Lessons Learnt Undertaking a Large-scale Systematic Literature Review

Mark Turner
School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG.
m.turner@cs.keele.ac.uk

Barbara Kitchenham
School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG.
B.A.Kitchenham@cs.keele.ac.uk

David Budgen
Department of Computer Science, Durham University, Durham, Science Laboratories, South Road, Durham City, DH1 3LE, UK.
david.budgen@durham.ac.uk

Pearl Brereton
School of Computing and Mathematics, Keele University, Keele, Staffordshire ST5 5BG, UK
o.p.brereton@cs.keele.ac.uk

Abstract. We have recently undertaken a large-scale Systematic Literature Review (SLR) of a research question concerning the Technology Acceptance Model (TAM). At the end of the study, we observed some anomalies during the analysis of the extracted data. In our attempts to identify the cause of the anomalies, we found a number of mistakes that had been made during the data extraction process. We discuss each of the mistakes in terms of why they occurred and how they might have been avoided. We suggest a number of ways in which the available guidelines for conducting SLRs should be amended to help avoid such problems occurring in future reviews.

Keywords: lessons learnt, systematic literature reviews

1. INTRODUCTION

Currently most large-scale systematic literature reviews (SLRs) in Software Engineering and IT have been mapping studies. A mapping study is a review that concentrates on identifying research trends and categorising research articles rather than addressing specific research questions. Kitchenham *et al.* [14] identified 20 SLRs published between 2004 and June 2007. Mapping studies were based on far more primary studies (i.e. between 63 and 1485) than SLRs that address more specific research questions (i.e. between 6 and 54).

In mapping studies, one of the major problems with data extraction and aggregation is determining an appropriate method of classifying the primary studies. As described by Jørgensen and Shepperd [10], the process of determining an appropriate classification is usually a bottom-up process by which the researchers read some of the primary studies, specify some categories based on the papers they have read and their own experience of the domain, then read some more papers and refine the categories as necessary.

In contrast, an SLR that addresses a specific research question should base its data extraction and aggregation on the specific research question (or questions). The major problems associated with data extraction and aggregation processes are therefore less generic and are

more related to the specific research question and the nature of the primary studies. However, we believe that a review of problems found during the conduct of a research question-oriented SLR can usefully identify some more general issues that might help other researchers to avoid repeating the mistakes of others. For this reason, we discuss the problems that arose during the process of data extraction and aggregation in a large SLR involving over 70 primary studies that addressed a specific research question related to the Technology Acceptance Model (TAM). We have discussed previously some of the problems we encountered undertaking this systematic literature review [1]. However, this paper concentrates on issues that were only detected late in the SLR process (i.e. during report writing) but which were introduced during the data extraction and data aggregation process.

Section 2 describes the background to our SLR. Section 3 discusses how the SLR was organized by the research team identifying some external factors that influenced the conduct of the SLR. Section 4 describes the data extraction problems we found late in the SLR process, we discuss each of the problems in terms of the underlying causes of the problem and how it might have been avoided. We discuss our observation in Section 5.

2. BACKGROUND

In this section we introduce the TAM model and explain the rationale for the research question we addressed in our SLR.

2.1 The Technology Acceptance Model

The Technology Acceptance Model (TAM) was proposed by Davis [2] and Davis *et al.* [3] as an instrument to predict the likelihood of a new technology being adopted within a group or an organisation. It is based on the theory of reasoned action (TRA) [7] and suggests that technology acceptance and use can be explained in terms of a user's internal beliefs, attitudes and intentions. As a result it should be possible to predict future technology use by applying the TAM at the point when the technology is introduced. The original TAM assessed the impact of four internal variables upon the *actual use* of the technology. The internal variables in the original TAM are: *perceived ease of use* (PEU), *perceived usefulness* (PU), *attitude toward use* (A), and *behavioural intention to use* (BI). Figure 1 illustrates the original TAM model.

