View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by UCL Discovery

Measuring the Success of Context-Aware Security Behaviour Surveys

Ingolf Becker, Simon Parkin and M. Angela Sasse
University College London
$\{$*i.becker, s.parkin, a.sasse*$\}$*@cs.ucl.ac.uk*

Abstract

Background. We reflect on a methodology for developing scenario-based security behaviour surveys that evolved through deployment in two large partner organisations (A & B). In each organisation, scenarios are grounded in workplace tensions between security and employees' productive tasks. These tensions are drawn from prior interviews in the organisation, rather than using established but generic questionnaires. Survey responses allow clustering of participants according to predefined groups.

Aim. We aim to establish the usefulness of framing survey questions around active security controls and problems experienced by employees, by assessing the validity of the clustering. We introduce measures for the appropriateness of the survey scenarios for each organisation and the quality of candidate answer options. We use these scores to articulate the methodological improvements between the two surveys.

Method. We develop a methodology to verify the clustering of participants, where 516 (A) and 195 (B) free-text responses are coded by two annotators. Inter-annotator metrics are adopted to identify agreement. Further, we analyse 5196 (A) and 1824 (B) appropriateness and severity scores to measure the appropriateness and quality of the questions.

Results. Participants rank questions in B as more appropriate than in A, although the variations in the severity of the answer options available to participants is higher in B than in A. We find that the scenarios presented in B are more recognisable to the participants, suggesting that the survey design has indeed improved. The annotators mostly agree strongly on their codings with Krippendorff's $\alpha > 0.7$. A number of clusterings should be questioned, although α improves for reliable questionsby 0.15 from A to B.

Conclusions. To be able to draw valid conclusions from survey responses, the train of analysis needs to be verifiable. Our approach allows us to further validate the clustering of responses by utilising free-text responses. Further, we establish the relevance and appropriateness of the scenarios for individual organisations. While much prior research draws on survey instruments from research before it, this is then often applied in a different context; in these cases adding metrics of appropriateness and severity to the survey design can ensure that results relate to the security experiences of employees.

1 Introduction

Engaging users is important to develop meaningful, effective security behaviour surveys. If studies are conducted out of context, reproduction of results is difficult [24]. Yet much of security awareness research examines individuals' abilities to internalise and enact knowledge of security risks and controls in an abstract setting. Efforts to measure security behaviour frequently assess individuals' competency in general security skills (see Section 2). Much of this research ignores the bounded effort of the individual [6, 16], and that employees in organisations have other responsibilities [3].

Here we run a validation exercise on scenario-based surveys conducted in two large organisations each with many thousands of staff. Scenarios are built on frictions between security and regular business tasks derived from prior exploratory interviews with a cross-section of employees. The core principles of the methodology underlying the two surveys are: determining attitudes toward security provisions and policy in the organisation, and; characterising how individuals act independently or with others to enact security-related behaviours. The differences between the surveys represent an evolution in survey design as lessons have been learned, where we develop measures which account for these differences and allow cross-comparison between survey deployments. We describe the framework for our scenario-based surveys in Section 3.

Here we explore the capacity to utilise additional types of questions to reflect on the survey design without further effort by the researchers. If the participants are given an opportunity to indicate the applicability of the scenarios to their environment, we can tailor the results not just to specific user groups, but also reflect on how a survey engages with diverse groups and their security needs.

From the analysis we formulate further metrics for measuring how aligned the security apparatus of an organisation is with the employees who are governed by the policies and controls that are in place. This is achieved through Likert-scale questions added to the existing questionnaire, as described in our methodology (Section 4), which serve as internal validity measures.

The surveys conducted with our two partner organisations contain questions structured in this way, and we discuss the results of our research in Section 5. We find that the appropriateness and applicability of the questions of the survey have improved from A to B. Similarly, the reliability of the clustering of the answer options has improved. Yet participants judge the answer options in B to be more severe and less balanced than in A.

This immediately available feedback allows researchers to continuously evaluate their survey design and discard unreliable questions from further analysis. In broader terms, we reflect on this approach in the discussion that follows in Section 6, and in the conclusions in Section 7.

2 Related Work

We consider scenario-based security behaviour survey research from two perspectives: Initially we examine the construction and motivation of these surveys, and in the second stage we focus on the reliability of survey analysis (given that surveys would be deployed in specific organisational contexts). Our review of related work highlights the need to situate scenarios in the participants' environment to build a reliable picture of how security provisions and workplace conditions interact.

Most of the works reviewed do not evaluate the external validity of the questionnaires applied but instead rely on additional prior work. Our methodology brings obvious benefits to survey designs in research and practice alike.

