown data, set-up times, tool changes and material handling data are all straightforward parameter classes, but production planning and organisational data contain some sub-types. Information needed for production planning incorporates data such as production schedules, arrival patterns of incoming parts, and sales data. Organisational information contains data about staffing plans, shift schedules and breaks. Note that the sum of the time-consumptions for all parameter classes is not equal to

100% since all classes are not applicable in every studied project.

Interview responses indicate that the reason for the heavy time-consumption for process data depends on problems with defining the process delimitations, e.g. when a cycle starts and stops.

For breakdown data, the corresponding problem is to sort out the stops of interest for the simulation study, among all other kind of logged process-disturbances in the IT-systems. Both process data and breakdown data often include large amounts of data since they are considered to be particularly important for model performance and dynamics.

#### 4.2 Data Sources and Availability of Information in DES Projects

The availability of data necessary for production analysis is not satisfying in most of the studied DES projects. Only one of the 15 cases had all data available when the project started, and combined with a study performed by Johansson et al. [17] it is obvious that insufficient work has been performed in order to

Parameter class	Time-consumption (percentage of the entire input data phase)
Process times	42%
Breakdown data	32%
Production planning data	19%
Material handling data	14%
Set-up times	12%
Tool-change times	8%
Organizational data	7%

**Table 2.** Required time efforts for collection of input parameter classes.

Parameter class	All data available	No data available	Combination of available and manually gathered data
Process times	33%	27%	40%
Breakdown data	64%	9%	27%
Production planning data	18%	55%	27%
Material handling data	0%	62,5%	37,5%
Set-up times	22%	44%	34%
Tool-change times	20%	80%	0%
Organizational data	40%	40%	20%

**Table 3.** Percentage of studied projects having all, none or parts of the needed input data available.

support analyses with proper input data. Two projects out of the 15 had no data at all to start with, and had to gather all data manually. Table 3 shows the data availability for each input parameter class, presented as the percentage of projects having all, none or parts of the required data available.

As seen in Table 3, breakdown data is the category that is most frequently collected and stored, followed by organisational data and process times. Contrarily, material handling data was not fully available in any of the projects. It is important to note that it is not relevant to directly compare time-consumption for different parameter classes, since the amount of raw data and importance for model performance varies significantly among the classes. Therefore, one should not draw the conclusion that data availability is insignificant for the time-consumption because breakdown data collection takes more time than to gather data for material handling equipment (Table 2), despite the fact that breakdown data has higher availability.

The study results rather show that a large share of available data has a positive correlation with rapidity of input data collection. One single example is that the only project having all needed data available is also the project with lowest percentage (12%) of time spent in the input data phase. Investigating the actual time for collection of input data in projects with full data availability compared to projects that fully or partly include manual gathering supports the same conclusion. To illustrate, the mean time required for collection of process times was less than one week when data was fully available and slightly more than three weeks when manual gathering was needed.

## 5 Discussion

The survey and the interview results clearly show the difficulty for companies to effectively manage their

data for use in production analysis tools like DES. It is obvious that no evident progress has been made to reduce the time-consumption for input data management in recent years. For instance, this study shows that the time-consumption for input data management in DES-projects is still 31% on average, which is a high percentage compared to older studies. The opinion is also supported by the fact that only 7% of the studied projects had all required input data available when the project started. This is almost the same availability ratio as Johansson *et al.* [17] found six years ago (6%).

Two of the top-three time-consuming input data activities both shed light on the same difficulty in input data management at present. Both problems with actual data collection (50%) and mapping of available data (10%) indicate a potential for reduced time-consumption by implementation of intelligently-designed computer-based data sources.

According to the findings presented in the results and analysis section, companies can gain a lot of time in production systems analysis by keeping track of data describing their processes. This in turn enables DES to be used more frequently; hence increased performance in production is achieved. There are several ways of continuously having up-to-date information available, some examples are automated PLC-logging or previously performed time studies stored in databases. However, it is very important to note that the design of the majority of existing databases is not developed with the needs of analysis tools like DES in mind. No less than 10 out of 13 projects in the study, having some available data at hand, reported problems with extracting relevant information from the databases due to problems with understanding the data structures, mapping relevant data for their specific application and sorting out the information needed among an often huge set of data. These findings are also supported by earlier research performed by Perera and Liyanage [5].

Moreover, companies often overestimate their ability to provide data for analysis tools like DES, which might be a result of the extensive information flow in present production systems. However, when the projects start they frequently lack important data or find that data is measured and stored in a way that is unsuitable for simulation models. Consequently, a lot of time needs to be spent on identification of relevant information and on recalculations or complementary measurements. This common statement of respon-

dents has resulted in problems with keeping the time plans for input data management in the studied projects. Only 20% of the projects reported that their input data phases were completed in time.

Requirements stated above are not just based on a DES perspective but also on the viewpoint of other production analysis methods. Companies could gain much productivity by keeping track of their production data more carefully. One way is to design future data systems having the viewpoints of production analysts in mind. But not to forget, today's purposes with the systems are also important to support, e.g. maintenance and process control.

There are some factors in the study that might affect the precision of each individual case study result. Since the exact number of hours responded in the study was not documented in all cases, the reported time-consumption is dependent on each respondent's perception and memory. However, the possible impact of this factor is reduced by the choice of recently performed projects, for example 13 of the 15 projects are performed within two years from when the questionnaires were completed. Moreover, it is important to remember that the purpose of the study is to identify time-consuming activities and serve as a guideline for future research, rather than presenting the exact number of hours needed to carry out the activities. To increase the precision of the study some more samples would have been favourable to add.