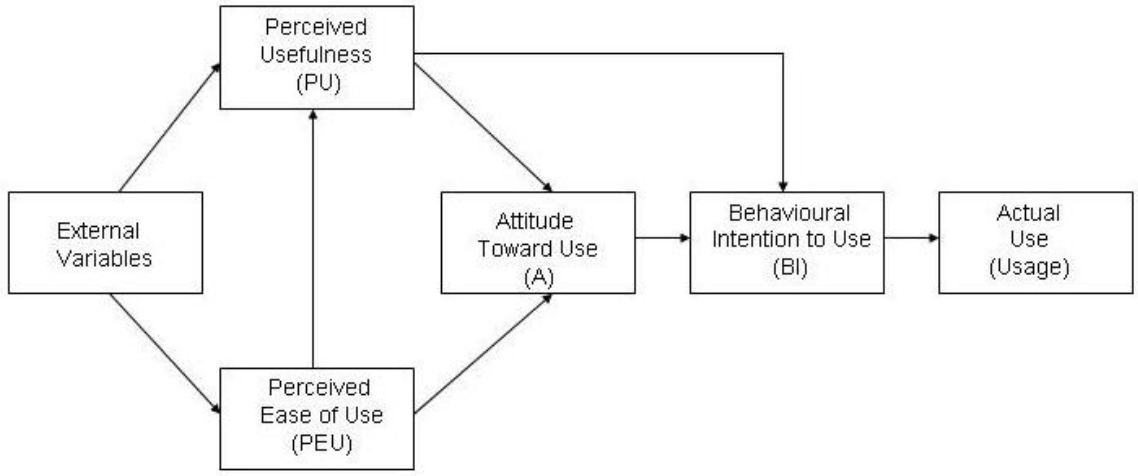

Figure 1 – The Original TAM Model

Venkatesh and Davis [21] subsequently proposed an extension of the TAM, referred to as TAM2, which dropped *attitude towards use* and incorporated additional variables such as *experience* and *subjective norm*. However, the basic model remained unchanged.

The variables within the TAM are typically measured using a short, multiple-item questionnaire. As an example, Figure 2 shows the questions used to measure the three TAM variables, PU, PEU and BI. When included, *actual usage* is usually measured in a similar way

through self-reported variables. Typical self-reported usage measures include frequency of use and intensity of use. Since its introduction, the TAM and its revisions have been applied to a variety of technologies, such as text editors [2], business intranets [9], and the Web [5]. Whenever the TAM is evaluated for reliability (i.e. internal consistency) and validity, it scores very highly [2, 20, 8]. Reliability, in this context, measures how strongly the set of questions related to a particular construct correlated with one another in contrast with their correlations with questions related to other constructs. Validity is usually assessed in terms of how well the relationships indicated in the model are confirmed by data analysis. In this context the analysis is usually based on an analysis of the partial correlations related to the specific model e.g. the PU construct should be associated with the Attitude construct, the Behavioral Intention construct which in turn relate to the Actual Usage construct. In terms of simple correlations this means that there need not be a direct correlation between PU and Actual Usage for the model to be regarded as valid. Recently King and He [12] undertook a major meta-analysis of the TAM and concluded that it was a valid and robust model.

The basic TAM Questionnaire

The specific name of the technology (e.g. the intranet) would replace "the technology" in a specific questionnaire.

Questions are measured on a Likert-like scale.

Perceived Usefulness Questions (PU)

Using the technology would improve my performance in doing my job
Using the technology at work would improve my productivity
Using the technology would enhance my effectiveness in my job
I would find the technology useful in my job

Perceived Ease of Use Questions (PEU)

Learning to operate the technology would be easy for me
I would find it easy to get the technology to do what I want it to do
It would be easy for me to become skilful in the use of the technology
I would find the technology easy to use

Behavioural Intention to use (BI)

I presently intend to use the technology regularly at work

Figure 2 – The Basic TAM Questionnaire

2.2 A Problem with the TAM

Although the TAM has been validated in many studies, the *actual usage* of the technology, as opposed to the *behavioural intention to use*, is rarely monitored. Furthermore, this may lead to a mistaken belief in the acceptability of a technology. For example, a study conducted by Keung *et al.* [11] found that the TAM predicted that a particular technology was likely to be adopted within the company in question. However, a year later the authors found that the technology was not being used. The TAM was re-applied at the later time and the results from this study were different from the initial TAM assessment. Therefore, there is a question as to whether the TAM can act as an accurate predictor of *actual usage* rather than *behavioural intention to use*.