Egelman et al. [14] developed the Security Behavior Intentions Scale (SeBIS) to predict security behaviours for common controls ('awareness', 'passwords', 'updating', and 'securement'). The SeBIS survey comprises of 16 items on a 5-point Likert scale. SeBIS was deployed on several occasions through Mechanical Turk (and in one case, PhoneLab). The goal of the work was to determine if self-reported, *intended*, behaviours translated into actual behaviour. To this end, tasks were set relating to each behaviour category (such as identifying fake login pages). The authors accepted that the designed tasks were targeted and narrow in scope, but with a focus on exploring SeBIS' predictive capabilities in this limited setting. Here we use scenarios and options based in real organisational settings to establish an individual's behaviour type and attitude toward the security apparatus around them; the focus is not on predicting behaviour, but rather to capture a snapshot of how effectively security provisions are perceived to be supporting the business.

Parsons et al. [26] sought to validate a survey tool for measuring information security awareness and awareness initiatives, the Human Aspects of Information Security Questionnaire (HAIS-Q).Two studies were conducted: in one study, participants completed the HAIS-Q and were tested for security skills (in this case, identifying potential phishing links amongst a range of fabricated emails); in the second study, engagement of participants in the survey was examined by establishing the level of *non-responsivity*. Here, we similarly seek to determine whether the scenarios and response options in our surveys resonated with participants, through examination of internal measures within our situated surveys. By doing so we identify repeatable measures for measuring engagement.

Rajivan et al. [27] propose a questionnaire for capturing users' level of *security expertise*, presented as being a critical factor in how well an individual can assess risk and use available security controls. The questionnaire seeks to separate respondents across the dimensions of skills, rules, and knowledge, toward understanding how individuals apply these in different situations. Here we discuss our survey methodology as a means to not only determine how employees use the tools available to them as individuals and groups, but also how they respond to specific risks which can potentially arise in their working environment. Rajivan et al. also included free-text questions to capture additional comments from participants, where we use a similar internal mechanism in our situated scenarios so that participants can further describe security experiences from their own perspective (further informing the picture of security *on the ground*).

Karlsson et al. [20] posit that in organisations, information security compliance must be evaluated relative to employees' work tasks (and with this, competing goals and their related *values* such as productivity and efficiency). The authors speak of there being "tensions and dilemmas" where one option is preferable to others that are available. While the argument is made that situational context is critical to understanding how tensions are resolved, the authors' questionnaire however is free from any contextual settings. This does allow it to be applied to any *"white-collar individual"*, but may limit how the questionnaire captures the *"tensions and dilemmas"* that exist in a specific organisation. Here we are assessing scenarios which are grounded in prior interviews with employees, specifically to identify those regular tensions and dilemmas which occur in the workplace.

We argue that surveys that are situated in scenarios that the participants can relate to will engage them and evoke genuine responses, which can inform efforts to improve the effectiveness of security solutions in an organisation. Some research has focused on basing scenario design in literature, where Blythe [9] argues that scenarios should *"[avoid] unusual events and characters but nonetheless resonate with the respondent in a way that they are readily understood while presenting multiple solutions."* Siponen and Vance argue that research needs to be practically relevant by ensuring contextual relevance [29]. Their five suggestions focus on studying information system policy violations, but are equally transferable to other behavioural research (and related to the principles derived by Krol et al. in [23] for studying usability in security and privacy). Many examples of security behaviour research using questionnaire instruments do not consider the role of task conflicts characterised by Karlsson et al. [20]. These works instead draw on existing questions from prior research [13, 19, 30, 31, 33]. The importance of scenario-based surveys is underlined by Wash et al.'s findings that individuals do not self-report security accurately [32]: it undermines much of the traditional self-reported constructs used for inferring personal security behaviour.

Sohrabi et al. for example argue that *"the lack of information security awareness, ignorance, negligence, apathy, mischief, and resistance are the root of users' mistakes"* [30]. Their questionnaire uses nine information security constructs from prior literature to model their interactions. These questions and mappings are adopted from four previous studies [13, 19, 31, 33]. However each of these articles in turn is standing on the shoulders of giants, citing a total of 29 prior works to support their survey design, which, in turn, cite over 100 other unique articles to support the quality of their survey design. The sources span the fields of sociology, education, criminology, information systems and medical research.

Of the literature referenced by Sohrabi et al. and their references in turn, the vast majority source their constructs and survey questions from further literature. A number of articles construct their own questions. The most rigorous of those papers in the chain to do any pre-testing validation of their question design is by Bulgurcu et al. [11], who conduct two rounds of card sorting by 11 students followed by two rounds of pilot testing by 110 individuals. Huang [17] and Chang and Chuang [12] also conduct some limited pre-testing.

