LANCASTER UNIVERSITY MANAGEMENT SCHOOL

WORKING PAPER NO: 2013:3

A SURVEY ON DATA IDENTIFICATION AND COLLECTION IN SIMULATION PROJECTS

Bhakti S. S. Onggo[1], James Hill[2], Roger J. Brooks[1]

1) Department of Management Science, Lancaster University Management School, Lancaster, LA1 4YX, United Kingdom

2) Accenture Ltd, Kingsley Hall, 20 Bailey Lane, Manchester Airport, Manchester, M90 4AN, United Kingdom

KEYWORDS

Data, Data Identification, Data Collection, Simulation, Project, Practice.

ABSTRACT

The research into the collection of data for use in simulation is lacking. This is rather unfortunate since data quality and availability are two of the most challenging issues in many simulation projects. We have conducted a pilot survey from simulation practitioners to understand the data collection process in simulation, its issues, solutions and impact on project outcomes. The result reveals interesting insights. Some of them confirm what we believe to be happening in practice. A few of them contradict what we may have assumed to be happening in practice.

For citation, please use:

Onggo, B.S.S., Hill J. and Brooks, R.J. (2013) 'A Survey on Data Identification and Collection in Simulation Projects', Proceedings of the 27th European Simulation and Modelling Conference, 23-25 October 2010, Lancaster, UK (ESM13). Ostend, Belgium: Eurosis-ETI.

INTRODUCTION

In the current business world where competition has become tougher, there is high demand for effective and useable simulation models to aid business decisions. Researchers have identified factors that affect the success of a simulation project (for example Robinson and Pidd, 1998). It is widely accepted that one of the main issues that has affected many simulation projects is the inefficient data collection (Perera and Liyanage, 2000; Trybula, 1994; Hill and Onggo, 2012).

This report presents the finding from our pilot survey that seeks to find out more about data problems faced by modellers in a simulation project, how they handle the issues and what the impact of data quality issues have on their projects.

METHODOLOGY

We used an on-line questionnaire to collect data from simulation modelers. After we designed the questionnaire, we tested the questionnaire in two stages. In the first stage, the questionnaire was tested by two PhD students who had been working in simulation modeling. The refined version was then tested by 10 simulation modelers. Based on the feedback, we constructed the final version of the questionnaire.

The questions in the survey were divided into four parts: (1) characteristics of respondents, (2) characteristics of the models, (3) characteristics of the projects and project related issues and (4) data collection and data issues. The questionnaire is given in the appendix.

SURVEY RESULT

The questionnaire was constructed using Qualtrics™ and the link was distributed to a number of simulation practitioners (industry and academics) through personal contact and a limited number of LinkedIn groups. At this stage, we wanted to ascertain that the questionnaire could help us achieve our research objectives, i.e. to find out typical data problems faced by simulation modelers, how simulation modellers handled the data problems and what impact data problems had on their projects. Hence, we limited our data collection period to three weeks between February and March 2012. We received 39 responses. After we validated the answers, we had to remove three invalid responses.

The characteristics of respondents

Practitioners from the industry and academics are represented equally in our survey (top chart in Figure 1). The majority of them develop simulation models for a client, either internally, externally or both (bottom chart in Figure 1). Most of them are experienced modellers with 11 years of experience on average. The distribution of the years of experience is shown in Figure 2. This is consistent with the number of models that they have developed (Figure 3). On average, each of them has been involved in 19 simulation projects.