If the TAM is *not* an accurate predictor of actual usage, then there is an obvious problem if organisations rely on the positive results from applying the TAM to justify the introduction of new technologies. This may be even more of a problem if the TAM is used to assess the acceptability

of a technology during the very early stages of design as proposed recently by Davis and Venkatesh [4].

This issue was also raised in an earlier review of the TAM undertaken by Legris *et al.* [17] which compared those studies that evaluated the TAM against *actual usage* with those that evaluated it against *behavioural intention to use*. In this study they found a direct relationship between *Perceived Usefulness* (PU) and *actual usage* in 8 out of the 13 studies in which it was tested, and a direct relationship between *Perceived Ease of Use* (PEU) and *actual usage (U)* in 4 out of the 9 studies in which it was measured. *Behavioral Intention to Use* (BI) was related to U in 10 of the 11 studies in which it the relationship was tested. Furthermore, in Legris *et al.*'s study all but one of the primary studies that measured actual usage employed self-reported usage rather than objective measures of actual usage. Straub *et al.* [19] investigated the relationship between the two types of actual usage measure and reported that self-reported measures of TAM variables (such as PU and PEU) appear to be related to self-reported measures of actual usage but show a much weaker relationship with objective measures of actual usage.

2.3 The Rationale for our SLR

The issues discussed in Section 2.2 raised questions concerning the validity of the TAM with respect to the relationship of TAM constructs with actual usage and whether any relationships depended on the type of actual usage measure (subjective or objective). Since the TAM is being used as a means for research students to validate their methods/tools [11] and a means of evaluating software technology during development [4], we concluded that it would be useful to undertake an extension of Legris *et al.*'s [17] study with the aim of investigating both the relationships between the constructs PU, PEU and BI, and actual usage and whether the type of actual usage influenced the relationship.

3. ORGANIZATION OF THE SLR

When we started the review, the review team comprised four senior researchers located at Keele University, and two research assistants (RAs) both located at Keele University. The RAs were both relative novices both to the SLR process and the topic. The SLR was conducted according to the process documented by Kitchenham [13]. In particular a planning phase was undertaken during which the study protocol was specified, and the protocol itself was reviewed by researchers external to the study team.

- However, some external events occurred that contributed to subsequent aggregation and analysis problems:
- During protocol development, one of the senior researchers (Professor Budgen) moved to Durham University.
- In a similar time frame one of the original research assistants (Dr Khalil) left the project to take up another position.

As a result of these events a new research assistant (RA) was appointed to the project who was located at Durham. This meant that the protocol construction was interrupted and the two RAs responsible for most of the conduct of the SLR were no longer located at the same university. In addition, the newly appointed RA had not been involved with some of the initial discussion associated with the SLR.

The two RAs were responsible for undertaking the search process which was organised as follows:

- Each RA was allocated a selection of digital libraries to search and developed search strings appropriate to that library
- The abstract and title of each study identified as a candidate primary study by the search process was reviewed by the relevant RA and papers that were clearly irrelevant were rejected.
- The full version of all the remaining papers were obtained and assessed against the specific inclusion/exclusion criteria. When there was any doubt about including a paper the RAs consulted one of the other members of the team.

- The references of all agreed primary studies were searched for other relevant papers.
- The protocol identified the need to manage multiple studies in a single paper and the data extraction process was designed to address this issue.
- The protocol also identified the need to remove duplicated studies but no specific instructions were included as to how they should be identified.

The data extraction process was organised as an extractor/checker process. The RA responsible for a specific primary study extracted the data and another member of the review team checked the extraction form.

We kept records of the deviations from the protocol which involved changes to the data extraction forms and data aggregation tables. These changes were necessitated because it became clear that the data available from different primary studies varied substantially from study to study. When we prototyped our data extraction and data aggregation forms we had found that:

- There might be multiple studies reported in a single paper.
- There might be multiple tests of the TAM variables at different times relative to the introduction of the technology or the first use of the TAM.
- We were interested in relationships between PU, PEU and BI and actual usage but not every paper reported every combination.
- Some of the studies reported the results of extensions to the TAM model in terms of the significance of the model parameters without reporting the basic correlations among all the variables.