While it is good scientific practice to rely on constructs that have been rigorously tested in prior works, only one of the papers ([19]) discussed above cites the primary literature which validates the constructs in their original setting ([11]). Further, as throughout the literature, the questions are taken out of context of their original research premise, where the validity of the original validation should be revisited in each new context. It is understandable that a full pre-study is infeasible for every new questionnaire, yet augmenting the survey with additional questions (as described subsequently) to support the measurement of validity post-hoc would be cheap and desirable.

The data of our research is grounded in competing goals in realistic scenarios, so that (i) security managers can better understand how employees' attitude and behaviour toward security policy and controls influence the approach to problem resolution, and (ii) researchers can gain further insight into the shape of tensions between security and productive tasks in an organisational setting.

3 Background

In this section we describe the two surveys that contribute the primary data in this paper.

The first organisation to be studied we shall call Company A[1] (which has many thousands of employees). In this organisation 118 semi-structured interviews were conducted, exploring conflicts between security and business processes lasting on average 40 minutes. These interviews were analysed with thematic analysis, and form the basis of workplace-based scenarios and possible solutions informed by approaches reported by employees. Scenarios are combined to create a scenario-based survey. A similar approach is described by Blythe et al. [10] for conducting interviews around security behaviours within organisations, presenting *dilemmas* as short stories with a central character in a specific context (and informed by an organisation's security policies). A small proportion of Company A's workforce was sampled through interviews, where a wider survey would attempt to capture the prevalence of the issues identified in interviews across the wider company. A total of 1486 participants provided complete responses. Further, the respondents gave 516 additional comments at various stages of the survey, through text-entry fields provided.

A second, similarly large organisation (which we call Company B) was studied subsequently, where lessons learned from the analysis of Company A improved the organisation of the survey process. These lessons have caused the progression in groupings as described in Table 1 and discussed in Section 4. The initial attitude and behaviour types [4] were derived from Adams [1].

[1] for brevity the companies will be referred to as 'A' and 'B' for the remainder of this paper

85 interviews were conducted in Company B of similar lengths to the interviews in A, where again these were used in a similar methodology as in A to develop a larger-scale survey. A survey was conducted using scenarios built upon themes emerging from the qualitative analysis of the interviews. 641 employees responded to this survey, including 195 free text responses. While the survey results have been analysed [4, 7] here we explore how the free-text responses can indicate the success of the survey in engaging with the organisation's employees. Further analysis of the interviews that informed the survey designs can be found in [22] for Company A, [21] for Company B, and [8] for a combination of both organisations.

3.1 Employee types

In each of the two surveys we attempt to position participating employees on two dimensions. In A these are Attitude and Behaviour types, whereas in B these have evolved to Maturity levels and Behaviour types. For the definitions of the types please see Table 1. Foremost, these two dimensions can be examined individually and in combination – across age groups, business divisions, and physical locations – to target interventions which reduce friction between security and productivity in the workplace.

The attitude types in A focus on individuals' interaction with security apparatus. In B, these have evolved to a scale of Maturity Levels, which are ranked levels of individuals' interaction with the organisation's policy (such that interventions would act to improve employees' working interactions with centralised security policy and security provisions). In both A and B, the participants were also asked to assign an appropriateness score for each answer option on a 5-point Likert scale, ranging from *not acceptable at all* to *very acceptable*.

The behaviour types in A are a measure of the individuals' likelihood to trade-off security for productivity. This evolved to the more abstract concepts in B where the answers are now mapped to four distinct behaviour types as defined by Adams [1], to better represent the role of teams and organisational culture in individual security behaviours. Additionally, participants were asked to assign a severity score to each answer option of the behaviour type questions, as well as give a general indicator as to how acceptable to the business it would be for the participant not to finish the task described in each scenario.

3.2 Survey design

Figure 1 shows one of the scenarios in B (note that participants did not see the Individual-

(a) Attitude Types for A

1 Discount suspicions, cause no bother, passive,
2 Report suspicions but take no direct action,
3 Take direct action through official channels,
4 Take direct personal action against the threat.

(b) Behaviour Types for A

1 Prepared to perform insecure acts to maximise productivity,
2 Show a minor priority for work over security when the two conflict,
3 Passive, expects others to take the initiative to ensure security,
4 Tries to remain secure wherever possible.

(c) Maturity Levels for B

1 Is not engaged with security in any capacity,
2 Follows security policy only when forced to do so by external controls,
3 Understands that a policy exists and follows it by rote,
4 Has internalised the intent of the policy and adopts good security practises even when not specifically required to,
5 Champions security to others and challenges breaches in their environment.

(d) Behaviour Types for B

Individualists rely on themselves for solutions to problems,
Egalitarians rely on social or group solutions to problems,
Hierarchists rely on existing systems or technologies for solutions to problems,
Fatalists take a 'naive' approach to solving problems, feeling that their actions are not significant in creating outcomes.