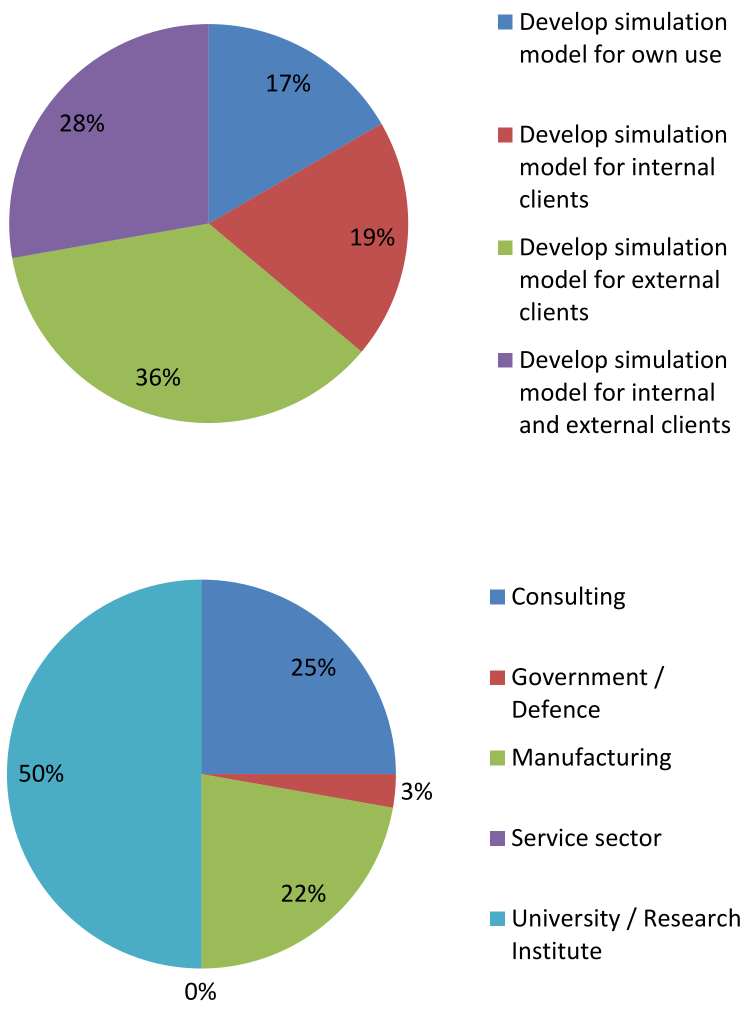

Figure 1: Characteristics of respondents

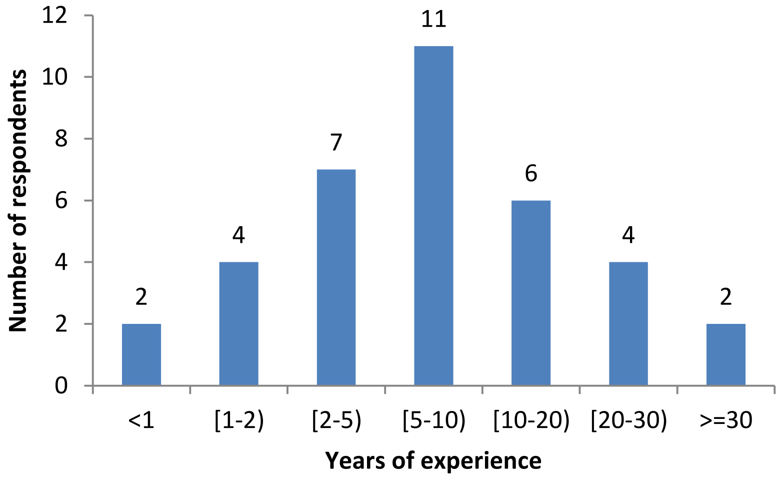

Figure 2: Years of experience of respondents

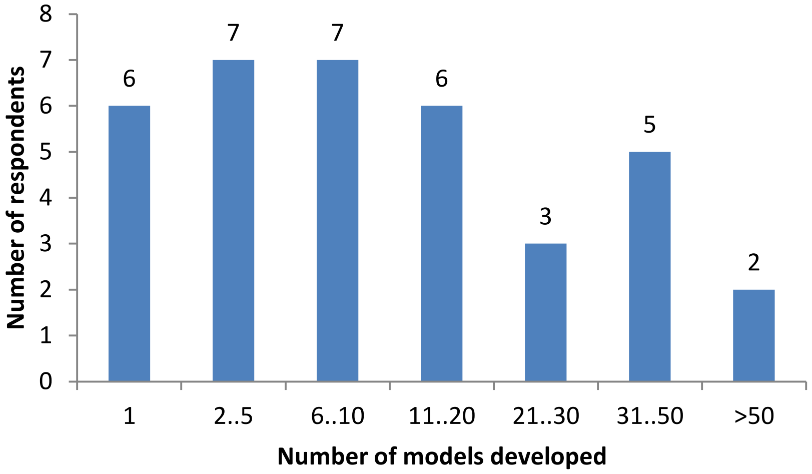

Figure 3: The number of models developed

The characteristics of projects

The projects in which our respondents have been involved are typically done in a team and the average team size is 3 or 4 people (Figure 4). The average project duration is around 8 months (Figure 5). If we look at the result further, the average project durations among industry practitioners and academics are 5 months and 10 months, respectively. There might be a number of reasons why the average project duration is shorter among industry practitioners. For example, many academics probably carry out consultation project outside their main duty at the university. However, we do not have any empirical data that explains the difference.