These conditions were catered for in the data extraction form, but when the full extraction activity started it soon became clear that there were other issues that needed to be addressed because the statistical analysis used in the TAM papers was very sophisticated and the RAs and some of the checkers were not familiar with the implications of the statistical analysis and reporting conventions. We held a project meeting to discuss the statistical analysis and the way in which data was reported. In particular, it became clear that data extraction needed to be refined to cater for studies investigating extensions to the TAM model. When additional variables are added to the model, some of the original model relationships might no longer appear significant, although the relationship would be significant if the additional variables were excluded. We, therefore, kept a record of occasions where a relationship between PU, PEU and BI and actual usage was not observed in the model parameters but additional variables were included in the model.

4 EXTRACTION AND ANALYSIS PROBLEMS

After the data extraction and aggregation process was completed, we began to encounter difficulties. The data was analysed using a vote counting process and the results were analysed to investigate whether the likelihood of finding positive relationships between PU and PEU and actual usage was affected by other factors e.g. whether the technology was mandatory, whether the technology had been in use for some time prior to the TAM questionnaire being administered, whether the presence of other variables and whether the sample size affected relationships.

At first we assumed the analysis was correct and sent our report to researchers external to the project for a review. One of the reviewers asked us to include the details of the logistic regressions performed to assess the impact other factors. The original logistic regressions had been done by one of the team (Kitchenham) based on extractions of the aggregated data provided by the RAs. While re-doing the logistic regressions with the data from the Technical Report, we obtained slightly different results. Attempts to pin down the reasons for these differences identified a number of problems:

- Some duplicated primary studies had been included.
- Some data extractions were incorrect.

- In some cases the two RAs had made different assumptions about how to deal with different types of primary study results when extracting data.

Below we discuss the specific problems, how we identified them, the underlying cause(s) of the problem and process refinements that would have avoided them.

4.1 Duplicated papers

When we reviewed the sample size data we found examples of studies with the same sample size. Further investigation identified that in some cases the papers reporting the studies had the same authors. A review of the primary studies then identified several of the papers as duplicate reports.

The RAs checked for duplicate papers in the lists they extracted from each digital library, so it is worth examining what went wrong. Two duplicates were among the additional papers that were found by checking the references of the primary studies. Other duplicates were found among the initial selection of papers. The causes for failing to find the duplicates were:

- Most duplicates were not obvious until the sample sizes were reviewed as well as the authorship. Papers had different titles and were published in different journals.
- The search and data extraction was split so the RAs were responsible for different parts of the overall search and no overall process was put in place to ensure that the different data sets obtained from each search were properly cross-checked at a more detailed level than authors, title and journal.
- The RAs used different coding styles to refer to individual primary studies which hindered the identification of duplicate papers.

The main problem here was that we had not specified a procedure to check for duplicates other than finding exactly the same paper by searching different digital libraries. We suggest the need for each protocol to detail the procedures that will be put in place to detect papers that report the same study:

- If the task of organising a large-scale research question-based SLR necessitates splitting the tasks among different researchers, the split should ensure that all papers by the same authors are handled by a single researcher.
- If primary studies are identified by codes, the coding standards must be set out in the protocol to ensure they are used consistently.
- Where possible some distinguishing feature(s) of each study should be recorded e.g. sample size. Kitchenham *et al.* [16] also needed to review sample sizes to detect studies that used the same data sets even though their SLR only included 10 studies.

It would also be helpful, if authors of papers that reported the same study made the duplication obvious. For example, it is customary to present a restricted report of a study in a conference paper and then an extended report in a journal. It would help the identification of duplicate studies, if authors kept the titles of such papers as similar as possible and referenced any related papers in subsequent papers.

With respect to extracting data from duplicate studies, standards advise using the most complete or the latest report [13, 15]. In fact we found it necessary to use different reports of the same study to extract all the necessary data. However, when the same data are reported in different reports, data must be extracted from the most recent one.

4.2 Incorrect data

When we started to review the extracted data after detecting duplicate papers, we found one case where data had been extracted from the covariance matrix not the correlation matrix. This happened in spite of the data being checked by another member of the research team. The extraction problem arose because the RAs were not experienced with statistics. It is likely that it was missed because the checker only confirmed that the value in the extraction form was found in the paper, not whether the correct data item had been extracted.