Table 1: The dimensions by which survey responses are measured in Company A and Company B

Question: File Storage (Behaviour-type)

Concerned about the safety of his current work, Shamal decides to back up his data, some of which is confidential. As he uses his own laptop under the 'bring your own device' scheme, he usually stores all his work on his drive on the central server but he wants to have a second copy just in case something happens or he loses connectivity to the company network. He thought about using one of the common drives but none of the ones he regularly uses have sufficient space.

Individualist: Create a local copy on the hard drive of your BYOD laptop, it is the only machine you work on so you know it will be safe and this ensures you will always have access to it if needed.

Egalitarian: Use a common drive that you used for an old project and still have access to, as your credentials were never revoked. It has enough space although you do not know who manages it now.

Hierarchist: Use an online service, such as Dropbox, to store the data as it is more under your control.

Fatalist: Back your work up onto a USB stick – you have ordered an encrypted one but while you wait for it to arrive you use a personal stick you have to hand.

Figure 1: Scenario 'File Storage' (QFS) in B

ist/Egalitarian/Hierarchist/Fatalist labels, as defined in Table 1). In each organisation, surveys were crafted for participants based upon their department to improve relevance, see [4, 7] for more details. For the example scenario in Table 1, we found through the interviews that data availability was a predominant issue in the organisations, where many interview participants mentioned the use of security *workarounds* [5] to guarantee that they reached their business goals. Question design attempted to offer the participants a number of options which would all be regarded as equally appropriate, based upon the themes identified from the preceding interviews with employees. The participants were asked to rank the four options in order of their preferences. Additionally, participants were allowed to offer additional comments, which included the following example:

> *"Shamal needs to find out who manages the common drive now, and whether the company authorises use of Dropbox and personal USB sticks, before using any of those options."*

As part of the work described here, two annotators coded the volunteered comments for the types (without reference to the alignment of responses to types already defined for each question). For example, the quotation above could be coded as a Hierarchist's point of view, as the individual falls back to existing structures for solutions to the problem.

> *"This scenario could easily be avoided by providing sufficient space on the common drives."*

Conversely, this statement has been coded as a Fatalist. The employee is frustrated that the natural solution to the problem is outside their reach.

4 Methodology

The methodology laid out in this section establishes three metrics to measure the quality of the survey design and its external validity retrospectively. The quality of the survey includes how engaged participants are in considering a scenario, and how relevant a scenario and its options are to their own experiences. If an organisation is committed to measuring how well its security provisions support the effective completion of business tasks towards identifying and removing frictions, decision-makers would have a natural interest in having a realistic picture of the current experiences of employees.

4.1 Appropriateness and applicability

For each of the answer options to Attitude and Maturity questions (see Tables 1a and 1c) participants were asked to specify the acceptability of that answer on a 5-point Likert scale ranging from "Not acceptable at all" to "Very acceptable". There are of course biases present here, namely that given a participant's type they may see some options as more acceptable than others. Indeed there is a statistically significant (at $p < 0.001$) correlation in our survey response data between the ranking of options and their associated behaviour types with Kendall's τ of 0.62. Yet the ideal scenario design would leave the participants with four objectively equally acceptable options, and allow the participant to freely rank the option.Hence a high appropriateness score is desired.

Similarly, for each of the answer options to Behaviour type questions (see Tables 1b and 1d) participants were asked to specify the severity of each option as well as the *acceptability of failing to complete the task* for each scenario on a 5-point Likert scale. Again, the severity scores are statistically significantly (at $p < 0.001$) correlated with the ranking of the answers (Kendall's $\tau = -0.20$), with less severe answers being ranked as more preferable. The severity scores of the different answers should be ranked equally by the participants (as questions are designed with no one 'right' answer), resulting in a low standard deviation throughout the questions. The ideal mean of the standard deviation of severity of options is 0, which would imply that all options given to the participants are perceived as equally severe.

The *acceptability of failing to complete the task* metric would ideally be identically distributed for all questions in order to allow for inter-question comparison. This is a metric that is difficult to establish through prior analysis. If participants think that for a scenario it is more acceptable for it not to be completed given the given consequences as in the scenario and its options, the participants do not fully commit to their choices of behaviour types, as in no scenario there is an option to do nothing (and in turn avoid side effects from the chosen solution).

4.2 Validation of ranked types by free-text responses

In the survey design for both A and B, participants are asked to rank four answer options according to their preferences. Participants are also invited to provide additional comments on the questions. We find that there are two common types of responses: those that further confirm a respondent's answer, or elicit suggestions / solutions that are not included in the question and the associated options. We code these according to the applicable mapping (Attitude or Behaviour) in each organisation as listed in Table 1, e.g., for each free text response the annotators have to choose from one of four options. While this is opportunistic (not all participants provided additional comments), we can validate the mappings by calculating inter-annotator agreement metrics as described in the following sections.