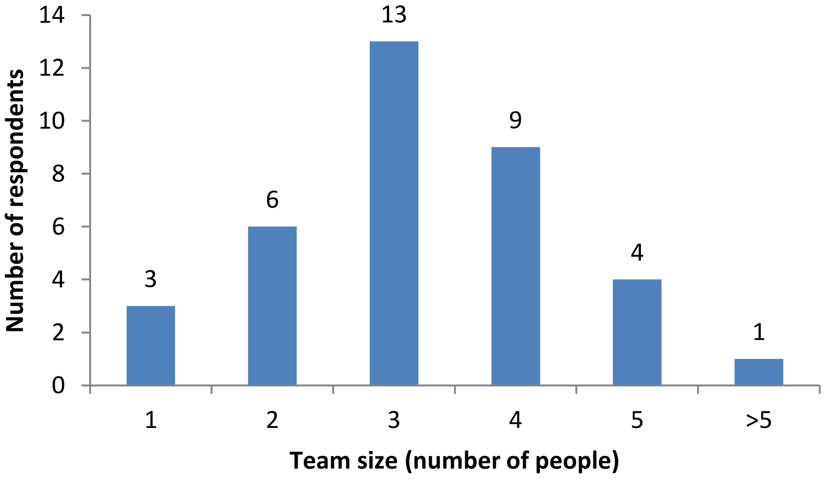

Figure 4: The team size of typical simulation projects

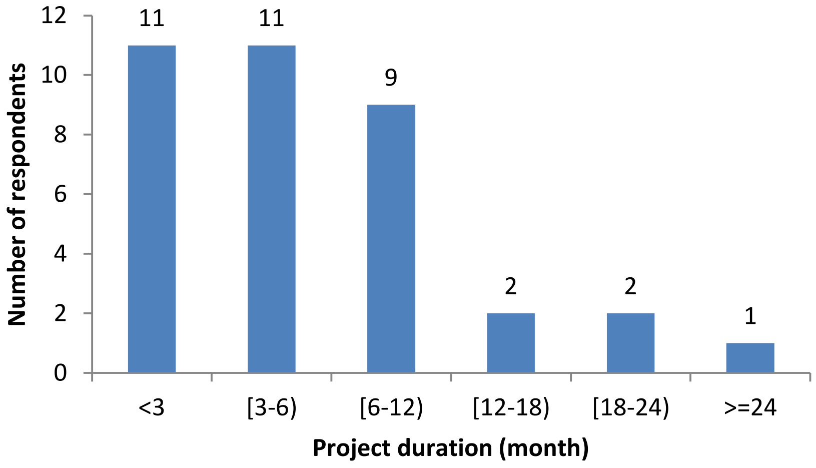

Figure 5: The typical project durations

The majority of the respondents (34 out of 36) have used discrete-event simulation in their projects (Figure 6). This indicates that most practitioners are familiar with discrete-event simulation. The result also shows that amongst 8 respondents who have developed agent-based simulation models, 6 of them are academics. This may suggest that the adoption of agent-based simulation among industry practitioners is relatively low. Figure 7 shows that almost 70% of the respondents have used one simulation modelling paradigm only. It is interesting that 30% of the respondents have used two paradigms or more and they include both industry practitioners and academics. This result shows that it is not true that only academics have used two paradigms or more.