Therefore, if the data being extracted are the outcome of some mathematical or statistical analysis and the extraction form is being checked by a different person, it is important to ensure that checkers understand the meaning of the data (i.e. how the data items are generated), or, at least, are aware that they need to check that the correct data have been extracted rather than simply checking that the figures in the extraction form match those in the paper. This should be set out clearly in the protocol so that checkers are aware of their responsibilities.

4.3 Inconsistent data extraction

When we checked some of the extracted data, we found that in some cases the structure of the study was more complex than the structure assumed by the data extraction form. There were two situations that the data extraction form did not cater for:

- In some cases, data from two separate sources was available and some relevant information was available from analyses performed on each data source separately, whereas some relevant information was available only from analyses based on the joint dataset. This type of study was sometimes treated as a single study, sometimes as two studies. This caused problems when analysing the effect of additional factors when the joint data set was mixed with respect to those factors (for example, in one source the technology was mandatory and in the other the technology was not mandatory). It appeared that the two RAs who performed the data extraction took different views as to how many studies to consider and how to assess the value of the additional factor variables for joint data sets.
- In some studies actual usage was measured in several different ways. In some cases the different usage measures were accumulated into a single metric for subsequent correlation analysis. This was the assumption made in the data extraction form. However, in other cases the TAM variables were correlated with each measure separately, leading to multiple tests per study; in this case the two RAs handled the data extraction differently.

In the problematic cases, the extraction results were not questioned by the checkers who simply confirmed that the data values in the extraction form could be found in the paper. This has two implications for the SLR process:

1. If a situation occurs when design of a study and/or the data analysis is more complex than is assumed in the data extraction form, the data extraction process must be halted until the problem is resolved. Furthermore, the resolution must be documented and all staff involved in the data extraction process need to be informed of the changes.
2. The use of an extractor and checker seems to be a means of reducing some of the overheads associated with performing an SLR, however, there is clearly a problem with this approach. We advise researchers undertaking an SLR to use two independent extractors whenever possible. The SLR guidelines [15] should be adapted to reflect this advice.

5. DISCUSSION

This paper has reported some of the detailed problems we encountered when undertaking a systematic literature review of the technology acceptance model. As with any lessons learnt paper, or a more formal case study, it is necessary to consider whether the conclusions drawn from the study are trustworthy [22, 18]. In particular, we need to consider whether there are other explanations for the events we observed (i.e. internal validity), and the extent to which the events are likely to affect other SLRs (i.e. external validity).

With respect to other influencing factors, some of the problems may have been influenced by changes to the research team personnel and location discussed in Section 3. It is also likely that our problems were exacerbated by lack of expertise. In the case of the TAM, the RAs responsible for the day-to-day research lacked practical experience of:

- The TAM itself.
- Statistical methods, in general, and the sophisticated methods used in TAM analysis in particular.
- The SLR process.

Furthermore the team member with most experience of statistical issues (Kitchenham) worked part-time, and was not always available to provide guidance.

We are currently investigating the problems faced by novices in a more formal case study as part of our EPSRC Evidence-based Practices Informing Computing (EPIC) project. Our initial results support the view that the SLR method is hard for novices to adopt. We believe the method is of value to novice researchers as well as more experienced researchers but it appears that supervisors or lead researchers need to provide a good deal of support and help when students and RAs first start to use the method.

With respect to the value of our observations to other SLRs, it is clear that many of our problems were influenced by the TAM itself, in particular, the use of extremely complex statistical analyses. Currently, we would not expect to have as many primary studies in a research-question based software engineering SLR, nor would we expect to see such sophisticated and complex statistical analyses in the primary studies. However, within the cost estimation field we certainly see the use of many different types of data analysis often involving fairly complex data preparation procedures, which require aggregation. Furthermore, if empirical methods become more central to software engineering we would expect the number of relevant primary studies included in SLRs to increase.

Thus, we believe that the problems we observed do have relevance to software engineering research, and software engineering researchers should be concerned about improving the SLR process by incorporating more detailed guidelines about the data extraction process. In addition, our recommendations should also help protocol reviewers to spot potential problems during the planning stage.