4.2.1 Inter-annotator agreement

	Coder A			
Coder B	T1	T2	T3	T4
T1	4	3	0	0
T2	0	36	2	15
T3	0	2	59	1
T4	0	0	1	32

Table 2: Confusion matrix for Question 1 in A between the coders' assignment of types to the free text responses

The calculation of the inter-annotator agreement between the two coders is straightforward. We first calculate a confusion matrix for each question (an example is shown in Table 2), and then calculate Krippendorff's chance corrected inter-annotator agreement metric α. Krippendorff's α ranges from -1 to $+1$, where 0 corresponds to chance agreement and $+1$ to perfect agreement. As the attitude types and maturity levels in Tables 1a and 1c are on a ranked scale, we weight the disagreement linearly. For the other two types described in Tables 1b and 1d the agreement is binary.

4.2.2 Validating the mapping

	Coding type			
Rank	T1	T2	T3	T4
1	2	6	84	3
2	5	26	22	10
3	2	19	12	34
4	2	43	6	34

Table 3: Confusion matrix for Question 1 in A between participants assigned ranks to the potential answers and the types assigned to the participants by the coders based on the coding of the free text responses

We validate the mapping of the survey answer options (an example is shown in Figure 1) by treating the participants as another annotator and calculating the inter-annotator metric α. However, the participants rank their options, but the coders annotate separate (but not independent) text statements. For example, a participant may provide a ranking of Type 3 $>$ Type 2 $>$ Type 4 $>$ Type 1 for a specific question, and from the coders we may see Coder X: Type 3, Coder Y: Type 2.

In this case the standard agreement table approach [15] for > 2 annotators cannot be used. Yet Krippendorff's α naturally extends to non-square weight matrices. In our case, this leads to a confusion matrix such as in Table 3. Here we tabulate the frequency that a coder has annotated a statement with a specific type with the rank that the participant gave that type. Perfect validation would therefore imply that all types chosen by the coders have rank 4; i.e. the only non-zero entries are in the bottom row of the confusion matrix. Given this matrix we can execute the calculation of Krippendorff's α with a weights matrix that treats numbers in the bottom row as perfect agreement, and linearly increases disagreement for lower ranked options.

4.2.3 Estimating confidence in α

In order to calculate the confidence in the calculated value of α we rely on α's standard deviation. As an analytic expression is not available, we bootstrap the calculation of α. In the following sections the confidence intervals are calculated using 1000-fold bootstrapping.

5 Results

In this section we present the application of the metrics defined in Section 4 to the datasets described in Section 3. There are four tables to consider in this section; Tables 4 and 5 for the analysis of the secondary coding of the free text responses, Table 6 for the analysis of appropriateness scores on attitude/maturity questions, and Table 7 for the analysis of the severity metrics on behaviour type questions.

5.1 Analysis of clustering

Table 3 is an example confusion matrix calculated based on the methodology presented in Section 4.2.2. The column headers list the four possible types assigned to the free-text responses by the coders. If a free-text response by the coders was judged to be type 1, but the participant ranked the answer corresponding to type 1 as rank 2, this would increment the number in row 1, column 2. Perfect agreement would be represented by the type assigned through coding of the free-text responses always being ranked highest (rank 4) by the participants. This would be a confusion matrix of non-zero entries in the bottom row only.

The strong disagreement between the coders and the assigned rank in question 1 can be identified by the strong mismatch in type 3: of the 124 statements assigned to type 3 by the coders, 84 were ranked least likely (rank 1) by the participants. This implies that the answer option assigned to type 3 *"Request that those with access share their (main log-in) account details and passwords with those without to allow them access to the information"*) does not match behaviour type 3 (as defined in Table 1) (*"Passive, expects others to take the initiative to ensure security"*).

Interestingly, this disagreement is not reflected in the coding of the free-text responses themselves. Table 2 shows the confusion matrix for Question 1 for the two coders. There is virtually no disagreement for types 1, 2 and 3; but some disagreement for type 4, where 15 statements assigned to type 4 by coder A were considered to be type 2 by coder B. The internal validity for the coding of free text responses for Question 1a can be accepted based on Krippendorff's α of 0.77 ± 0.00 as shown in Table 4, but we are unable to validate the mapping of answer options to types.

Tables 4 and 5 list the number of free-text responses coded and Krippendorff's α for both the validation of the mapping as well as the coders agreement.