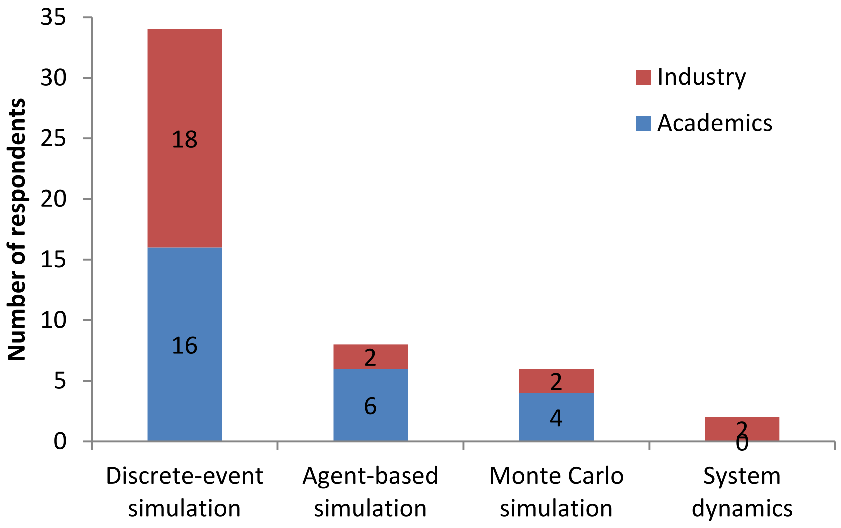

Figure 6: Simulation paradigms used

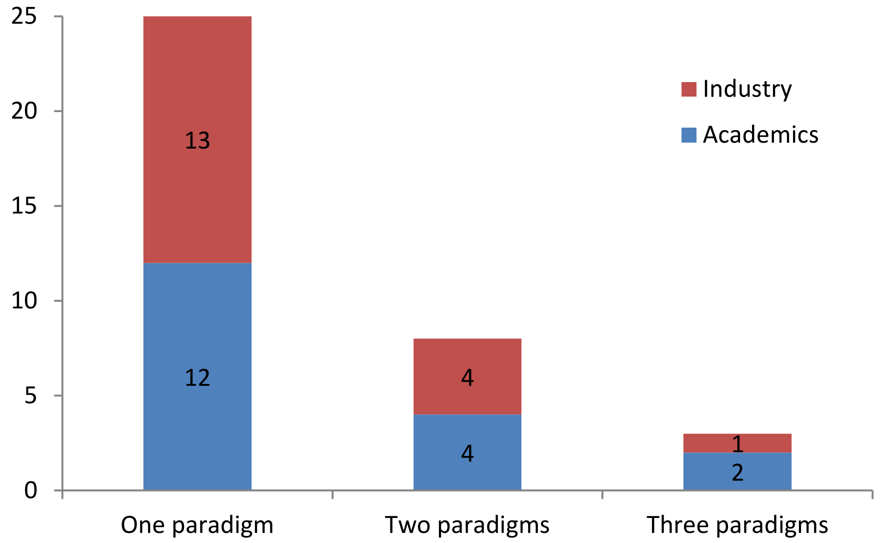

Figure 7: Number of paradigms used

The survey shows that around 90% of the models are used at least once (Figure 8). It is also interesting to see that 60% of the models have been used more than once. This result shows that the belief about most simulation models are ad-hoc and only used once is likely to be incorrect.