In addition, our previous suggestion of using an extractor and a checker [1] is definitely unsound for SLRs where the data extraction process is particularly complex. Furthermore, its value has not been proven for simple SLRs. We suggest researcher teams revert to the use of two independent extractors and single researchers (such as post-graduate students) use the procedure of re-extracting the data from a subset of papers and checking for internal consistency [6].

REFERENCES

[1] Brereton O. P., Kitchenham, B., Budgen D., Turner M. and Khalil M. (2007) Lessons from applying the Systematic Literature Review process within the Software Engineering domain, Journal of Systems & Software, 80(4), pp. 571–583.

[2] Davis F. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology, MIS Quarterly, 13(3), pp. 319-340.

[3] Davis F.D., Bagozzi R.P., and Warshaw P.R. (1989). User acceptance of computer technology: a comparison of two theoretical models, Management Science, 35(8), pp. 982-1003.

[4] Davis F.D. and Venkatesh V. (2004) Toward Preprototype User Acceptance Testing of New Information Systems: Implications for Software Project Management, IEEE Transactions on Engineering Management, 51(1), pp. 31-46.

[5] Fenech T. (1998). Using perceived ease of use and perceived usefulness to predict acceptance of the World Wide Web, Computer Networks and ISDN Systems, 30(1-7), pp. 629-630.

[6] Fink, A. (2005) Conducting Research Literature Reviews. From the Internet to Paper, Sage Publication, Inc.

[7] Fishbein M., and Ajzen, I. (1975). Belief, Attitude, Intention, and Behaviour: An Introduction to Theory and Research, MA, USA: Addison-Wesley.

[8] Heijden, H.V.D. (2003) Factors influencing the usage of websites: the case of a generic portal in The Netherlands, Information and Management, **40**(6), pp. 541-549.

[9] Horton, R.P., Buck, T., Waterson, P.E., and Clegg, C.W. (2001). Explaining intranet use with the technology acceptance model, Journal of Information Technology, **16**(4), pp. 237-249.

[10] Jørgensen, M., and Shepperd, M. (2007) A Systematic Review of Software Development Cost Estimation Studies, IEEE Transactions on SE, 33(1), pp. 33-53.

[11] Keung J., Jeffery R., and Kitchenham B. (2004) The Challenge of Introducing a New Software Cost Estimation Technology into a Small Software Organisation, in Proceedings of the 2004 Australian Software Engineering Conference (ASWEC'04), IEEE Computer Society Press.
[12] King W.R., and He J. (2006) A meta-analysis of the technology acceptance model, Information and Management, 43(6), pp. 740-755.
[13] Kitchenham, B.A. (2004) Procedures for Undertaking Systematic Reviews, Joint Technical Report, Computer Science Department, Keele University (TR/SE-0401) and National ICT Australia Ltd (0400011T.1).
[14] Kitchenham, B.A., Brereton, O.P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. (2008), Systematic Literature Reviews in Software Engineering – A Systematic Literature Review, submitted to Information and Software Technology.
[15] Kitchenham, B.A. and Charters, S. (2007) Guidelines for performing Systematic Literature Reviews in Software Engineering Technical Report EBSE-2007-01.
[16] Kitchenham, Barbara A., Mendes, E., and Travassos, G.H. (2007) Cross versus Within-Company Cost Estimation Studies: A Systematic Review, TSE, 33(5), pp 316-329.
[17] Legris, P., Ingham, J., and Collerette, P. (2003) Why do people use information technology? A critical review of the technology acceptance model, Information & Management, ACM Press, 40(3), pp. 191-204.
[18] Shaddish, W.R., Cook, T.D., and Campbell, D.T. (2002) Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Houghton Mifflin Company.
[19] Straub D., Limayem M., and Karahannaevaristo E. (1995). Measuring System Usage - Implications for Is Theory Testing, Management Science, 41(8), pp. 1328-1342.
[20] Szajna, B. (1994). Software Evaluation and Choice: Predictive Validation of the Technology Acceptance Instrument, MIS Quarterly, 18(3), pp. 319-324.
[21] Venkatesh V., and Davis F.D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies, *Management Science*, 46(2), pp. 186–204.
[22] Yin, R.K. (2003) Case Study Research: Design and Methods, 3rd Edition, Sage Publications.

Acknowledgements
This study was funded by the UK Engineering and Physical Sciences Research Council project EP/E046983/1.