Question	#	Mapping α	Coder's α
Q4	40	0.21 ± 0.02	0.29 ± 0.02
Q5	34	0.35 ± 0.02	0.02 ± 0.03
Q6	2	-0.33 ± 0.76	0.00 ± 0.67
Q8	29	0.30 ± 0.04	0.94 ± 0.02
Q10	37	0.23 ± 0.03	0.73 ± 0.02
Q1	155	-0.03 ± 0.01	0.77 ± 0.00
Q2	137	0.43 ± 0.01	0.91 ± 0.00
Q3	12	0.33 ± 0.08	0.38 ± 0.10
Q7	25	0.24 ± 0.03	0.13 ± 0.04
Q9	45	0.13 ± 0.02	0.76 ± 0.02

Table 4: Krippendorff's α measures for compA with 95% confidence intervals.

Question	#	Mapping α	Coder's α
QID	33	0.27 ± 0.02	0.85 ± 0.02
QCDP	53	0.31 ± 0.01	0.38 ± 0.02
QT	22	0.24 ± 0.04	0.34 ± 0.05
QSD	27	0.53 ± 0.02	0.47 ± 0.04
QRM	12	0.27 ± 0.07	0.42 ± 0.07
QVPN	23	-0.09 ± 0.03	0.37 ± 0.04
QFS	18	0.19 ± 0.05	0.46 ± 0.05
QCC	7	0.38 ± 0.13	0.75 ± 0.21

Table 5: Krippendorff's α measures for compB with 95% confidence intervals.

5.1.1 Suitable values for α

Before discussing this data further we must delineate the boundaries for which we consider Krippendorff's α to be reliable. From a statistical perspective we can conduct a t-test where the null hypothesis is $\alpha = 0$, i.e. the data is equivalent to chance. This t-test is represented in our tables through the use of 95% confidence intervals. Indeed all rows that are statistically significant at the 95% confidence interval are also significant at the 99% confidence interval. However the literature [15] is clear that primary data is only sufficiently reliable for further analysis at $\alpha > 0.667$.

It is clear that most of the coder's agreement values in A satisfy this criteria. There are a number of exceptions: *Q5*, *Q3* and *Q7*. The inter-coder agreement is not as strong in B, where only *QID* satisfies this criteria. When focusing on the validation of the mapping/clustering however, none of the scenarios satisfies this stringent criteria.

Considering the difficulty the coders have to establish agreement on the free-text responses in B, the low mapping α values are not surprising: the coding is a difficult task (given the brevity of comments and potential lack of contextual information). Yet rather than discarding the results at this stage, it may be more important to identify the scenarios which are indistinguishable from random data: scenarios *Q1* in A and *QVPN* in B. Apart from these two scenarios, our data allows the focus of further investigations and policy decisions to be guided by data with known uncertainty.

5.2 Appropriateness

		Mean		1st choice	
Question	#	mean	std	mean	std
Company A					
Q4	374	0.626	0.120	0.923	0.195
Q5	820	0.570	0.110	0.925	0.161
Q6	137	0.427	0.138	0.821	0.321
Q8	364	0.529	0.085	0.983	0.084
Q10	903	0.483	0.082	0.917	0.185
Company B					
QID	152	0.488	0.122	0.778	0.316
QCDP	456	0.508	0.108	0.893	0.220
QT	164	0.499	0.095	0.873	0.252
QSD	292	0.546	0.118	0.939	0.181

Table 6: Appropriateness scores for each attitude question in compA, the higher, the more appropriate. As each answer option is assigned an appropriateness score, the mean represents the mean appropriateness score of all answer options irrespective of that answer's ranking. The *1st choice* only considers the appropriateness assigned by the participants to their top choice.

Table 6 shows the appropriateness scores the participants have given the answer options for specific questions. The scores vary from 0 (not appropriate) to 1 (very appropriate). The mean appropriateness score is more varied in A than in B, although it is close to 0.5 for all questions, indicating that the average answer option is balanced. This is desirable as it offers participants the option to swing to both extremes as necessary. The appropriateness score given by participants to their highest ranked choice is very high, confirming the participant's stance that they view their preferred choice as most appropriate.

5.3 Severity

Table 7 compares the distribution of severity scores and *acceptability of failing the task* scores across the different scenarios and organisations. There are a number of variations: Scenarios in A are considered less acceptable to be left undone, however the standard variations of the severity scores across the different scenario options are higher. According to Table 7 the scenarios in B are therefore believed by the participants to be more applicable to their environment (particularly *QRM* and *QCC*), but the answer options are more balanced in severity in A, implying that options represent potential solutions that may be seen in everyday work in A compared to the more contrived answer options in B.