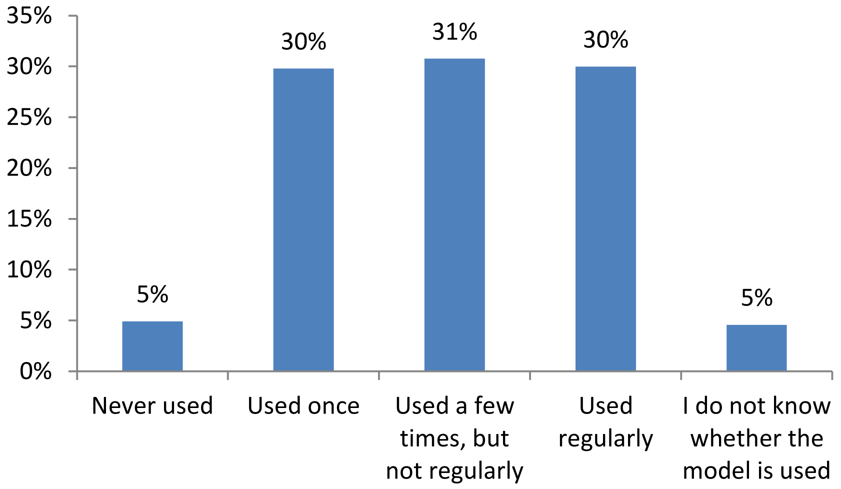

Figure 8: Model usage

For projects involving a client (i.e. we exclude model for personal use), we would like to know more about the client's involvement in the project. The result is shown in Table 1. This result confirms that clients are more involved in the early stage of a simulation project, i.e. setting problem definition and objectives. The involvement of client is less in the model design stage but they are still involved. The result for the data collection and analysis give an indication that both modelers and clients are actively involved in this project stage. As expected, clients' are not usually involved in the model development stage. However, the clients are usually involved in the verification and validation of the model. Similarly, the clients are usually interested in the output analysis and experimentation. Hence, they are likely to be involved.

Table 1: Client's involvement in projects

Stages	None	A little	A lot	Most of the work
Problem definition and setting objectives	0	4	19	7
Deciding model design and content	4	20	5	1
Data collection and analysis	4	12	13	1
Model development	17	9	4	0
Model verification and validation	3	15	11	1
Output analysis and experimentation	2	14	14	0

Data collection in simulation projects

We have dedicated one section in the questionnaire on data problems in simulation project. First, we ask the respondents if data problems are common in their simulation projects. Specifically, we ask them to rank five problems that we have experienced in our past projects with the most occurring problem taking rank 1. The average rank is shown in the second column of Table 2. Most respondents agree that data problems are the most common in simulation projects. This is followed by the high complexity of the systems being modelled and the lack of clear objectives which may lead to incorrect problem definition. It should be noted that most of our respondents are an experienced modeller; hence most of them agree that technical skill is not an issue in comparison to the other four problems.

As expected, the respondents agree that the lack of clear objectives have the most significant impact on project performance (column 3 on Table 2). This is consistent with a well-known knowledge that mistakes in the early stage of a project are likely to be more costly than those found in the later stages. Interestingly, most respondents agree that the next problem that has a serious impact on project performance is data problems. It supports the motivation of our research into the data identification and collection in simulation because most respondents agree that data problems are common and they have a significant impact on project performance.

Table 2: Issues in simulation project – frequency and impact

Stages	Frequency	Impact
Data problems	2.1	2.4
High system complexity	2.3	3.1
Incorrect problem definition and lack of clear objectives	2.5	2.0
Project management issues	3.6	3.6
Technical skill of the project team	4.6	3.9