Another factor that has been hard to determine in every specific project is the input data precision and quality. Consequently, it's hazardous to exclude the quality dimension's influence on time-consumption from the survey results. However, all projects managed to validate their models according to the real world system, which indicates that the data quality was satisfying in all cases. Many of the projects (73%) also validated the input data separately to production follow-ups or to process expert knowledge.

## 6 Conclusions

To summarise the findings from this study, some results deserve to be highlighted:

- The work to increase the support of input data to production analysis has not yet resulted in successful implementations in industry. The time needed for input data management in DES projects is still around 31% of the total project duration. Moreover, the percentage of companies

having all data available for DES projects is as low as 7%.

- The three most time-consuming input data activities are *data collection*, *mapping of available data*, and *data analysis and preparation*, respectively.
- One major reason for the heavy time-consumption is the need for manual gathering due to insufficient data availability.
- Another reason is the complex design of many computer based data systems, which slows down the identification of available data as well as the extraction of information from the systems.

There is also a newly published paper related to this contribution [15], which proposes a methodology for increased efficiency in input data management. It aims to improve the present working procedures (mapped above) by describing good practice guidelines for each activity.

## Acknowledgements

The funding for this research is granted by VINNOVA (Swedish Governmental Agency for Innovation Systems), and SSF (Swedish Foundation for Strategic Research). The authors would also like to thank all members of the DES projects included in the study. They shared a significant amount of time by taking part in the interviews. Additionally, Cecilia Berlin (PPU, Chalmers) has provided valuable suggestions for enhancing the presentation of this paper.

## References

- [1] H. Driva, K.S. Pawar, U. Menon: *Measuring product development performance in manufacturing organisations*. International Journal of Production Economics, vol. 63, 2000, pp. 147-159.
- [2] N.H. Robertson, T. Perera: *Feasibility for automatic data collection*. In: B.A. Peters, J.S. Smith, D.J. Medeiros, M.W. Rohrer (eds.): *Proceedings of the 2001 Winter Simulation Conference* (Washington, DC), IEEE, 2001, pp. 984-990.
- [3] J. Banks, J.S. Carson, B.L. Nelson: *Discrete-Event System Simulation* (2nd ed.), Prentice-Hall, 1996.
- [4] W. Trybula: *Building simulation models without data*. IEEE International Conference on Systems, Man, and Cybernetics. Humans, Information and Technology, vol. 1, 1994, pp. 209-214.
- [5] T. Perera, K. Liyanage: *Methodology for rapid identification and collection of input data in the simulation of manufacturing systems*. Simulation

Practice and Theory, vol. 7, 2000, pp. 645–656.

- [6] A.M. Law, W.D. Kelton: *Simulation Modelling and Analysis* (3rd ed.), McGraw-Hill, 2000.
- [7] M. Rabe, S. Spieckermann, S. Wenzel: *A New Procedure Model for Verification and Validation in Production and Logistics Simulation*. In: S.J. Mason, R.R. Hill, L. Mönch, O. Rose, T. Jefferson, J.W. Fowler, (eds.): *Proceedings of the 2008 Winter Simulation Conference* (Miami, Florida), IEEE, 2008, pp. 1717-1726.
- [8] T.H. Davenport, L. Prusak: *Working Knowledge: How Organisations Manage What They Know*, Harvard Business School Press, 1998.
- [9] R. Van der Spek, A. Spijkervet: *Knowledge Management: Dealing Intelligently with Knowledge*. In: J. Liebowitz, L. Wilcox (eds.): *Knowledge Management and its Integrative Elements*, CRC Press, 1997.
- [10] J. Bernhard, S. Wenzel: *Information Acquisition for Model Based Analysis of Large Logistics Networks*. In: Y. Merkurjev, R. Zobel, E. Kerckhoffs (eds.): *Proceedings of the 19<sup>th</sup> European Conference on Modelling and Simulation* (Riga, Latvia), ECMS, 2005, pp. 37-42.
- [11] Dassault Systèmes: *DELMIA* <http://www.3ds.com/products/delmia/>
- [12] SIEMENS: *Teamcenter* [http://www.plm.automation.siemens.com/en\\_us/products/teamcenter/](http://www.plm.automation.siemens.com/en_us/products/teamcenter/)
- [13] L.G. Randell, G.S. Bolmsjö: *Database Driven Factory Simulation: a Proof-of-concept Demonstrator*. In: B.A. Peters, J.S. Smith, D.J. Medeiros, M.W. Rohrer (eds.): *Proceedings of the 2001 Winter Simulation Conference* (Washington, DC), IEEE, 2001, pp. 977-983.
- [14] M. Denscombe: *The Good Research Guide – for small-scale social research projects*, Open University Press, 1998.
- [15] A. Skoogh, B. Johansson: *A Methodology for Input Data Management in Discrete Event Simulation Projects*. In: S.J. Mason, R. Hill, L. Moench, O. Rose (eds.): *Proceedings of the 2008 Winter Simulation Conference* (Miami, Florida), IEEE, 2008, pp. 1727-1735.
- [16] L.W. Schruben: *Establishing the Credibility of Simulation*. *Simulation*, vol. 34, 1980, pp. 101-105.
- [17] B. Johansson, J. Johnsson, A. Kinnander: *Information structure to support Discrete Event Simulation projects*. In: S. Chick, P.J. Sánchez, D. Ferrin, D.J. Morrice (eds.): *Proceedings of the 2003 Winter Simulation Conference* (New Orleans, Louisiana), IEEE, 2003, pp. 1290-1295.

**Corresponding author:** Anders Skoogh  
 Chalmers University of Technology.  
 Dept. of Product and Production Development  
 Hörsalsvägen 7A, 41296 Gothenburg, Sweden  
[anders.skoogh@chalmers.se](mailto:anders.skoogh@chalmers.se)