		Failing		Std of Severity	
Question	#	mean	std	mean	std
Company A					
Q1	903	0.281	0.307	0.270	0.128
Q2	893	0.270	0.296	0.239	0.123
Q3	137	0.394	0.340	0.271	0.122
Q7	291	0.458	0.393	0.296	0.144
Q9	374	0.668	0.449	0.274	0.123
Company B					
QRM	152	0.196	0.312	0.377	0.101
QVPN	152	0.439	0.370	0.323	0.120
QFS	164	0.430	0.318	0.297	0.114
QCC	292	0.240	0.410	0.182	0.163

Table 7: Acceptability of failing to complete the task (higher more acceptable) and standard deviation of severity of options for behaviour type scenarios in compA.

6 Discussion

This research supports a process of continuous improvement to organisational security, by providing measures for (i) typical workaround to regular frictions with security in the workplace (by analysing the perceived suitability of solutions derived from interviews), and (ii) how the interactions employees have with security apparatus can be designed to minimise the demand on their 'compliance budget' [6]. Employee's willingness to expend effort for the security of not only themselves but those around them can be explored by articulating embodied security cultures which may arise in any number of situations in the workplace where security controls can be applied. Both the survey results and the free-text responses can inform targeted interventions as part of incremental improvement, an approach advocated by Renaud and Goucher [28]. Unfortunately in striving for internal validity for security behaviour constructs it is easy to overlook the need to establish the applicability of the results to the real world, that is, to measure the quality of *engagement* with employees (where tensions can arise with local demands on effort and capacity). The related works discussed in Section 2 demonstrate this well.

Security managers ought to identify the non-divisible security behaviours in their own organisations, and equally deploy information security surveys that shine light on previously unseen *workaround* or *compromise* behaviours by engaging employees. To do this, available options (and ideally, additional feedback from users) must point to clear responses to security-related challenges that employees see as acceptable given the pressures they perceive in a particular situation. Where respondents imply confusion about what is being asked of them in a scenario-based survey question – or indeed, see two or more behaviours as one and the same – this implies that more can be done to clearly separate candidate behaviours. In turn, this can be achieved if security managers act to grow their understanding of how security manifests for employees who have other competing demands for their attention (see also Ashenden and Lawrence [2], Herley [16] and Parkin et al. [25]).

Our proposed survey methodology and validity measures address these challenges. This is achieved both internally (by way of inter-annotator agreement), and externally (by way of appropriateness and severity scores). We are able to highlight strengths and shortcomings in the survey design which not only inform the design of realistic scenarios by researchers, but also inform the investment in security by policy managers when designing interventions. Organisational environments are complex, and researchers cannot assume that they have a full understanding of security behaviours prior to deploying a survey. This research helps to identify these known unknowns. Security practitioners considering potential investments may do well to understand the quality of the data they base their decisions upon [18].

7 Conclusions

In this paper we have described a methodology for post-hoc assessment of the quality of situated security behaviour survey designs. We utilise free-text responses and reflective metrics to measure the surveys' external validity. We have demonstrated this approach on two surveys in two large organisations, drawing on 711 free text responses and over 7000 reflective scores in the process. This has allowed us to quantify the evolution of our scenario-based surveys through clearly-defined and repeatable metrics, and partially validate the mapping from survey responses to constructs. This knowledge will allow security managers to tailor future improvements to their organisation's security policy and behavioural interventions more accurately to the local working environment, relative to the demonstrable strengths and weaknesses of the survey design.

We strongly advice researchers designing surveys in future to include open questions that can be answered by the participant without being biased. Survey designers should not assume they know everything about the responders even if the survey is grounded in qualitative research, and continually look for ways to involve respondents to gather more context-specific information, such as by including the reflective questions described in this research.

Dataset

The participant's assigned categories and the two sets of coding for both organisations as well as the analysis code can be found at DOI `10.14324/000.ds.10038283`.

Acknowledgements

We would like to thank the LASER program committee and workshop attendees for their feedback, as well as the participating organisations. The authors are supported in part by UK EPSRC grants, no. EP/G037264/1.

References

1. Adams, J. Risk and morality: three framing devices. *Risk and morality*, 2003: 87–106.
2. Ashenden, D., and Lawrence, D. Can we sell security like soap?: a new approach to behaviour change *Proc. NSPW '13*, 87–94.
3. Ashenden, D., and Lawrence, D. Security Dialogues: Building Better Relationships between Security and Business. *IEEE Security & Privacy*, 14(3), 2016: 82–87.
4. Beautement, A., Becker, I., Parkin, S., Krol, K., and Sasse, M. A. Productive Security: A Scalable Methodology for Analysing Employee Security Behaviours *Proc. SOUPS ' 16*.
5. Beautement, A., and Sasse, A. The economics of user effort in information security. *Computer Fraud & Security*, 2009(10), 2009: 8–12.
6. Beautement, A., Sasse, M. A., and Wonham, M. The compliance budget: managing security behaviour in organisations *Proc. NSPW '08*, 47–58.
7. Becker, I., Parkin, S., and Sasse, M. A. Finding Security Champions in Blends of Organisational Culture. *Proc. USEC*. 2017, 11.
8. Beris, O., Beautement, A., and Sasse, M. A. Employee Rule Breakers, Excuse Makers and Security Champions: Mapping the risk perceptions and emotions that drive security behaviors *Proc. NSPW '15*.