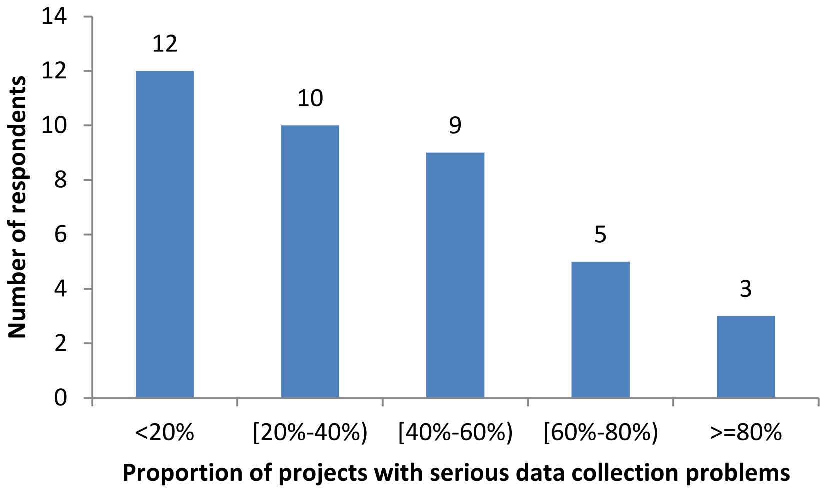

Figure 9: Proportion of projects with serious data problems

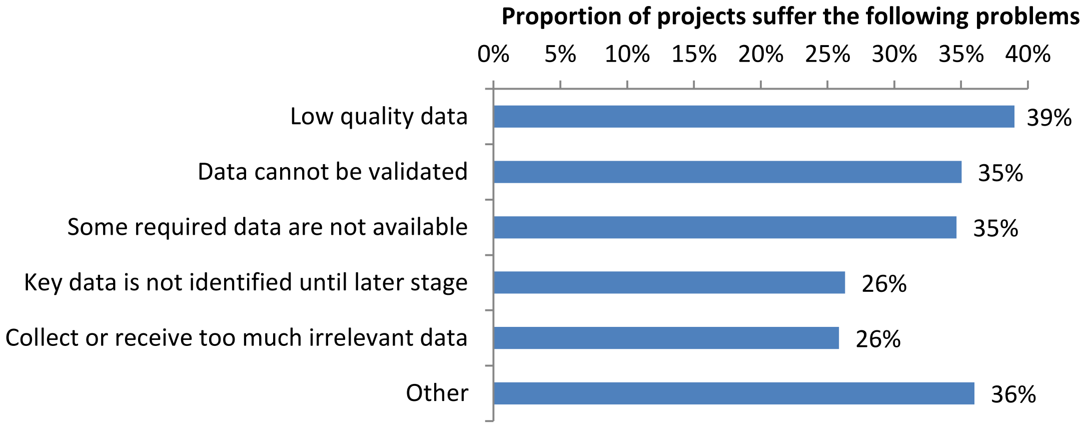

Figure 10: Proportion of projects with specific data problems

We ask our respondent how often they have experienced serious data problems. The result is shown in Figure 9. On average, 43% of the projects have had serious problems during the data identification and collection. From our experience, we have identified five common data problems: quality, availability, validation, relevance and late identification. Our respondents confirm that they have also experienced the same problems. Some respondents have also added problems such as clients changing the data several times, clients often not understanding their data and clients not giving enough priority to giving the data to modellers. The detailed result can be seen from Figure 10.

Since data problems are common and have a significant impact on simulation projects, we are interested in understanding how modellers handle the data problems. First, we focus on projects that involve clients. Based on our experience we have identified a few actions that we have done in Table 3. We ask the respondents if they have done the same actions and let us know if they have other actions not listed in the table. One respondent mentions the possibility of finding data from published literature. The result shows that all respondents have done the listed actions. Some actions are done more often than others. For example, requesting more data from the client, editing/cleansing data, using client's estimate and validating data with clients are done relatively more frequently than the remaining actions.

Table 3: Actions in response to data problems (excluding model for own use)

Actions in response to data problems	Never	Rarely	Sometimes	Often	Most occasions
Request more data from client	0	3	9	11	6
Collect data yourself	1	9	12	5	3
Edit or clean data	0	1	5	16	7
Use client estimate	1	6	9	11	3
Use modeller estimate	3	4	15	7	1
Ask client to validate data	1	3	12	10	4
Validate data yourself	2	9	10	7	2
Validate data with client	1	2	10	14	3
Other	0	1	0	0	1

Although the number of respondents who develop model for personal use is very limited in our sample, it is still useful to gather initial information on how they handle data problems. The result is shown in Table 4. They often collect more data, edit/cleanse the data or use own estimate to deal with the data problems. However, given the very limited number of samples, we cannot make any general conclusion from this result.