9. Blythe, J. Information security in the workplace: A mixed-methods approach to understanding and improving security behaviours. PhD thesis. Northumbria University. 2015.
10. Blythe, J. M., Coventry, L. M., and Little, L. Unpacking Security Policy Compliance: The Motivators and Barriers of Employees' Security Behaviors. *SOUPS*. 2015, 103–122.
11. Bulgurcu, B., Cavusoglu, H., and Benbasat, I. Information security policy compliance: An empirical study of rationality-based beliefs and information security awareness. *MIS Quarterly: Management Information Systems*, 34, 2010: 523–548.
12. Chang, H. H., and Chuang, S.-S. Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. *Information & Management*, 48(1), 2011: 9–18.
13. Cheng, L., Li, Y., Li, W., Holm, E., and Zhai, Q. Understanding the violation of IS security policy in organizations: An integrated model based on social control and deterrence theory. *Computers & Security*, 39, 2013: 447–459.
14. Egelman, S., Harbach, M., and Peer, E. Behavior ever follows intention?: A validation of the security behavior intentions scale (SeBIS) *Proc. SIGCHI '16*, 5257–5261.
15. Gwet, K. L. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC, 2014.
16. Herley, C. More Is Not the Answer. *IEEE Security Privacy*, 12(1), 2014: 14–19.
17. Huang, C.-C. Knowledge sharing and group cohesiveness on performance: An empirical study of technology R&D teams in Taiwan. *Technovation*, 29(11), 2009: 786–797.
18. Hubbard, D. W. How to measure anything: finding the value of intangibles in business. John Wiley & Sons, 2014.
19. Ifinedo, P. Information systems security policy compliance: An empirical study of the effects of socialisation, influence, and cognition. *Information & Management*, 51(1), 2014: 69–79.
20. Karlsson, F., Karlsson, M., and Åström, J. Measuring employees' compliance-the importance of value pluralism. *Information & Computer Security*, 25(3), 2017:
21. Kirlappos, I., Parkin, S., and Sasse, M. A. Shadow security as a tool for the learning organization. *ACM Computers and Society*, 45(1), 2015: 29–37.
22. Kirlappos, I., Parkin, S., and Sasse, M. A. Learning from "Shadow Security": Why understanding non-compliance provides the basis for effective security. *Proc. USEC*. 2014.
23. Krol, K., Spring, J. M., Parkin, S., and Sasse, M. A. Towards robust experimental design for user studies in security and privacy. *Learning from Authoritative Security Experiment Results (LASER) Workshop*. 2016.
24. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015: aac4716.
25. Parkin, S., Krol, K., Becker, I., and Sasse, M. A. Applying Cognitive Control Modes to Identify Security Fatigue Hotspots *Proc. SOUPS '16*.
26. Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A., and Zwaans, T. The Human Aspects of Information Security Questionnaire (HAIS-Q): Two further validation studies. *Computers & Security*, 66, 2017: 40–51.
27. Rajivan, P., Moriano, P., Kelley, T., and Camp, L. J. Factors in an end user security expertise instrument. *Information & Computer Security*, 25(2), 2017: 190–205.
28. Renaud, K., and Goucher, W. The Curious Incidence of Security Breaches by Knowledgeable Employees and the Pivotal Role of Security Culture. *Human Aspects of Information Security, Privacy, and Trust*. 2014, 361–372.
29. Siponen, M., and Vance, A. Guidelines for improving the contextual relevance of field surveys: the case of information security policy violations. *Eur J Inf Syst*, 23(3), 2014: 289–305.
30. Sohrabi Safa, N., Von Solms, R., and Furnell, S. Information security policy compliance model in organizations. *Computers & Security*, 56, 2016: 70–82.
31. Tamjidyamcholo, A., Bin Baba, M. S., Shuib, N. L. M., and Rohani, V. A. Evaluation model for knowledge sharing in information security professional virtual community. *Computers & Security*, 43, 2014: 19–34.
32. Wash, R., Rader, E., and Fennell, C. Can People Self-Report Security Accurately?: Agreement Between Self-Report and Behavioral Measures *Proc. SIGCHI '17*, 2228–2232.
33. Witherspoon, C. L., Bergner, J., Cockrell, C., and Stone, D. N. Antecedents of organizational knowledge sharing: a meta-analysis and critique. *J of Knowledge Management*, 17(2), 2013: 250–277.