Table 4: Actions in response to data problems when the model is for own use

Actions in response to data problems	Never	Rarely	Sometimes	Often	Most occasions
Collect more data	0	1	3	2	0
Edit or clean data	0	2	2	2	0
Use own estimate	0	0	2	2	2
Validate the data	1	0	0	0	0
Other	0	0	0	0	0

We are also interested to know the impact of data problems on simulation project. We have listed five possible consequences as shown in Figure 11. The respondents agree that data problems are likely to cause a project to be delayed. This is followed by the reduced confidence in the model and the limitations in carrying out experiments. It is interesting to know that some projects and models have to be abandoned due to the data problems. This may indicate the seriousness of the impact of data problems on simulation projects.

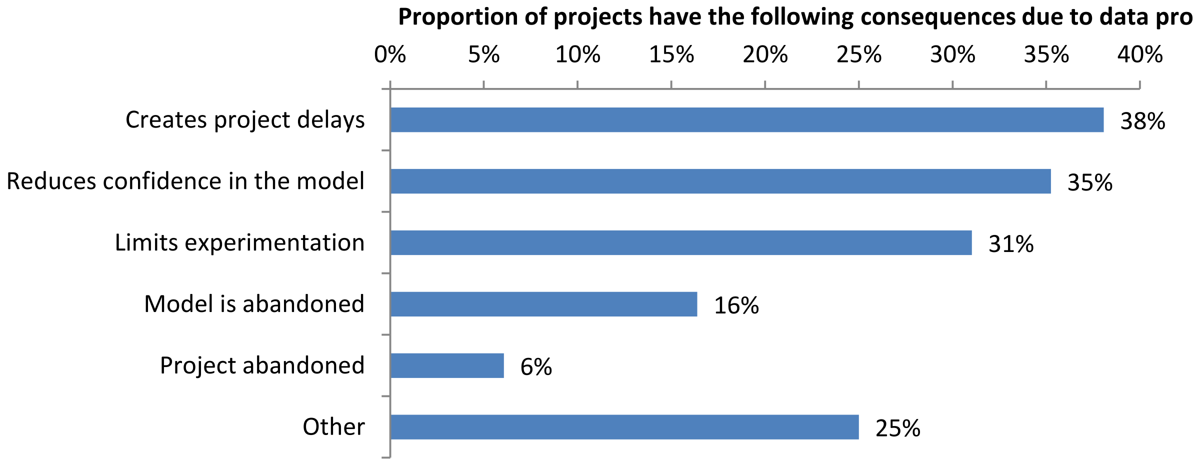

Figure 11: Impact of data problems in simulation projects

Trybula (1994) stated that up to 40% of project time was spent on data gathering and data validation. We ask our respondents the percentage of time spent on data identification and collection in a typical simulation project. The distribution is shown in Figure 12. On average, the time spent on data identification and collection is 32%. The result is not significantly different from one reported in Trybula (1994) but our result shows that cases where more than 40% of project time is spent for data collection may happen more than what we expect.

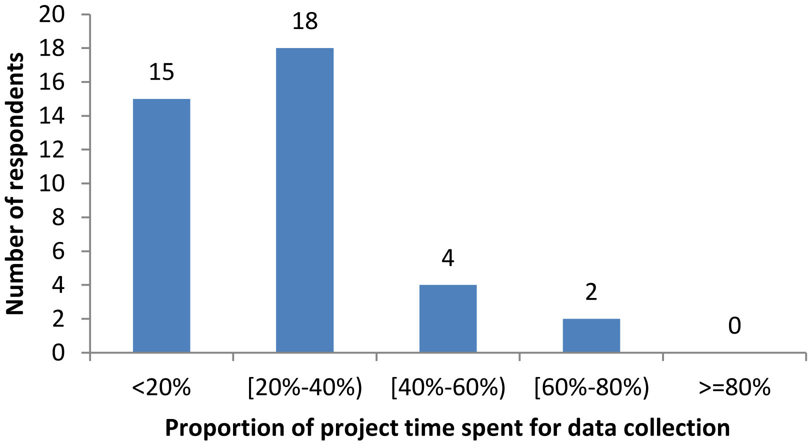

Figure 12: Distribution of project time spend on data identification and collection

The comparison between Trybula's finding in 1994 and ours may give an indication that we have not done enough to make the data collection stage more effective and efficient since the result has not significantly changed since 1994. We further ask our respondents whether they have come across best practice document for data collection in simulation project. Half of them have not seen one (Figure 13). Since our respondents have been in the industry for an average of 11 years, this confirms our observation that research into data collection in simulation has been lacking.

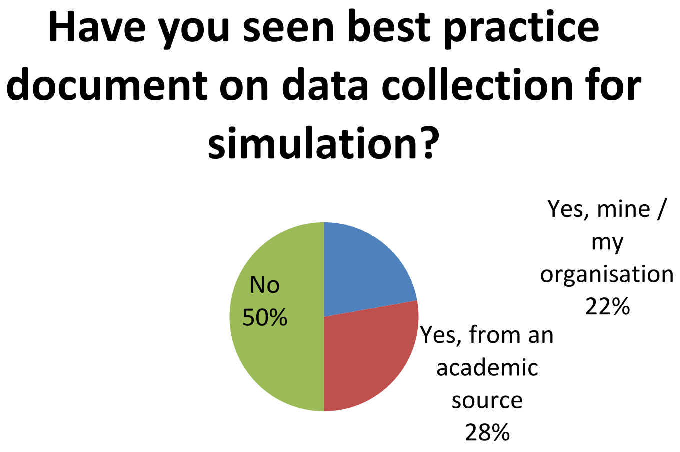

Figure 13: Best practice document on data collection for simulation

Finally, on hindsight, close to 40% of our respondents think that they spent more time than necessary on data collection (Figure 14). For those who think they have spent more time than necessary, on average they think 20% of the project time have been wasted unnecessarily (presumably by unexpected serious data problems).

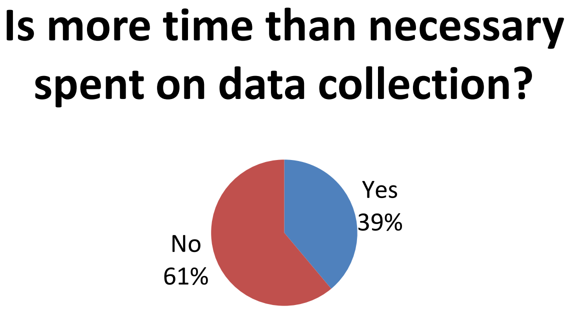

Figure 14: Estimated time spent on data collection

SUMMARY

This survey has been designed as a pilot. Hence, it was aimed to get some feedback and initial understanding on simulation modelling practice from representative samples (the industry and academics with varying modelling experience). We have limited the number of samples with an intention to carry out a larger scale study later. Hence, some of the findings cannot be generalized for all practitioners. Nevertheless, our results have shown evidences that:

- research into data collection in simulation is lacking
- data problems is common and have a significant impact on project
- data identification and collection requires a significant portion of project time
- significant majority of modelers work as a team in a simulation project
- most modelers have used discrete-event simulation
- the number of modelers who use one simulation paradigm is significantly more than those who use two or more paradigms
- it is not true that only academics have used two simulation paradigms or more
- although significant number of models are used once, the number of models that are used more than once are significantly more; hence, it is not true that most simulation models are used only once

The result of this survey is useful because the result gives us an insight into what has been happening in practice in relation to data collection in simulation and simulation projects in general. In the future, we plan to carry out a larger scale study on the same topic based on the feedback and knowledge obtained from this pilot survey.

REFERENCES

Hill J and Onggo B S S. 2012. Data identification and collection methodology in a simulation project: An action research. In: Tjahjono B, Heavey C, Onggo B S S and van der Zee D-J (eds). Proceedings of the Operational Research Society Simulation Workshop. Operational Research Society: Birmingham, UK.

Perera T and Liyanage K. 2000. Methodology for rapid identification and collection of input data in the simulation of manufacturing systems. Simulation Practice and Theory, 7, 645-656.

Robinson S and Pidd M (1998). Provider and customer expectations of successful simulation projects. Journal of Operational Research Society, 49(3): 200–209.

Trybula W J. 1994. Building simulation models without data. Proceedings of the IEEE International Conference of Systems, Man and Cybernetics, pp. 209-214.